



Katja Wengler, Dirk Reichardt, Dennis Pfisterer (Eds.)

DHBW AI Transfer Congress 2023

Proceedings

Imprint

Stuttgart, September 2023

DHBW AI Transfer Congress
ISSN: 2976-1735

Published by:
Duale Hochschule Baden-Württemberg
Baden-Wuerttemberg Cooperative State University
Friedrichstraße 14, 70174 Stuttgart
represented by: Prof. Dr. Martina Klärle

PREFACE



This volume contains the conference proceedings of the 2nd AI Transfer Congress organized by the Cooperative State University Baden-Wuerttemberg (DHBW) and held on September 29th, 2023 in Heilbronn, Germany. The conference brought together academic and industrial researchers as well as users and practitioners in the field of Artificial Intelligence. As the name of the conference clearly states, transfer is in the focus. The conference provides two main tracks to cover the academics' and practitioner's perspective.

In the conference track, current findings in the field of Artificial Intelligence focusing applications, are discussed. AI applications have reached numerous domains and research areas like industrial applications in predictive maintenance, quality control, recommender systems, medical support systems, personalized nutrition, autonomous driving, stock market prediction or the interdisciplinary field of sustainability.

The workshop track provides a platform for discussions on selected application areas, especially with practitioners. Topics of the workshops range from business economics and insurance as well as executive level discussion on AI potentials via quality control and digital care to future skills in profession and education.

The review process involved an internationally staffed program committee consisting of experienced practitioners and scientists. We would like to thank the authors and reviewers for their excellent work. And last but not least we thank the multitude of members of the DHBW university administration and communication departments for their extensive work which made the AI Transfer Congress possible.

29. September 2023

Katja Wengler
Dennis Pfisterer
Dirk Reichardt



Sehr geehrte Damen und Herren, liebe
Teilnehmer:innen,

herzlich willkommen zum zweiten DHBW AI Transfer Congress (AITC) der Dualen Hochschule Baden-Württemberg auf dem Bildungscampus in Heilbronn. Tauchen Sie ein in die faszinierende Welt der Künstlichen Intelligenz. Als Transferhochschule am Puls der Zeit schaffen wir mit diesem Kongress eine einzigartige Plattform zur Vernetzung und Weiterbildung für unsere Dualen Partner, Wirtschaft, Gesellschaft, Politik und natürlich für die Hochschule. Wir spüren unsere Verantwortung, die rasante Entwicklung der Künstlichen Intelligenz verantwortungsbewusst mitzusteuern. Der AITC bringt unsere wissenschaftliche Expertise mit den Entwicklungsfragen unserer Dualen Partner zusammen und stößt damit AI-Innovationen an.

Künstliche Intelligenz ist der Treiber für Innovation und Fortschritt. Sie durchdringt nahezu jeden Aspekt unseres Lebens, von der Medizin über die Logistik bis hin zur Bildung. KI-Technologien eröffnen immense Möglichkeiten und werden die Art und Weise, wie wir arbeiten, studieren, forschen aber auch leben und kommunizieren verändern. Mit dem AITC eröffnen wir die Gelegenheit aktuelle Entwicklungen der AI interdisziplinär zu erkunden, Forschung zu erleben und selbst neue Erkenntnisse zu erlangen. Dieses Jahr bieten wir neben einer spannenden Keynote auch zwei übergeordnete Tracks. Track A setzt den Schwerpunkt auf wissenschaftliche Forschungsarbeiten zu Künstlicher Intelligenz insbesondere im Transfer- und Anwendungsbereich. Die hier präsentierten Arbeiten wurden von einem international und interdisziplinär besetzten Programm Komitee peer reviewt. Track B besteht aus interaktiven Weiterbildungsangeboten in Form von Workshops zu Themen der Künstlichen Intelligenz.

Als größte Transferhochschule im Land bringen wir so Wissenschaftler:innen, Duale Partner und Expert:innen aus verschiedenen KI-Bereichen zusammen, um Ideen auszutauschen, innovative Ansätze zu diskutieren und potenzielle Projekte und Partnerschaften zu entwickeln und anzustoßen.. Wir sind stolz darauf, ein solches Forum für den Austausch und die Zusammenarbeit zu bieten und setzen uns dafür ein, Innovation durch Forschung und Transfer für die Gesellschaft zugänglich zu machen und die Chancen Künstlicher Intelligenz zu nutzen.

Mein Dank gilt vor allem auch den Program Chairs Prof. Dr. Dennis Pfisterer, Prof. Dr. Dirk Reichardt und Prof. Dr. Katja Wengler. Ihr Engagement, das Thema KI im ganzen Land und mit allen Dualen Partnern zu vernetzen, verleiht dem Kongress das, was uns als DHBW ausmacht: wissenschaftlich praxisnah, regional und dual.

Ich wünsche Ihnen allen einen impulsreichen zweiten DHBW AI Transfer Congress, spannende Diskussionen und neue Einblicke in die faszinierende Welt der KI.

Ihre

Prof. Dr. Martina Klärle

Präsidentin der Dualen Hochschule Baden-Württemberg

Table of Contents

Invited Papers	
Leveraging Artificial Intelligence for Sustainability.....	1
<i>Christiane Plociennik</i>	
Conference Session	
ChatGPT – Are the students ready for the AI age?.....	2
<i>Ulrich Bucher</i>	
Predictive Maintenance for Distributed Systems Using Data Science.....	10
<i>Ferdinand Koenig and Karlo Kraljic</i>	
Patterns in the Context of Small Data - Lessons learned from Data Science Projects.....	20
<i>Friedemann Schwenkreis</i>	
Model-Agnostic Overconfidence Reduction for Tabular Data.....	27
<i>Alexandros Nanopoulos</i>	
An approach to implement user-based recommendation systems with small-sized data sets	34
<i>Gerhard Götz</i>	
A stacking approach for vehicle loan fraud detection.....	41
<i>Jan Wolf</i>	
Behavioral Biases in Human-Machine-Interactions in Machine Learning.....	48
<i>Theresa Scheutzow</i>	
How to prepare students for the AI era?.....	58
<i>Ulrich Bucher and Kai Holzweißig</i>	
Digital Twins von Organisationen (DTO) für die Personalarbeit: KI als Ersatz oder Unterstützung des HR-Bereichs	70
<i>Benedikt Hackl and Joachim Hasebrook</i>	
AI Glucose Prediction For An Insulin Recommendation System Based On Smartwatch Activity And Glucose Monitoring Data For Type 1 Diabetics.....	82
<i>Stefanie Neumann, Armin Zundel and Kay Berkling</i>	
A Language Training Approach to Improve Children’s Expression Skills Using AI Methods for Text-to-Picture Conversion.....	92
<i>Elisa Schäfer, Armin Zundel and Kay Berkling</i>	
”Senior Consultant ChatGPT” - a Model of Collaboration Between Generative AI and Consultants.....	102
<i>Friedrich Augenstein</i>	
Poster Session	
Enhancing Quality Control through Computer Vision: A Comprehensive Study.....	109
<i>Shobhit Agarwal, Bozena Lamek-Creutz and Rami Mochaourab</i>	

Analyzing news articles on the COVID-19 pandemic regarding the timeline, vaccination, and sentiment	118
<i>Linus Eickhoff, Florian Kellermann and Monika Kochanowski</i>	
Classification algorithms in database management systems	119
<i>Alicia Dietrich and Olaf Herden</i>	
Künstliche Intelligenz im Handel: Online-Kurs für Betriebswirte	120
<i>Johannes Kolb, Daniela Wiehenbrauk, Oliver Janz and Armin Mueller</i>	
KI-basierte Punktwolkenanalyse und webbasierte VR-Visualisierung (KIP-VR)	121
<i>Dominik Ruoff and Tim Jansen</i>	
Steered Training Data Generation for Learned Semantic Type Detection	123
<i>Sven Langenecker, Christoph Sturm, Christian Schalles and Carsten Binnig</i>	
A pilot study comparing the performance of deep learning model (LSTM) and statistical models (ARIMA and SARIMA) for demand forecasting of an automotive spare parts	124
<i>Nehalben Ranabhatt and Wilhelm Ruckdeschel</i>	
Künstliche Intelligenz in der Optimierung personalisierter Ernährungsstrategien: Das Individual Nutrition Advisory Tool" (INAT)	125
<i>Kathrin Friedrichs, Hande Gagali, Timo Sievernich, Cornelia Klug, Katja Lotz and Alexandr Parlesak</i>	
Fahrspurerkennung für ein autonomes Modellfahrzeug mit Convolutional Neural Networks	127
<i>Anton Utz, Benjamin Arp and Matthias Drüppel</i>	
Re-existing inside of the world of the algorithmic determinism	128
<i>Renato Silva Guimaraes</i>	
Quantum Computing for Feature Selection in Machine Learning	129
<i>Gerhard Hellstern, Vanessa Dehn and Martin Zaefferer</i>	
Parallel classification algorithms for big data applications	130
<i>Jannik Duerr and Olaf Herden</i>	
CITAI: Building Bridges or Breaking Barriers? Unveiling the Secrets of Citizen Trust in AI Innovations	131
<i>Sinu Thirukketheeswaran, Marc Kuhn, Lars Meyer-Waarden, Gonser, Lay, Österle and Yuras</i>	
Realization of an environment for event based vision - Abstract for publication	132
<i>Felix Ehret and Erik Langner</i>	
Multi Object Tracking using Machine Learning	133
<i>Christian Holz, Christian Bader and Matthias Drüppel</i>	
Stock Market Prediction System using Hybrid Model	134
<i>Akshat Singh and Virrat Devaser</i>	
Towards Learned Cost Estimation for Streaming Queries on Heterogeneous Hardware	139
<i>Roman Heinrich, Manisha Luthra, Harald Kornmayer and Carsten Binnig</i>	

Birdstrike - Laufende Forschungskooperation mit dem Zentrum für Geoinformationswesen der Bundeswehr	140
<i>Herbert Neuendorf, Alexander Auch, Christian Rohlf, Sebastian von Massow and Moritz Deininger</i>	
Abgleich der Kompetenzen aus dem Curriculum des Bachelorstudiengangs Data Science und Künstliche Intelligenz mit DASC-PM	141
<i>Stephan Daurer and Martin Zaefferer</i>	

Leveraging Artificial Intelligence for Sustainability

Christiane Plociennik
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
christiane.plociennik@dfki.de

Abstract—Artificial Intelligence can contribute to our living within the planetary’s boundaries. There are many AI applications that can bring sustainability forward. However, it is important to consider the environmental effects of AI as well.

Index Terms—AI, sustainability, Circular Economy, Digital Product Passport, resource consumption

I. AI FOR SUSTAINABILITY

As the world population grows, resources become more and more scarce and CO₂ emissions are on the rise, humanity faces the challenge to make societies and industries truly sustainable. Artificial Intelligence (AI) can play a key role in this context. It is estimated that AI may help achieve many of the UN’s Sustainable Development Goals (SDGs): AI helps combat the pollution of our rivers and oceans, for instance. It thus contributes to clean water supply for every human being. Another area that can benefit from AI is our economy – the way we make, use, and dispose of products. Collecting and analyzing data using machine learning algorithms can facilitate the transition from a linear to a circular economy that focuses on reuse, refurbishment, recycling and similar strategies rather than throwing products away as trash.

II. SUSTAINABILITY OF AI

However, the relationship between AI and sustainability is complex. Being an inherently resource-consuming technology itself, AI may have a negative environmental impact when used carelessly. Large AI models consume significant amounts of energy during training and, to a lesser extent, during their use phase. For example, training GPT-3 once is estimated to consume as much energy as 500 German households consume in a year. Additionally, there are more environmental effects to consider: The resource usage for producing, running and later disposing of the hardware associated with the usage of AI must also be taken into account. At present, however, it is hard to estimate these effects and even harder to compute CO₂ equivalents for them. This is due to the fact that supply chains in the electronics industry are complex and global. Information about the environmental impacts of all the parts of an electronic device is not readily available. In the future, Digital Product Passports can help create more transparency here.

Funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection, project ReCircE, grant number 03EN2353B.

III. THE PATHWAY TO A GREENER FUTURE

Techniques like transfer learning or reducing a model’s size can mitigate some of the environmental effects of AI. This, however, should only be the first step towards a more resource-aware usage of AI. Traditionally, AI developers have been focusing largely on performance. It is now crucial to shift that focus such that sustainability considerations are also taken into account. Do we need, for instance, always the newest and largest model, even if that means that models become ever more larger and more resource-consuming and must be frequently retrained? For some applications, this may not be necessary. Equally important is a broad societal debate: Which are the areas where AI can have definite positive impacts on sustainability? These are the areas where we definitely want to employ AI. But are there also areas where we should refrain from AI usage because the benefits simply do not merit the resource consumption? For a greener future, we must all join forces to make AI part of the solution, not part of the problem.

ChatGPT – Are the students ready for the AI age?

Bucher Ulrich
Study Centre Service Management
DHBW Stuttgart
Stuttgart, Germany
ulrich.bucher@dhbw-stuttgart.de

Abstract— AI is continuously changing work content and ways of working and thus also the demands on the skills of its users. The survey presented here raised the question of the extent to which students have the necessary critical thinking skills and are thus sufficiently prepared to use ChatGPT. ChatGPT currently has a number of weaknesses and limitations. Through this study, the aim was to investigate for the first time whether students have the necessary critical thinking competencies to deal with them. For this purpose, various tasks were developed and an online survey was used to record the competencies of 307 DHBW students. The study revealed considerable deficits. It was found that students often adopt the output of ChatGPT without reflection. Inconsistencies between the task and the output of ChatGPT are largely ignored. In particular, the group of students who actively use ChatGPT seems to be especially vulnerable to ChatGPT's weaknesses and limitations. This is because this group ascribes to itself competencies that it does not possess adequately. The results of the study underscore the need to better prepare students for the AI era.

Keywords— AI Literacy, Artificial Intelligence, ChatGPT

INTRODUCTION

In the context of artificial intelligence, numerous reasons can be found for the need for critical thinking. These include the limitations of AI, filter bubbles and doomscrolling, user manipulation and cybersecurity. In addition, language models may learn negative generalisations, stereotypes or misrepresentations of certain social groups [1]. For example, women's higher likelihood of being associated with nursing professions may perpetuate undesirable stereotypes [1]. Furthermore, the information a chatbot disseminates may be false or biased, or it may discriminate against individuals or groups. For example, Hudson reports that she had ChatGPT answer the question of whether a person can change their gender 30 times [2]. In three cases ChatGPT answered her that this was not possible, in seven cases the answer was yes, and in 20 cases the answer referred to "gender reassignment" instead. According to Hudson, it took ChatGPT eighteen attempts to output the correct answer for the first time [2].

Borij provides numerous categories and examples of ChatGPT's failures [3], including in the areas of argumentation, logic, mathematics and arithmetic, factual errors, and bias and discrimination. Shen et al. find in a broad test of ChatGPT that its reliability varies and is especially below average in scientific questions [4].

Beautifully packaged, ChatGPT gives users the impression that it can understand language or grasp meaning. According to Sulmont et al., many students outside of computer science assume that computers think like humans [5]. However, according to Bender / Koller, ChatGPT does not have an understanding of meanings because they always have a reference to something outside of language [6]. In this respect, the models can indeed conduct a dialogue similar to that of a human being. However, this ability is not

based on an understanding of meanings, but on the fact that statistical regularities were found in the training data [6].

The aforementioned weaknesses and limitations highlight the need for a critically reflective use of AI technologies [7]. The need for critical thinking skills in the context of using AI tools arises in particular from the fact that students generally overestimate the complexity and intelligence of IT systems and have difficulty recognising their limitations and identifying constraints [5].

However, quantitative empirical research on students' critical thinking skills in the context of ChatGPT is still in its infancy. Although numerous empirical studies have been published in the context of generative AI tools, most of them deal with critical thinking only in their conclusions without making it the actual object of investigation. Thus, numerous studies emphasize the need for critical thinking in the context of ChatGPT without investigating the status quo of students' competencies [8, 9, 10].

In contrast, the use of ChatGPT by students has been widely studied [11, 12, 13]. ChatGPT is by far the most frequently used AI tool [11]. While it is used by almost half (49%) of the students in Germany in mid-2023, the second most used AI tool DeepL is used by only 12% of the students [11]. The use of generative AI tools takes place for a wide range of tasks. In particular, they are used for research, clarification of comprehension questions, text generation, and problem and decision making [11]. In this context, more than half of the respondents were critical of the scientific nature as well as the error avoidance of the output [11]. Overall, trust in the accuracy of ChatGPT's answers is low. Four out of five users of the service do not trust its answers or at best trust them only partially [12]. As a result, a large proportion of users (73.3%) verify ChatGPT's answers [12]. Shoufan investigated perceptions of ChatGPT following its use with a group of 56 students and found that several students indicated problems with accuracy [8]. Choudhury / Shamszare state in their study that trust is a central factor for the use of ChatGPT. However, overreliance on its advice could lead to misinformation and the risks associated with it [14].

To the author's knowledge, an empirical investigation of students' critical thinking competencies in the context of ChatGPT has not yet taken place. However, such a survey would be significant to identify any skill deficiencies and adapt the curricula accordingly. This is especially true against the backdrop of rapidly increasing diffusion and growing importance of generative AI tools such as ChatGPT.

METHODICAL APPROACH

Determining students' competencies and behavioral dispositions regarding critical thinking in the context of ChatGPT requires that the construct of critical thinking be clearly determined. To date, however, this construct remains elusive [15]. For this reason, it is important to first be clear about the purpose of critical thinking. This consists of

avoiding wrong and hasty decisions [16]. Therefore, critical thinking also aims at a conscious mental processing of information. This also reflects the definition of critical thinking by O'Hare and McGuiness, according to which critical thinking consists of questioning statements or opinions in order to find out what one should believe or how one should behave [25].

The distinction between a hasty (wrong) decision and conscious mental information processing echoes the duality of thought processes as viewed by dual process theorists (such as Kahneman) in cognitive psychology. Bonnefon (2016) pointed out the parallel between critical thinking and the analytic system [24]. While type 1 thinking is characterized by fast, implicit, and automatic processes, type 2 is characterized by analytic processes that are goal-directed, self-regulating, conscious, and effortful [22].

Kahneman argues that we need to sensitize ourselves to recognize the situations in which quick thinking leads to poor results in order to slow down thinking in these situations [17]. Therefore, in assessing students' competencies in critical thinking, the study starts from the weaknesses that ChatGPT has shown in the past. For it is in this context that critical thinking plays out its benefits.

In order to determine which type of thinking takes place, the students were confronted with different tasks, whereby they were presented with an output of ChatGPT in the form of a screenshot for each task. Here, the question arose whether the subjects took over the output of ChatGPT without reflecting on it (which corresponds to type 1 thinking) or whether they analyzed it and thus used their critical thinking skills (= type 2 thinking).

Since critical thinking is not a one-dimensional construct [15], but requires different competencies and behavioral dispositions such as interpretation, analysis, evaluation, reasoning, explanation, and self-regulation [19], a total of six different tasks were developed. All tasks were formulated in an open-ended manner, as is common in critical thinking surveys [19]. In evaluating the students' free-text responses, we examined whether they were aware of deficiencies in the output of ChatGPT. The adoption of incorrect statements from the output of ChatGPT was seen as an indication that critical thinking had not sufficiently taken place.

The literature points out that critical thinking is not independent of the subject area and therefore must be taken into account when measuring it [19]. Therefore, the survey of critical thinking must be situated [20] and connected to tasks that students are confronted with in their studies.

For this reason, the students were confronted with the situation of using ChatGPT in the context of writing their bachelor thesis. As a topic for the bachelor thesis, the students were given the task of creating a market entry concept for an English beverage company for the German beer market. This topic was chosen because it is very general and not very specific and therefore a large part of the students (especially the business students) can identify with such a topic.

In addition, the topic of creating a market entry concept allowed for the construction of a story. Building a story is also used in other studies to measure critical thinking [19]. This has several advantages, such as a greater interest of the respondents and thus a higher motivation when answering the tasks [18]. In particular, the subject matter of a bachelor thesis

should signal to students the demands of the tasks and thus the requirement for critical thinking.

The tasks were inserted along the storyline, which proved to be a more effective strategy than placing the tasks only at the end of the story [21]. To avoid spillover effects between the different tasks, they were clearly delineated from each other within the context of the bachelor thesis being written. For example, the question about the number of breweries in the first task should not affect the characterization of the target group in the fourth task, which asked about the "abropause."

Students were told that not only the output of ChatGPT, but also other sources could be used to answer the assigned task. The answers to the questions should then reflect what the students would write in their bachelor thesis. Each task therefore first took up the background of how the question arose in the context of the bachelor thesis.

In addition to determining the Critical Thinking competencies, the study should answer whether the students who actively use ChatGPT have higher competencies than the students who do not actively use it. Because if differences are found here, then this would be an indicator that the competencies emerge with the use of ChatGPT.

In addition, the question arose as to whether active users ascribe to themselves a higher level of self-competence than non-users. For this purpose, a self-assessment of competencies and behavioral dispositions regarding critical thinking was collected in the study. The latter is based on an item set for the survey of AI literacy by Laupichler / Raupach, which was adapted to the context of ChatGPT [23]. Furthermore, the self-assessment survey should allow a comparison between the students' self-perception and their performance on the competency tests.

The survey took the form of an online survey based on a standardized questionnaire. A pre-test was conducted prior to the survey and the feedback was incorporated into the questionnaire. Subsequently, 756 students of business administration and business informatics at DHBW Stuttgart, DHBW Heilbronn and DHBW Mosbach were invited to participate by e-mail. In addition, a call for participation in the study took place in various lectures. In these, students were given the opportunity to participate in the study without any time pressure built up. Students were informed that participation was on a voluntary basis. The data collection took place in February and March 2023. A total of 307 students took part in the survey. A high response rate of 40.6% was thus achieved.

RESULTS

As expected, most students were aware of ChatGPT. 260 respondents (85%) were aware of ChatGPT prior to the survey. 41 students (13.4%) were unaware of ChatGPT, 5 could not answer this with certainty and one respondent refused to answer this question.

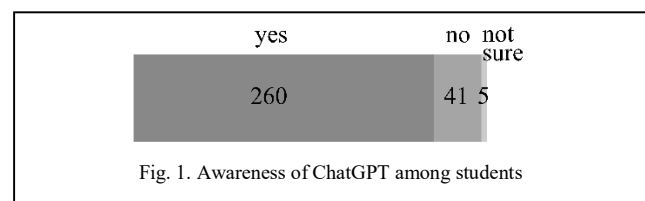


Fig. 1. Awareness of ChatGPT among students

The majority of respondents can draw on experience with the internet service. Only 62 out of 265 students have neither used nor tested ChatGPT so far. 89 students (33.6%) have tested the service but do not actively use it. 99 students (37.4%) have included the internet service in their tool set. 11 students (4.2%) were disappointed with the service and therefore stopped using it. 4 students (1.5%) classified themselves as "other" with regard to the use of the service. The widespread use of ChatGPT among the students means that any weaknesses that the students show in completing the tasks can only be attributed to a limited extent to a lack of experience with the internet service.

A. Type 1 thinking - unreflective adoption of the output

Part of critical thinking is based on evaluating different sources and checking them for inconsistencies [34]. This part of critical thinking was picked up by the following experiment. In this experiment, the test persons were asked to answer the question of how many breweries there are currently in Germany. The subjects were randomly assigned to one of three groups, which were presented with different information (see Fig. 2):

Group 1: A supposed output of ChatGPT stating that there are over 900 breweries in Germany.

Group 2: The actual edition of ChatGPT: Here it is said that there are over 1,300 breweries.

Group 3: A supposed issue of ChatGPT stating that there are over 1,700 breweries. In addition, an excerpt of a Google results page was shown, which refers to 1,492 breweries according to the Federal Statistical Office.

The statement that there are more than 900 or more than 1300 breweries in Germany is not fundamentally wrong (even if it is misleading). However, the statement becomes false if only the number of 900 or 1300 breweries is mentioned without adding "more than". This was the case in the vast majority of the students' answers. A large part of the students in the first two groups took the number from the output ChatGPT without the addition of "more than", which distorts the statement. In these two groups, the task is only solved correctly by about one in five students (see Fig. 3). How much the output deviates from the truth does not seem to play a significant role. In the first group, the proportion of incorrect answers is hardly lower than in the second group, although this group was given a significantly lower number of breweries.

The response behavior of the third group makes it clear that the students are quite capable of identifying the correct answer if they are also presented with the data from the Federal Statistical Office in parallel to the output of ChatGPT. Since Google shows the data of the Federal Statistical Office on the first result page, the effort to verify the output of ChatGPT would also have been manageable.

	<i>false answer</i>		<i>correct answer</i>	
Group 1	69	79.3%	18	20.7%
Group 2	79	82.3%	17	17.7%
Group 3	17	19.8%	69	80.2%

Fig. 3. Evaluation of the answers to the question about the number of breweries in Germany

Comparing the first two groups with the third group shows that in the latter the probability of a correct answer is 17.1 times higher than in the first two conditions (95% confidence interval: 9.0 to 32.7). The Z-test for two proportions shows significant differences between the groups ($z = 9.6$, $p < 0.001^{***}$). This makes it clear that the majority of the test persons do not take the trouble of researching information themselves, but rather adopt the contents of the ChatGPT output without reflection.

This result is surprising in that students in other studies had considerable doubts about the truthfulness of the output, and 73.3% of respondents said they verified ChatGPT's answers [12]. This behavior was not exhibited by students in the present study. A variety of reasons may be responsible for this, such as low motivation, the credibility of the output, or the different timing at which the studies were conducted and the students' awareness that had occurred in the meantime.

B. Ability to think through a statement logically

Critical thinking is closely related to logical thinking. One of the essential skills is to assess the logic of an argument and to process the information in a logical form in such a way that the correct conclusions are drawn in the end [28]. The following task was designed to investigate whether students are able to critically reflect on the logic of reasoning through

Group 1

According to one estimate, there are over 900 breweries in Germany. The exact number can vary, however, as there are always new breweries being founded and others closing down. Germany is known for its rich brewing tradition and its diverse beer styles, ranging from pilsner to wheat beer and dark beer varieties.

Group 2

According to one estimate, there are over 1,300 breweries in Germany. The exact number can vary, however, as there are always new breweries being founded and others closing down. Germany is known for its rich brewing tradition and its diverse beer styles, ranging from pilsner to wheat beer and dark beer varieties.

Group 3

According to one estimate, there are over 1,700 breweries in Germany. The exact number can vary, however, as there are always new breweries being founded and others closing down. Germany is known for its rich brewing tradition and its diverse beer styles, ranging from pilsner to wheat beer and dark beer varieties.

On Google you have seen the following information regarding the number of breweries:

Currently, 1492 breweries are operating nationwide. That is 82 establishments more than last year and over 150 more than five years ago. This is reported by the German Brewers Association, the umbrella organization of the brewing industry, with reference to new data from the Federal Statistical Office.

<https://brauer-bund.de/uploads/2023/07/18022...>

Number of German brewers continues to rise

Fig. 2. Output presented to the groups on the question about the number of breweries in Germany (the output of ChatGPT as well as Google were translated from German)

ChatGPT and whether they can draw the correct conclusions based on this.

This experiment goes back to ChatGPT's limited abilities in calculating mathematical expressions. In the past, the service had great difficulty in multiplying large numbers, taking roots, adding or subtracting irrational numbers and calculating powers (especially for fractions) [3].

In experimental condition A (shown on the right side of Figure 4), the process for determining the probability is output correctly and the probability is calculated correctly. In experimental condition B (shown on the left), the calculation process is also described correctly, but here a calculation error was introduced that leads to a result that is 0.92, which is 0.50 higher than the true value. This is due to the fact that $5/63$ is not resolved into $125/216$ but into $12/216$. The presentation of the calculation process does not require leaving the questionnaire. The respondents have all the information available that is necessary to check the result. The description of the calculation process should encourage them to check it, especially for those respondents who are not yet familiar with the tool's capabilities. Since the test persons participate in the survey via smartphone, tablet or PC, the effort for checking is comparatively low.

The probability of a correct answer to the question was 37.0 times more likely for the output that contained the correct solution than for the output with the incorrect solution (95% confidence interval: 17.56 to 77.95). The Z-test for two proportions shows significant differences between the experimental conditions ($z = 10.9$, $p < 0.001^{***}$).

What is surprising about the results is the considerable proportion of students who thought independently about the solution to the task. Thus, 18.75% of the respondents answered the task incorrectly although the output contained the correct solution. At the same time, 10.48% of the respondents answered the task correctly despite an incorrect output. These results are encouraging in that they show a certain degree of critical reflection (even though in numerous cases this resulted in an incorrect answer - even though the output contained the correct solution). Considering the results of the previous experiment, this suggests that while motivation is not sufficient to conduct an information search outside of ChatGPT, reflection then occurs within the logic of the output.

C. Context - up-to-dateness of the language model

The third task takes up the finding from the literature that critical thinking cannot be considered independently of the context and the characteristics of the medium used [29]. Thus,

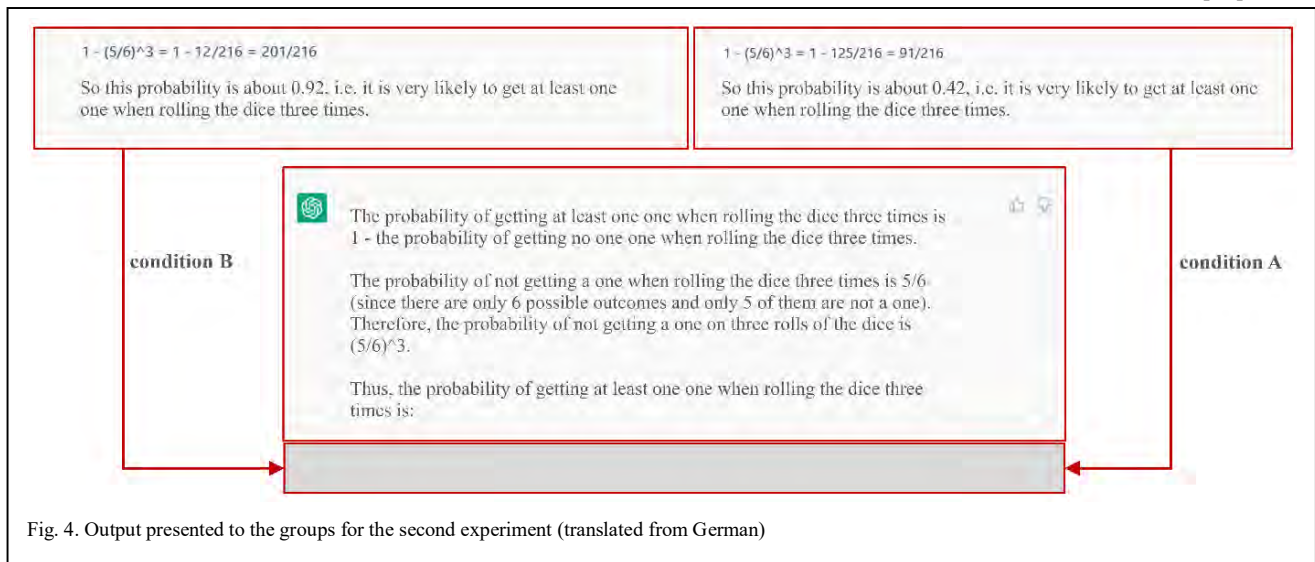


Fig. 4. Output presented to the groups for the second experiment (translated from German)

The background story for this task was that the students had thought of a lottery as part of the development of the market entry concept. A random mechanism was to decide whether participants would receive a prize or not.

The results show significant differences in response behavior between the two experimental conditions (see Fig. 5). If the procedure was correctly described in the output of ChatGPT and the correct result was output (condition A), then the answers were also generally correct. However, it was obviously difficult for the students to identify the error introduced in the second experimental condition and to correct the result accordingly.

	<i>false answer</i>		<i>correct answer</i>	
Condition A	21	18.75%	91	81.25%
Condition B	111	89.52%	13	10.48%

Fig. 5. Evaluation of the answers to the question about the number of breweries in Germany

Kahneman establishes a connection between the thinking processes and the context [30]. Depending on the context, the thought processes run differently and are either more strongly influenced by type 1 thinking or by type 2 thinking.

The feature addressed in this task is the development date of the language model used in March 2023, which dates back to 2021. Events that have occurred since 2021 were not considered by ChatGPT. This includes that in October 2022 there was a change at the top of the government in the United Kingdom and Rishi Sunak was elected as the British Prime Minister. Instead, ChatGPT gave Boris Johnson as the answer to the question about the British Prime Minister in March 2023.

As part of this task, students were told that due to Brexit, the administrative burden of exporting beer from England to the EU had increased significantly. Therefore, their bachelor's thesis would address the question of what course the current British government is pursuing. This in turn led to the question of who the British Prime Minister is. This question was

intended to test whether the respondents are aware of the limitations of ChatGPT and take them into account when answering.

Although general knowledge was asked at this point, about one third of the respondents are still not able to give a correct answer. This group of students seems to be unaware of the limitations of ChatGPT. Seven students gave neither Boris Johnson nor Rishi Sunak as their answer, but named former prime ministers such as Liz Truss (see Figure 6).

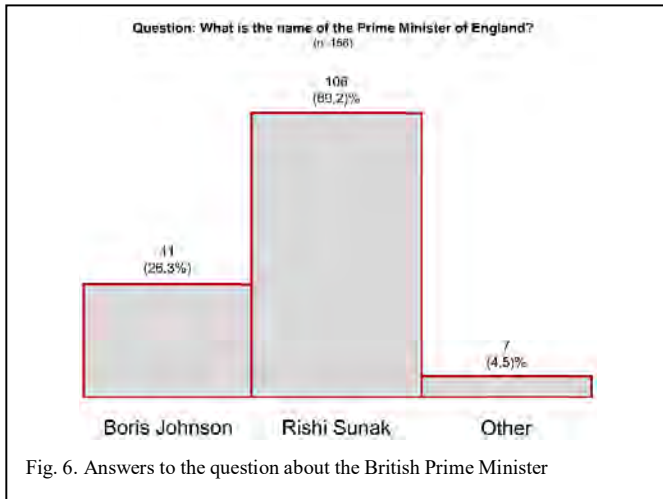


Fig. 6. Answers to the question about the British Prime Minister

D. Truth seeking - dealing with discrepancy

Truth seeking is an essential component of critical thinking and was therefore the subject of another experiment. This is of particular importance in that truth seeking is considered a predictor of dispositions with respect to critical thinking [26].

Therefore, when measuring critical thinking, many empirical studies construct problems that deal with finding the truth [27]. One of these problems is how to deal with discrepancy or contradictory information [27].

The following task examined how students deal with a discrepancy. The context of the task was the examination of the target group in the development of the market entry concept. Since the target group considered in the bachelor thesis is men over the age of 45, the question was raised as to when the “abropause” begins in men. The term “abropause” is fictitious. Since ChatGPT searches for terms that are as similar as possible to unknown terms, the answer does not refer to the “abropause” but to the andropause, which leads to a discrepancy (see Fig. 7).

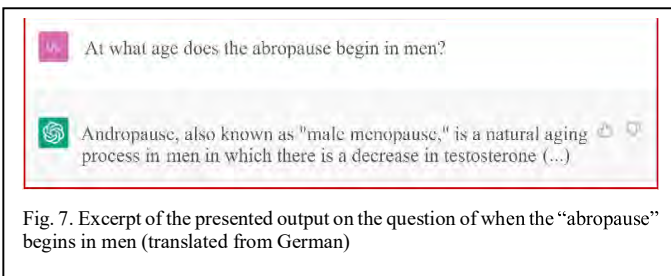


Fig. 7. Excerpt of the presented output on the question of when the “abropause” begins in men (translated from German)

In order to reduce the resulting cognitive dissonance, the subjects have various options:

- Ignoring the discrepancy

- The procurement of information
- The reference to the discrepancy

The results make it clear that the discrepancy between the question about the “abropause” and the output of ChatGPT regarding the “andropause” are largely ignored (see Fig. 8). Only 3.3% of the respondents answered that there is no abropause and only 2.6% stated that the answer refers to the andropause.

Category	Frequency	Percent
40-55 years	79	52.0%
Other age statement	64	42.1%
There is no “abropause”	5	3.3%
Note: Answer reference andropause	4	2.6%

Fig. 8. Answers to the question at what age in men begins the “abropause”

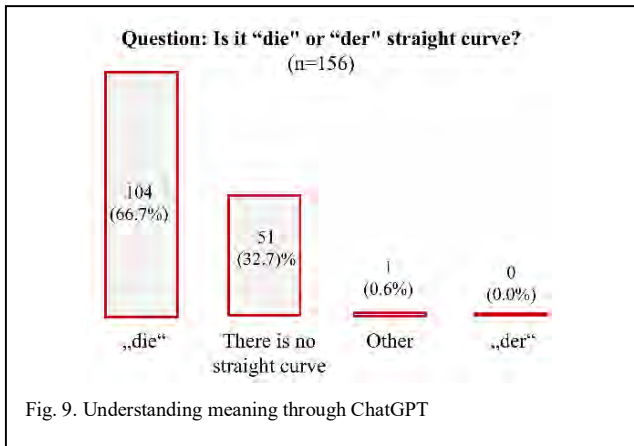
E. Understanding of the world / of an issue

The ability to think critically cannot be separated from the understanding of an issue and the task to be accomplished in that context [31]. ChatGPT does not have an understanding of the world [32]. Ultimately, the service is based on an evaluation of the statistical regularities in the use of language [33]. However, whether these regularities reflect reality or not cannot be answered by the service. In this respect, the service can also generate answers that sound plausible, but have nothing in common with reality.

This was taken up with the question whether in German when asking about the straight curve as an article "die" or "der" should be used. When you ask ChatGPT this question, the service does not recognize the nonsense of the question. Since the service has no idea about the world, it is also not aware that in general understanding there is no straight curve (even if this phenomenon does exist in physics). Thus, ChatGPT can determine on the basis of the training data that the correct article for the noun "curve" is "die". However, it obviously does not understand the contradiction between the term curve and the adjective "straight". Therefore ChatGPT outputs that it is called “die” straight curve.

The question whether it is called "die" or "der" straight curve was the only task that was not self-developed. This one rather has an origin going back long into the past. This question is a trap in the form of an oxymoron. For the hint that the noun curve must be connected with the feminine article in German is quite correct. However, the sole hint overlooks the fact that the statement makes no sense in general understanding.

Students also show significant weaknesses in critical thinking on this task (see Fig. 9). Two-thirds of the students do not develop sufficient understanding of the context in which the task takes place and, like ChatGPT, answer the question with "die". One possibility is that the response to the task was reduced to determining the correct adjective.



F. Evaluation over all tasks

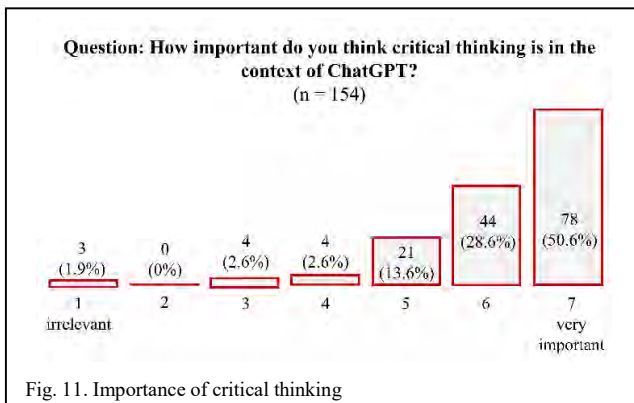
Looking at the response behaviour across all five tasks, the students were able to solve an average of 42.4% of the tasks correctly. There is a strong scatter in the data ($sd = 0.288$). The lower third of the test persons only managed to solve a single task at most (see Fig. 10). The broad majority of the subjects ranged between 40% and 60%, and about one in eight students had no difficulties and were able to answer at least four out of five tasks correctly.

Proportion of correctly answered tasks	Frequency	Percent
0	51	18.8
0.2	33	12.2
0.25	2	0.7
0.4	54	19.9
0.5	46	17
0.6	46	17
0.67	2	0.7
0.75	2	0.7
0.8	11	4.1
1	24	8.9

Fig. 10. Proportion of correctly answered tasks (n = 271)

G. Assessment of the importance of critical thinking

There is widespread agreement among the students about the importance of critical thinking. The majority of students consider critical thinking to be very important or important in the context of ChatGPT (see Fig. 11). Hardly any student



assigns no or only little importance to critical thinking. An awareness of the importance of critical thinking thus seems to be clearly present.

H. Self-assessment of competences and behavioural dispositions related to critical thinking in the context of ChatGPT

The students consider themselves surprisingly well equipped with regard to their own competences and behavioural dispositions (see Fig. 12). They only see deficits with regard to their own ChatGPT literacy in individual aspects. For example, many students are unclear about how machine learning models are trained, validated and tested. Accordingly, it is also not clear to some of the students how biases are created in AI systems. The wide spread of answers regarding the black box character of many AI systems is also striking. While some of the students are not able to explain the concept of the black box, another part of the students has dealt with it.

I. Comparison of active users with students who have never used ChatGPT before

In order to check whether the necessary competences are built up through the use of ChatGPT and thus whether it is possible to do without promoting AI literacy in teaching, the group of students who actively use ChatGPT was compared with those who have never used ChatGPT. Here, the question arose in particular as to whether these two groups differ in terms of self-assessment of competences. For this purpose, a competence score was formed from the twelve items of the self-assessment of one's own competences and behavioural dispositions in relation to critical thinking. This was calculated additively from the twelve items.

Comparing the group of active users with those who have never used ChatGPT, the first group attributes significantly higher competences to itself. They achieved a competence score of 59.15 ($n = 46$). The group for whom ChatGPT is new territory was only 52.74 ($n = 27$) in comparison. Between the two groups, there were statistically significant differences between the self-assessment of competences, $t(71) = -2.491$, $p = .008$.

Although the group of active users assesses themselves as significantly higher in terms of their competences, this group does not perform better in answering the tasks. Thus, when tested using the T-test, there are no significant differences between the two groups, $t(132) = -1.273$, p (two-sided) = .205. The active users in particular therefore appear to be particularly at risk, as they attribute a higher level of self-competence to themselves, which, however, is not evident when using ChatGPT to solve tasks.

CONCLUSION

AI tools such as ChatGPT have spread and gained acceptance among students at a rapid pace. With the use of these tools, the ways of working as well as the demands on students' critical thinking skills are changing.

In the study, students demonstrated significant weaknesses in critical thinking competencies and behavioral dispositions. The weaknesses and limitations of ChatGPT included in the five tasks would have required Type 2 thinking, but this is not sufficiently demonstrated by the students. Students adopt a flawed output of ChatGPT unreflectively in many cases

ChatGPT Critical Thinking Literacy

Question: How do you assess your own competences and behavioural dispositions in relation to critical thinking in the context of ChatGPT?

Statement	very low						very high	Overall
I can explain the differences between human and artificial intelligence.	4 2.4%	4 2.4%	20 11.98%	21 12.57%	54 32.34%	47 25.15%	22 13.17%	167 100%
I can explain the difference between general (or strong) and narrow (or weak) artificial intelligence.	18 11.76%	24 15.69%	22 14.38%	19 12.42%	38 24.84%	18 11.76%	14 9.15%	153 100%
I can name weaknesses in artificial intelligence.	2 1.31%	11 7.19%	10 6.54%	19 12.42%	39 25.49%	49 32.03%	23 15.03%	153 100%
I can name the strengths of artificial intelligence.	1 0.66%	2 1.32%	1 0.66%	17 11.18%	44 28.95%	56 36.84%	31 20.39%	152 100%
I can describe risks that can occur when using artificial intelligence systems.	2 1.31%	2 1.31%	7 4.58%	29 18.95%	41 26.8%	49 32.03%	23 15.03%	153 100%
I can describe the benefits that can arise from the use of artificial intelligence systems.	1 0.65%	3 1.96%	2 1.31%	19 12.42%	40 26.14%	61 39.87%	27 17.65%	153 100%
I can describe how machine learning models are trained, validated and tested.	18 11.76%	35 22.88%	36 23.53%	26 16.99%	19 12.42%	15 9.8%	4 2.61%	153 100%
I can explain why data plays an important role in the development and application of artificial intelligence.	3 1.96%	4 2.61%	9 5.88%	25 16.34%	42 27.45%	35 22.88%	35 22.88%	153 100%
I can explain what the term "black box" means in the context of artificial intelligence systems.	40 26.32%	26 17.11%	16 10.53%	19 12.5%	22 14.47%	14 9.21%	15 9.87%	152 100%
I can describe how biases arise in AI systems.	19 12.5%	14 9.21%	18 11.84%	28 18.42%	29 19.08%	23 15.13%	21 13.82%	152 100%
I can think critically about the potential impact of artificial intelligence on individuals and society.	2 1.32%	3 1.99%	9 5.96%	21 13.91%	45 29.8%	46 30.46%	25 16.56%	151 100%
I can explain what an algorithm is.	4 2.63%	5 3.29%	19 12.5%	24 15.79%	43 28.29%	34 22.37%	23 15.13%	152 100%

Fig. 12. Self-assessment of competences and behavioural dispositions related to critical thinking in the context of ChatGPT.

without sufficiently verifying its truthfulness (Task 1). Errors in the logic of ChatGPT's reasoning are not identified and the correct conclusions drawn from them (Task 2). The limitations of ChatGPT regarding the up-to-dateness of the language model are not perceived in view of the black box character of the service and are therefore not taken into account in the decision making process (Task 3). Even obvious discrepancy between the question expressed in a prompt and the content of the output of ChatGPT are often simply ignored (Task 4). Considerable parts of the students also did not acquire an understanding of the problem although the service has no understanding of the world (task 5).

In particular, the group of students who actively use ChatGPT seems to be particularly vulnerable to the weaknesses and limitations of ChatGPT. This is because this group ascribes competences to itself, which it, however, only insufficiently possesses. This raises the question of whether the use of ChatGPT has a positive effect on the self-assessment of competences, but these do not arise from the use alone. In this way, a gap could emerge that leads to a special vulnerability situation for this group.

In order to promote a competent use of AI tools such as ChatGPT and to avoid negative effects, it is important to train both critical thinking and the competence of meaningful use among students.

However, it is possible that the study overstates the students' deficits. Solving the tasks had no consequences for the students. Since critical thinking or type 2 thinking is exhausting, it may well be that the students avoided it. In

another situation where critical thinking has a concrete benefit, students may behave differently. Moreover, the subject of this study is a dynamic field. Both the weaknesses and limitations of ChatGPT are constantly changing, as is students' awareness of them. This dynamic needs to be taken into account both in the measurement of competencies and behavioral dispositions regarding critical thinking in the context of generative AI and in the design of curricula.

ACKNOWLEDGEMENT

This study was conducted with the support of Nicole Klein (DHBW Stuttgart).

REFERENCES

- [1] H. R. Kirk, Y. Jun, H. Iqbaly, E. Benussi, F. Volpiny, F. A. Dreyer, A. Shtedritskiy, Y. M. Asano, "Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models," in *Advances in Neural Information Processing Systems*, 2021, pp. 2611–2624. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/file/1531beb762df4029513ebf9295e0d34f-Paper.pdf>.
- [2] V. Hudson, Perspective: ChatGPT and the dawn of the new Dark Ages. [Online]. Available: <https://www.deseret.com/2023/1/23/23562681/chatgpt-artificial-intelligence-critical-thinking-dark-ages> (accessed: Feb. 10 2023).
- [3] A. Borji, "A Categorical Archive of ChatGPT Failures," Feb. 2023. [Online]. Available: <http://arxiv.org/pdf/2302.03494v2>.
- [4] X. Shen, Z. Chen, M. Backes, and Y. Zhang, "In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT," Apr. 2023. [Online]. Available: <https://arxiv.org/pdf/2304.08979>

- [5] E. Sulmont, E. Patitsas, and J. R. Cooperstock, "Can You Teach Me To Machine Learn?," in Proceedings of the 50th ACM Technical Symposium on Computer Science Education, Minneapolis MN USA, 2019, pp. 948–954.
- [6] E. M. Bender and A. Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5185–5198, 2020, doi: 10.18653/v1/2020.acl-main.463.
- [7] P. C. Su-Yeon, K. Haejoong, L. Sangmin, "Do Less Teaching, Do More Coaching: Toward Critical Thinking for Ethical Applications of Artificial Intelligence," Journal of Learning and Teaching in Digital Age, vol. 6, nr. 2, pp. 97–100, 2021.
- [8] A. Shoufan, "Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey," IEEE Access, vol. 11, pp. 38805–38818, 2023, doi: 10.1109/ACCESS.2023.3268224.
- [9] A. Tlili et al., "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," Smart Learn. Environ., vol. 10, no. 1, 2023, doi: 10.1186/s40561-023-00237-x.
- [10] X. Zhai, "ChatGPT User Experience: Implications for Education," SSRN Journal, 2022, doi: 10.2139/ssrn.4312418.
- [11] J. von Garrel, J. Mayer, and M. Mühlfeld, "Künstliche Intelligenz im Studium Eine quantitative Befragung von Studierenden zur Nutzung von ChatGPT & Co," Gesellschaftswissenschaften, 2023.
- [12] Uni Duisburg-Essen and Civey, Kurzstudie Uni Duisburg-Essen & Civey. [Online]. Available: <https://app.civey.com/dashboards/kurzstudie-uni-duisburg-essen-civey-14024> (accessed: Jul. 25 2023).
- [13] K. Bhattacharya, Lernen mit KI. Einsatz von ChatGPT & Co. beim Lernen. IU Studie. [Online]. Available: <https://www.iu.de/forschung/studien/kurzstudie-lernen-mit-ki/> (accessed: Jul. 25 2023).
- [14] A. Choudhury and H. Shamszare, "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis," Journal of medical Internet research, vol. 25, e47184, 2023, doi: 10.2196/47184.
- [15] R.E. Landrum and M. A. McCarthy, "Measuring Critical Thinking Skills," in "A compendium of scales for use in the scholarship of teaching and learning", pp. 74–86, 2015. [Online]. Available: <https://psycnet.apa.org/record/2018-41234-000>.
- [16] W. Sumarni, K. I. Supardi, and N. Widiarti, "Development of assessment instruments to measure critical thinking skills," IOP Conf. Ser.: Mater. Sci. Eng., vol. 349, p. 12066, 2018, doi: 10.1088/1757-899X/349/1/012066.
- [17] S. Cargas, S. Williams, and M. Rosenberg, "An approach to teaching critical thinking across disciplines using performance tasks with a common rubric," Thinking Skills and Creativity, vol. 26, pp. 24–37, 2017, doi: 10.1016/j.tsc.2017.05.005.
- [18] C. F. John, Storytelling and market research: A practical user guide. New York, NY: Routledge, 2022. [Online]. Available: <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2937908>.
- [19] R. H. Ennis, "Critical Thinking and Subject Specificity: Clarification and Needed Research," Educational Researcher, vol. 18, no. 3, pp. 4–10, 1989, doi: 10.3102/0013189X018003004.
- [20] D. Gelerstein, R. d. Río, M. Nussbaum, P. Chiuminatto, and X. López, "Designing and implementing a test for measuring critical thinking in primary school," Thinking Skills and Creativity, vol. 20, pp. 40–49, 2016, doi: 10.1016/j.tsc.2016.02.002.
- [21] P. van den Brook, P. Kendeou, S. Lousberg, and G. Visser, "Preparing for reading comprehension: Fostering text comprehension skills in preschool and early elementary school children," IEJEE, vol. 4, no. 1, pp. 259–268, 2011.
- [22] E. J.N. Stuppel, F. A. Maratos, J. Elander, T. E. Hunt, K. Y.F. Cheung, and A. V. Aubeeluck, "Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking," Thinking Skills and Creativity, vol. 23, pp. 91–100, 2017, doi: 10.1016/j.tsc.2016.11.007.
- [23] M. C. Laupichler, A. Aster, and T. Raupach, "Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy," Computers and Education: Artificial Intelligence, vol. 4, p. 100126, 2023, doi: 10.1016/j.caeai.2023.100126.
- [24] J.-F. Bonnefon, "The Pros and Cons of Identifying Critical Thinking with System 2 Processing," Topoi, vol. 37, no. 1, pp. 113–119, 2018, doi: 10.1007/s11245-016-9375-2.
- [25] L. O'Hare and C. McGuinness, "Measuring Critical Thinking, Intelligence, and Academic Performance in Psychology Undergraduates," The Irish Journal of Psychology, vol. 30, 3–4, pp. 123–131, 2009, doi: 10.1080/03033910.2009.10446304.
- [26] D. Kurniati, Purwanto, A. R. As'ari, Dwiyan, Subanji, and H. Susanto, "Development and Validity of Problems with Contradictory Information and no Specified Universal Set to Measure the Truth-Seeking of Pre-Service Mathematics Teachers," 2019.
- [27] E. R. Wulan and N. F. Ilmiyah, "Prospective Mathematics Teachers' Critical Thinking Processes in Dealing Truth-Seeking Problem with Contradictory Information," 2nd National Conference on Mathematics Education 2021 (NaCoME 2021), pp. 90–100, 2022, doi: 10.2991/assehr.k.220403.013.
- [28] J. A. Moon, Critical thinking: An exploration of theory and practice. London: Routledge, 2008.
- [29] W. Condon and D. Kelly-Riley, "Assessing and teaching what we value: The relationship between college-level writing and critical thinking abilities," Assessing Writing, vol. 9, no. 1, pp. 56–75, 2004, doi: 10.1016/j.asw.2004.01.003.
- [30] Kahneman, D. (2011), Thinking, Fast and Slow, Penguin Books, London.
- [31] S. Bailin, R. Case, J. R. Coombs, and L. B. Daniels, "Common misconceptions of critical thinking," Journal of Curriculum Studies, vol. 31, no. 3, pp. 269–283, 1999, doi: 10.1080/002202799183124.
- [32] M.-J. Kolly and M. Flügel, Chatbots wie GPT können wunderbare Sätze bilden. Genau das macht sie zum Problem. [Online]. Available: <https://www.republik.ch/2023/04/11/chatbots-wie-gpt-koennen-wunderbare-saetze-bilden-genau-da> (accessed: Jun. 7 2023).
- [33] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, 2023, doi: 10.1016/j.iotcps.2023.04.003.
- [34] S. Cottrell, Critical Thinking Skills: Effective Analysis, Argument and Reflection, 4th ed. London: Bloomsbury Publishing Plc, 2023.

Predictive Maintenance for Distributed Systems Using Data Science

Ferdinand Koenig

Baden-Wuerttemberg Cooperative State University (DHBW) Stuttgart
Hewlett Packard Enterprise
Boeblingen, Germany
0000-0003-3311-7355

Karlo Kraljic

Technical University of Munich
Hewlett Packard Enterprise
Muenchen, Germany
0000-0002-8832-3132

Abstract—Preventing failures in largely-scaled and distributed data centers’ system stacks in real-time is impossible for humans. This paper addresses this issue by creating a system that is able to detect anomalies. It collects data from the monitoring application Nagios, creates groups of hosts (cliques) with similar behavior based on the correlation of time series data, and prepares data per clique by imputing *NaN* values and scaling. An LSTM autoencoder is created with TensorFlow and a 9:1 train-validate split for each clique. The solution is able to detect all artificially generated anomalies regarding CPU load, memory occupation, and network traffic. It is hence able to perform condition-based monitoring. Using this framework, we can reduce the operations team’s workload significantly.

Index Terms—predictive maintenance, automation, data center, data science, machine learning, autoencoders

I. INTRODUCTION

In largely-scaled data centers, there is a huge number of devices, applications, and events that are most times monitored manually. However, this process is expensive and error-prone which might lead to unexpected, expensive down times.

The objective is to reduce the effort and costs of manually crawling through the log files and avoid the penalties of breaching service-level agreements. To achieve this goal, we implement a system that performs condition-based monitoring to enable predictive maintenance employing data science. To accomplish predictive maintenance, normal behavior will be learned and deviant behavior, i.e., anomalies, will be detected. The anomalies can be examined and remedied before symptoms arise.

The structure of this article is as follows. First, the literature is researched. Then, the method is described and an overview of the solution is given. Data from numerous system components is collected and aggregated. Groups of hosts with similar behavior (cliques) are created. We build a pipeline to prepare the data and, in the subsequent chapter, develop a machine-learning model. Finally, the model will be evaluated. The final chapter summarizes the article, outlines future work, and gives a conclusion.

II. RESEARCH BACKGROUND

In this section, papers are presented that apply predictive maintenance not limited to data centers. Log-based solutions are of particular interest. In general, three types of solutions

exist: classical machine learning, supervised deep learning, and unsupervised reconstruction models.

Relevant research is discussed in the order of those categories.

A. Classical Machine Learning

Su and Huang [1] built a predictive maintenance system for hard disk drives in data centers via supervised learning using random forest classifiers.

The following papers use one-class classifiers where the class represents normal behavior. Points, that are not assigned to it, are anomalies. Wang et al. [2] create an anomaly detection in data centers based on the estimation of a Gaussian density distribution using sensor data, e.g., from the Central Processing Unit (CPU), the Graphics Processing Unit (GPU), and memory. The authors of [3] apply a Support Vector Machine (SVM) to predict if an engine will fail. They used accuracy as a metric while the dataset was imbalanced with a ratio of positives to negatives of 996 : 12,100. A hypothetical classifier that labels all samples negative would have an accuracy of 92.4%. Their result was 95%. Pereira et al. [4] implemented a One-Class Support Vector Machine (OCSVM) – here, a Support Vector Data Description (SVDD) [5] –, a Gaussian Mixture Model [6, pp. 430ff.], and reconstruction methods for fault detection of hard disk drives. For them, reconstruction models gave the best results. The authors of [7] use a classical OCSVM [8] to detect anomalies in time series data. [9] from Wang et al. is a log-based approach to predict failures of automated teller machines and compared eXtreme Gradient Boosting (XGBoost) [10], random forest [11], Adaptive Boosting (AdaBoost) M1 [12], and an SVM. Their recall is below 50%. Optimizing for the recall is essential for anomaly detection in imbalanced datasets. Liu et al. [13] created independent fault detection models for each type of fault by using a method based on K-Nearest Neighbors (KNN) and minimal spanning trees. This solution might not be always applicable because domain knowledge and labeling of the fault types are required. Additionally, unknown faults and their sources will probably stay undetected. Sipos et al. [14] demonstrated a log-based prediction of equipment failures. They created a model with Multi-Instance Learning (MIL). In MIL, multiple instances from an interval, here, one week,

are grouped in bags. An L1-regularized SVM optimization problem was defined and solved by Liblinear [15]. For the model to be explainable and can provide insights, L1 was chosen. Their models use stratified subsampling and take 300 to 400 features. A PR-AUC of 0.73 for a dataset with 11,238 features was reached, where 0.14% of the samples were known failures. This methodology poses the challenge that there is not necessarily labeled data of occurring faults. The authors of [16] feed log data of woodworking machines labeled by service reports to calculate their Remaining Useful Life (RUL). Gradient Boosting Machine (GBM), XGBoost, and Distributed Random Forest (DRF) were compared to other ML techniques such as SVM. They favored tree-based models due to the prediction depending on a set of rules and therefore their interpretability. The results are 98.2% accuracy, 98.6% recall, and 98.3% precision. The problems with this solution are that service reports often do not have an accurate timestamp and they used a prediction period of 24 hours which might be too large for the prediction of some types of faults, e.g., unexpected reboots or out-of-memory exceptions. Giommi et al. [17] consider two time frames of log data: a four-day-long period of normal behavior and a subsequent problematic interval of equal length. The detected anomaly was reported by a service ticket on the first day of the problematic time frame. The issue was a high rate of transfer and deletion errors. Among others, the scikit-learn implementation [18] of AdaBoost [19] is applied to predict the anomalies. The results are not validated by unseen data and the problem statement was narrow as just one instance of an anomaly was considered. Decker et al. [20] consider log messages as an event stream and derive five features from the log activity rate. The rate is a normally distributed random variable $u \in \mathbb{N}$. This poses the challenge of not being able to capture all anomalies and it is not justified to interpret the log activity rate to be normally distributed. The solution's quality was measured with accuracy using a recursive formula. Again, the measure is not suitable for imbalanced datasets. The authors of [21] use the Scikit implementation of OCSVM [22]. As with the previous two papers, it derives the volatility from the log activity. Fluctuating volatility could occur frequently, e.g., less traffic on weekends that might occur in industry Information Technology (IT) environments. This is not captured.

B. Supervised Deep Learning

The authors of [23] use models based on Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) for the predictive maintenance of turbofan engines. The team behind [24] utilizes a mix of LSTM and CNN to estimate the RUL of aircraft engines. Both projects take sensor data as input.

Canizo et al. [25] use one CNN per sensor. Therefore, removing or adding sensors does not require retraining the remaining CNNs. The article on hand is inspired by this sensor-wise approach and creates models on a per-data-center basis and for groups of similar hosts.

C. Unsupervised Reconstruction Methods

The idea of unsupervised reconstruction models in predictive maintenance is that the model is trained on data showing mostly correct behavior. That means it learns to correctly reconstruct the training data. The model is then not able to correctly reconstruct the data of outliers, i.e., anomalies. Those are candidates for predictive maintenance cases. The following papers use this class of models.

GANomaly was introduced by Akcay et al. [26]. It is a Generative Adversarial Network (GAN)-based *encode-decode* model with convolutional layers. Ahn et al. [27] compared *GANomaly* to variational autoencoders for solving an anomaly detection problem for spacecraft control systems with simulated time-series data and found that *GANomaly* performed better.

Bouabdallaoui et al. [28] use two LSTM layers per encoder and decoder. They use the Root Mean Squared Error (RMSE) as the reconstruction error and call it the *anomaly score*. If it is above a threshold, an alert is raised. The threshold was chosen to limit the number of notifications below an acceptable number in accordance with the user. Limiting the number of notifications that way is a good start. To improve the model, the sensitivity of the model can be adjusted such that the number of alerts matches the number of (future) failures.

Breux et al. [29] adapted the multi-sensor approach of [25] to create a multi-autoencoder approach, meaning they are using one LSTM-autoencoder per family of sensors. The data is min-max-scaled to $[0, 1]$. The model is unsupervised but the available data is labeled. They used the data labeled as 'normal' for training. The latent vector is set to be half of the length of the input window size. The encoder was built using three layers (64, 32, and $\lfloor \frac{1}{2} \text{ of the window size} \rfloor$ units). The decoder is composed of two LSTM layers (32 and 64 units) followed by a Time Distributed Dense layer with a linear activation function. The window size is a hyperparameter and was selected by trying discrete values and choosing the one that optimizes the F1 score. As a result, they use a window size of 6, which is the lowest value of their tests. Between the LSTM layers, a dropout layer with a dropout parameter of 0.2 is used to prevent overfitting. This time, the autoencoder does not output a scalar anomaly score but some features of the squared difference sequence of the original vector and reconstructed vector. These feature vectors are used to train a Random Forest. The classifier is operated in "balanced_subsample" mode where weights are used to balance the data set. The choice of the model allows them to set the decision threshold of whether a state is normal or abnormal by minimizing the entropy of the predictions. For the training, sensor data was collected in 5-minute batches. The sensors, i.e., data sources, were selected such that if a second sensor has a high correlation with a given first sensor, it was neglected. The data is min-max normalized and multiple batches are merged to fit a parameter that determines the time windows of interest. The merged batches, i.e., sequences, are labeled by the use of alarm history. If an alarm occurred during the covered time,

the sequence is labeled as an anomaly. The model scored an F1 score of 83% and 65% for a 30 respectively 60-minute time window and outperformed in this test, i.e., a model based on [4], [28], [30]. The hyperparameters window size, number of trees of the random forest, and the threshold on the probability of the normal class were optimized. Other sensors that are highly correlating are neglected. It is possible to use them as additional training data from which ANNs tend to benefit by better learning the underlying distribution and avoiding overfitting [31].

III. METHOD AND OVERVIEW

This section will first define a framework adapting the Data Science Trajectories model (DST) [32]. The chosen framework reveals the methodology. Then, an overview of the solution is given.

A. Adaption of DST in this project

DST is chosen as the structure as it meets all requirements.

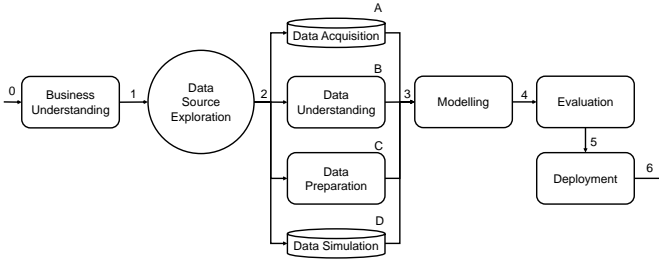


Fig. 1: Data Science Trajectories model (DST) structure of this project

The focus of this article is on the technical part, hence, *Business Understanding* (0) and business needs are only briefly examined in section I. The other steps are considered in the following sections. The start is the *Data Source Exploration* (1) to get an overview of the possible sources of data. The second action will be to obtain the data. This is *Data Acquisition* (2A). It is strongly connected to *Data Understanding* (2B) – which includes the familiarization with the data and discovery of first insights – and *Data Preparation* (2C) to transform and cleanse the data because an understanding of the data is required to comprehend which data must be collected. To imply the strong interconnectivity of these three parts, they are modeled as parallel tasks. Still, the *Data Acquisition* must be already started when beginning with *Data Understanding*. This underlying order is indicated by the letters. To add some anomalies to the evaluation, there exists a block of *Data Simulation* (2D). After creating models (*Modelling* (3)) and possibly going back to the previous steps, an *Evaluation* (4) is conducted. Finally, the model is deployed (*Deployment* (5)).

B. Architectural Overview

Figure 2 provides an overview of the solution.

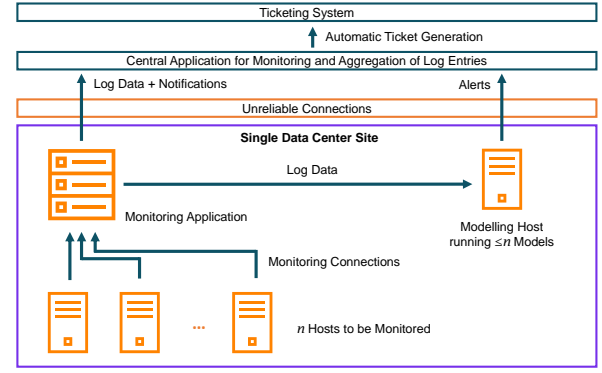


Fig. 2: Physical overview

The overall architecture consists of: Hosts to be monitored, Monitoring application, Modelling Host, Central Application for Monitoring, and Ticketing System.

Figure 2 is a simplified representation that focuses on the parts that are important for the operation of the ML models. The general data flow is bottom-up. In the existing system, n hosts are monitored by the monitoring application. The central application for monitoring fetches log data from the monitoring application. Per a ruleset, the monitoring application creates notifications and sends them to the central application that writes the notifications to a database and creates tickets in the ticketing system. The connection to and from a single data center site might be unreliable. To not be reliant on this connection, the solution will be part of each single data center site. Therefore, the prediction still works even if the data center is cut-off. For that, a modeling host is introduced that can be part of the monitored hosts. Figure 3 shows the steps that are performed by the modeling host.

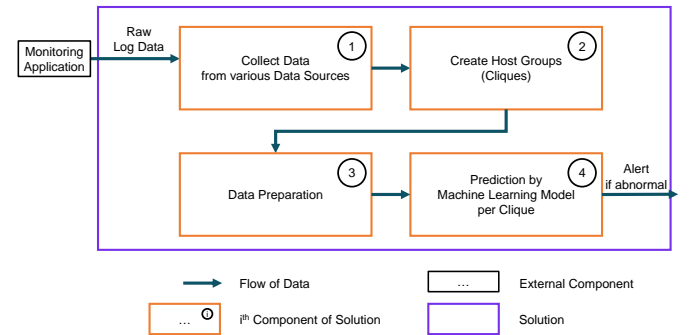


Fig. 3: Logical overview of the modelling host

Generally, raw time series log data is collected from the monitoring application since the data is not present. Each site has its own monitoring instance. Then, hosts are grouped to build specific models for hosts with similar behavior. These groups of hosts will be called *cliques*. The data is processed and prepared to be fit and predicted by a machine learning model. If it detects abnormal data, an alert is raised.

IV. DATA COLLECTION

The purpose of the first component in the solution is data collection. This section includes its design, the details of implementation, and the data understanding, i.e., a discussion of the data.

A. Design

At first, the component for data collection is planned. The intention is to get the available time series log data from each host. For that, the potential data sources are examined. Subsequently, the algorithm for fetching the data is introduced.

1) *Data Source Exploration*: The monitoring solution provides the following data: performance and capacity planning data. Both data types will be discussed in detail below. Further potential sources include incident tickets and transition acceptance tests. They are not used due to redundancy with the elaborated sources and inaccuracies. Therefore, a deep dive is out of the scope of this article.

a) *Performance Data*: The raw time series log data is available via a performance data endpoint in the monitoring application. The log data includes but is not limited to memory, disk, swap usage, ping times, CPU usage, and network bandwidth. It is logged via a circular buffer approach. Fetching this data is possible via a separate database that copies data or with Application Programming Interface (API) calls on the REpresentational State Transfer (REST)ful interface directly to the monitoring instance. The database is populated by a script that is manually triggered. In addition, there is a mask to define, which hosts or services should be included. Hence, historical data in the database might not be complete and the solution should rely on API calls. The performance data is used to build the solution.

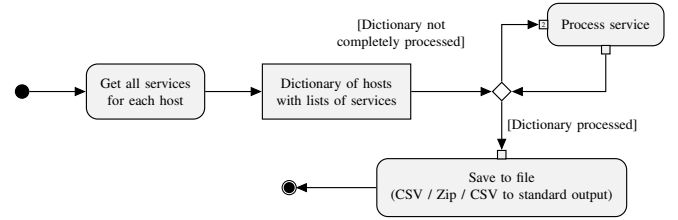
b) *Capacity Planning Data*: The capacity planning data is only accessible through the API. It is intended to estimate the future load of a system based on historical performance data. The response contains statistical values, defined levels of when the values are on a {fit, warning, critical} level, historical data of the service – which covers twice of the specified period –, and the predicted data. This endpoint will not be used since the data is redundant to the performance data. The statistical values can be easily derived from the data and the underlying prediction model could be applied by another program.

2) *Component to Fetch Data*: This paragraph introduces the algorithm of the component that collects data. It is capable to get data from various sources. Eventually, performance data, i.e., raw time series log data from the monitoring application per data center, is used.

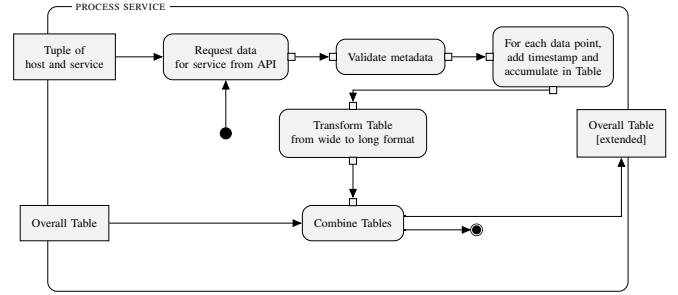
The monitoring application provides the data in so-called *measurands*, which are scalars. If it provides historic values, it is a time series. The measurands are the *features* of the machine learning task. Multiple measurands are organized in *services*.

The overall algorithm is shown in Figure 4a.

First, the application gets a list of services and measurands per host to iterate over them. Services that do not provide



(a) Overall algorithm



(b) Action "Process service"

Fig. 4: Activity diagram of fetching component

valid time series are skipped. These services are intended to just give a binary status, e.g., whether the service is running.

Figure 4b shows the procedure that is executed per service. Firstly, the time series data is requested for this service over the API. The response is validated. It contains all time-series values of all measurands of this service. For each time series value, the time is derived from the metadata. The time series are accumulated in a table. There is one column that contains the time and one column per measurand. This is called a wide format. The table is melted into a long format to get a unified format. With that, the number of columns does not change with the number of measurands. Tables in the long format have columns for: monitoring instance, host, service, timestamp (ts), measurand, and value.

The table in the long format of each service is concatenated to the overall table which is then eventually saved to a file. There is an option to write the CSV file to the standard output. Therefore, the program can be used in a pipeline. Data during the development phase, where data from every host is moved, is compressed. This helps with a cost-efficient transfer of data that is highly redundant as, e.g., the hostname, is repeated for every value of a host. The output table of this application will be called raw data.

B. Implementation

A Python application with Command-Line Interface (CLI) and argument parser is used to fetch data from the Nagios¹ monitoring application via its RESTful API. The process involves requesting and processing data from Nagios instances, with the option to add more instances. The application retrieves a list of services per host, loads the data into a Pandas²

¹<https://www.nagios.org/>

²<https://pypi.org/project/pandas/>

DataFrame with timestamps, and converts the DataFrame to a unified format. The resulting DataFrame can be saved to a file or written to the standard output for pipeline usage.

C. Results and Evaluation

In this paragraph, the two data sets which are the output of the data collection are described for Data Understanding. For the description and later for the development, only the data from one monitoring instance, i.e., one Nagios system, is kept. Recall that the measurands from the services are the features. The five-minute-resolved, eleven-day dataset that was available in the course of the project has 260 features. When removing features that do not contain values, i.e., columns with all values being *NaN*, the dataset has then 258 features.

The raw data (in the long format) has 2.8% missing values (*NaN*). Let TH be the feature matrix. After removing TH 's features with no values, 8.4% of vectors have only *NaN* as values. Removing them results in TH covering 111,278 of 121,536 vectors.

As described before, a service consists of one or more measurands. In total, there are 134 services eleven-day data set. The services include but are not limited to:

- bandwidth of various network interfaces and ping
- CPU usage (hosts and virtualization layer)
- load, i.e., utilization, of the whole data center
- disk and datastore usage
- virtual machine Input / Output
- memory usage (hosts and virtualization layer)
- Uninterrupted Power Supply (UPS) related metrics

V. CLIQUE CREATION

To balance the specificity of the model with its abstraction, and therefore keeping the number of models to a minimum, hosts are grouped by similar behavior. These groups will be called ‘cliques’. The origin of the name is covered by the design section which introduces the method and discusses its use. Furthermore, it shows the implementation of the algorithm. Then, the results of the clique creation of both datasets are examined.

A. Design of the clique-creation mechanism

The solution leverages the creation of groups of hosts. The property of a group is that every member is similar to every other member. The similarity is measured with the correlation of hosts. To measure the correlation between hosts, a host feature matrix H is introduced. Each feature vector corresponds to one host. As defined in subsection IV-A2, a unique or specific measurand is called a feature. The columns of H are then a feature at a point in time. Take Equation 1 as an example of a host feature matrix. There are two hosts $h\{0,1\}$ with two features and three points in time. Hence, H has eight columns. Column $f1, t0$ describes some feature $f1$

at time $t0$. The pairwise correlation of the hosts $h\{0,1\}$ is then 0.85.

$$H = \begin{matrix} & \begin{matrix} f1, t0 & f1, t1 & f1, t2 & f2, t0 & f2, t1 & f2, t2 \end{matrix} \\ \begin{matrix} h0 \\ h1 \end{matrix} & \begin{bmatrix} 10.0 & 12.0 & 15.0 & 5.0 & 11.0 & 12.0 \\ 1.0 & 1.3 & 1.7 & 1.0 & 1.3 & 1.5 \end{bmatrix} \end{matrix} \quad (1)$$

The threshold γ , where two hosts (x, y) are seen as similar and have a high correlation $\rho_{xy} > \gamma$, is arbitrarily set to $\gamma = 0.7$. With this value, we cover the hosts with a high correlation, which is bounded by 1, without being too strict. A graph is created where the nodes are the hosts and an edge between them indicates similarity. The complete subgraphs are then the groups of hosts that are similar by correlation. Since the algorithm to find these, solves the clique problem [33], the host groups are called *cliques* from now on.

A model per clique is created to follow a modular approach. This results in a balance between specific models, as a solution with one model per host has, and one overall model per data center site which might not be able to learn the all present behaviors. Using this technique, the idea is that when new hosts are added and they can be assigned to an existing clique, the model of this clique can be applied to the host without the requirement of training. If hosts cannot be assigned to a clique, data could be collected over a period of time, e.g., seven days. This might be a good trade-off of maximizing the amount of training data to cover different phases of normal behavior, e.g., less traffic on weekends, and minimizing the initial time with no prediction model. This training data must be normal behavior. Else, an abnormal behavior would be learned as normal.

The time complexity is exponentially dependent on the number of nodes in the graph. The number of hosts is expected to not get exorbitant wherefore this algorithm is suited.

B. Implementation

In this section, the implementation and its details are presented.

After loading the data, it is present in the long, raw data format. Pandas DataFrames are used as a data structure. For correlation and graph generation, Pandas functions are used. To manage the graph, the package networkx³ is used. It constructs the graph and comes with the function `find_cliques` which is able to find the cliques.

C. Results and Evaluation

For the evaluation, the property ‘compact’ is introduced for cliques.

There are two kinds of cliques, which we will name compact cliques and non-compact cliques. Groups whose hosts only appear in this very group and not in a second clique will be called *compact cliques* henceforth. Consequently, groups that share members with other cliques are *non-compact cliques*.

For the case on hand, the algorithm groups similar hosts into 18 cliques. For instance, one clique combines hosts that run virtualization software, e.g., bare-metal hypervisors.

³<https://pypi.org/project/networkx/>

Let TH be a feature matrix where, as denoted before, the features are the measurands and a feature vector is essentially the value of a measurand for a point in time (for a host). It will be formally defined in the next chapter. When focusing on one clique c , some features might be dropped since they only yield NaN values. The size of the new feature matrix TH_c of clique c has usually fewer features.

VI. DATA PREPARATION

The task of the third component is to prepare the data for the use of Machine Learning (ML) models. At first, the pipeline is drafted and the feature matrix TH is introduced. Then, the implementation in Python is examined. The output and the results are summarized.

A. Design

In this section, the data preparation pipeline is introduced which is illustrated in Figure 5.

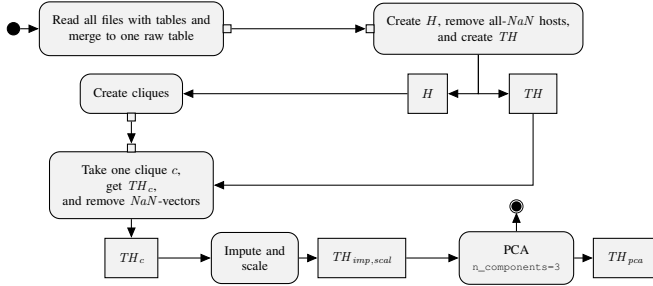


Fig. 5: Activity diagram of Machine Learning pipeline

In the narrow sense, the clique generation is in the logical overview (Figure 3) not part of the preparation. For an overall view of the data flow during the processing of new data, it is still included. Additionally, it provides context for this step.

The pipeline starts by reading all files with raw data that is collected by the first component. The raw data tables are iteratively merged such that a new junk of data will update the old data. The mechanism is simple: The data junks overlap since each file covers two days but is created every day. For the latest values of a file, they can be NaN because the log entry was already created by the monitoring application but not yet filled with an actual value. This value will be covered in the next table with data. The output of the first activity is therefore the merged raw performance data table. Then, cliques are created as already described in the previous section. As an additional step of data preparation, all hosts with no values, i.e., only NaN as values, are removed from H .

The current feature matrix H has observations, i.e., rows, representing hosts. The time dimension is part of the features. But the prediction model should learn and predict hosts at a given time to capture their characteristics w.r.t. time and therefore the current status. Hence, the matrix is transformed such that an observation or row corresponds to the characteristics of hosts at a specific time. The resulting matrix will be called TH as it is the representation of $(time, host)$ feature vectors, and therefore, time is not part of the features anymore. The

information content is equal to H . Equation 2 shows the TH feature matrix of the previously used example for an H matrix in Equation 1. It describes two features for two hosts at three different points in time.

$$TH = \begin{matrix} & \begin{matrix} f1 & f2 \end{matrix} \\ \begin{matrix} t0, h0 \\ t1, h0 \\ t2, h0 \\ t0, h1 \\ t1, h1 \\ t2, h1 \end{matrix} & \begin{bmatrix} 10.0 & 5.0 \\ 12.0 & 11.0 \\ 15.0 & 12.0 \\ 1.0 & 1.0 \\ 1.3 & 1.3 \\ 1.7 & 1.5 \end{bmatrix} \end{matrix} \quad (2)$$

In Figure 5, the process regarding only one clique c is shown, but the following steps can be repeated for every clique. TH_c is derived from TH by only selecting and keeping the hosts that are a member of c . Missing values (NaN) are imputed by replacing them with 0. The features are then *min-max scaled* to generate $TH_{imp,scale}$. Finally, PCA is applied and three components are kept. All data objects are saved and made accessible for the development of the models.

B. Implementation

The solution reads and processes compressed Comma-Separated Values (CSV) files into Pandas DataFrames, combining older and newer, updated data. It creates an overall DataFrame with raw time series log data. The data is transformed into a TH feature matrix using Pandas' `pivot`. NaN values are imputed and the DataFrame is scaled using scikit-learn's `SimpleImputer` and `MinMaxScaler`. Principal Component Analysis (PCA) is applied, and the resulting imputed and scaled data ($TH_{imp,scale}$) is used for the LSTM autoencoder.

C. Results

The input data is numerical. Because of this, no transformation from other formats to numerical data is required. The pipeline deals with NaN values by removing all- NaN hosts, as well as all- NaN features and feature vectors regarding one clique. Other NaN values are replaced by 0. The data is scaled.

For the implementation of one of the models, clique 9 is used. With 86,1%, the three first Principal Components (PCs) of the PCA cover the most of the variance (48.5%, 29.3%, 8.3% respectively for $PC\{1,2,3\}$).

VII. MACHINE LEARNING MODEL

The core component is the machine learning model. An LSTM autoencoder is considered.

First, the prerequisites and appropriate options for solving the unsupervised learning problem on hand are presented. The design, the implementation, and the results are shown. The latter includes the learning curve of the autoencoder.

A. Design

The topic of this section is the design of models to detect anomalies in the time series data TH or its processed forms.

There is the requirement of a health metric, that gives each TH feature vector, i.e., each host at each point in time, a measure of how normal the behavior is. It is used to define

a threshold from when behavior is abnormal. In addition, it enables ordering and prioritizing anomalies.

The three solution types were covered in section II. This project only provides unlabeled data. Therefore, the third type (unsupervised reconstruction methods) is suitable and is chosen. LSTMs can learn the temporal dependency of time series data and are therefore suited to learn the data on hand that is sequential, too. When creating an encoder and a decoder neural network consisting of LSTMs, this is called an LSTM autoencoder.

An LSTM autoencoder works under the assumption that normal data in the training data largely outnumber the anomalies. The idea is that the model is good at reconstructing normal data but not as good for abnormal data. Hence, the reconstruction error for anomalies is higher. A threshold for the reconstruction error will be set to label feature vectors as an anomaly.

The LSTM autoencoder of this project should assign an interpretable *health score* to each new feature vector of a host h at time t in the feature matrix TH_{pred} for prediction, to measure the normality of an observation. Consider a feature vector $v \in \mathbb{R}^p$ where p is the number of features. To construct such a health score, all feature vectors of TH are first min-max-scaled ($v \in [0, 1]^n$). Then, the RMSE between v and reconstructed v' is calculated. $v' \in [0, 1]^n$ if an activation function is chosen with an image of $[0, 1]$ or a subset of it (like *sigmoid* with an image of $(0, 1)$). In that case, the RMSE's image is also $[0, 1]$. RMSE is easy to interpret since it preserves the dimensional space. The health score is then $1 - RMSE$ and represents a value to measure the health between 0 and 100%, where 100% means the host is completely working as expected. This is the reason why *sigmoid* is used in the final layer.

The batching creates $[x_{t-(timesteps-1)}, x_{t-(timesteps-2)}, \dots, x_{t-1}, x_t]$ sequences per host x . Therefore, the LSTM is able to capture the temporal dependencies on a per-host basis.

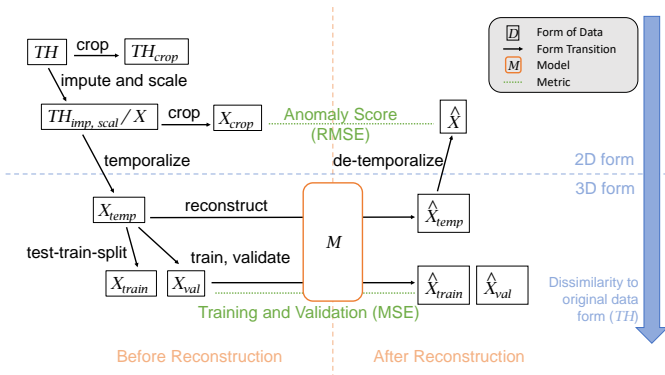


Fig. 6: Data flow of LSTM autoencoder

In Figure 6, the black rectangles are different forms of the data which processing is depicted by arrows. The deeper a data object is, the higher its distance to the original data. If two data objects are on the same level, they represent a similar degree of processing. Data above the horizontal

dashed line is two-dimensional since it is in the form of *samples* \times *features*. TH and $TH_{imp,scal}$ have been discussed before. $TH_{imp,scal}$ will have the alias X from this point on. Now the data is batched or *temporalized*. Recall, that the LSTM expects input in the shape of (*batch* \times *timesteps* \times *feature*). In consequence, the data will be grouped in *batch* batches with temporal sequences of *timesteps* steps: $[x_{t-(timesteps-1)}, x_{t-(timesteps-2)}, \dots, x_{t-1}, x_t]$. The result is X_{temp} . Hence, data below the horizontal dashed line is three-dimensional. As a next step, the data is split in an arbitrarily chosen ratio of 9 : 1 into training data X_{train} and validation data X_{val} to maximize observations for training but remaining a sufficient amount of validation samples. The training data is put into model M . X_{val} is utilized to validate its reconstruction ability. For an error metric, the MSE is used since minimizing the MSE optimizes the RMSE. After the training, the entirety of the data is used to investigate for anomalies. For production, M is saved and can classify new vectors in their temporal context.

When reconstructing X_{temp} , the model will return \hat{X}_{temp} . The data is then de-batched or *de-temporalized*. Let $[\hat{x}_{t-(timesteps-1)}, \hat{x}_{t-(timesteps-2)}, \dots, \hat{x}_{t-1}, \hat{x}_t]$ be a time sequence in the reconstructed set of batches \hat{X}_{temp} . When de-temporalizing that time sequence, all items of the sequence that are there for the temporal context ($[\hat{x}_{t-(timesteps-1)}, \hat{x}_{t-(timesteps-2)}, \dots, \hat{x}_{t-1}]$) are removed, thus, \hat{x}_t remains. Repeating this for every time sequence in \hat{X}_{temp} yields \hat{X} which is de facto the reconstruction of X . Due to the de-temporalization, \hat{X} might have a couple fewer (*timestamp*, *host*) vectors. This is due to (the temporally) first *timesteps* - 1 vectors of a host having no *timesteps* - 1 predecessors to create a batch with the required temporal context, i.e., of length *timesteps*. Those first *timesteps* - 1 vectors are removed from TH and X . The outcome is TH_{crop} and X_{crop} . Since X_{crop} and \hat{X} have the same shape, the difference or anomaly score per host at time t can be calculated between them. For a good autoencoder and normal data, it holds $X_{crop} \approx \hat{X}$. The RMSE is used and the health score is added as presaged.

B. Implementation

This section reveals the implementation details.

The LSTM Autoencoder part is subdivided into two paragraphs. First, the construction and training are shown. Thereafter, the deployment and operation are described. It includes a prediction pipeline.

1) *Construction and Training*: To create an LSTM autoencoder, the general ML pipeline from Figure 5 is extended in Figure 6. The model is built using TensorFlow 2.9.1⁴ and scikit-learn 1.1.2⁵. For reproducibility, the seeds are arbitrarily set to 42. To be able to verify the model and run tests on the hosts, a clique must be chosen where anomalies can be simulated. Clique 4, the clique where the Nagios instance is running, is used.

⁴<https://pypi.org/project/tensorflow/2.9.1/>

⁵<https://pypi.org/project/scikit-learn/1.1.2/>

In the code, data below the line in Figure 6 is stored in Numpy arrays and data above Pandas DataFrames as data structures. As a next step, the data is split into 90% training data X_{train} and 10% validation data X_{val} using scikit-learn's `train_test_split`.

The architecture is inspired by Breux et al. [29] and is illustrated in Figure 7. The window size is arbitrarily set to 5. The idea is that the window size is big enough to learn the short-to-medium-term ‘normal’ behavior. Hence, each batch consists of the feature vector of interest plus four additional feature vectors as a temporal context.

The autoencoder will be symmetric with two layers each, for the encoder and decoder. The size of the layers, i.e., the number of units of the LSTM or the layer's output dimensionality, is doubled to be able to learn more complex dependencies. Therefore, the layers have 128 and 64 units. For the first LSTM layer, `return_sequences` is set to `True` which means that it returns a full sequence, i.e., the processed batch. If set to `False`, only the last output, i.e., the processed $x_{t,processed}$ is returned. The latent vector has the shape $1 \times units = 1 \times 64$. This is a bigger latent vector compared to [29]. Instead of learning the behavior of one sensor, the more complex behavior of a whole host should be captured by the latent vector, hence, the increase in size.

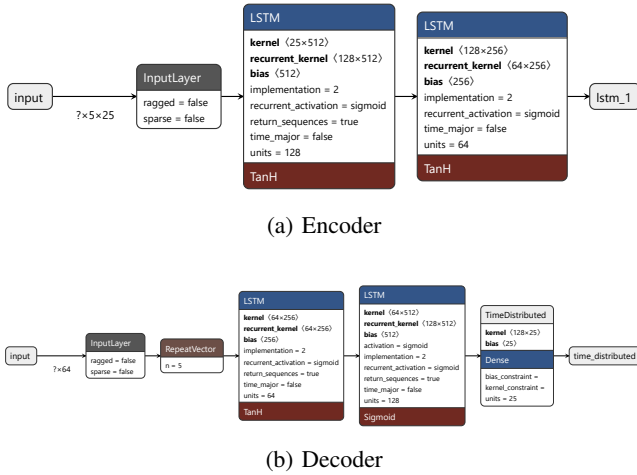


Fig. 7: LSTM autoencoder architecture

The decoder works mainly analogous to the encoder. All LSTM layers have the activation function *tanh* except for the last one that uses *sigmoid*. The model is compiled with the optimizer ‘Adam’ and is trained to minimize MSE which optimizes RMSE. The health score threshold is arbitrarily set to 70%.

2) *Operation*: Each site has its own models that can continue predicting even if the data center is disconnected. The number of models depends on the formed cliques and the number of hosts to be monitored. An alert is raised when an anomaly is detected, either addressed to a recipient at the site or through notifications by creating tickets. Transmitting only new data on-site and generating outgoing traffic only for

anomalies is cost-efficient. When predicting, only the required subset of features is used, and the data is imputed and scaled using a serialized scaler. If the health score falls below a threshold θ , an alert is triggered.

C. Results and Evaluation

This section demonstrates the results of the model.

The model is then trained with X_{train} with the shape $8540 \times 5 \times 35$ and validated with X_{val} .

It is evident that the model learns well. There is a steep drop at the beginning. The reconstruction capabilities of M for X_{val} keep improving until around $epoch = 200$. After that, the performance is stagnating. The model in this form with this amount of training data cannot improve more.

Now, the model is ready for detecting anomalies. For an analysis, X_{temp} is reconstructed and de-temporalized such that the Anomaly Score between X_{crop} and \hat{X} can be calculated.

VIII. RESULTS AND EVALUATION

For evaluating the prediction capabilities of the model, anomalies were created on one host. This belongs to Data Simulation in our framework, i.e., the DST model. Figure 8 shows the context of CPU and memory usage, as well as outgoing network traffic. The time frame with artificially created abnormal behavior is marked in black and shown in (b).

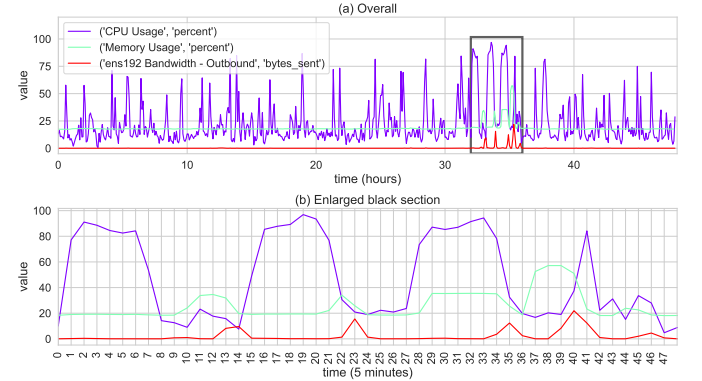


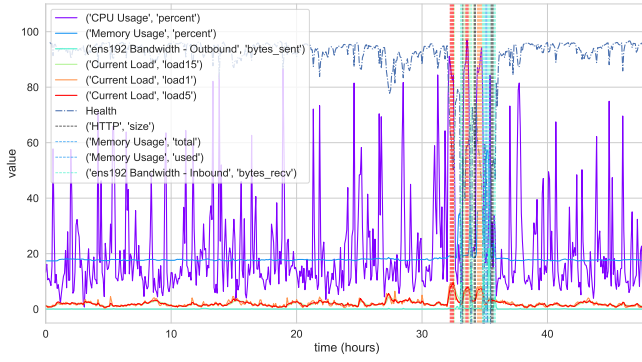
Fig. 8: Generated anomalies

The first produced behavior was a high CPU load of 80% with the FIRESTARTER benchmark⁶. The next anomaly was a high utilization of memory which wrote five gigabytes of zeros to memory and kept it for 1800 seconds. The anomaly during $37 \leq t \leq 40$ used twelve gigabytes whereas the small bump at $t = 44$ was two gigabytes. For the generation of network traffic, *trafgen*⁷ was run.

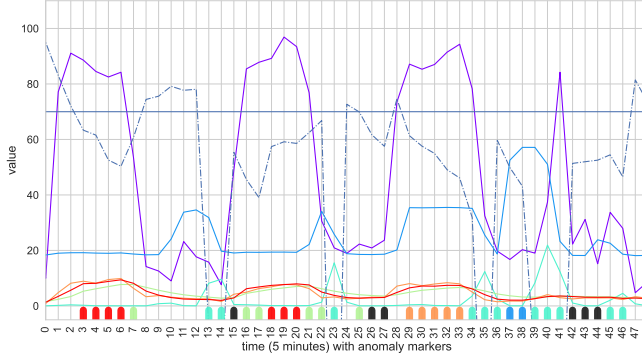
Figure 9a shows the same period as Figure 8 (a). The dash-dotted line provides the health score returned by the model. The predicted anomalies are illustrated by dashed vertical lines. The three tracks of the service *Current Load* are added since they have a high correlation with the CPU usage and

⁶<https://github.com/tud-zih-energy/FIRESTARTER>

⁷<https://manpages.ubuntu.com/manpages/xenial/man8/trafgen.8.html>



(a) Overall



(b) The snippet with generated anomalies and markers set by the solution

Fig. 9: Anomalies detected by the solution

alerts are raised for those three measurands instead. If an anomaly is predicted for a track that is plotted, the anomaly is not added to the legend as the color identifies it. If a detected anomaly shares the same service as a service from a track, the same color is used. The anomaly period is detected and separated by the machine learning model.

Figure 9b provides a detailed view of the window where anomalies are created, hence, it has the same period as Figure 8 (b). The legend of Figure 9a applies. In this image, anomalies are marked along the time axis. The line at $value = 70$ is the health threshold. It turns out that this is a reasonable point for a start. When the FIRESTARTER benchmark is running, the model does not raise an alert for load or CPU usage for the first one to two points in time ($t \in \{1, 2, 15, 28\}$). Then, when the problem persists and the load tracks are inclining, anomalies are predicted. When the memory occupation is raised by five gigabytes, no alert is created ($t \in [10, 12]$). With an increase of twelve gigabytes, the memory anomaly is identified ($t \in \{37, 38\}$). The most significant drops in health are seen when the network traffic is generated. The generation via `trafgen` loops the packages back. Therefore, the anomaly is created for the received bytes. Here, one can see that the health score can drop below zero even though it was designed to be between zero and one. The reason is, that the new maximal values exceed the maximum seen by the min-max-scaler during the

training. For the user, those values could be displayed as zero with a warning flag to signal the difference to a normal zero.

IX. SUMMARY AND CONCLUSION

A. Summary

This paper demonstrated a solution for anomaly detection to accomplish predictive maintenance in a large-scale, distributed, and unreliable data center. DST was chosen as a framework for the data science project. The solution is modular and was implemented in Python. Architectural-wise, the models are part of each data center site and hence, not reliant on the connection to a site. An application was designed and written to fetch the time series log data from the monitoring application Nagios and data understanding was conducted. Hosts were grouped by behavior by correlating the time series, constructing a graph, and finding maximal cliques, i.e., groups with mutually high similarity. It was implemented using `networkx`. Models were created on a per-clique basis. A pipeline was designed and implemented with `scikit-learn` and `pandas` to cleanse and min-max scale the data. An LSTM autoencoder was designed, constructed with `Tensorflow`, trained, and tested to detect anomalies. The LSTM autoencoder could also learn new behavior without supervision. Anomalies were simulated by creating CPU load, memory occupation, and network traffic. The solution was able to spot them.

B. Future Work

The health score threshold was set arbitrarily. The model should be tuned by adjusting the health score threshold such that the detected anomalies optimally match the generated tickets or are proved if they are hits, i.e., true positives. Since there is a precision-recall trade-off, a PR curve should be constructed to set a health score that maximizes the recall but still provides reasonable precision to not create too many alerts which necessarily results in false positives. With too many false positives, the user might get overwhelmed by the number of alerts and might not pay attention to all predicted positives. Hence, true positives might be missed.

When considering the tickets from the system or a constant check of predicted anomalies whether they are true positives or false positives, real positives, i.e., cases when predictive maintenance was necessary, and real negatives, i.e., cases when no predictive maintenance was necessary, can be determined. With that, the metrics of interest (F1 score, recall, and PR-AUC) can be calculated to assess the quality of the model.

C. Conclusion

The solution can detect anomalies in a time series while being trained unsupervised. The training data is not required to be labeled manually and the indication of abnormal behavior does not rely on inaccurate incident data. In addition, the created models by the presented solution are specific since they are created per clique, i.e., on a per-host basis. Further, the models have the ability to detect unseen anomalies.

There are limitations. The LSTM autoencoder learns the behavior that is presented in the training data. Hence, the

normal behavior in the training data must exceed the number of abnormal samples. If the normal behavior of a host or a whole clique changes due to, e.g., other configurations or tasks, the model could rate this as anomalies. In this case, the model can be retrained with training data that captures the new behavior.


The overall goal was to create a system that can perform condition-based monitoring to enable predictive maintenance. This goal is achieved as the detected anomalies can be presented to a human to investigate the root cause and the symptoms that could arise. The solution will be used to filter the number of log entries and events.

REFERENCES

- [1] C.-J. Su and S.-F. Huang, "Real-time big data analytics for hard disk drive predictive maintenance," *Computers & Electrical Engineering*, vol. 71, pp. 93–101, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0045790617328409>
- [2] L. Wang, J. Yu, L. Shi, S. Lu, H. Pang, H. Chen, Z. Mei, M. Xu, and L. Qian, "Anomaly monitoring in high-density data centers based on gaussian distribution anomaly detection algorithm," in *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, 2020, pp. 836–841.
- [3] H. A. Gohel, H. Upadhyay, L. Lagos, K. Cooper, and A. Sanzetenea, "Predictive maintenance architecture development for nuclear infrastructure using machine learning," *Nuclear Engineering and Technology*, vol. 52, no. 7, pp. 1436–1442, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1738573319306783>
- [4] F. L. F. Pereira, D. N. Teixeira, J. P. P. Gomes, and J. C. Machado, "Evaluating one-class classifiers for fault detection in hard disk drives," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 2019, pp. 586–591.
- [5] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.
- [6] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4.
- [7] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, X. Zhao, and L. Dong, "Detecting anomalies in time series data via a meta-feature based approach," *IEEE access : practical innovations, open solutions*, vol. 6, pp. 27 760–27 776, 2018.
- [8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001. [Online]. Available: <https://doi.org/10.1162/089976601750264965>
- [9] J. Wang, C. Li, S. Han, S. Sarkar, and X. Zhou, "Predictive maintenance based on event-log analysis: A case study," *IBM Journal of Research and Development*, vol. 61, no. 1, pp. 11:121–11:132, 2017.
- [10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: <http://arxiv.org/abs/1603.02754>
- [11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [12] Y. Freund and R. E. Schapire, "Schapire R: Experiments with a new boosting algorithm," in *In: Thirteenth International Conference on ML*, 1996, pp. 148–156.
- [13] Y. Liu, W. Yu, T. Dillon, W. Rahayu, and M. Li, "Empowering IoT predictive maintenance solutions with AI: A distributed system for manufacturing plant-wide monitoring," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1345–1354, 2021.
- [14] R. Sipos, D. Fradkin, F. Moerchen, and Z. Wang, "Log-based predictive maintenance," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 1867–1876. [Online]. Available: <https://doi.org/10.1145/2623330.2623340>
- [15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, Jun. 2008.
- [16] M. Calabrese, M. Cimmino, F. Fiume, M. Manfrin, L. Romeo, S. Ceccacci, M. Paolanti, G. Toscano, G. Ciandrini, A. Carrota, M. Mengoni, E. Frontoni, and D. Kapetis, "SOPHIA: An event-based IoT and machine learning architecture for predictive maintenance in industry 4.0," *Information-an International Interdisciplinary Journal*, vol. 11, no. 4, 2020. [Online]. Available: <https://www.mdpi.com/2078-2489/11/4/202>
- [17] L. Giommi, D. Bonacorsi, T. Diotallevi, S. R. Tisbeni, L. Rinaldi, L. Morganti, A. Falabella, E. Ronchieri, A. Ceccanti, and B. Martelli, "Towards predictive maintenance with machine learning at the INFN-CNAF computing centre," in *Int. Symp. on Grids & Clouds (ISGC). Taipei: Proceedings of Science*, 2019, pp. 1–17.
- [18] scikit-learn developers, "Sklearn.ensemble.AdaBoostClassifier." [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>
- [19] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [20] L. Decker, D. Leite, F. Viola, and D. Bonacorsi, "Comparison of evolving granular classifiers applied to anomaly detection for predictive maintenance in computing centers," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2020, pp. 1–8.
- [21] F. Minarini and L. Decker, "Time-series anomaly detection applied to log-based diagnostic system using unsupervised machine learning approach," in *Conference of Open Innovations Association, FRUCT*, no. 27. FRUCT Oy, 2020, pp. 343–348.
- [22] scikit-learn developers, "Sklearn.svm.OneClassSVM." [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html>
- [23] H. V. Dündükçü, M. Taşkıran, and N. Kahraman, "LSTM and WaveNet implementation for predictive maintenance of turbofan engines," in *2020 IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI)*, 2020, pp. 000 151–000 156.
- [24] A. P. Hermawan, D.-S. Kim, and J.-M. Lee, "Predictive maintenance of aircraft engine using deep learning technique," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 1296–1298.
- [25] M. Canizo, I. Triguero, A. Conde, and E. Onieva, "Multi-head CNN-RNN for multi-time series anomaly detection: An industrial case study," *Neurocomputing*, vol. 363, pp. 246–260, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231219309877>
- [26] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," 2018. [Online]. Available: <https://arxiv.org/abs/1805.06725>
- [27] H. Ahn, D. Jung, and H.-L. Choi, "Deep generative models-based anomaly detection for spacecraft control systems," *Sensors*, vol. 20, no. 7, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/7/1991>
- [28] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, and B. Bennadji, "Predictive maintenance in building facilities: A machine learning-based approach," *Sensors*, vol. 21, no. 4, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/4/1044>
- [29] V. Breux, J. Boutet, A. Goret, and V. Cattin, "Anomaly Detection in a Data Center with a Reconstruction Method Using a Multi-Autoencoders Model," *International Journal of Mechanical and Industrial Engineering*, vol. 16, no. 3, pp. 46–53, 2022. [Online]. Available: <https://publications.waset.org/vol/183>
- [30] F. L. F. Pereira, I. Castro Chaves, J. P. P. Gomes, and J. C. Machado, "Using autoencoders for anomaly detection in hard disk drives," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.
- [31] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: A primer," p. 6, 2018.
- [32] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, 2021.
- [33] P. M. Pardalos and J. Xue, "The maximum clique problem," *Journal of Global Optimization*, vol. 4, no. 3, pp. 301–328, Apr. 1994. [Online]. Available: <https://doi.org/10.1007/BF01098364>

Patterns in the Context of Small Data - Lessons learned from Data Science Projects

Friedemann Schwenkreis
Business Information Systems
Baden-Wuerttemberg Cooperative State University Stuttgart (DHBW Stuttgart)
Stuttgart, Germany

<https://orcid.org/0000-0003-4072-0582> 

Abstract—This paper summarizes aspects of pattern recognition in the context of empirical study projects. It reflects the lessons learned from several real-world projects that represent the set of projects which have to discover knowledge based on relatively small amounts of data. The paper introduces into two descriptive methods that can be applied in the described context. The main challenges when applying the methods are identified and approaches are introduced to address these challenges. Furthermore, as a result of an evaluation of programming environments for parallel co-occurrence grouping, the properties of certain programming environments will be presented.

Keywords— *Data Science, Pattern Recognition, Machine Learning, Clustering, Co-Occurrence Grouping*

I. INTRODUCTION

Big Data Analytics and Data Science have become two of the most hyped fields of computer science these days. With the availability of artificial-intelligence-based services like ChatGPT [1], there seems to be the perception that machine learning is the universal solution for any real-world problem. However, the toolbox of data mining includes much more than machine learning and for many real-world application cases it is suboptimal to generally prefer machine learning approaches over other non-machine-learning methods. Based on real-world data science projects, this paper will describe patterns of cases for which non-machine-learning approaches are better suited to extract valuable information from data.

Section II will introduce into the terminology used in this paper to avoid misunderstandings. Furthermore, the most important concepts in the context of pattern recognition will also be introduced in section II. Section III will describe specifics of clustering methods that need to be considered when applying these methods in real-world projects. In section IV the paper will present how co-occurrence grouping can be used to find rules particularly in the case of empirical studies. The paper will be concluded by section V which will summarize the major aspects of this paper.

II. TERMINOLOGY

A. Data Science and Data Mining

Data Science denotes the application of principles, processes, and procedures to develop an understanding of phenomena by (at least partially automated) analyzing data [2]. This includes mechanisms for data collection and data preparation, as well as methods to analyze data as statistical methods and data mining methods. Furthermore, it includes the very important step to evaluate the results of the analysis to determine the quality and value on the application level.

Data Mining denotes a process which includes a set of steps to extract useful patterns from data (as for instance

CRISP-DM) [3]. The process particularly includes a pattern extraction step that uses one or more *pattern recognition* methods but also steps to prepare data (before the pattern extraction) and to calculate quality indicators for the extracted patterns. From this perspective, data mining is a part of data science and in some cases the two are identical. However, data science may go beyond data mining and can include multiple data mining processes.

B. Pattern Recognition and Machine Learning

In the context of this paper, we will differentiate between pattern recognition and machine learning. In the following the differences will be described.

1) Pattern Recognition

There are three basic families of so-called *pattern recognition* methods that are currently differentiated [4]:

a) Descriptive methods

Methods that extract patterns from data which describe or summarize the content of the data. The patterns can be interpreted as a *summary* of the original data or as the extracted *important aspects* of the data. Descriptive patterns are valid for the given input data but might not be projected on other (particularly future) data.

b) Predictive methods

Methods which extract patterns from data that are intended to be used as the basis to “estimate” the value of a certain target attribute of other data. Therefore, the input data for predictive models must contain the target attribute. Furthermore, extracted patterns can be validated using test data (which was ideally not used during the extraction of patterns) for which the target attribute is also known. The validation based on test data allows to evaluate whether the extracted patterns can be used to estimate the value of the target attribute for data that has not been used to extract the patterns.

c) Prescriptive approaches

Prescriptive analytics uses the predictive methods as a foundation to estimate the effect of a set of possible actions. Furthermore, there needs to be a notion of “application-level value” for raw values of the target attribute. By calculating the application-level value of the target attribute for the possible actions, prescriptive methods determine a ranking of the possible actions with respect to the application-level value of the estimated target attribute.

2) Machine Learning

The basic idea of machine learning is to automatically extract one or more models based on so-called training data which can then be used in an application to estimate target values by using the model and a set of input attributes (which have been a subset of the attributes of the training data).

Furthermore, the model can be enhanced by training it with additional data.

Today, machine learning is almost a synonym for the usage of predictive methods – regardless of whether they are used in the context of prescriptive analysis or “stand-alone”. The set of descriptive methods differentiates machine learning from the overall set of pattern recognition methods.

C. Big Data and Small Data

With the identification of the so-called Fourth Paradigm in 2009, a new view on the value of large quantities of data was introduced [4]. With the availability of analysis technology, it was possible to extract previously unknown knowledge from large amounts of data which was not possible based on lower amounts of data. Large denotes a quantity of data that was previously not processable by humans even when using computerized systems.

However, the term *Big Data* that was “coined” by McKinsey and Gartner in 2011 (also known as the three Vs) goes beyond “large” [5]. It summarizes a family of methods and techniques to handle data that cannot be processed using “conventional” (database / data science) methods and techniques. In a nutshell, methods and techniques of Big Data introduce mechanisms that allow to focus on the important data while ignoring the other non-important data which is necessary because sufficient resources or abilities to process all data are not available.

Today there are a lot of marketing messages of products that use the label Big Data or Big Data Analytics. Most of them do not actually focus on the Big Data aspect but data science and sometimes large data. There is also a significant number of cases in which the size of data seems to be “big” from an application perspective but turns out to be rather small from a computer science perspective – in a nutshell: “everything that fits on a USB-stick cannot be big”.

Predictive data mining methods assume that the data used to train models contain all “cases” that might occur when later applying the model to estimate the target attribute. Since the set of possible cases is usually unknown when training models, the number of possible cases is estimated based on the potentially occurring combinations of the number of different values of the attributes of the input data. This results usually in a huge number of possible combinations. For instance, in a project to determine the different effects of menstruation on the performance of female athletes, more than 50 different attributes have been observed. If we just assume that each attribute has two potential values (which is an oversimplification in most cases), at least 2^{50} (more than 10^{15}) records are needed as training data to cover all theoretically possible cases.

Empirical studies based on non-automated data observation or recording, generate significantly less records because each case needs a significant amount of time. In the example mentioned above, the study generated approximately 700 records of data. Compared to the size of the theoretical “feature space” this is far from the needed number to cover all cases. However, we still want to benefit from data mining methods even in case of such *Small Data* situations.

The *small data problem* has been identified since the beginning of data mining methods but was named as such just recently as a counterpart to the term big data [6]. Whenever we are applying data mining methods based on small data, we

must be aware that the results cannot simply be projected on future cases. On the contrary, it must be assumed that the results are specific for the used data which is somehow a contradiction to the intention of the previously introduced predictive methods.

Even in case of small data there is always “hope” that the amount of available data is representative to some extent and thus the results can be interpreted accordingly. Many analysis results of projects have been published based on that hope, lacking a comprehensive analysis of the data representativity as well as an in-depth investigation whether the results can be applied in general.

D. The Overfitting Phenomenon

The fact that deep-learning applications are sensitive to the problem of small data is common knowledge. Since deep-learning methods tend to extract models that are very specific to the data used to train the deep-learning model, using the approach based on small data results in non-generalizable analysis results. Even if this *overfitting effect* is reduced by introducing “artificial noise”, the resulting models do not really help with the high-quality prediction of “unseen” cases.

There is a more general problem with all predictive approaches independently from any specific approach when we have a small data context: The available data does not cover a significant or even sufficient amount of the feature space. Thus, any approach will extract a model that is specific to the used training data and hence will fail to extract an accurate, generalizable model. The effect of this variant of overfitting might be “softened” in terms of improving the quality indicator values of the model by using specialized methods in the context of certain test data, but this does improve the prediction quality for unseen cases,

Simply speaking, predictive approaches cannot have an overfitting effect if the used input data completely covers the feature space because “completely covered” means, that all possible cases have been represented in the training data to extract the model. Thus, distinguishing predictive methods based on their “tendency” for overfitting means, that, given the same set of incomplete training data, less-overfitting methods extract a more “blurry” or “fuzzy” model that might handle more so-far unseen cases coincidentally correctly compared to methods with a higher tendency for overfitting. However, this does not mean that methods with a lower tendency for overfitting generally extract a better model with a higher overall quality.

III. CLUSTERING INSIGHTS

There is a significant amount of cases in real-world scenarios where a target attribute is not given or can only be provided at relatively high cost or with unacceptable effort. A similar problem arises if only small data can be used as input data. In both cases it does not make sense to apply predictive approaches. Nevertheless, there is a significant interest of identifying similarity groups in the available data. The family of methods that search for groups of similar records is denoted by the term *clustering* methods in the context of this paper. These methods do not require a target attribute to define similarity as in case of predictive approaches but use a similarity or distance function of the records to find similar records.

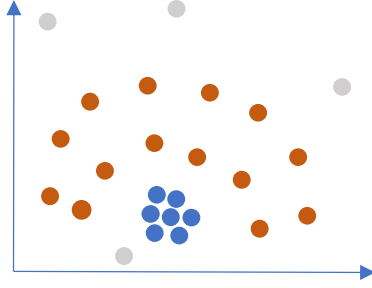


Fig. 1. Varying Cluster Densities

A. Number of clusters

There are two extreme “results” when searching for clusters:

- All records are grouped into a single cluster.
- Each record is treated as a separate cluster.

Both cases will not help on the application-level because they both simply mean that no *relevant* groups could be found. As in case of scoping a microscope, clustering methods allow to influence the granularity of the clustering result by using some method-specific sort of parameter settings. This means, that depending on the parameter settings, more or less clusters are identified.

Unfortunately, it cannot be absolutely determined which clustering model is the better, if there is no data available to verify whether the grouping of a record together with other records is correct. This so-called “ground-truth” is generally not available. Thus, indirect indicators are usually used (see section III.D).

B. Noise or Non-Groupable Records

There are several clustering methods like the popular k-means approach, that simply assume that all records of the input data can be or should be assigned to a similarity group [7]. However, this is not the case in many real-world applications. There are either records in the data that are erroneous, or there are “individual” cases which should not be grouped together with other cases. In case of the former, the erroneous records are considered to be actual noise, meaning records containing wrong or corrupted information and might be identified and discarded before applying a clustering method.

In the latter case the situation is more difficult. If we have small data, individual records might be just an indication of not having enough data to identify further records that are similar to these individual records. However, it might also be a property of the application space that there are non-groupable records which are only similar to noise from a clustering perspective but not from an application perspective. To decide whether an individual record belongs to a group that has been underrepresented in the data or whether it is a single case from an application perspective a manual inspection is necessary. Hence, in case of an application scenario having records outside of similarity groups, a clustering method must be used that allows to identify the “non-groupable” records explicitly.

Density-based spatial clustering of applications with noise (aka *DBSCAN*) is one of the few clustering methods that explicitly handle the mentioned ungroupable records [8]. Recent extensions of spectral clustering have introduced a

similar feature but still have the limitation that they can handle only a small amount of noise records compared to the number of groupable records [9]. DBSCAN uses a density criterion based on a notion of distance of records to identify groups of similar records. Based on the number of records in a certain distance of a record DBSCAN distinguishes between direct neighbors in a group (core points) and transitive neighbors (boundary points).

Points that do not have enough neighboring points in the user-provided distance are treated as noise points. These points are assigned to a special group that can be investigated separately. The critical parameter settings of DBSCAN are the distance threshold and the minimum number of direct neighboring points that are needed to form a group. If the distance threshold is selected too small, then all points are treated as noise points. If the distance is selected too large, then all points end up in the same cluster.

C. Distribution of Points with Varying Density

As DBSCAN shows, density, the number of points in a certain region, is a valid criterion to find groups of similar records if there is a meaningful distance criterion to express similarity. By using the concept of transitive distance or transitive neighbors DBSCAN can even identify concave regions of similar records which is also an advantage to other clustering methods.

A shortcoming of DBSCAN is the need to specify the distance threshold for identifying the neighboring points. Since there is only a single threshold it is equally applied in case of all points. Thus, specifying such a “uniform” threshold assumes that the density of similarity groups in the data is uniform as well. Unfortunately, there are a significant number of applications for which this is not true. Fig. 1 shows an example of two-dimensional data containing two similarity groups of different density (blue and orange) and some noise points (grey). Specifying a low distance threshold would identify only the blue points as a similarity group. Increasing the threshold to detect the similarity between the orange points would result in one big cluster of the orange and the blue points including even noise points.

To avoid the need for specifying a fixed distance threshold, Shared-Nearest-Neighbor-clustering (SNN-clustering) defines density differently: A notion of density based on the number of common nearest neighbors between records [10]. The basic idea is to define the similarity of two records by comparing the sets of their nearest neighboring records rather than directly evaluating their distance. The number of points contained in the intersection of the two sets is defined as the similarity of the two points.

With this definition of similarity between points, the need for specifying a fixed distance to find similar points is avoided. However, the number of nearest neighbors that are considered to calculate the similarity, needs to be specified. Based on the calculated similarity and the total number of considered nearest neighbors a distance can be calculated which allows to use DBSCAN to search for similarity groups.

A shortcoming of SNN-clustering is the fact that the resulting shared-nearest-neighbor-based distance depends on the size of the sets of nearest neighbors. This results in changing similarity values depending on the number of points considered. Substituting the original notion of similarity and the derived notion of distance respectively with the Jaccard

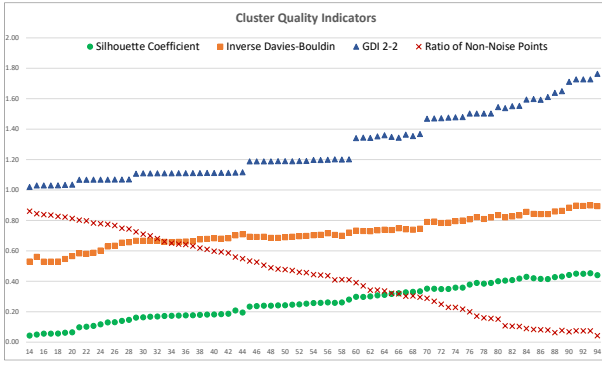


Fig. 2. Quality Coefficient

distance based on the two sets of nearest neighbors of two records, helps to overcome this problem because it is independent of the size of the compared sets [11].

D. Quality Indicators

Several quality coefficients for clustering have been defined in the past [12]. Most of these indicators are based on a notion of distance and some distance-based density of the original records. When using SNN-clustering it is important to adapt the definitions to the specific notion of distance introduced in section III.C rather than using the original distance of the records.

Since the quality coefficients of clustering are a means for comparing cluster model quality rather than an indicator for the absolute model quality, they need to be either maximized or minimized to find the “optimal” cluster model. To be specific, when varying the parameter settings used to extract a cluster model from given data, the quality coefficients can be used to determine whether the variation of the parameter settings resulted in a better or worse clustering model based on the value of the selected quality coefficient.

The selection of a specific quality coefficient depends on the available data and the computational complexity of the quality coefficient. Some coefficients are defined using the distance to cluster centers which usually assume convex clusters. These approaches should not be used when clustering methods like SNN-clustering are used.

It is important to notice that quality coefficients assume a constant set of input data of two cases to be comparable. Clustering methods that group noise records in separate clusters usually violate that assumption because the noise records are not used in the calculation of the quality coefficients. Thus, they are removed from the data from the quality coefficients’ perspective. Hence, it is very unlikely that two different cluster models of noise handling methods have the same set of noise records. Therefore, the usual cluster coefficients do not help to identify the optimal parameter settings in case of noise handling clustering methods.

Fig. 2 shows an example of a comparison of 3 different clustering coefficients varying over the same parameter settings of SNN-clustering. Additionally, the ratio of clustered points and noise points is depicted as red crosses. It can be clearly seen that the values of the clustering coefficients are steadily increasing. No maximum can be identified which would be an indication for the best parameter settings.

The reason for that is the steadily decreasing number of clustered points. In a nutshell: the clustering quality gets better and better from the perspective of the quality coefficients by

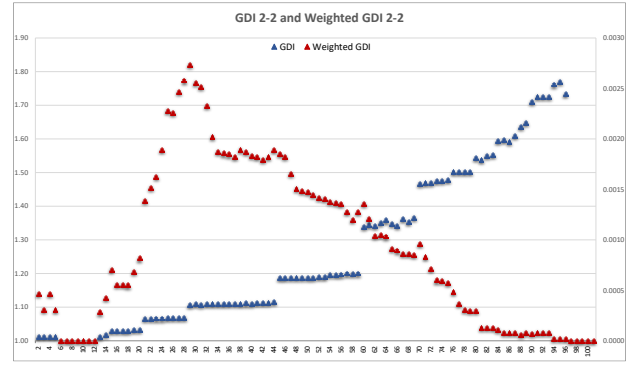


Fig. 3. Weighted Quality Coefficient

reducing the number of considered records and potentially not because of a real quality improvement of the model.

To address this phenomenon a simple consideration helps: since the used methods do not know the ground-truth (a given association with groups), they “compromise” between adding records to groups and treating them as noise. Hence, quality indicators should take the number of clustered records into account: models with the same quality coefficient but more records are better than models based on less records. This can be expressed by weighting the quality coefficient with the number of clustered points.

Fig. 3 shows an example for the so-called *Weighted GDI-22* quality indicator based on the original GDI-22 quality coefficient. The red triangles depict the weighted value while the blue triangles depict the original value of the GDI-22 coefficient. A global maximum can be clearly identified in case of the weighted coefficient which is a perfect indication for the best compromise between the number of records included and the relative quality of the resulting clustering model.

IV. CO-OCCURRENCE GROUPING REVISITED

A. Basics of Co-Occurrence Grouping

Co-occurrence grouping, also known as the search for association rules, is a data mining method that belongs to the set of descriptive methods. The basic idea is to have set of transactions and each transaction consists of a set of so-called items. Co-occurrence grouping identifies groups of items that occur frequently together in transactions. These so-called *frequent itemsets* are then used to generate rules like “if itemset A is contained in a transaction, then itemset B is also contained in the same transaction”.

Three main properties are associated with each rule:

- The (relative) *support* of the rule: How many transactions contain itemset A and itemset B?
- The *confidence* of the rule: The ratio of the number of transactions containing itemset A and the number of transactions containing itemset A and itemset B.
- The *lift* of the rule: The ratio of the support of the rule divided by the product of the support of itemset A and the support of itemset B.

Co-occurrence grouping has been introduced in the context of shopping basket analysis to determine which products are usually bought together. However, the concept can be generalized to classification-like application areas as well.

The advantage of co-occurrence grouping is the concept to generate and evaluate all possible rules. This helps particularly in the context of empirical studies because co-occurrence grouping can be used to generate hypotheses based on the collected data rather than having to “manually” develop hypotheses which are then verified (or falsified) based on standard statistical methods.

B. Decision Trees and Co-Occurrence Grouping

1) Classification with Decision Trees

As previously mentioned in section II.B.1)b), classification as a predictive data mining approach, has a categorical target attribute [2]. The basic task of classification is to predict the value of the target attribute depending on the other attribute values of input records. There are special application cases, where the prediction itself is not the main objective but rather to find the minimal set of attributes which determine the value of the target attribute. Particularly decision trees are commonly used as the classification method in these cases because decision trees can be converted into rules that “describe” the influence of input attribute values onto a certain value of the target attribute.

A shortcoming of decision trees is the fact that the underlying model is a tree. This means that all rules that can be derived from the tree start with a condition based on the same attribute – the most determining attribute according to the selected split criterion and thus root of the tree. This might be a good approach when the objective is to maximize the overall prediction quality of a tree but not when the application is interested in all rules with a high confidence or in the identification of unusual dependencies. This shortcoming is addressed by a classification approach called *random forests* which generates multiple trees with varying roots [7] to cover more cases than single decision trees. However, the extraction of *interesting rules* from a random tree model is very time consuming, because all rules need to be extracted first and then the previously mentioned properties of the rules need to be calculated separately.

2) Mapping onto co-occurrence grouping

The rules that are extracted from decision trees are a subset of the rules that are identified by co-occurrence grouping. The right-hand side of the rules from classification, the so-called *consequent*, is a single attribute value rather than a set of values as in case of co-occurrence grouping. The left side of the rules, the *antecedent*, are sets of values as in case of co-occurrence grouping. Hence, co-occurrence grouping conceptually discovers more rules than decision trees or random forests.

Two mapping steps are needed to be able to apply co-occurrence grouping in case of a classification problem:

- The input data needs to be transformed into transaction format.
- The consequent of the rules needs to be restricted to values of the target attribute of the classification.

Generating transaction data from the classification input data is almost a straight-forward process. All attributes need to be categorical and numerical attributes need to be discretized as in case of decision tree algorithms that use for instance the information gain as a split criterion.

The basic difference between classification data and transaction data is the encoding. Classification data allows

multiple attribute values while transaction data reflect the information whether an *item* is contained in the transaction or not. All attribute values of a record of the classification data are represented by items of a transaction. To be able to identify the source attribute of an attribute value, the attribute name needs to be encoded in the corresponding value of the item. For example, consider an attribute named “gender” with the value “female” in the classification data, then the corresponding transaction might contain an item with the value “gender:female”.

It might sound trivial that rules that are extracted from a set of transactions can only contain items that were contained in the transaction data. However, there is a consequence resulting from that: rules based on the non-existence of certain attribute values can only be found if the non-existence is represented explicitly as an item. This is a difference to rules extracted from decision trees. The branches in decision trees use the negation of a condition of other branches which results in rules that contain a term that represents the “absence” of certain values. In case of co-occurrence grouping, it is necessary to explicitly generate items that represent the absence of attribute values, if the application is interested in rules that are based on this information.

Handling the consequent restriction for the rules of co-occurrence grouping can be done in two ways. Either the usual search for all itemsets is done with a subsequent filtering on rules that contain a consequent of interest or, the generation of itemsets is restricted upfront based on the same condition. The restriction speeds-up the computation because only itemsets are considered that contain one item which represents a value of the target attribute. Then, in the rule generation phase, only rules are generated that have the special items representing the value of the target attribute as the consequent.

C. Performance and Programming Aspects

Although the number of possible itemsets grows over-exponentially, experiments with real-world data from empirical studies show that co-occurrence grouping can be used successfully if the minimum support parameter is chosen carefully. The minimum support parameter is used to filter out itemsets that are not supported by enough transactions to be relevant.

However, the search for itemsets as well as the generation of rules from frequent itemsets is a time -consuming process which can benefit substantially from parallel computing on modern multi-core CPUs. If the *Apriori* approach is used to generate the frequent itemsets [13], then the computation can be accelerated using straight forward parallelization. The candidates on level $n+1$, denoted by C_{n+1} can be generated from the candidates of the previous level C_n by splitting C_n into subsets and subsequently generating the candidate sets of all subsets using the frequent items of level one.

Therefore, the overall set C_n is split into subsets of size k denoted as the (user-defined) partition size. Then the set of next-level candidates of each partition can be generated independently by a separate thread. The same approach works for counting the support of candidates. To be able to implement this concept of parallelization, the programming environment needs to support the definition and parallel execution of user-defined functions. The following paragraphs introduce a bit into some popular programming environments and point out some aspects worth noting in this context.

Python and Perl have become popular programming languages in the context of data mining [14] [15]. Mainly because they are easy to learn and because of the available library functions for data mining. Furthermore, the availability of the web based Jupyter notebooks have given Python a boost because they provide an interactive programming environment without the need for a local setup. Both, Python and Perl, are “general” programming languages that allow to program functions. Additionally, they both support the parallel execution of functions.

There are also several specialized data science programming environments like IBM SPSS Modeler™, Mathworks Matlab®, and Rapid Miner® (which was recently acquired by Altair) [16] [17] [18]. While SPSS Modeler and Rapid Miner introduce a graph-based programming approach based on built-in operators following the concept of data-flow programming, Matlab comes with its own programming language on top of C/C++ libraries. Both SPSS Modeler as well as Rapid Miner are implemented based on Java in their core engine and extending the available operators is rather complicated. However, both allow to use parallelism on the process-level to execute the available operators in parallel. The problem of parallelizing the search for frequent itemsets in this context is, that the search cannot be parallelized as a whole, but only parts of it. Consequently, the complete search for frequent itemsets needs to be implemented based on user-defined operators which then need to be embedded in processes using the parallel execution.

Matlab is a data mining programming environment that comes with a complete IDE which supports the Matlab-specific programming language. Since Matlab has been originally introduced in statistics and matrix computing, it contains very performant operators for vector and matrix computations. The Matlab-specific language is, as in case of Perl and Python, an interpreted language that is mapped onto the underlying C/C++-libraries. The language allows the definition of arbitrary functions which can be executed in parallel using a special parallel execution function (*parfeval*), or the so-called parallel toolbox of Matlab that needs to be purchased separately.

Since SPSS Modeler as well as Rapid Miner do not directly support the embedding of parallel user-defined functions, it has not been evaluated whether the use of any of the two would result in a real benefit from the point of view of overall performance. When taking a closer performance look at recent versions of Python and Perl, then sometimes Perl outperforms Python and vice versa. Since recent developments of Python in the area of parallel processing solved some shortcomings compared to Perl (without the usage Jupyter notebooks!) and Python became very popular in the last years, a Python implementation of parallelized co-occurrence grouping was compared to a Matlab implementation.

Even though Matlab has kind of a bad reputation regarding performance, the Matlab implementation outperforms the Python implementation by an order of magnitude. Only a few Matlab specific optimizations have been used, but they seem to be important. For instance, for-loops should be avoided when iterating over a vector or a matrix, to filter the elements of a vector, or when searching for elements satisfying a certain condition. The Matlab language has special expressions for these cases resulting in a direct mapping to underlying C/C++-functions, thus avoiding the evaluation of the loop in

interpretation mode. This results in an execution that is a hundred times faster compared to the for-loop approach.

V. SUMMARY AND CONCLUSION

Using predictive data mining methods in the context of small data underlies significant restrictions. Furthermore, there are many empirical studies that are rather interested in knowledge or the ability to explain a phenomenon than being able to predict something without knowing why or how the prediction works. Descriptive data mining approaches like clustering and co-occurrence grouping help in these cases and provide much more insights in collected data than for example popular artificial intelligence methods based on deep neural networks.

Current data science tools enable end-users to apply all kinds of data mining functions in any data context. However, the results will be arbitrary without considering some important prerequisites. First, it needs to be evaluated whether the data contains noise or not. Depending on this criterion the clustering method must be selected. Furthermore, clustering will only produce sensible results if a meaningful notion of similarity or distance can be found. Even if such a distance has been found, it needs to be considered whether the density of the data is distributed uniformly across all data points. In general, this is not the case, and an indirect notion of similarity helps to handle this situation. Since there is no direct quality criterion in case of clustering, the comparison of quality indicator values needs to be used to find optimal results. It has been shown that classical indicator values do not really help in case of noise identifying clustering algorithms. However, the concept of weighted indicator values helps in this situation.

When searching for rules in data, co-occurrence grouping is a descriptive approach that outperforms even decision trees. The necessary mapping of the data format is simple and well-known mechanisms can be used to map numerical data if needed. Unfortunately, the computational complexity of co-occurrence grouping is demanding which results in the need to use parallel processing in case of real-world problems. However, the computational power of a recent gaming PC is sufficient to produce meaningful results when performant platforms like Mathworks Matlab® are used. Note that other programming environments might be significantly slower even if they are very popular.

REFERENCES

- [1] OpenAI, May 2023. [Online]. Available: <https://openai.com/product/chatgpt>.
- [2] F. Provost and T. Fawcett, Data Science for Business, Sebastopol, CA: O'Reilly and Associates, 2013.
- [3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. P. Reinartz, C. Shearer and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," 2000.
- [4] T. Hey, The Fourth Paradigm, Microsoft Research, 2009.
- [5] B. Brown, M. Chui and J. Manyika, "Brown, Brad, Michael Chui, and James Manyika. "Are you ready for the era of 'big data'," *Mc Kinsey Quarterly*, vol. 4, no. 1, pp. 24-35, 2011.
- [6] R. Kitchin and T. P. Lauriault, "Small data in the era of big data," *GeoJournal*, no. 80, pp. 463-475, 2015.
- [7] J. Han, J. Pei and H. Tong, Data Mining: Concepts and Techniques, Cambridge, MA, USA: Morgan Kaufmann Publishers, 2022.
- [8] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with

- noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, USA, 1996.
- [9] A. Bojchevski, Y. Matkovic and S. Günnemann, "Robust Spectral Clustering for Noisy Data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, New York, 2017.
 - [10] L. Ertöz, M. Steinbach and V. Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data," in *Proceedings of the 2003 SIAM International Conference on Data Mining (SDM)*, San Francisco, CA, USA, 2003.
 - [11] P. Jaccard, "The Distribution of the Flora in the Alpine Zone," *New Phytologist*, vol. 11, no. 2, p. 37–50, 1912.
 - [12] B. Desgraupes, "Clustering Indices," Lab Modal'X, University Paris Ouest, 2017.
 - [13] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proceedings of the 20th VLDB Conference*, Santiago Chile, 1994.
 - [14] Python Software Foundation, "Welcome to Python.org," 2023. [Online]. Available: <https://www.python.org/>.
 - [15] Perl.org, "The Perl Programming Language," Perl.org, 2023. [Online]. Available: <https://www.perl.org/>.
 - [16] IBM, "SPSS Modeler | IBM," 2023. [Online]. Available: <https://www.ibm.com/products/spss-modeler>.
 - [17] The MathWorks Inc., "MATLAB - Mathworks," 2023. [Online]. Available: <https://www.mathworks.com/products/matlab.html>.
 - [18] Rapid Miner, "Rapid Miner | Amplify the Impact of your People, Expertise & Data," 2023. [Online]. Available: <https://rapidminer.com/>.

Model-Agnostic Overconfidence Reduction for Tabular Data

Alexandros Nanopoulos
Business Informatics
Baden-Wuerttemberg Cooperative State
University
Mosbach, Germany
alexandros.nanopoulos
@mosbach.dhbw.de

Abstract—Overconfidence is the problematic situation when a machine learning model predicts incorrectly albeit with high confidence. This situation becomes more severe when models are confronted with out-of-distribution data, which often emerge in real-life applications and can lead to catastrophic mistakes. In this paper, we focus on tabular data, which is the most widely used data format in real-life use-cases of machine learning. We identify experimentally how overconfidence affects machine learning models in this application domain. We also propose and evaluate a model-agnostic method that constructs a training set, which can lead to models that drastically reduce overconfidence, without significantly affecting prediction accuracy.

Keywords—overconfidence, out-of-distribution, label smoothing, outlier exposure, tabular data

I. INTRODUCTION

Overconfidence is the problematic situation when a machine learning model predicts incorrectly albeit with high confidence; or to put it more simply: when the model does not know that it does not know.

Overconfidence becomes more severe for models in production, especially when they are confronted with *out-of-distribution* (OOD) data that are radically different from the training data with which models have been trained. Notice that models with high prediction accuracy, evaluated with *in-distribution* data, which are similar to the training data (e.g., using cross validation), can still be extremely overconfident when predicting for OOD data, which emerge in real-life applications due to various types of distribution shifts [1].

Due to overconfidence, users of machine learning models may receive incorrect predictions with misleadingly high confidence. This can affect negatively their decision-making process and reduce their trust to the model altogether. In many applications of autonomous machine learning systems (e.g. self-driving cars), overconfidence, no matter how infrequent, is considered as a major risk that can lead to critical errors with high costs [2].¹ This situation has motivated recent research to investigate possibilities to reduce overconfidence (please refer to Section II for a review).

In this paper, we focus on classification models for tabular data (i.e., data in which features are structured as columns of

a table), which is the most widely used data format in real-life use-cases of machine learning. As it will be shown, overconfidence is also present in models that are typical in applications with tabular data, such as Extreme Gradient Boosting (XGBoost), which compete favorably with deep learning models in this application domain [3,4]. However, the effect of overconfidence in the case of standard models for tabular data has not been studied as thoroughly as it has been done in the case of deep learning models for computer vision and natural language processing. For this reason, we first identify experimentally how overconfidence affects models that are representative for tabular data. Additionally, we propose a *model-agnostic* method, which achieves an enrichment of the training set by a combination of two techniques: *label smoothing* and *outlier exposure*. Our experimental results indicate that this method can help train models that have drastically reduced overconfidence without significantly affecting their prediction accuracy.

In the rest of this paper, we first review related work. Next, we present the examined method, which is then evaluated experimentally. Finally, we present the conclusions and directions of future work.

II. RELATED WORK

Several deep learning models have been identified as prone to overconfidence when faced with OOD data, because they extrapolate wildly, due to loose regularization in areas of the feature space that are unobserved in the training set [5,6]. Although overconfidence is also expected to be present in models for tabular data, to our knowledge, it has not been thoroughly investigated in this application domain.

Approaches to reduce overconfidence in deep learning models mainly examine *epistemic uncertainty*. An effective approach are ensembles [5,7]. The reason is that models in an ensemble tend to disagree in their predictions for OOD instances. Since we focus on models for tabular data, which typically are themselves ensembles (for example, XGBoost and RandomForest), it is therefore not meaningful to consider ensembles of other ensembles.

¹ The criticality of the 0.1% errors in predictions with OOD data can fatally undermine the safety of self-driving cars even with 99.9% accurate vision systems.

Calibration methods follow another approach towards reducing overconfidence, by mapping predictions to well calibrated probabilities [8]. As, however, explained in recent research [6], calibration methods are not guaranteed to work for OOD, since they are restrained to learn the mapping to calibrated probabilities only from *in-distribution* data.

Label smoothing is another technique, which transforms the (one-hot encoded) labels to a probability distribution. Each element of the label gets a non-zero probability, trying to reduce the amount of confidence that is assigned to instances of the training set. As already shown [9], label smoothing is an effective regularization technique that can reduce both overfitting and overconfidence. Label smoothing has been investigated primarily for deep learning models. Concerning machine learning models for tabular data, to our knowledge, the impact of label smoothing has not been studied thoroughly. In this work, we focus on: a) understanding whether label smoothing can reduce overconfidence in various model types for tabular data, and b) evaluating experimentally the impact on the degree of label smoothing, i.e., the trade-off between reducing overconfidence, especially for OOD data, and prediction accuracy for in-distribution data.

The work of [6] shares the same motivation with our approach, i.e., the reduction of overconfidence. However, this work focuses on methods applicable to deep learning models for computer vision, whereas we focus on (model-agnostic) approaches for tabular data.

Finally, there exist several works in the area of detecting outliers (a.k.a. anomaly or novelty) so as to avoid making predictions on spurious instances; see [10] for a more recent approach in this area. To this end, a method for outlier exposure (a technique that we also consider in our work) has been examined in the context of anomaly detection [11]. However, such approaches are not considering the quality of confidence estimates in the predictions and require the development of additional models for detection. In our work, we focus on improving the estimated confidence in predictions of machine learning models directly (i.e., not on constructing additional models) and especially on ways to reduce overconfidence therein.

III. METHOD

Aiming at being model-agnostic, the examined method focuses solely on the enrichment of a given training set, which can then facilitate the training of various machine learning models (under a mild set of preconditions explained at the end of this section). This enrichment procedure comprises two steps, which are illustrated with the example of Figure 1 and are analyzed in the rest of this section.

The original training set of this example is depicted in Figure 1(a). It consists of three instances (i.e., rows), each having three numerical features (i.e., first three columns), called x_1, x_2, x_3 . Each instance has also a binary label (last column), called y . Notice that the binary labels have been one-hot encoded. In this way, the first and third training instances in Figure 1(a) belong to the ‘positive’ class and their labels are, thus, one-hot encoded as a vector with two elements: (0, 1). The second instance, on the other hand, belongs to the ‘negative class’ and its label is encoded as (1, 0).

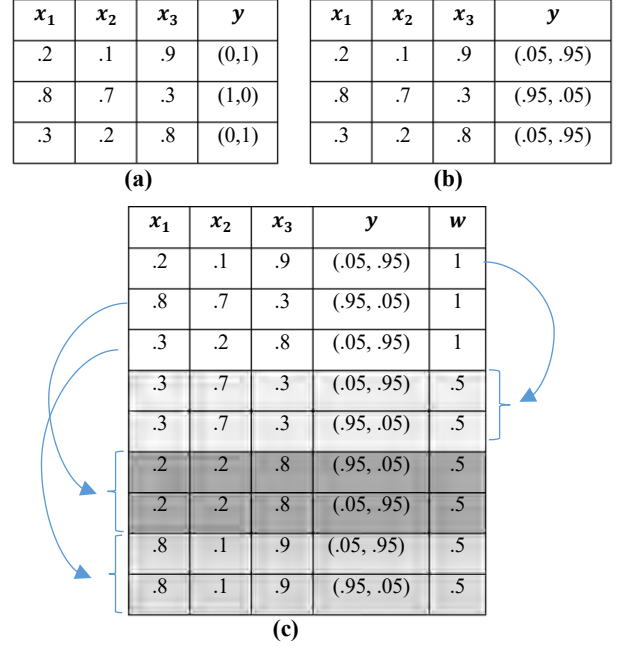


Fig. 1. Example of the examined method.

A. Label Smoothing

The first step of the examined method applies the existing technique of label smoothing [9], which is briefly described in the following for the sake of completeness.

Label smoothing transforms each element, y_i (for binary classification: $1 \leq i \leq 2$), of the one-hot encoded label, y , of each training instance according to the following formula:

$$y_i = y_i(1 - s) + \frac{s}{C}$$

where s is called *label-smoothing factor* and C is the number of classes.

Figure 1(b) depicts the result of applying label smoothing with $s = 0.1$ to the training set of Figure 1(a).² In this example, $C = 2$ (binary classification). The label of the first instance is (0, 1). Thus, its first element, y_1 , is transformed from 0 to .05. Its second element, y_2 , is transformed from 1 down to .95. The smoothed labels represent probabilities that map training instances to the labels with less than 100% certainty, a fact that can reduce overconfidence [9].

B. Outlier Exposure

The second step of the examined method extends the training set, by applying outlier exposure that is performed as follows:

- We replicate all instances of the training set, called T , in an identical copy of it, called T' . (We assume that T has already been transformed by label smoothing in the first step.)
- The values of each feature x_i across all instances of T' are reshuffled randomly. In this way, each feature (i.e., column) of T' is a random permutation of the values in the corresponding feature (i.e., column) of T . As a result,

² To save space (e.g., in tables), the leading zero of positive decimal numbers smaller than 1 is sometimes omitted.

T' comprises OOD instances, which implement the intended *outlier exposure*, since the scrambled feature values of the instances in T' differ substantially from those of the instances in T (nevertheless, their scale remains the same).

- Next, we duplicate each instance of T' . In each resulting pair of duplicates, the first instance maintains its original label, whereas the other instance is assigned to the opposite class and thus gets the inverse label (transformed by label smoothing, too).³
- Finally, we append T' to T and form a single training set. Additionally, we assign weights to each instance of this combined training set. The instances originating from T receive a weight equal to 1 and those originating from T' receive a weight equal to 0.5.

Figure 1(c) exemplifies outlier exposure, by depicting the resulting combined training set as a single table. The shadowed part of this table (i.e., the last six rows) represents the instances originating from T' , whereas the rest (i.e., the first three rows) represents the instances originating from T . The instances of T' comprise pairs with identical feature values but opposite labels (after being transformed by label smoothing). The feature values in T' are randomly permutations of those in T . The mapping between instances of T and T' is presented in Figure 1(c) with side arrows. For example, the first instance originating from T (first row in table of Figure 1(c)) corresponds to the first two instances originating from T' (rows 4 and 5 in the same table). Finally, we see that both T and T' form a single training set in which all instances originating from T' receive a weight equal to .5 (column w in table of Figure 1(c)), whereas those from T have weight equal to 1.

What is the purpose of outlier exposure? It produces a training set that is able to achieve both the following goals:

- Train a model that predicts accurately for new, *in-distribution* instances, since the original training set (T) is preserved, albeit transformed by label smoothing, which in this case acts as a form of regularization.
- Reduce overconfidence, since the model is exposed both to smoothed labels and, especially, to outliers in the form OOD data in T' . The generated outliers expose the model to – by design – *contradicting* labels within each pair of instances in T' . That is why both labels are used instead of, e.g., selecting one of them at random. In this way, the model has the chance to learn that predictions for this kind of instances should be uncertain. To compensate for the fact that the outlier instances are duplicated, we assign them half the weight of the original training instances, so that they get half the importance during training (although this weight might be considered, in general, as a hyperparameter).

What are the preconditions that a machine learning model has to fulfill, so that the examined method can be used to generate a training set for it? First, such a model has to be able to be trained for the task of classification by using smoothed labels. Any model that transforms its outputs with the Softmax function and uses Cross-Entropy as loss function satisfies this precondition. In this sense, a large variety of machine learning models can utilize the examined method, for instance Logistic Regression, XGBoost, as well as Neural Networks for tabular data, such as Multilayer Perceptrons. Other classification models for tabular data, such as k -nearest neighbors, can be easily adapted to predict continuous (i.e., smoothed) labels within the range between 0 and 1 (i.e., transforming them to a regression task). A second precondition is the ability for a model to be trained with instances that have weights. This is a common functionality in most existing machine learning frameworks and the vast majority of algorithms in them can utilize such weights.

IV. PERFORMANCE EVALUATION

In this section, we investigate experimentally the following questions:

- Can the examined method help reduce overconfidence in predictions for OOD data?
- Can the examined methods help preserve the accuracy in predictions for new but in-distribution data?
- What is the impact of the label-smoothing factor, s , on a possible trade-off between reduction of overconfidence for OOD data and preservation of prediction accuracy for in-distribution data?

To this end, we focus on the problem of (binary) classification for tabular data and we examine two representative types of machine-learning models for this purpose: Multilayer Perceptron (henceforth referred to as MLP) and XGBoost (henceforth referred to as XGB).

In the rest of this section, we first describe the experimental setup and then we present the results of the empirical performance evaluation.

A. Experimental Setup

Datasets:

We use 14 datasets from the “tabular data learning benchmark”⁴, which represent diverse classification tasks for tabular data. These datasets have been used to evaluate the performance of tree-based models (such as XGBoost) compared to deep learning on tabular data [4]. Table I shows the characteristics of the datasets.

It is worthwhile to mention that all 14 datasets have numerical features and require no further preprocessing other than standardization, which is necessary in the case of MLP. Moreover, the datasets contain two balanced classes, i.e., the number of instances per class is almost equal.

³ In case of $C > 2$ classes, the same method can be used with the common *one-vs-rest* schema that results to C binary classification problems.

⁴ Downloaded from <https://github.com/LeoGrin/tabular-benchmark> (SUITE_ID = 337)

TABLE I. DATASETS

Dataset-ID	Dataset Name	Number of samples	Number of features
1	credit	16 714	10
2	electricity	38 474	7
3	pol	10 082	26
4	house_16H	13 488	16
5	MagicTelescope	13 376	10
6	bank-marketing	10 578	7
7	MiniBooNE	72 998	50
8	eye_movements	7 608	20
9	Diabetes130US	71090	7
10	jannis	57 580	54
11	default-of-credit-card-clients	13 272	20
12	Bioresponse	3 434	419
13	california	20 634	8
14	heloc	10 000	22

Models:

We implemented XGB based on the XGBoost Python Package.⁵ In order to use label smoothing, we developed a custom loss function based on (Binary) Cross-Entropy. MLP was implemented in Keras (<https://keras.io/>), which supports directly the possibility to use label smoothing. Both implementations can directly use weights for their training instances.

XGB contains a default number of 1000 estimators with early stopping as option, whereas MLP contains one hidden layer consisting of 100 to 500 neurons (tuned separately per dataset) with Rectified Linear Unit (ReLU) activation and L1 regularization. Investigation of different settings for both model types have indicated comparable results.

The application only of label smoothing (LS), considered as a baseline, on the two examined model types is denoted as MLP(LS) and XGB(LS), respectively. The application of the combination of label smoothing and outlier exposure (LS+OE) is denoted as MLP(LS+OE) and XGB(LS+OE), respectively. The original models, without LS or OE, are denoted as MLP(Original) and XGB(Original), respectively. It is also noted that OS as standalone method is outperformed by LS, thus for brevity we examine only their combination.

Metrics:

We evaluate the performance of the examined method according to the following metrics:

- *Maximum (Softmax) Probability, abbreviated as MaxProb*: The maximum value of the predicted class probabilities. Both examined model types, i.e., MLP and XGB, compute a probability per class based the Softmax function. For binary classification, the closer is MaxProb to 1, the

more confident is a classifier in its prediction. In contrast, when MaxProb tends to 0.5, then the prediction has less confidence. Therefore, higher MaxProb values indicate higher overconfidence. MaxProb is considered as simple, yet effective metric of overconfidence in prediction, especially in case of OOD detection [5,12].

- *Prediction accuracy*: The fraction of correct predictions for a test set. Since all 14 examined datasets contained balanced classes, prediction accuracy is an adequate metric for binary classification and there is no need for other metrics, such as area under the curve (AUC).

Experimental protocol:

Performance is evaluated separately for each dataset. We apply 10-fold cross validation. Similar to the generation of additional instances by *outlier exposure*, we create a copy of each test fold and reshuffle randomly the values within each feature (i.e., column). We then measure:

- Prediction accuracy *only* for the original (not reshuffled) test folds; thus these predictions have to be accurate and the expected maximum value for prediction accuracy is 100%. By measuring accuracy for the original test folds, we can assess whether the examined method can help preserve the accuracy in predictions for new but in-distribution data (our second research question in this evaluation).
- MaxProb *only* on the erroneous predictions of the reshuffled test folds. Since these predictions are – by design – for OOD test data and also erroneous⁶, the larger MaxProb is for such data, the more the overconfidence. Therefore, MaxProb with value equal to 0.5 is the expected optimum for these predictions, indicating that we can have no confidence in them. In this way, we can assess whether the examined method can help reduce overconfidence in predictions for OOD data (our first research question in this evaluation).

For simplicity, we examined MLP with a single hidden layer. The tuning of the hyperparameters of the classifiers, both for MLP (e.g., learning rate and L1-regularization coefficient) and XGB (e.g., depth of decision trees) is performed with a separate holdout set that is extracted from each training set. The default label-smoothing factor for XGB(LS) is 0.4 and for MLP(LS) is 0.5.

B. Experimental Results

Results on MaxProb:

Table II compares, for each dataset, the resulting average and standard deviation of MaxProb for MLP(Original),

⁵ <https://xgboost.readthedocs.io/en/stable/>

⁶ We verified that, all predictions for the reshuffled test folds have a 50% accuracy, i.e., they are random.

MLP(LS), and MLP(LS+OE). Table III shows the corresponding results for the case of XGB.

Clearly, Label Smoothing (LS) substantially reduces MaxProb both for MLP and XGB. Additionally, joint Outlier Exposure with Label Smoothing (LS+OE) achieves a further reduction.

TABLE II. MAXPROB FOR MLP

<i>Dataset-ID</i>	<i>MLP (Original)</i>	<i>MLP (LS)</i>	<i>MLP (LS+OE)</i>
1	.75 ± .14	.62 ± .08	.56 ± .04
2	.76 ± .14	.63 ± .08	.57 ± .05
3	.95 ± .11	.69 ± .09	.62 ± .07
4	.88 ± .14	.67 ± .09	.60 ± .06
5	.85 ± .15	.66 ± .09	.60 ± .06
6	.79 ± .14	.64 ± .08	.57 ± .04
7	.89 ± .13	.67 ± .08	.59 ± .05
8	.64 ± .11	.61 ± .09	.56 ± .05
9	.62 ± .09	.55 ± .04	.52 ± .02
10	.85 ± .15	.63 ± .09	.57 ± .06
11	.75 ± .14	.62 ± .08	.56 ± .05
12	.88 ± .14	.65 ± .10	.58 ± .06
13	.86 ± .14	.67 ± .09	.59 ± .06
14	.79 ± .14	.64 ± .09	.57 ± .05

TABLE III. MAXPROB FOR XGB

<i>Dataset-ID</i>	<i>XGB (Original)</i>	<i>XGB (LS)</i>	<i>XGB (LS+OE)</i>
1	.80 ± .14	.66 ± .12	.56 ± .07
2	.81 ± .15	.68 ± .12	.59 ± .09
3	.94 ± .10	.74 ± .09	.60 ± .08
4	.90 ± .13	.71 ± .12	.61 ± .09
5	.86 ± .14	.69 ± .12	.60 ± .09
6	.82 ± .14	.67 ± .12	.56 ± .05
7	.93 ± .11	.75 ± .11	.66 ± .09
8	.64 ± .10	.59 ± .10	.53 ± .05
9	.62 ± .09	.57 ± .07	.52 ± .03
10	.78 ± .14	.66 ± .12	.59 ± .09
11	.76 ± .14	.64 ± .12	.57 ± .08
12	.78 ± .14	.63 ± .11	.57 ± .08
13	.88 ± .14	.71 ± .13	.59 ± .08
14	.78 ± .15	.64 ± .12	.57 ± .07

Remarkably, the standard deviation of MaxProb is also reduced by LS and, especially, by LS+OE, both for MLP and XGB. This fact indicates that the examined methods additionally help avoid extremely overconfident erroneous predictions in the worst cases.

To aid comprehension, Figure 2 summarizes these results for both model types, by illustrating aggregated MaxProb values that are averaged across all 14 datasets. The error bars show the aggregated values of the standard deviation.

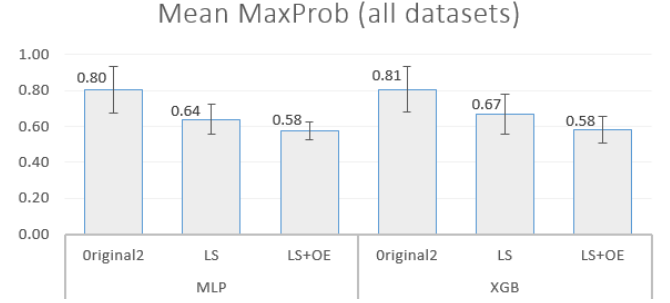


Fig. 2. Results on MaxProb.

XGB(Original) tends to lead to a slightly higher overconfidence compared to MLP(Original). Both MLP(LS) and XGB(LS) attain a large reduction in overconfidence. However, MLP(LS+OE) and XGB(LS+OE) are able to further reduce overconfidence. Overall, it can be mentioned that LS+OE is beneficial for both examined model types and can drastically reduce overconfidence that is originally present in them, as expressed both by the mean and standard deviation of MaxProb aggregated over all datasets.

Results on Prediction Accuracy:

We now turn to the second question in our evaluation concerning prediction accuracy for new, in-distribution data. Figure 3 presents the results of mean prediction accuracy aggregated, for brevity, over all datasets (error bars show the standard deviation).

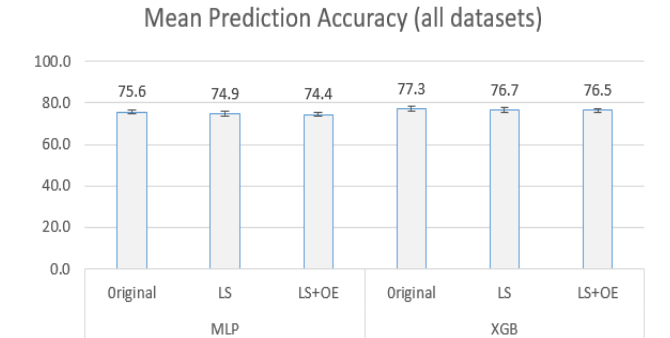


Fig. 3. Results on Prediction Accuracy.

As expected, XGB(Original) tends to have higher prediction accuracy compared to MLP(Original), since it is an established model type for tabular data. LS causes a negligible reduction (less than 1%) for both MLP and XGB, which is within the range of the standard deviation (which is about 1% in all cases). The additional use of OE in the combination LS+OE leads practically to no additional notable reduction in

prediction accuracy. Overall, it can be mentioned that LS+OE is able to reduce overconfidence for OOD data (as shown previously) without significantly affecting the quality of predictions for new, in-distribution data.

Results on the impact of Label-Smoothing Factor:

We have to recall that the aforementioned results depend also on the label-smoothing factor, s , that we use in the examined method. To understand the impact of s , and thus the degree to which we can apply smoothing to the original labels, we measure both MaxProb and Prediction Accuracy for varying values of s .

Figures 4 shows for MLP(LS+OE) the mean values of MaxProb (left vertical axis) and of prediction accuracy (right vertical axis) aggregated over all datasets for varying values of s , ranging from 0.0 (to the left of the horizontal axis) to 1.0 (to the right of the horizontal axis).

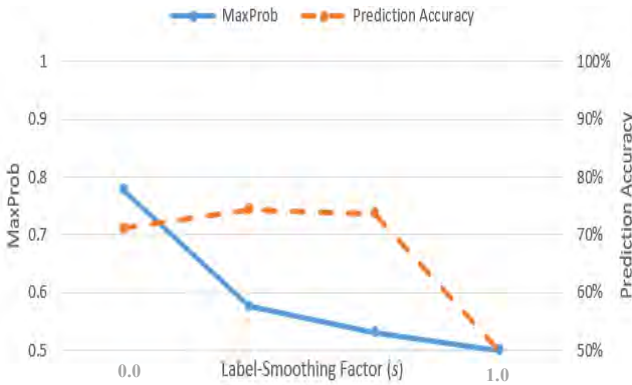
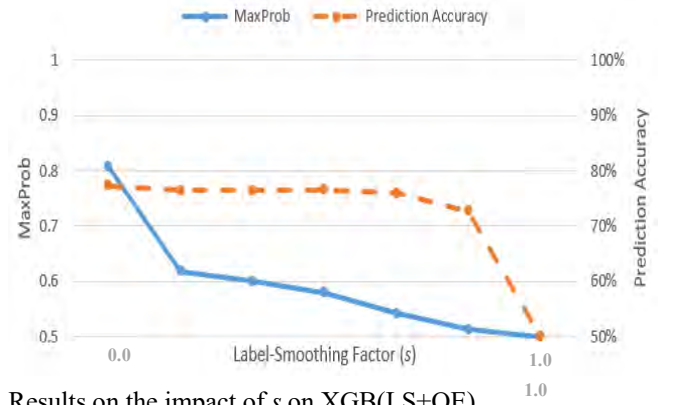


Fig. 4. Results on the impact of s on MLP(LS+OE).

As expected, MaxProb reduces steadily with increasing values of s , and thus reaches the minimum possible value of 0.5, indicating total uncertainty in predictions for OOD data. For intermediate values of s , prediction accuracy is not affected significantly. However, for larger values of s , close to 1.0, prediction accuracy drops suddenly to the level of 50%, i.e., predictions for new, in-distribution data become random.

The reason for this deterioration is that too much of label smoothing “contaminates” the original labels and thus the model cannot effectively learn anything useful out of the training set. Nevertheless, as already mention, when avoiding extreme high values of s , MLP(LS+OE) attains a substantial reduction in MaxProb without harming the prediction accuracy.

The corresponding result for the case of XGB(LS+OE) is depicted in Figure 5. There is a similar tendency as in the case of MLP(LS+OE): increasing values of s lead to a reduction of MaxProb down to 0.5. When s approaches 1, then prediction accuracy drops also drastically, again due to the “contamination” of the label. Nevertheless, for intermediate values of s , XGB(LS+OE) is able to reduce overconfidence and also to preserve prediction accuracy almost at the levels of XGB(Original).



Results on the impact of s on XGB(LS+OE).

V. CONCLUSIONS

In this paper, we focused on the problem of overconfidence in the case of standard machine learning models for tabular data. We have demonstrated that representative classification models in this domain are prone to overconfidence when predicting for OOD data. This result complements similar observations for deep learning models in application domains of computer vision and natural language processing.

We have examined a model-agnostic method that utilizes the techniques of label smoothing and outlier exposure. The examined method has been evaluated experimentally, by exploring two representative classification model for tabular data that are trained with data generated by the examined method. Our results indicate that the examined method can help substantially reduce overconfidence in predictions for OOD data, whereas it preserves prediction accuracy at its original levels in predictions for new, in-distribution data.

As future work, we plan to include the following directions:

- Examination of the performance of the proposed method for imbalanced classification problems.
- Consideration of overconfidence in regression problems for tabular data. It is noticed that the problem of uncertainty estimation in the case of regression models faced with OOD data has started to be examined only recently [13]. Since regression is a task that is frequently used with tabular data (e.g., forecasting problem), it is a promising direction to investigate.
- Another direction of future work is the examination of the applicability of the examined method for specialized deep learning models for tabular data, which are currently gaining attention [14] as well as extensions of tree ensemble methods that generate probabilistic predictions [15].
- Finally, we would like to consider alternative methods for outlier exposure, which can examine various patterns of OOD data.

REFERENCES

- [1] S. Rabanser, S. Günnemann, Z. C. Lipton, “Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [2] K. R. Varshney, H. Alemzadeh, “On the safety of machine learning: Cyber-physical systems, decision sciences, and data products,” *Big data*, vol. 5(3), pp. 246–255, 2017.
- [3] R. Shwartz-Ziv, A. Armon, “Tabular data: Deep learning is not all you need,” *Information Fusion*, vol. 81, pp 84–90, 2022.
- [4] L. Grinsztajn, E. Oyallon, G. Varoquaux, “Why do tree-based models still outperform deep learning on tabular data?” *arXiv:2207.08815 [cs]*, 2022.
- [5] B. Lakshminarayanan, A. Pritzel, C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [6] Z. Li, D. Hoiem, “Improving Confidence Estimates for Unfamiliar Examples,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [7] A. Kendall, Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?,” *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [8] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, “On calibration of modern neural networks,” *International Conference on Machine Learning (ICML)*, 2017.
- [9] R. Müller, S. Kornblith, G. Hinton, “When does label smoothing help?,” *Advances in Neural Information Processing Systems (NIPS)*, 2019.
- [10] S. Liang, Y. Li, R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *International Conference on Learning Representations (ICLR)*, 2018.
- [11] D. Hendrycks, M. Mazeika, T. Dietterich, “Deep Anomaly Detection with Outlier Exposure,” *International Conference on Learning Representations (ICLR)*, 2019.
- [12] D. Hendrycks and K. Gimpe, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *International Conference on Learning Representations (ICLR)*, 2017.
- [13] F. K. Gustafsson, M. Danelljan, T. B. Schön, “How Reliable is Your Regression Model’s Uncertainty Under Real-World Distribution Shifts?,” *arXiv: 2302.03679 [cs]*, 2023.
- [14] V. Borisov et al., “Deep Neural Networks and Tabular Data: A Survey,” *arXiv:2110.01889 [cs]*, 2021.
- [15] T. Duan et al. “NGBoost: Natural Gradient Boosting for Probabilistic Prediction,” *International Conference on Machine Learning (ICML)*, 2020.

An approach to implement user-based recommendation systems with small-sized data sets

Gerhard Götz
Institute for Higher Education and
Educational Research at DHBW
Cooperative State University
Baden-Wuerttemberg (DHBW)
Mosbach, Germany
Gerhard.Goetz@mosbach.dhbw.de

Abstract—A hybrid approach from a content-based to a user-based recommendation system within four steps is proposed. It allows implementing recommendation systems in cases where only small-sized data sets are available, like in educational settings. The hybrid approach is carried out for a recommendation system for mathematical items, which is used in online trainings at university entry level. On the way to build this system, both deep learning simulations and automated extraction of representation knowledge of the recommended items play an important role. After the presentation of some results for the first steps of this approach, a roadmap on, how to proceed is given as well as how these insights may be generalized to other use cases and applications.

Keywords— Recommendation systems, deep learning, AI in education, representation knowledge, collaborative filtering, content-based recommendations, user-based recommendations, small-sized data sets, hybrid system

I. INTRODUCTION

In first instance, recommendation systems are developed and deployed for product recommendations in online retail or in the entertainment industry where finding the right recommendations may have a strong revenue impact. In the last years they appear as well more and more often in the field of education [1, 2, 3, 4, 5, 6], because online learning platforms play an increasingly important role in the mix of learning approaches at universities and other higher educational institutions. Nevertheless, their potential in these educational settings is by far not completely developed. Often purely linear paths [e.g. 7] and processes with low level of variations for the recommendations are used. Especially, the opportunity of individualization of these recommendation systems with respect to the diversity of students is highly desirable, but still in its infancy.

In recent years, deep learning dominates the field of recommendation systems, leading to an improved recommendation quality [8]. Similar to other fields of interest, there are two main types of algorithms used for recommendation systems in educational settings: content-based, user-based (e.g. collaborative filtering) and hybrids thereof. As both of them play an important role in this paper, a small comparison of these two types of algorithmic approaches is provided in table 1. The desirable collaborative filtering requires a reasonable high number of user interactions and hence of participating users, in order to be able to provide high-quality recommendations. As there are

usually a comparably low numbers of users and therefore of user interactions with a recommendation system in the field of education, there turns out the issue of not being able to simply and directly transfer the insights and models which are successfully used e.g. in online retail [8]. As a result, in educational settings, high-quality recommendations may in the first instance not purely be determined by user data and its analysis. Instead, there is the need for integrating additionally insights about the content by e.g. embeddings of language models, knowledge graphs or ontologies in order to improve the quality level of the recommendations.

TABLE I.

Comparison of Collaborative Filtering and Content-based Recommendations		
	Collaborative Filtering	Content-based Recommendations
Basic working principle	Similarities between users are calculated (e.g.: Which users have rated many items equally?)	Generation of categories based on additional prior meta data on items (e.g. pedagogical or didactic models)
Assumption	Similarities between users and items are transferable: Users evaluate items the same way as their neighbors	Interests of members within the categories is similar
Advantages	No prior knowledge on items and users is needed and the system is self-adapting	Reasonably good recommendations are already possible for the first users
Disadvantages	A huge number of users is needed and no good solution for the first users	Categories do not always represent reality and recommendations are based on static models

Comparison of different approaches for recommendation systems

In addition, there remains the so-called cold start problem, which is one of the biggest challenges in machine learning approaches to overcome. It is the dilemma of not having enough data as a starting point at hand in order to be able to use user-based machine learning algorithms appropriately. This challenge increases even more, once one is aiming for implementing algorithms belonging to the field of deep learning, were in general most attempts demand a large initial data set in order to perform well. For an educational recommendation system in particular, this is, apart from

massive open online courses, due to a rather low number of users even more difficult to handle. In addition, caused by rather strict interpretations of the data protection laws at universities, the ethical restrictions on algorithms being used with students are - for good reasons - often much higher than in the private sector. In particular, it is not only important to proceed the collected user data anonymously but much interesting usage information cannot be used as its individual character may allow to identify particular users afterwards. And from a pedagogical teaching perspective, it is as well often automatically imposed, to avoid that first users may not be offered a similar quality of recommendations as later ones. This avoids the situation that the first users might be reduced to the role of pure data deliverers to support mostly the improvement of the algorithm and hence the quality of recommendations for later users.

One common solution of surrounding this challenge is the approach of using an artificial training data set in the beginning. However, there remains the question of how well such a training data set fits a given real problem, which has to be solved. In addition, it needs to be considered how big the risk is of starting to train the recommendation system with a biased training set and ending up with a system showing this intrinsic bias as well [e.g. 9]. For all these reasons, we sketch here a different path by proposing a hybrid solution, in order to employ the combined power of the two integrated algorithm types collaborative filtering and content-based recommendations. In the end, our main goal is to implement a user-based, self-adapting recommendation system. It shall be integrated into an already existing course in basic mathematics to support the students' individual training and learning process at the level of short, elementary exercise items. Based on deep learning, this recommendation system shall at the final stage provide users individually recommendations for each upcoming training item based on their proficiency and the history of their processed items.

Let us emphasize, that there are several reasons why this recommendation system shall not provide the same order of items within a training session to all users like it is the case in e.g. a classical textbook. First of all, a predefined order may on average be good for the majority of users, but it will definitely not be the best one individually. Secondly, the training sessions need to be finetuned with respect to the knowledge and proficiency of each user in order to avoid to train the wrong aspects. Thirdly, the attention span of students is definitely not increasing within the last years so that training sessions need to be as efficient as possible. Therefore, a training session should not take too long time and is hence restricted to a maximum duration of one hour. And students should only train aspects of a topic on which they show an individual improvement potential. As a consequence, not only the individual appropriate aspects need to be selected by the recommendation system, but both, the items which are individually too easy and the ones which are too difficult need to be excluded as they are less likely to be good training items regarding students' motivation. To sum it up from a pedagogical point of view, the trainings should select a next item one which is expected to demand the maximal proficiency without being overwhelming.

II. FOUR-STEP APPROACH TOWARDS AN USER-BASED RECOMMENDATION SYSTEM FOR SMALLER DATA SETS

A. General idea of the four-step approach for a hybrid recommendation system

Even though the goal seems to be building a rather specific recommendation system, the proposed hybrid approach may be understood and used in a much wider sense of applications. Hybrid approaches have been already used for many years in different settings and use cases [10, 11, 12]. There are many situations, where, even though recommendations are strongly desired, no big data sets are available. Given no or a rather small-sized initial data set does in general not allow to start immediately with a collaborative filtering approach. Therefore, one may first begin with a content-based approach which uses the fact that there is a basic understanding and knowledge of the recommended items in many applications available. For an educational scenario, this is likely to be a pedagogical or a didactic model, for novel products in online retail instead, this ought to be key features or insights from former, similar products. As a first step, these insights may result in an initial classification, which one may implement into a first content-based prototype rather easily. Such a first prototype can then be modified within three further steps towards a user-based version. Secondly, the collected user data of this first prototype can be used for AI-simulations leading to additional insights, which may afterwards be implemented into the next version of the prototype in order to refine the underlying recommendation system. Thirdly, it seems to be very promising to understand the recommended items on a more fundamental level by gaining insights in the underlying, structural representation knowledge of the items via deep learning algorithms. This may either refine the defined, initial categories of the first step or add an extra structural level to implement a better clustering. In addition, it allows the system to annotate and cluster the items by itself which supports the automatic integration of new items. As a fourth and final step, all these ingredients are combined to build up a user-based recommendation system being based on collaborative filtering. In the following, we will present more in detail how these four steps look like in practice for a recommendation system inside a training tool for mathematical items, how far this is already realized and what the next steps are. More details on already existing recommendation systems for STEM subjects (science technology engineering mathematics) in general and for mathematics in particular can be found in the literature [13, 14, 15].

B. Implementation of the four-step approach at a recommendation system for mathematical items

The final goal is the implementation of a user-based, self-adapting recommendation system on item level into a training tool for mathematical items. Embedded into an existing learning goal-oriented online course in basic mathematics, its key role is the support of the user's individual training and learning process to ensure relevant and needed proficiencies before the start of their studies. In its final version, based on deep learning algorithms, this system shall provide individually recommendations for each upcoming training item based on the knowledge and the history of the already finalized items of a user considering as well the behavior of similar ones who participated earlier. This will be implemented within four steps via a hybrid approach starting

with a rule-based algorithm based on ontologies. The latter have already been widely used to support algorithms [16, 17], but most of these approaches are rather generic and not at a didactic level.

In our use case, a promising first step consists in the integration of manually generated didactic ontologies into a first prototype of a rule-based recommendation system [18]. This allows providing the recommendation system additional information in hierarchically ordered didactic categories for each fundamental subject like e.g. arithmetic [16, 19, 20]. This step is in general to some extent laborious [21] and has the disadvantage, that this type of rule-based recommendation system does not automatically adapt to new mathematical subjects. Therefore, it implies the need of a refinement of the underlying didactic model to new subjects, which is another good reason for the final goal of a user-based recommendation system. It should be emphasized, that the combination of existing ontologies with user-based approaches provides for certain domains promising solutions [22]. Such a rule-based recommendation system has already been implemented for mathematical training items based on didactic ontologies and they have been evaluated to some extent [18, 23]. Their ontologies are based on two-dimensional models of knowledge and proficiency [19, 20, 24] which allow a classification of all items within a particular mathematical subject like e.g. arithmetic. They are currently available for six basic subjects used in these online courses and their fundamental structure is similar, so that they can be transferred to different subjects. Nevertheless, there is some effort needed, because the exact number of aspects and the specific aspects within the ontologies differs from subject to subject.

The second step consist in the performance of simulations with the collected data set of the first step. This helps both to understand the weaknesses of the first prototype and to detect the improvement potential for an upcoming one. One very interesting and important perception is the number of unfinished trainings and the reasons thereof. As the training sessions can be interrupted and continued at any time, unfinished trainings must be understood as unsuccessful terminations of the training process which need to be avoided. As there can be many particular reasons for unfinished trainings identified, it is very important to understand, why and for which reason they appear, as well as which of them may be directly influenced by the training system itself and how this rate may be reduced. One reason lies in the fact that the online training is offering a path of individual training items which do not provide the right difficulty level. As a consequence, students can get either easily frustrated or bored. To prevent this issue, it is inevitable to predict the probability of a user being able to solve a certain next training item knowing his history of already processed items. Being performed with a certain precision, the result of this simulation can then be implemented into the recommendation algorithm to suggest next items, which offer the right level of difficulty. This means, users receive a path of training items that cover all different aspects of the subject with respect to the given models of knowledge and proficiency. And this path of items increases with its difficulty level, such that the user is just at the limit being able to solve them.

Even if it turns out later, that the difficulty level needs to be chosen differently in order to generate the best training effect, one needs to know the probability of a user to be able

to solve a certain training item anyway. Both, the network structure for this type of simulation and the insights from the very first run of the prototype are shown in figure 1 [25]. There, both a convolutional neural network (CNN) and a collaborative filtering (CF) approach were used. Both were better than the majority baseline with the collaborative filtering approach showing overall slightly better results. Currently, modified simulations are performed with respect to several different aspects on the collected data set. The insights are implemented into the second, refined prototype.

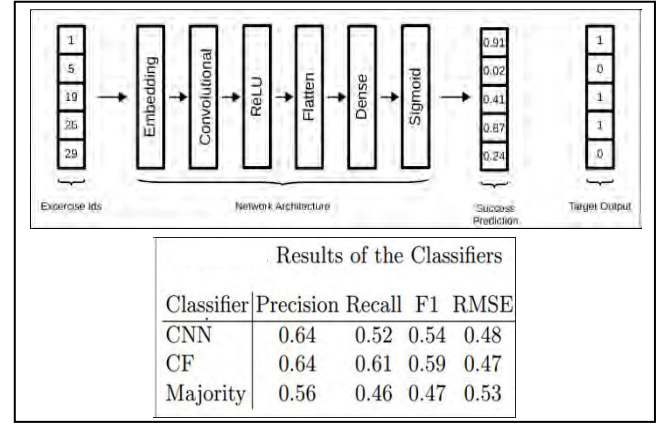


Fig. 1. Structure of the artificial network and results of the classifiers for simulations on the data of the initial prototype [25]

Contemporaneously one can extract representation knowledge of the mathematical items at hand as a third step. This is strongly related to the research question on how semantic representations for recommendation systems may automatized be extracted from items via deep learning approaches. In natural language processing, deep learning has been successfully used to learn language models [e.g. 26]. These types of language representations can be adapted and transferred to more specific domains, like e.g. the language used in mathematical items or in scientific publications [e.g. 27]. Hence, these representations are ideal candidates to improve content-based recommendation systems without a huge domain specific data set, as has been shown for less small data sets in online retail [28]. In particularly for educational applications the deployment of such types of representations in order to calculate a topic-specific classification of items seems to be encouraging. Even though they have not yet been proven applicable to smaller data sets, the recent results on representation knowledge look promising [29]. There, tree recurrent neural networks (Tree RNN), tree long short-term memory (Tree LSTM) and tree smooth activation function (Tree SMU) were performed on mathematical transformations within a set of mathematical items. All of these approaches deliver very good results in comparison to the majority baseline [29]. Moreover, currently, several generalizations extent this analysis to a wider class of transformations.

A second interesting challenge is to find out, if solely looking at the equations within mathematical items allows to tag them automatically towards a certain subtopic [30]. The outcome of different models like bag of words (BoW), sequence-to-sequence (seq2seq), long short-term memory (LSTM), embeddings from language models (ElMo) and bidirectional encoder representations from transformers (Bert) were compared. One of the conclusions was that visual representations showed better result than textual ones as

indicated in table 2. All these results show the potential of these proposed methods: Once, these representations are large enough to cover most of all types of mathematical items, they also allow an automatized annotation of new items, by which the system will become much more flexible than before.

TABLE II.

Model	Precision (Prc), Recall (Rcl) and Accuracy (Acc) ^a			
	Majority Baseline	Tree RNN	Tree LSTM	Tree SMU
All	0,307	0,754	0,832	0,838
Unrelated Prc		0,735	0,735	0,722
Unrelated Rcl		0,648	0,718	0,754
Unrelated F1		0,689	0,726	0,740
Equal Prc		0,679	0,853	0,874
Equal Rcl		0,659	0,860	0,850
Equal F1		0,669	0,857	0,860
Derivative Prc		0,842	0,853	0,862
Derivative Rcl		0,799	0,810	0,820
Derivative F1		0,820	0,830	0,840
Permutation Prc		0,732	0,900	0,921
Permutation Rcl		0,924	0,984	0,971
Permutation F1		0,817	0,940	0,950

^a. Results on representation knowledge on mathematical formulas [30]

Fourthly, the final goal is the implementation of a user-based recommendation system, which is developed from the original content-based to a user-based machine learning approach. It includes all insights from the three steps before. In detail this means, the final system is supported by automatized generated representations of mathematical items explained in the third step above, so that the ontologies are refined and new items can be integrated automatically. These modified user-based ontologies allow to provide didactic subtle suggestions even for already smaller numbers of users and hence shows a hybrid character. This hybrid approach shall on the other hand as well embed the insights of the collected user data like average success rate in solving items, individual evaluation of the level of difficulty of items, similarities between different users and the time being used for solving a certain item as well as the prediction results from the simulations. Before this final step four can be implemented, the other three steps need to be finalized first.

III. SOME RESULTS OF THE FIRST PROTOTYPES IN ACTION

A. Description of the data set of the prototype for a recommendation system for arithmetic items

The rules-based prototype of step one has been used for three years for freshman students in order to collect user data. It should be kept in mind, that these users received already reasonably well recommendations due to the use of didactic ontologies. This data was discussed for each year separately within the first years and was compared with each other [31, 32, 33] showing no big difference in the student cohorts. As

the target group of users is similar each year, we start with a different approach by analyzing +the collected data for the last three years grouped together in a single, larger, combined data set. The recommendation system is used for the six fundamental subjects: arithmetic, equations, powers/roots/logarithms, functional connections, geometry and trigonometry. In order to keep it more manageable, we focus herein solely on one of them, namely arithmetic, where most items have been processed. In the corresponding 795 training sessions for arithmetic, of which 526 were completed, there were in total 19518 processed mathematical items. Let us emphasize here that finalized trainings provide between 26 and 33 answered items in total. This means, the exact number of items to solve within a training session is not fixed, instead it depends on the user's performance, on average its exact number was 30.49. This reason behind this is rather simple: If the performance is very good, a user does not need a maximum of items to solve. The same is true for very weak users: instead of just continuing to train it is more suitable to repeat the topic more profoundly first before they restart solving training items.

The 269 not finalized training sessions are stopped after an average of 12.93 processed items, which corresponds to roughly 40% of a finalized training. One may raise the question, if there should be a lower limit of processed questions within a training session in order to count them as a performed training and consider them for the evaluation. For this reason, we compared the basic statistical parameters for the different user groups with a minimum of respectively two, three, four and five processed questions. It turned out that the differences, which appeared after taking out the trainings with less than three, four or five items of the full set were comparably small. Therefore, the 269 trainings sessions which have not been finalized contain the ones consisting of only two processed items and more. Understanding the full set of reasons, why students stop training sessions earlier without finishing them later, is definitely one very interesting question to answer. However, finding a satisfying answer in full detail would need the evaluation of quantitative and qualitative student surveys, which goes beyond the scope of this work.

B. Some user statistics of the prototype for a recommendation system for arithmetic items

Due to a strict interpretation of the law of data protection, the data set consists of the following restricted information being gathered during the use of the recommendation system: sequence of processed items, time needed per item and the precise answer to all processed items. In addition, the users were asked to choose their subjective perception of the level of difficulty of each item on a four-point Likert-scale, where a higher number is interpreted as more difficult. Next to the consideration of the average success rate of users (proportion of correct solved items) and the information if a particular item was solved correctly, it seemed to be interesting to gather as well the individual user's perception on the level of difficulty. This allows to verify, if this subjective perception is related to any of the other collected data. The detailed results are presented in table III. Please note that it turns out at this stage to be important to cluster the training sessions into different categories. One, rather simple and artificial split of all finalized trainings is done with respect to the average success rate of 70% of solving the items. Even though, normally one would consider different values like e.g. 50% as a reasonable number, nevertheless these are fundamental

aspects of mathematics, which play a crucial role for further lectures, so 70% seems to be more reasonable to expect.

TABLE III.

	Training sessions for arithmetic			
	<i>Finalized^b</i>	<i>Successfully finalized</i>	<i>Not successfully finalized</i>	<i>Not finalized</i>
Number of training sessions	526	345	181	269
Average success rate in %	72.0	81.1	54.7	57.9
Perception of the item difficulty (4-point Likert-scale)	1.87	1.78	2.04	2.05
Average time needed per item in seconds	99.10	101.21	95.10	98.05

^b. The set of finalized training session has been split into two subsets, successful (average success rate bigger than 70%) and not successful ones (average success rate smaller than 70%)

First, considering that these trainings are mainly used as additional voluntary training options for students, they are reasonably well used. Not surprisingly, it turned out in the past that the usage rate is the better, the higher the involvement degree of a teaching person is. The average time having spent on one item is between 95 and 101 seconds, which is a reasonable time spent on elementary items. Nevertheless, the values for all different groups are so close to each other that it may not be a good parameter for understanding the users better. Comparing the average success rate, it turns out that there seems to be a difference in the mean value between 72.0 percent for finalized and 57.9 percent for not finalized trainings. Unfortunately, in both groups the variety is so high, that the difference is statistically not significant. So due to the fact that these clusters of users turn out to be not refined enough, the correlation between average success rate and average subjective perception of the level of difficulty was unfortunately not as high as expected. This indicates the need of a refinement of the user clusters. As this data set is rather new, there remain many simulations to be done with, which cannot be presented in this work yet. The current goal will be an extension of the analysis of the prediction of students' success [25] to this much larger data set.

C. Summary of the current status of the recommendation system

A four-step approach towards a hybrid user-based recommendation system was presented. The first step of a content-based system with prior ontologies based on didactic models of proficiency is already finalized and used to generate user data. For the subject arithmetic, there is currently a data set of more than 19.000 processed items within 795 training sessions available, which is still rather small for deep learning applications. Nevertheless, for an educational application this is not such a low number which will be even increased this autumn. In a second step, the available data is currently analyzed. The goal is an improved understanding of the performance of the system and the optimization of the impact of the implied ontologies. One important step is definitely to

find appropriate clusters of users showing very similar behavior. In addition, the data is used for further simulations e.g. on predicting next items within a training path. This step demands some further analysis, in order to gain the maximum of information and needs some refinement of the corresponding algorithms, which is currently done. Methods being used hereby are LSTM, transformer architectures, CNN and BoW which are evaluated against the random and majority baselines. Regarding the third step, consisting in automatically generated representation knowledge, we gained interesting new insights on the level of the formulas being used within the items to be solved but would love to add representations with respect to the language being used as well. In addition, there has been good progress in automated classification of items by the embedded formulas.

IV. NEXT STEPS AND APPLICATION TO OTHER USE CASES

A. Next steps for the prototype

In addition to an extensive data analysis there will be a focus on gaining novel insights via simulation (LSTM, transformer architectures, BoW and CNN) with this new data set. The first trials being performed on a subset look so far promising but indicate that the full set is needed to get good results. With respect to the representation knowledge, the next step is finding additional representations by using existing language models and adapting them to the particular type of mathematical language. We are confident, that this will provide additional insights on the items, which may afterwards be used for the final hybrid recommendation system. This leads to the fourth step, implementing all insights of the user data and the representation knowledge into a hybrid system, which is still some way to go. One promising and interesting further development on the didactic side would be the precise analysis of the item trajectories of users. This allows a more fundamental understanding of the individual learning process. As a result, the recommendation system may be adapted accordingly. Especially from an educational point of view, there have been pioneering insights generated concerning the design of learning trajectories within the last years. This includes in particular as well the diagnosis of learning success of learning trajectories in order to optimize the later [34, 35].

B. Using this approach for different applications within and outside the field of education

At first, this proposed approach seems to be very particular for a dedicated educational setting. However, this is not true, because this four-step approach is adaptable step by step. Indeed, rather close applications would be training tools for different educational settings. This could be learning and training tools at universities, other higher educational institutions, courses for other stem subjects or online trainings for employees at companies. But it goes even beyond: any use case where a recommendation system with a small data set and a certain pre-knowledge about the recommended items is suitable. In most of these applications, the initial situation looks very similar to the one being presented here. On the one hand, there is few data available at the very beginning. And from a pedagogical point of view, neither a randomly acting recommendation system to primarily collect user data nor the use of an artificially generated data set with possible biases

would be a good option to choose in an educational setting. On the other hand, often there are already pedagogical or didactic models at hand, which could build a reasonable foundation of an ontology to be implemented in a very first prototype. Collecting and analyzing data of such a prototype would work in a similar way as for the presented system: data may be gathered while early users are already pedagogical reasonable recommendations are provided. On the level of additional insights via representation knowledge, there might be the biggest difference to the system being presented, because they are domain-specific. Nevertheless, for many subjects, at least either formulas or text play an important role, so that similar approaches to the ones being explained above may be used.

Once one leaves the field of education, life may look more challenging. Nevertheless, there are some good reasons, why this approach would be desirably helpful in this case as well. Firstly, any automatized learning system has to start and to cope with the cold start problem. Big online retail platforms may easily come quickly to the point where artificial intelligence can be deployed to full extent, but what if this is not the case for very specific platform with less customers and the number of interactions is not large enough. This indicates that there are definitely many applications with no big data set being available, especially in the case that a company enters a new market or offers a new type of product online. Or imagine all the recommendation systems which are needed for a digital factory to run. Given these circumstances, one has then to verify if there is any prior information about the recommended items that may allow a pre-clustering which is at the same time easily implementable as an ontology into a prototype. For online retail this would be information about the products and for the digital factory the knowledge and the insights on the machines being used. This procedure corresponds to the first step of implementing an ontology. Once this content-based prototype is running and recommending items, data analysis and simulations on the thereby collected data is then rather similarly performed to the educational case and corresponds to step 2. In this scenario there might be even the advantage, that the number of user interaction increases in these settings much faster than for the usual educational systems.

The third step, to deal with representation knowledge may sound rather abstract for applications in industry and online retail, but visual and language models are already very well developed for these use cases. These models often still have to be trained, refined and adapted to the specific given situation. As a concluding remark one could say, that this proposed four-step process is rather close to the natural learning process which is often based on basic rules in the beginning and gets stepwise modified with increasing experience.

ACKNOWLEDGMENT

The author would like to thank Sebastian Wankerl for many fruitful discussions on recommendation systems and Dr. Myriam Hamich and Prof. Dr. Guido Pinkernell for a very interesting didactic exchange on models of knowledge and proficiency.

REFERENCES

- [1] H. Drachsler, K. Verbert, O.C. Santos and N. Manouselis: "Panorama of recommender systems to support learning". In: *Recommender systems handbook*, pp. 421–451. Springer (2015).
- [2] O. Zawacki-Richter et al., "Systematic review of research on artificial intelligence applications in higher education—where are the educators?" In: *International Journal of Educational Technology in Higher Education* 16 (1), 3. 2019.
- [3] F. Ouyang, L. Zheng and P. Jiao, Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, 27(6), 7893-7925. 2022.
- [4] X. Chen, D. Zou, H. Xie, G. Cheng and C. Liu, Two decades of artificial intelligence in education. *Educational Technology & Society*, 25(1), 28-47. 2022.
- [5] X. Chen, H. Xie, D. Zou and G. J. Hwang, Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100002. 2020.
- [6] M. C. Urdaneta-Ponte, A. Mendez-Zorrilla and I. Oleagordia-Ruiz, Recommendation systems for education: Systematic review. *Electronics*, 10(14), 1611. 2021.
- [7] P.A. Henning et al., *Learning pathway recommendation based on a pedagogical ontology and its implementation in moodle* (pp. 39-50). 2014.
- [8] S. Zhang et al., „Deep learning based recommender system: A survey and new perspectives.“ *ACM Computing Surveys (CSUR)* 52.1 (2019): 5.
- [9] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang and X. He, Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3), 1-39. 2023.
- [10] P. B. Thorat, R. M. Goudar and S. Barve, Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4), 31-36. 2015.
- [11] W. Chen, Z. Niu, X. Zhao and Y. Li, A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, 17, 271-284.2014.
- [12] A. Klačnja-Milićević, B. Vesin, M. Ivanović and Z. Budimac, E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & education*, 56(3), 885-899. 2011.
- [13] T. Kabudi, I. Pappas and D. H. Olsen, AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2, 100017. 2021.
- [14] W. Xu and F. Ouyang, The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *International Journal of STEM Education*, 9(1), 1-20. 2022.
- [15] G. J. Hwang and Y. F. Tu, Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics*, 9(6), 584. 2021.
- [16] J. K. Tarus, Z. Niu and G. Mustafa, Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. *Artificial intelligence review*, 50, 21-48. 2018.
- [17] N. W. Rahayu, R. Ferdiana and S. S. Kusumawardani, A systematic review of ontology use in E-Learning recommender system. *Computers and Education: Artificial Intelligence*, 3, 100047. 2022.
- [18] G. Götz and S. Wankerl, „Adaptives Online-Training für mathematische Übungsaufgaben“. In: *Pinkernell, G., & Schacht, F. (Hrsg). Digitale Kompetenzen und Curriculare Konsequenzen. Tagungsband der Herbsttagung des Arbeitskreises Mathematikunterricht und digitale Werkzeuge vom 27. Bis 28. September 2019 an der Pädagogischen Hochschule Heidelberg*. Franzbecker Verlag. (S. 85-96), 2020.
- [19] G. Pinkernell, C. Düsi and M. Vogel, „Aspects of proficiency in elementary algebra.“ In *Proceedings of CERME 10*. 2017, pp 464–471
- [20] D. Schönwälder, „Grundlegendes Wissen und Können im Bereich der Sekundarstufenarithmetik am Übergang Schule – Hochschule“. Verlag Franzbecker Hildesheim, 2022.
- [21] S. Staab and R. Studer (Eds.), *Handbook on ontologies*. Springer Science & Business Media. 2010

- [22] S. E. Middleton, H. Alani, N. R. Shadbolt and D. C. De Roure. 2002, "Exploiting synergy between ontologies and recommender systems". In *Proceedings of the 3rd International Conference on Semantic Web*.
- [23] G. Götz, M. Hamich, G. Pinkernell, D. Schönwälder, D. Ullrich and S. Wankerl, „Adaptives Üben, adaptive Aufgabentrainings, Modelle grundlegenden Wissens und Könnens“. In *Selbststudium im digitalen Wandel* (pp. 93-126). Springer Spektrum, Wiesbaden; 2020.
- [24] D. Schönwälder, G. Pinkernell and G. Götz, „Aspects of basic knowledge and comprehension in secondary school arithmetic“. In: Jankvist, U. T.; Van den Heuvel-Panhuizen, M. & Veldhuis, M. (Eds.): *Proceedings of the 11th Congress of the European Society for Research in Mathematics Education*. CERME 11: Utrecht University, the Netherlands, 6.–10.02., 2019, 4833–4834.
- [25] S. Wankerl, G. Götz and A. Hotho, "Solving Mathematical Exercises: Prediction of Student's Success". In: Jäschke R. & Weidlich M. (Eds.) *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen" (LWDA 2019)*, Vol 2454, 190-194.
- [26] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019
- [27] L. Hettinger, A. Zehe, A. Dallmann and A. Hotho (2019). „EClaiRE: Context Matters! – Comparing Word Embeddings for Relation Classification“. In: *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft*. pp. 191-204.
- [28] G. Cenikj and S. Gievska, "Boosting Recommender Systems with Advanced Embedding Models". In *Companion Proceedings of the Web Conference 2020*.
- [29] S. Wankerl, A. Dulny, G. Götz and A. Hotho, "Learning Mathematical Relations Using Deep Tree Models," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 2021, pp. 1681-1687.
- [30] S. Wankerl, G. Götz and A. Hotho, "f2tag - Can Tags Be Predicted Using Formulas?" In: *19th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 565-571, 2020.
- [31] G. Götz, „Automatisierte, adaptive Aufgabentrainings“. In *Hans-Stefan Siller, Wolfgang Weigel & Jan Franz Wörler (Hrsg.), Beiträge zum Mathematikunterricht 2020 (S. 341–344). Münster: WTM-Verlag, 2020*.
- [32] M. Brüstle, G. Götz and M. Hamich, "Virtual Inverted-Classroom Courses vs. Face-to-Face Courses in German Higher Education: Comparing Students' Learning Progress and Perspectives on Mathematics Preparatory Courses". *Handbook of Research on Teacher and Student Perspectives on the Digital Turn in Education*, pp. 49-72, 2022.
- [33] G. Götz, „Automatisierte Aufgabentrainings – Unterstützung des Lernprozesses durch ergänzende Onlinetrainings?“ In: *Digitales Lernen in Distanz und Präsenz: Herbsttagung 2021 des Arbeitskreises Mathematikunterricht und digitale Werkzeuge in der Gesellschaft für Didaktik der Mathematik* am 24.09.2021. S. 49–56, 2022.
- [34] R. C. I. Prahmana, Y. S. Kusumah and Darhim, "Didactic trajectory of research in mathematics education using research-based learning". *J. Phys.: Conf. Ser.* 893 012001. (2017)
- [35] M. Fahlgren, M. Brunström, F. Dilling, B. Kristinsdóttir, G. Pinkernell and H. G. Weigand, „Technology-rich assessment in mathematics“. In *Mathematics Education in the Digital Age* (pp. 69-83). Routledge. 2021.

A stacking approach for vehicle loan fraud detection

Jan Wolf

Wirtschaftsinformatik – Data Science,
DHBW, Stuttgart,
Stuttgart, Germany,
janrwolf@icloud.com

Abstract— Nowadays it is common for consumers to finance expensive goods using loans. Especially so in the vehicle industry. Simultaneously, vehicle thefts are high and the resulting damages are tremendous. Thus, it is important for financial institutions offering vehicle loans to secure their assets and minimize their exposure. While there are numerous studies about the prevention of fraud in other domains, there has been little research on the detection of vehicle loan fraud. In order to address this problem, this paper proposes a stacking approach for detecting vehicle loan fraud. The stacked model consists of ten different base learners and uses a Balanced Random Forest as its meta-model. The results of this paper show, that using the proposed method, one can expect an increase in classification performance of around 1.4 percent.

Keywords— machine learning, stacking, super learner algorithm, fraud detection, vehicle loan fraud, imbalanced data

I. INTRODUCTION

In 2021 there were 30,952 reported cases of vehicle thefts. Around half of them are classified as permanently non retrievable [1]. According to the German Insurance Association this led to damages of 187 Mio. Euros [2]. Simultaneously the current vehicle market is heavily dependent on consumer loans. In Germany around 47% of all newly bought vehicles are financed via loan or leasing [3]. Thus, the theft of such vehicles leads to serious economic damages for financial institutions [4]. Furthermore organizations, specifically financial institutions, are often obligated to have appropriate measures against financial crimes in place. Having legal regulations and wanting to ensure long-term sustainability, it is therefore essential for organizations to employ appropriate fraud prevention measures [5].

The usages of machine learning (ML) are very broad and consequently modern business use-cases often build upon data mining techniques. One frequently researched use-case is the detection of credit card fraud [6]. While in general being closely related to the detection of loan fraud, the underlying data and concepts often differ. There have been numerous approaches to identify fraudulent transactions using machine learning techniques such as neural networks, support vector machines, Naïve Bayes Algorithms or stacking approaches [7], [8], [9], [10].

However, there are multiple issues hindering the successful detection of fraud. One such issue being the skewed data distribution. In financial transactions usually only a small percentage of all transactions are fraudulent [11]. Having pre-labeled data, this results in very imbalanced class distributions. For example, out of a hundred transactions only one transaction might be fraudulent. There are multiple challenges to overcome with imbalanced data such as the measurement of the actual performance or the model underfitting the minority class. Another issue in the fraud domain is concept drift. Concept drift refers to a change in the underlying

concepts represented by the data [6]. While these issues will be addressed later, the focus of this paper will be on the evaluation of stacking in the fraud prevention domain as well as the comparison with different ML models for detecting fraud in vehicle loans.

Stacking is the process of combining multiple models on at least 2 levels. All models on level-1 will predict the outcome of a given sample, whereas the level-2 model will use the level-1 predictions to get a final prediction [12]. The underlying idea is, that stacked models are prevalent in giving accurate predictions as they are able to compensate for the weaknesses of their base learners.

II. RELATED WORK

A. Fraud Detection

To the best of the author's knowledge research specifically relating to vehicle loan fraud is relatively scarce [9], [10]. One proposed method specifically focusing on vehicle loan fraud prediction, was able to achieve an accuracy score of 95% using Naïve Bayes Algorithm [9]. While these results seem to be promising, it is important to mention that measuring classification results via accuracy score for highly imbalanced datasets might be misleading. This problem will be addressed in detail in section II-E. The analyzed study uses 34 input features out of which a large selection is also present in the dataset used for this paper. However, none of the identified papers researches the application of stacked models for vehicle loan fraud prevention.

Most of the identified literature, focusses on related areas such as fraud detection for consumer loans or credit card transactions. It is worth noting that some of them use similar data, hence some of the features for identifying fraud seem to be widely the same across different fraud domains. Currently there are multiple broad literature reviews such as Abdallah, Maarov, Zainal [6] and Omair, Alturki [14], targeting various financial fraud domains and methods to detect fraud within those. There are numerous effective approaches to detect fraudulent transactions, with the vast majority of them being supervised learning techniques. For example, Wen and Huang examined various algorithms for personal loan fraud detection resulting in a variation of XGBoost to be the most accurate one [8]. Having various ML models applied, two approaches also assessed stacked models [11], [15]. In both proposed methods the stacked model was showing promising classification results.

B. Concept Drift

Concept drift describes a practical problem, in which underlying concepts represented in data are changing over time [6]. For example, some fraudsters might realize their attempts do no longer yield any fruits e.g. a deployed model always identifies them. Thus, they change their strategies, resulting

in a behavioral shift and therefore in a change of concept. The formal definition for concept drift between two points in time t_0 and t_1 can be described as $p_{t_0}(X, y) \neq p_{t_1}(X, y)$, while p represents the joint distribution between the input variables X and the target variables y [16].

The majority of all ML-Algorithms are based on the assumption that data is static [17]. Furthermore, it can be very tedious to identify and handle concept drift in practice as the model needs constant monitoring. In general, there are two methods to deal with this issue [16], [18], [19].

1. Active methods: Active methods refer to all methods, where the model is constantly monitored to identify any changes. In case any changes are detected the model needs to be updated.

2. Passive methods: In a passive approach the data as well as the model are constantly updated. This means in case there is any concept drift, the model should automatically pick up the behavioral change.

C. Imbalanced Data

Imbalanced data refers to highly different class distributions. For datasets with binary labels, this results in the minority class being much sparser than the majority class. Due to its frequent occurrences in various domains such as health care, anomaly detection and fraud detection, imbalanced data is a widely researched area on its own [20], [21].

Imbalanced datasets lead to a variety of problems when performing modeling tasks on them. These include but are not limited to algorithms that aren't capable to correctly fit both minority and majority class, high dimensionality as well as small data volumes and misleading metrics being used for evaluation [22]. Mainly, there are three different approaches to tackle these issues – data preprocessing methods, algorithmic methods, and hybrid methods [23].

Data preprocessing methods focus on reducing class imbalances before training any models on the selected dataset by changing the data distribution. As such, using resampling techniques a dataset can either be oversampled (e.g. by randomly duplicating instances of the minority class) or under-sampled (e.g. by randomly deleting instances of the majority class). There have been numerous variations of resampling techniques such as SMOTE [24] and ADASYN [25]. It is also possible to combine different resampling techniques, which is often referred to as hybrid data preprocessing methods. If data resampling methods are used, it is important to only resample the training data to avoid bias in the test set and the evaluation.

Algorithmic methods do not alter the data but instead adapt the learning algorithm [26]. Often cost sensitive error functions are used to change an algorithm to be more sensitive towards the minority class [23], [27].

Lastly, it's possible to combine data preprocessing methods and algorithmic methods into **hybrid methods**. For example, one could resample the data to be less imbalanced and then use cost sensitive algorithms to further compensate any imbalances.

D. Stacking

The original method of stacked generalization was proposed by Wolpert in 1992 and later extended for regression by Breiman in 1996 [28], [29]. Stacked models are a type of

ensemble models. However, contrary to most ensemble methods, instead of combining weak learners in stacking a selection of strong learners is used. To combine the predictions of the selected models a separate metamodel is used. The fundamental idea behind stacking is, that a metamodel can learn from the weak spots of its base learners. In stacking meta-learning is enabled by aggregating all base learners (first level predictors) as well as their predictions and using those predictions as input data for a second level predictor [29]. Using the first level predictions as training data for the meta-model, it is also possible to include additional features into the meta-data [30]. Furthermore, it can be beneficial to have vastly different level-1 predictors, creating a heterogeneous set of classifiers [15]. Stacking is also known as Super Learning, as in 2007 Van der Laan, et al. proved that a meta-learner performs asymptotically as well as the oracle selector amongst the initial set of candidate learners [31].

Formally in stacking by applying K-fold cross validation a finite set M of predictors c_0, \dots, c_M generates predictions for an entire dataset of length n . This process results in a matrix of shape $n \times M$ including the predictions a_t of each base learner. After appending the target variable to the matrix turning it into a matrix of shape $n \times (M + 1)$, the matrix can be used to train the meta-model. Formally the meta-learner needs to minimize the Loss \mathcal{L} , across all possible meta-parameters θ , with respect to the input values of the base learners and, if applicable, any additional values x_t forwarded to the meta-learner.

$$\min_{\theta} \mathcal{L}(a_t, x_t) \quad (1)$$

Instead of using discrete class predictions or binary values as features in the meta-data one might consider using continuous probabilities [32]. Many models offer the possibility to output such class probabilities by default. As such a prediction Δ of any classifier can be represented as a vector of shape $1 \times c$, where c denotes the number of classes and δ_c represents the predicted probabilities.

$$\Delta_m = [\delta_{1_m}, \dots, \delta_{c_m}] \quad (2)$$

In the case of binary classification, it is sufficient to use either one of those probabilities, as $\delta_{total} = 1$. Thus, $1 - \delta_1 = \delta_2$, which results in no change in shape for the original meta-data. For selecting the best meta-model, to the best of the authors knowledge, there is currently no sophisticated method.

E. Evaluation metrics

To ensure a bias-free evaluation it is important to compensate for the class imbalances in the test dataset [33]. While it is important to run tests on a large enough portion of the original dataset, many evaluation metrics can induce bias when applied on imbalanced datasets. One such metric is the accuracy score. It is calculated by the sum of sensitivity (recall) and specificity.

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (4)$$

As most metrics are based on combinations of the confusion matrix, TP denotes true positives, TN true negatives, FP false positives and FN false negatives. Suppose a dataset has 1000 samples with 990 instances of class A and 10 instances of class B. If an algorithm now classifies all samples to be class A, the resulting accuracy score would be 99%. In this case class B would be completely underfitted by the algorithm, rendering evaluation via accuracy score useless [34].

Consequently, numerous attempts were made to evaluate algorithms on imbalanced datasets. One such metric is balanced accuracy. Balanced accuracy (BA) calculates the arithmetic mean of the accuracy scores for all classes [35].

$$\text{Balanced Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2} \quad (5)$$

Using the previous example BA would return 50%, which is a much more accurate representation of the actual performance of the classifier. Similarly, it is possible to calculate the geometric mean using G-Mean [36]. As both metrics represent classification performance using a mean both can be applied to highly imbalanced datasets.

$$G - \text{Mean} = \sqrt{\text{sensitivity} * \text{specificity}} \quad (6)$$

Two graphical measures are the receiver operating characteristics curve (ROC) and the Precision-Recall-Curve (PRC). The ROC shows the relationship between sensitivity and specificity for different thresholds, while the PRC displays different thresholds for sensitivity and precision [37], [38]. To summarize the ROC the area under the ROC (AUC) can be calculated by using integration as follows.

$$AUC(f) = \int_0^1 ROC(k) dk \quad (7)$$

Lastly, in this paper Matthew's correlation coefficient (MCC) will be used, which measures the correlation between the predicted and the actual values. Therefore using MCC a high score, can only be obtained, whenever a classifier achieves good results in all four confusion matrix categories [35]. While all other metrics range from 0 to 1, MCC ranges from -1 to 1.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (8)$$

By default, it is reasonable to evaluate any classifier using multiple methods to cover various aspects of the given classifier.

III. PRELIMINARIES AND METHODOLOGY

The following section describes the most important specifications of the used dataset as well as the data preprocessing and resampling methods being used. Furthermore the development methodology is broadly described.

A. Fraud Dataset

This paper is based on a dataset provided by a third entity. It contains 569,488 samples of loan contracts with each sample having 241 features. The feature mostly consists of contract related information, such as type of vehicle, amount of

down payment or exposure. Furthermore, additional information related to the borrower were included, which was mostly provided by credit agencies. Within the dataset, contracts in whose term a fraud was committed are already labeled fraudulent. As such 5,601 fraudulent cases were recorded, while the remaining 563,887 contracts are non-fraudulent. This results in a highly imbalanced datasets with class distribution between minority and majority class of roughly 0.984% to 99.016%.

Taking the high dimensionality into account, it was necessary to reduce the number of input features. To accomplish this, relevant features were extracted with the help of p-value analysis as well as under careful consideration of various scientific articles such as [9], [39]. This resulted in 35 relevant features, which were used in modelling.

B. Data preprocessing

Most classifiers available require variable encoding before modeling is possible. Thus, categorical variables were encoded via one-hot-encoding. However, to avoid the curse of dimensionality, all features with a feature level above 10 were encoded binary. This is especially important as the number of samples required to train an algorithm will grow exponentially with respect to the number of features. After transforming the input features the entire dataset is split into 80% training data and 20% test data.

In section II-C, methods to address the imbalanced data problem were discussed. Consequently, random under-sampling, random oversampling as well as hybrid resampling techniques were tested on the training dataset using all classifiers, later integrated into the stacking architecture. It is important to only resample the training data to ensure a bias-free evaluation. The results of the resampling process indicated that most classifiers performed best when undersampling the majority class at a 0.1 rate. Nevertheless, it should be noted that further investigation could potentially yield additional performance gain.

C. Methodology

The aim of this paper is to develop a stacking architecture for detecting vehicle loan frauds as well as to analyze whether stacking models can bring benefit to the classification results in the domain of vehicle loan fraud prevention. To accomplish this goal in section IV the proposed method is described. The architecture of the stacking model is derived from the literature and tweaked within an evolutionary prototyping process [40]. After describing the proposed method, evaluation will take place using the metrics described in section II-E. To decide which algorithm performed the best, a majority vote is taken based on the defined metrics as well as the graphical measures.

IV. PROPOSED METHOD

In this section, firstly the selection as well as configuration of the applied base learner is discussed. Afterwards the stacking architecture in its entirety is described, including the meta-model choice.

A. Base learner selection

Creating stacked ensemble models, it is important to choose different base learner algorithms to optimally exploit the advantages of meta-modeling. That being said, there is no optimal way of choosing the best set of candidate learners. To

have some indications previous methods were analyzed and following algorithms were identified to be used frequently: neural networks, decision tree-based algorithms, logistic regression, SVM, Naïve Bayes as well as k-nearest neighbors [6], [12], [15], [41]. Consequently, the list of classifiers used contains the following seven algorithms based on scikit-learn implementations: Multi-layer Perceptron (MLP), Support Vector Machine, Gaussian Naïve Bayes (GNB), Gradient Boosting Classifier (GBM), Logistic Regression (LR), Random Forest (RF) as well as k-nearest neighbors (KNN). Additionally, two algorithms specifically developed for learning on imbalanced data are integrated to further increase diversity amongst the candidate learners. Those are implementations of the imbalanced-learn library, namely Balanced Bagging Classifier (BBC) and Balanced Random Forest (BRF). Lastly, XGBoost (XGB) will be integrated as the final base learner. This results in the first level consisting of ten different algorithms. After selecting the algorithms, it is possible to increase the classification performance by optimizing hyperparameters. While thorough hyperparameter optimization was beyond the scope of this paper, different class weight settings were tested for LR, SVM, RF and XGB. In this process each model was evaluated using 3-fold cross validation. The average classification performance is measured by balanced accuracy and shown in Table I.

TABLE I. INFLUENCE OF DIFFERENT CLASS WEIGHTS

model	class weights				
	1:99	10:90	20:80	30:70	50:50
LR	0.54	0.71	0.64	0.58	0.53
SVM	0.5	0.7	0.59	0.54	0.51
RF	0.57	0.56	0.57	0.57	0.58
XGB	0.74	0.74	0.74	0.74	0.63

It can be observed that most algorithms perform best when the class weight setting was equivalent to the class distribution present in the training data. The only exception is the RF algorithm, which performs slightly better having equal weights. Similar to the class weights, the k-nearest neighbors were derived in the same way. Resulting in the highest BA of 0.57 for $k = 1$. After defining said parameters, all models were trained on the entire training set and afterwards evaluated on the test set to establish a performance base line. To do so, all models were measured on BA, G-Mean, AUC and MCC. Table II shows the results of this process. It can be derived that BRF, GBM and RF are outperforming each other model in at least one metric.

TABLE II. PERFORMANCE BASELINE

model	metric			
	BA	G-Mean	AUC	MCC
SVM	0.71	0.709	0.777	0.095
LR	0.724	0.724	0.796	0.1
RF	0.581	0.407	0.854	0.213
MLP	0.678	0.678	0.724	0.076
GNB	0.615	0.54	0.476	0.076
GBM	0.609	0.476	0.865	0.187
KNN	0.588	0.487	0.588	0.064
XGB	0.74	0.735	0.833	0.01
BRF	0.784	0.783	0.858	0.131
BBC	0.783	0.747	0.838	0.139

B. Stacking architecture

After the model selection the stacking architecture can be defined. The fundamental design will emerge along the lines of the Super-Learner algorithm by Van der Laan et al. [31]. Therefore, the selected models are incorporated into one layer (level-1). All level-1 models will generate a prediction, which can later be used as an input for the meta-model. To enable the meta-model to generalize over its input features, a meta-dataset is created in advance. For doing so, the level-1 models are each trained on the training dataset using five-fold cross-validation.

If resampling techniques are applied on the training data, it might be necessary to shuffle the training data before splitting it. The reason for this is, that some libraries will order the data by classes when resampling. This may drastically reduce classification performance when applying cross validation. In the proposed architecture each model is being trained five times on four partitions of the training dataset. A prediction can be generated for the remaining partition. For this paper, this leads to a 49203×10 matrix. This matrix can be used as a meta-dataset after appending the target variable. The meta-dataset serves as the basis for training the meta-model. Following the meta-training, each individual level-1 model is trained again using the entire training dataset. The entire architecture can be referred to as super learner or stacked model and is shown in Fig. 1.

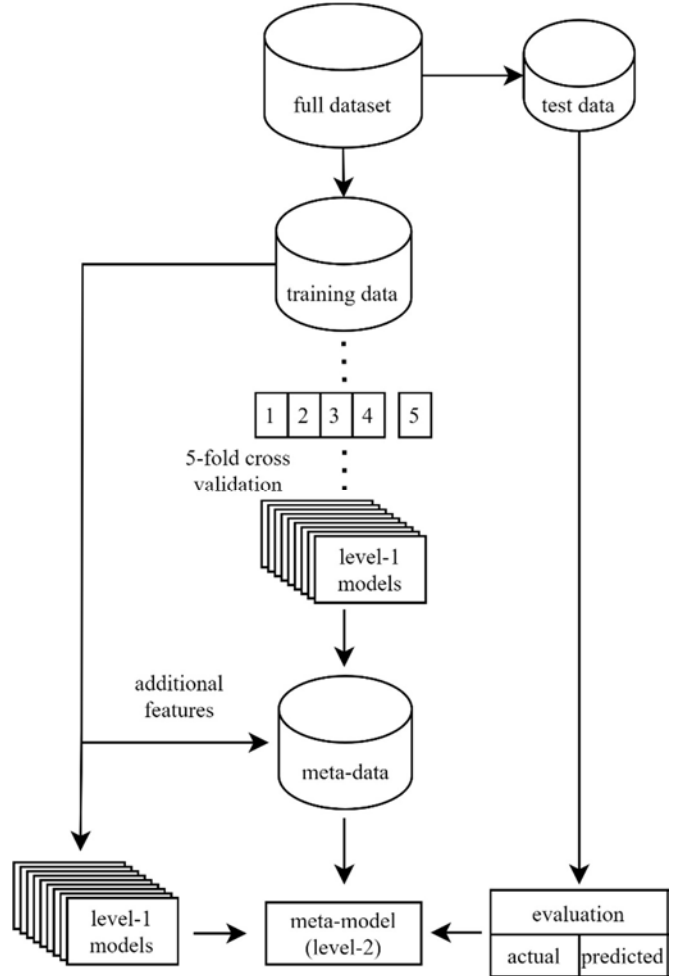


Fig. 1. Stacking architecture.

Using binary predictions as output on level-1 negatively affected the overall classification performance. Leading to the stacked model not being able to outperform its base learners. However, as mentioned in section II it is possible to not only use binary predictions as meta-data, but rather continuous probabilities allowing the meta-model to learn from more granular data. In the given use-case this proved to increase the model's classification performance.

Furthermore, the stacked model could be improved by using the original features as additional meta-data features. This means, the meta-model is trained on the entire training dataset extended by the level-1 predictions. On the contrary, using only the most important features (derived from the level-1 models by majority voting) as a selection to enrich the meta-data proved to be futile.

In the current architecture there is no active measure to tackle the issue of concept drift. Hence it is essential to frequently update the underlying data of the algorithm, to avoid loss due to concept drift. This means the architecture relies on passive methods to handle concept drift.

V. EVALUATION

The super learner can be evaluated on the predefined test data. For this purpose, predictions can be generated using the trained level-1 model. Afterwards, those predictions can be used as input values for the metamodel, which generates the final prediction used for evaluation. The final meta-model is chosen according to the highest performance achieved. To allow for easier comparison three level-1 models were chosen as benchmark. The benchmark is based on the level-1 models which outperformed every other level-1 model in at least one metric.

After training the different meta-models, they were evaluated using the test data. The results of this process are recorded in Table III. It can be observed, that both BRF and GBM as meta-model manage to outperform each model in two metrics. Using BRF as meta-model, it is possible to achieve BA and G-mean scores of 79.1%. This is an increase by roughly 1%, compared to the highest benchmark score, which was also achieved by a BRF classifier. Additionally, it is also possible to increase its AUC and MCC scores regarding to its stand-alone version. Secondly, the GBM as meta-model exceeds the benchmark in terms of AUC (0.87) and MCC (0.219). However, its classification results measured by BA, and G-Mean fall far short of the benchmark.

TABLE III. SUPER LEARNER PERFORMANCE

Meta-model	metric			
	BA	G-Mean	AUC	MCC
XGB	0.627	0.515	0.856	0.184
RF	0.625	0.488	0.858	0.211
GBM	0.629	0.515	0.87	0.219
BRF	0.791	0.791	0.863	0.135
BBC	0.755	0.747	0.833	0.144
Benchmark				
RF	0.581	0.407	0.854	0.213
GBM	0.609	0.476	0.865	0.187
BRF	0.784	0.783	0.858	0.131

As all models performed similarly in terms of AUC a graphical analysis of their ROC could not offer additional insights. In terms of a PRC-analysis (Fig. 2) XGB manages to

achieve high precision for relatively low recall. GBM shows the highest precision for recall thresholds between 0.15 and 0.45. Nevertheless, besides BRF and GBM all other examined meta-models perform similar to their stand-alone versions, with the exception of XGB performing arguably worse, in terms of BA and G-Mean. It is therefore evident that for the given dataset RF, BBC and XGB are unfit to serve as a meta-model, even though some are strong candidate learners.

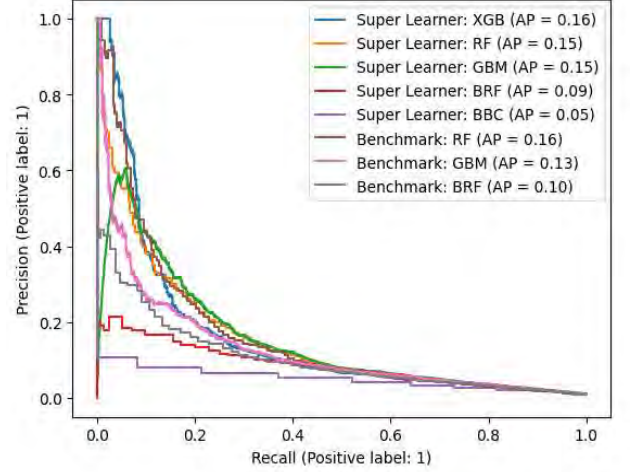


Fig. 2. Precision-Recall-Curves.

Based on this paper the meta-model, which achieves the highest overall performance and hence the proposition is to use BRF as meta-model. Depending on the metric used for measurement a performance increase between 0.7% and 3% compared to the meta-models' base learner can be expected.

VI. DISCUSSION

The architecture and choice of meta-model are closely related to the focus of the application domain and thus highly dependent on the actual use-case. There are numerous methods not examined in the scope of this paper, which are likely to increase the classification performance. One such method is thorough hyperparameter optimization, which the limitations of this paper didn't allow for. Another way to increase performance could be to use a larger selection of level-1 models increasing the variety and enabling the meta-model to learn from broader range of patterns. Lastly, it might yield performance gains to further customize the data preprocessing techniques for each model.

Interesting is that within the development process, no meta-model was able to exceed its candidate learners using binary predictions as inputs. Therefore it can be concluded that the classification capabilities of a super learner are closely related to its architecture. Furthermore, it opens up the question on where the practical as well as the theoretical limitations of stacking actually are.

In practical applications a stacking model is not necessarily better in every aspect when measured on multiple metrics but is likely to outperform its stand-alone models. Nevertheless, in the area of fraud detection for automotive finance, the theoretical paradigm that super-learners outperform their baseline classifiers can be confirmed. However, it should be noted that both the architecture of the stacking model and the choice of meta-model are of significant importance.

One segment open for further discussion is the difference in computing time and resource requirements necessary to train a stacked model, compared to its actual benefit. A stacked model usually exceeds the training time and thus the required resources by a large margin compared to stand-alone models [11]. Having a performance increase of around 1.4% across all metrics, the practical benefits of stacking are highly dependent on the circumstances. In the case of fairly limited resources stacking might not be the best possible solution available.

VII. CONCLUSION

In this paper a stacked model was proposed to identify fraud within a vehicle loan dataset. Given the above it can be concluded that using the proposed method an increase in performance can be expected. In general, the field of fraud prevention is an extremely fluctuating problem area. Fraud patterns may change over time and adapt to existing prevention measures. Consequently, subsequent generations of models usually have a more difficult time achieving better results using the same dataset.

Nevertheless, stacking is a valid option for improving classification performance. Future work could investigate the influence of hyperparameter optimization and individual data preprocessing on the classification results of stacking-based models. It could also be promising to apply different weights to the transactions, resulting in high value transactions being better protected against fraud. Furthermore, lessons learned in fraud prevention may be transferable to other areas such as the prevention of money laundering.

REFERENCES

- [1] Bundeskriminalamt, "Kfz-Kriminalität, Bundeslagebild", 2021.
- [2] Gesamtverband der Versicherer, "Diebe stehlen Autos im Wert von mehr als 200 Millionen Euro," Okt. 21, 2021. [Online]. Available: <https://www.gdv.de/gdv/themen/schaden-unfall/diebe-stehlen-autos-im-wert-von-mehr-als-200-millionen-euro--71506>. [Accessed Apr. 27, 2023].
- [3] Statista, „Anteil der per Kredit oder Leasing finanzierten privaten PKW in Deutschland von 2021 bis 2022,“ [Online]. Available: <https://de.statista.com/statistik/daten/studie/1065113/umfrage/anteil-der-per-kredit-oder-leasing-finanzierten-privaten-pkw-in-deutschland>. [Accessed Apr. 28, 2023].
- [4] H. N. Pontell, W. K. Black and G. Geis, "Too big to fail, too powerful to jail? On the absence of criminal prosecutions after the 2008 financial meltdown," *Crime Law and Social Change*, vol. 61, pp. 1-13, 2014.
- [5] C. Free, "Looking through the fraud triangle: a review and call for new directions," *Meditari Accountancy Research*, vol. 23, no. 2, pp. 175-196, 2015.
- [6] A. Abdallah, M. A. Maarof and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, S. 90-113, 2016.
- [7] J. J. Xu, D. Chen, M. Chau, H. Zheng, H. and L. Li, "Peer-To-Peer Loan Fraud Detection: Constructing features from transaction data," *MIS Quarterly*, vol. 46, no. 3, pp. 1777-1792, 2022.
- [8] H. Wen and F. Huang, "Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), Xiamen, China, 2020, pp. 339-343, doi: 10.1109/ICBDA49040.2020.9101277.
- [9] H. Patel, K. N. Mohana Sai, N. Sai Vishal Devulapalli, V. V. Kumar Mudunuru, B. Mantri and V. Kaushik, "Vehicle Loan Fraud Prediction using Data Science And Machine Learning Techniques," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1288-1291, doi: 10.1109/ICICCS53718.2022.9788394.
- [10] J. Błaszczyński, A. T. de Almeida Filho, A. Matuszyk, M. Szeląg and R. Słowiński, "Auto loan fraud detection using dominance-based rough set approach versus machine learning methods," *Expert Systems with Applications*, vol. 163, Art. no. 113740, 2021.
- [11] R. Soleymanzadeh, M. Aljasim, M. W. Qadeer and R. Kashaf, "Cyberattack and Fraud Detection Using Ensemble Stacking," *AI*, vol. 3, no. 1, 2022, pp. 22-36.
- [12] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. -S. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," *IEEE Access*, vol. 7, pp. 93010-93022, 2019, doi: 10.1109/ACCESS.2019.2927266
- [13] X. Bao, L. Bergman and R. Thompson, "Stacking recommendation engines with additional meta-features," *Proceedings of the third ACM conference on Recommender systems*, Oct. 2009, pp. 109-116.
- [14] B. Omair and A. Alturki, "A Systematic Literature Review of Fraud Detection Metrics in Business Processes," *IEEE Access*, vol. 8, pp. 26893-26903, 2020, doi: 10.1109/ACCESS.2020.2971604.
- [15] I. D. Mienye and Y. Sun, "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection," *IEEE Access*, vol. 11, pp. 30628-30638, 2023, doi: 10.1109/ACCESS.2023.3262020.
- [16] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1-37, 2014.
- [17] S. Wares, J. Isaacs and E. Elyan, "Data stream mining: methods and challenges for handling concept drift," *SN Applied Sciences*, vol. 1, no. 11, pp. 1-19, 2020.
- [18] S. Disabato and M. Roveri, "Learning Convolutional Neural Networks in presence of Concept Drift," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851731.
- [19] Z. Yang, S. Al-Dahidi, P. Baraldi, E. Zio and L. Montelatici, "A Novel Concept Drift Detection Method for Incremental Learning in Nonstationary Environments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 309-320, Jan. 2020, doi: 10.1109/TNNLS.2019.2900956.
- [20] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, „Analysing the classification of imbalanced data-

sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-based systems*, vol. 42, pp. 97-110, 2013.

[21] M. Denil and T. Trappenberg. "Overlap versus imbalance," *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence*, Ottawa, Canada, May 31–Jun. 2, 2010, pp. 220-331.

[22] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, „Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220-239, 2017.

[23] H. Kaur, H. S. Pannu and A. K. Malhi, „A Systematic Review on Imbalanced Data Challenges in Machine Learner," *ACM Computing Survey*, vol. 52, no. 4, pp. 1-36, 2020.

[24] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[25] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[26] I. Almomani, R. Qaddoura, M. Habib, S. Alsoghyer, A. Al Khayer, I. Aljarah, and H. Faris, "Android Ransomware Detection Based on a Hybrid Evolutionary Approach in the Context of Highly Imbalanced Data," *IEEE Access*, vol. 9, pp. 57674-57691, 2021, doi: 10.1109/ACCESS.2021.3071450.

[27] F. Cheng, J. Zhang, and C. Wen, „Cost-sensitive large margin distribution machine for classification of imbalanced data," *Pattern Recognition Letters*, vol. 80, pp. 107-112, 2016.

[28] D. Wolpert, "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.

[29] L. Breimann, "Stacked regressions," *Machine learning*, vol. 24, pp. 49-64, 1996.

[30] F. Hutter, L. Kotthoff, J. Vanschoren, *Automated Machine Learning*, Springer International Publishing, 2019.

[31] Mark J. Van der Laan, Eric Polley, and Alan Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, pp. 1-20, 2007.

[32] R. Sikora and O. H. Al-laymoun, "A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms," *Journal of International Technology and Information Management*, vol. 23, no. 1, pp. 43-53, 2014.

[33] D. V. Carvalho, E. M. Pereira and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, no. 8, pp. 832-866, 2019.

[34] V. Plakandaras, P. Gogas, T. Papadimitriou and I. Tsamardinos, "Credit Card Fraud Detection with Automated Machine Learning Systems," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1585-1599, 2022.

[35] D. Chicco, N. Tötsch, and G. Jurman. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData mining*, vol. 14, no. 1, pp. 1-22, 2021.

[36] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono and P. Lestantyo, "Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data," 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 2019, pp. 14-18, doi: 10.1109/IC3INA48034.2019.8949568.

[37] T. Yang and Y. Ying "AUC maximization in the era of big data and AI: A survey," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-37, 2023.

[38] J. Miao, W. Zhu, "Precision–recall curve (PRC) classification trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545-1569, 2022.

[39] V. Khattri, D. K. Singh, "Parameters of automated fraud detection techniques during online transactions," *Journal of Financial Crime*, vol. 25, no. 3, pp. 702-720, 2018.

[40] L. J. Heinrich, A. Heinzl and R. Riedl, *Wirtschaftsinformatik Einführung und Grundlegung*, 4th ed., Berlin: Springer Berlin Heidelberg, 2011.

[41] D. Choi, K. Lee, "An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation," *Security and Communication Networks*, vol. 2018, pp. 1-15, 2018.

Behavioral Biases in Human-Machine-Interactions in Machine Learning

Theresa Scheutzow
*Behavioral Economic Engineering and
Responsible Management
Heinz Nixdorf Institute
Paderborn, Germany
theresa.scheutzow@hni.uni-
paderborn.de*

Abstract—The use of machine learning is leading to an uncontrolled number of human rights violations. Machine learning experts are predominantly white and male, and most of the data for model training originates from this group. Accordingly, the inherent problem of discrimination mostly becomes visible later when the machine learning model interacts with other groups of people. Structural discrimination can be reproduced and reinforced by machine learning based on data that has inherent biases. Sources of bias are not only in the data, but also in the choice of learning model, algorithms, model evaluation, and feedback effects in the actual application. It has been shown that biases and heuristics in human-machine-interactions, such as automation bias or the illusion of control, can trigger inadequate model control. Therefore, this conceptual work examines supervised, unsupervised, and reinforced learning in the different life cycle stages to identify potential sources of bias. With the help of existing empirical evidence, insights from behavioral economics are transferred to machine learning. It presents human-machine-interactions in the context of fairness and its limitations due to the intersectional nature of discrimination. This is one of the first papers to address the interrelationships and interactions of behavioral biases in the different life cycle stages of machine learning. In particular, the perspective on the perpetuation and reinforcement of structural discrimination through the interaction between humans and machines in the ML process are highlighted.

Keywords—Machine learning, human-machine-interaction, behavioral bias, discrimination, fairness

I. INTRODUCTION

The advancement of machine learning (ML) can help humans to prevent deadly diseases, detect environmental pollution, or overcome language barriers. However, many researchers have shown how individuals and groups are structurally discriminated against ML. In most cases, human decision errors based on ML put non-white people at a disadvantage [1], [2], [5], [6], among others. In particular, misclassification of ethnicity and/or gender based on visible characteristics, facial recognition, crime prediction and the use of war drones harm many people [5], [6]. An extremely high number of human rights violations by ML remain without consequences due to the current lack of effective control mechanisms [1]–[8]. Decolonization, intersectionality, and power are in close interplay in AI research and application, the reasons for which are manifold. On the one hand, the reasons can be examined from a political, sociological, but also behavioral economic perspective. In the context of a decolonial theory, [56, p. 664] emphasizes how Eurocentric research, especially in AI, perpetuates and reinforces

colonization structures. Also, the AI Index Report 2023 demonstrates that there is little diversity in AI research and business [9].

The current state of research partially examines the influence of biases on the ML process by using behavioral economics experiments that focus mainly on technology acceptance. Researchers point out that technology acceptance research has so far mainly examined technical aspects and not deviations from rational behavior such as cognitive and behavioral biases [3], [7], [10], [11], [17]. This is shown in theoretical contributions such as the further development of the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) [11], [12]. From the perspective of the intersection of computer science and behavioral economics, these theoretical models are insufficient to explain the interaction between ML applications and behavioral patterns along the entire life cycle of ML and its application. Since technology acceptance and prejudice can occur reciprocally throughout the entire ML life cycle, a comprehensive process view is required.

Other researchers focus mainly on specific application areas. In particular [13] have conducted experiments to describe the behavior of humans with decision support systems, including the influence of interpretability, but not from an overall process perspective. Based on 90 experimental studies with computer gamers from it was shown that people behave more selfishly and rationally in the presence of computer gamers [14]. According to other studies, evidence-based algorithms also have higher predictive accuracy than human forecasters [15]. Humans who have influence over algorithms are more likely to forgive mistakes, no matter how small the influence, which can be attributed to the desire for control. This relationship with algorithm aversion has been studied by [16]. In behavioral engineering, experiments can help investigate the robustness of the theory compared to actual behavioral complexity [18, p. 672]. Related to this, AI can be used to identify the variables that influence behavior, the limits of human cognition in relation to AI implementation, and which cognitive limits AI can transcend but also exploit [17]. As pointed out in [19, p. 8], there is an increasing research interest in combining ML and behavioral economics, especially since it can be observed that humans judge machines in similar patterns to how they judge humans.

Since humans not directly involved are also affected by ML-based decisions, it is even more important to distinguish at which point in the ML life cycle humans interfere to make the consequences of the use of ML more transparent,

controllable, explainable, and predictable. It is particularly relevant to examine it from a behavioral economics perspective because it assumes that a human's decisions are often influenced by biases, heuristics, and social norms [20, p. 453]. Therefore, the paper examines the relationship between behavioral biases and discrimination through ML in the context of fairness.

Firstly, the theoretical foundations of ML, human-machine-interaction (HMI), behavioral biases, and fairness are highlighted, including the ML life cycle as well as supervised, unsupervised and reinforcement learning. In addition, the relationship between intersectionality, discrimination, and fairness is highlighted in the theoretical foundations. This is followed by the results section, in which the theoretical insights from behavioral economics are applied to ML in terms of HMI in a conceptual way. Especially the several types of interaction and touch points of ML experts and users in the ML process are discussed. The subchapters highlight corresponding differences among the three ML types, as some of them differ fundamentally in their susceptibility to bias. This paper concludes by examining behavioral biases and discrimination from the perspective of fairness metrics and their potential impact on making the ML process less discriminatory. This research provides the basis for empirical studies through which the interaction of behavioral biases of different roles in the HMI can be examined in different machine learning life cycles. Accordingly, the research question is: What forms of behavioral biases can occur in HMI during the ML life cycle?

II. METHOD

The main purpose of this paper is to highlight the research gap and the conceptual search for connections between the findings of behavioral economics and AI research. Due to the novelty of the research area, the theoretical background mainly refers to the behavioral economics foundations of cognitive and behavioral biases and the existing experiments in the intersection area. To this end, a literature review was conducted covering the area of HMI in ML development. A structured literature search of literature databases and review of research papers from institutes and projects like Algorithmic Justice League was conducted using the key terms "HMI", "ML", "AI", "ML expert", "bias", "discrimination", and "fairness". Relevant research on HMI in ML was identified, evaluated, and classified into the phases of the (generic) ML life cycle. Subsequently, further research on ML fundamentals was added, deepening the phases from a computer science and systems engineering perspective. In this way, implicit HMI were identified and used as a basis for the second literature review. With respect to the ML life cycle, behavioral economic theories and empirical results on behavioral biases and heuristics were structured. These were sorted by familiarity and relevance to HMI to apply the existing theoretical foundations of behavioral economics to the new field of HMI in ML. Possible sources of bias were identified and related to the individual steps and decision situations of ML experts. This results in starting points for further empirical investigations on the question of what role behavioral biases may have on the flow of the ML life cycle. The openness of the chosen conceptual approach provides a basis for further empirical investigations, especially behavioral experiments with ML experts.

III. THEORETICAL BACKGROUND

A. Machine Learning

ML is a subfield of AI and an important element in computer and data science. It is trained to recognize patterns and to make decisions without having to be explicitly programmed for each task [21]-[23]. The process involves providing a large amount of data to a ML model, which then uses statistical techniques to learn patterns and relationships in the data [21]-[23]. The three most used types of ML are supervised, unsupervised and reinforcement learning [22], [24], [25]. Supervised learning is a type of ML that involves training a model on labeled data to make predictions on new, unseen data [22, p. 687]. The model is trained by minimizing a loss function that measures the difference between the predicted output and the true output to find the optimal parameters of the model that minimize the loss function on the training set [22, p. 687]. Unsupervised learning is a type of ML to discover patterns and structures in unlabeled data and the objective is often less well-defined and concretized [22, p. 687]. This can be achieved through techniques such as clustering, dimensionality reduction, and anomaly detection and the choice of algorithm depends on the nature of the data and the problem [22, p. 687]. Reinforcement learning deals with learning from interactions with an environment to achieve a particular goal [24, p. 3]. The goal is typically framed as a maximization of a cumulative reward signal received from the environment, and the agent learns to select actions that maximize this reward signal over time [25, p. 25]. The agent's goal is to learn a policy, which is a mapping from states to actions, which maximizes the expected cumulative reward over time [25, p. 25]. The agent uses trial and error to learn this policy, and it can adapt to changes in the environment over time [24, p. 3].

The **ML life cycle** is a framework that provides a structured approach to these stages, from data collection and preprocessing to model development, evaluation, deployment, and monitoring. There are no clear or established instructions or guidelines on how a ML life cycle should be developed [22, p. 691]. As this is an iterative life cycle, there are numerous repetitions of training, evaluation and tuning prior to deployment possible [27, p. 293]. The individual steps of the three learning types have similarities and differences along the ML life cycle. As shown in figure 1, in all three types of ML, data collection is a decisive stage and the quality and quantity of data collected impacts the accuracy and performance of the model. Once the data is collected, it needs to be preprocessed to make it suitable for ML. This involves tasks like data preparation, cleaning, labeling, enrichment, filtering, and aggregation. In supervised learning, data is labeled with the correct output, whereas in unsupervised learning, data is not labeled. Furthermore, the data is labeled with a reward signal in reinforcement learning [22], [27], [28]. In supervised learning, the objective function is to minimize the difference between the predicted output and the actual output [22], [27], [28]. Whereby in unsupervised learning, the objective function is to learn a representation or pattern of the data. In reinforcement learning, the objective is to maximize the reward signal. Due to the completely human-independent pattern recognition in unsupervised learning the so-called **black box** is existing during model training and in the field [29, p. 206]. Black box models often use complex algorithms that can have many parameters and layers, making it difficult to understand how the model arrives at its predictions field [29, p. 206].

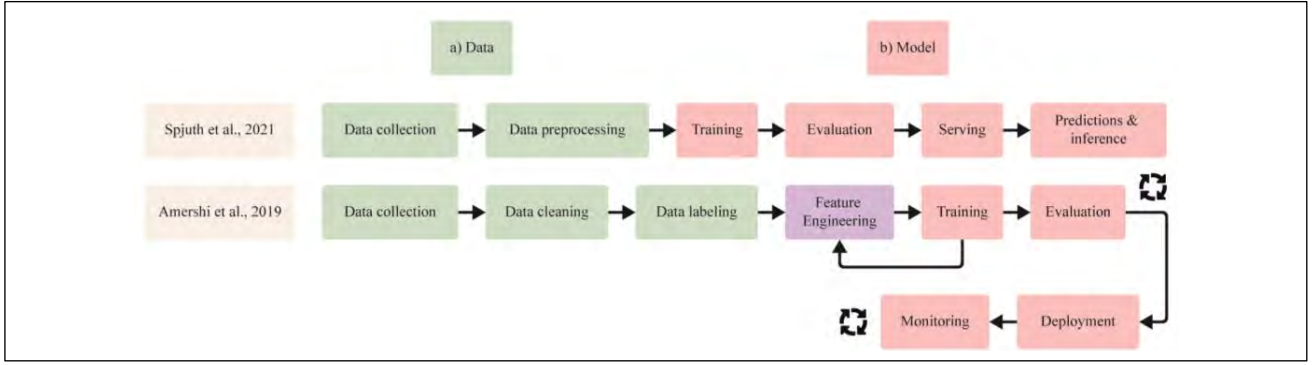


Fig. 1. Structuring of the machine learning life cycle based on [27], [55].

B. Human-Machine-Interaction in Machine Learning

HMI is an interdisciplinary research field in the context of the direct and indirect interaction between involved and impacted humans and machines. The humans involved in HMI can take on different roles and often have a controlling role attribution to prevent as much as possible unwanted results and distortions in the process of HMI [30, p. 23]. Historically, Norman's interaction theory in particular shapes research on HMI. The theory proposes that the design of a system or interface affects the way users interact with it [31].

In **supervised learning**, HMI consists of humans providing labeled data on which ML models train. Humans exert significant influence in deciding which types of data are relevant and which labels are applied [32, p. 10]. Accordingly, they can change the results at the end through a corrective role or at the beginning through data labeling [32, p. 10]. In **unsupervised learning**, humans provide unlabeled data, in **reinforcement learning**, they provide feedback to the ML model in the form of rewards or punishments. The process of data collection and aggregation can be an indirect HMI, as data is collected from humans and used as input for the ML process using automated data mining techniques. The HMI of programmers, ML architecture specialists, testers and validators is a more obvious interaction, as is the deliberate application in the use of the ML application. However, the data that is reused as learning input by the application is also part of the HMI in ML.

C. Behavioral Biases and Fairness in Machine Learning

The interdisciplinary research field of behavioral economics and the related behavioral economic engineering offers a new perspective to study biases in the ML process [17], [18], [33], [34]. AI, its applications, and especially ML, are not solely responsible for creating behavioral distortions and thus any negative impacts on society and individuals. They reflect behavioral distortions of the world that surrounds them [35, p. 2]. Behavioral biases in ML refer to the HMI that have a negative impact on humans. They apply to the systematic patterns of deviation from rational judgment or decision-making that individuals exhibit, often leading to suboptimal outcomes [21, p. 94]. These biases arise from the way individuals process information, form judgments, or make decisions [21, pp. 94-97]. Behavioral biases in ML refer to the tendency of ML algorithms to perpetuate or amplify biases in human decision-making [35, p. 2]. ML models are vulnerable to various biases throughout their life cycle. The types of biases that occur during the ML life cycle can vary depending on the type of learning that is being used. An important distinction in the term bias is whether it means the statistical, necessary bias in ML programming or biases in the sense of behavioral economics. Continuously, the term bias is

used in the sense of the definition of cognitive psychology. Behavioral biases can have a significant impact on perceptions of fairness and the relationship between behavioral biases and fairness is crucial for the design and assessment of a ML application that does not harm humans.

Fairness as a notion of justice provides important guidance in the study of behavioral biases in the ML life cycle. Fairness can be defined as "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [7, p. 1]. There is no clear consensus on what can define, measure, and ensure fairness in decision making, especially in ML [3], [7], [21], [23], [35]. References [40] and [41] both proved that "it is statistically impossible to ensure fairness across three base fairness metrics" [3, p. 47]. Considering the mathematical impossibility of using fairness as a guide, especially in the statistical section of the ML life cycle, it is particularly important to ensure fairness where it is possible [32, pp. 37-38]. Furthermore, discrimination is the unfair treatment of individuals or groups based on their perceived or actual characteristics such as gender, race, age, religion, nationality, or disability [7, p. 10]. **Discrimination** and behavioral biases are closely related as biases can often lead to discriminatory behaviors and decisions [7, p. 10]. **Intersectional bias** in ML refers to the unfair treatment or discrimination that can occur when an algorithm produces biased outcomes based on multiple aspects of a person's identity, such as their (socially constructed) race, gender, age, or socioeconomic status [38], [39].

IV. RESULTS

A. Behavioral Biases in the Machine Learning Environment and Data-Related Process Steps

At the beginning of the ML life cycle, data is needed with which to start the learning process. The common thread is that historical data is used as the basis for ML model input causing **historical bias** to have a negative impact on the learning process and caused disadvantages for certain groups or individuals in the past through distortions [35, p. 2], [42, p. 309]. Historical biases in the context lead to the transfer of structural discrimination into the data. They thus reflect societal and historically shaped biases that mostly arise from or maintain intersectional discrimination [42, p. 309]. Whether directly or indirectly involved, humans have priority for confirmation of their own biases, which distorts the actual representation of reality and the search, perception, assessment, and reaction to ML-related interfaces.

The first cause of bias is that in most cases the entire target group cannot be included, but is taken from a so-called "subset", "development sample" or existing data sets are used [23, p. 2]. The **choice** of size, quality and source of data significantly affects the ML model's susceptibility to bias,

with many researchers emphasising that data is the most common source of bias besides algorithm development [36]. Confirmation bias, ingroup bias, outgroup bias, representativeness bias, self-serving bias, and social desirability bias in combination with groupthink bias can influence the choice of data to a great extent by ML experts. This makes the data quality and accuracy dependent on the subjectivity of the responsible data collector. If the behavioral biases are consciously or unconsciously introduced in the context of data collection, it can lead to one's own prejudices and biases being recursively confirmed through the choice of data. This can lead to biases that later combine with statistical biases and distort the entire ML model. The biases mentioned above, including confirmation bias, also can have an influence, illustrated by the interaction between the two bias streams shown in fig. 2. Especially in data selection, anchoring bias, availability bias, recency bias, representativeness bias and survivorship bias can play a major role by leading to inaccuracy and incorrect representation of reality through biased data selection.

The whole **data collection process** is also theoretically subject to the framing effect, as the decision about certain data set properties and collection methods may depend on how the information is presented to them about it. Furthermore, if the ML expert overestimates their own power of control over automated data collection this poses a great risk in data collection and related biases. It is also a potential source of bias if the person relies too much on automatic data collection or overestimates their general abilities, e.g. to operate data mining software. Representation bias during the data collection process step can also result in inconsistent ML-applications, which is not applicable on the real world. This can happen if the data is influenced too strong through a “one-time phenomena” [21, p. 99]. Additionally, [23, pp. 4–5] point to the resulting bias in representation when the definition of the target group does not reflect the population. If the data set collected then contains underrepresented groups, the model is less robust as it learns about the underrepresented data based on less data. The same can happen if the sampling method is limited [23, pp. 4–5].

Sampling bias is also related to this [7, p. 6]. The researchers describe the sampling bias as the data set and the resulting model not being transferable to new populations.

In the **data processing** steps of data preparation, data cleaning, data labeling, data enrichment, data filtering and data aggregation, the behavioral biases can also have a major impact on data quality. Especially [44] show that embeddings link men's and women's occupations with stereotypical gender roles. These refer to the risk of bias, as in data collection in terms of amplification of bias. In the individual process steps in which the data is processed, cleaned, and filtered, the data processor brings his own biases into the objective and rational process.

Feature engineering also carries a large potential risk for behavioral bias effects on data and model quality. Based on a data analysis, the features that have the greatest influence on the target variable are selected [27]. This step carries a comparatively high risk of being a trigger of structural discrimination. By checking the correlation of the feature with the target variable, the responsible person or group evaluates whether the feature engineering is good enough [27]. If an error becomes visible in the feature evaluation, a critical examination of one's own responsibility could lead to a shifting of the error to external factors due to the influence of the self-serving bias. If a constructive error handling fails, the errors are located on existing data collection or process steps or the technical support along the steps.

Due to the **data splitting** at the end, the bias effects are theoretically passed on. If the data sets contain corresponding biases, the ML model is checked in the evaluation step with the likewise biased test data set. The biases probably originate in behavioral biases and are not or hardly tangible. Accordingly, data splitting can be a danger but also an opportunity for bias identification. Related to this is the confirmation bias, whereby humans involved in the ML process can feel confident that the model is accurate because the test data set confirms their hypothesis, since it theoretically contains the same biases as the training data set.

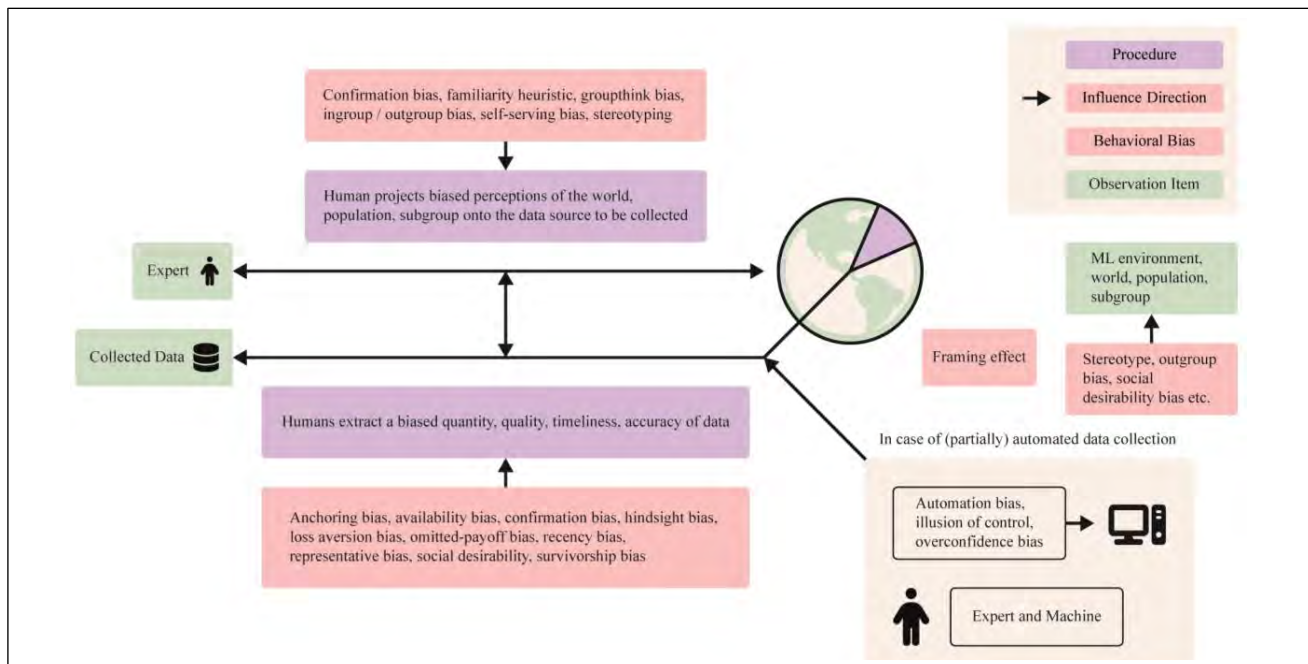


Fig. 2. Influence of behavioral biases during data collection process.

Since labeled data is necessary for **supervised learning** models, there is another leverage of the named bias forms in this learning method. Due to the fact that in supervised learning the correlations between the input and (searched) output data in the data set are already known, there is theoretically also a greater vulnerability to biases. Through the manual choice of assignment or the human decision about the matching, the distortions are theoretically introduced to a greater extent, e.g. if the pattern recognition is passed automatically as in unsupervised learning. As **unsupervised learning** models use unlabeled data, a particularly large data set is necessary. Due to the possible use of different data types, the partly mutual effects of biases are mixed. It can be assumed, however, that the behavioral biases in supervised learning are theoretically easier to control, since they are mostly clear data sets that are directly or indirectly (automated) labeled under human supervision, depending on the method. Behavioral biases in the ML environment of unsupervised learning models are often more difficult to identify. The large, mixed data sets and the fully automated and therefore less direct HMI in the run-up to model development make the bias effects less visible in unsupervised learning. In reinforcement learning, the environment is the input and feedback source for the agent. When the problem definition contains described bias types in the reinforcement environment, it will also be part of the reinforcement learning process and later it can lead to unintended discrimination or unusable applications.

B. Behavioral Biases in Model Development

The ML model design decisions can equally produce biased and potentially unfair ML models and results, and the influence should not be underestimated [45, p. 1]. Therefore, not only incomplete and biased data can be seen as the factor for biased ML output. There are overlapping sources of bias and interactions in the four identified model development phases. If the tasks are completed in teamwork, groupthink bias can have an impact on the correctness of the developing model. There, especially social desirability bias can have an effect, since the responsible model and algorithm experts may behave in the way that is most desired in the group. Furthermore, framing effect can have an impact and is also influenced by what the overall goal of the model development is. Extreme caution and iterative reflection of the chosen algorithms and (hyper)-parameters is appropriate, especially for such sometimes life-critical ML models. In the **selection** of the learning algorithm, there is a great deal of decision-making freedom on the part of the ML expert and thus an HMI in which the human usually makes a judgement about which learning algorithm is suitable for the modelling problem without the support of machine decision-making bases. In many cases, ML experts fall back on common learning algorithms, so-called best practices. Here, the automation bias can strongly influence the quality of the ML process.

On the one hand, an exaggeratedly positive perception of the ML model can lead to an underestimation of error-proneness in the context of automation bias [37]. On the other hand, algorithm aversion can influence the design of the model training through a biased evaluation of the learning algorithm [46]. By preferring human decisions over algorithmic decisions, ML experts and other decision-makers may fail to realise the potential of learning algorithms [47, pp. 1–4]. Automation bias and algorithm bias have been investigated by some research, showing robust evidence that they can strongly bias HMI [46]. Also, [48, p. 340] shows in

a controlled user study with 116 participants that when humans involved in the initial interaction with technical systems recognise strengths first, they are more susceptible to automation bias. If they encounter weaknesses earlier, this usually results in fewer errors because they rely less on the technical system. However, they also clearly underestimate the model competencies. This is also closely related to the anchoring bias, in which ML experts rely on the first solution or information.

If the ML experts make the **choice of learning algorithm** based solely on their experience and knowledge, confirmation bias, overconfidence bias and self-serving bias also can have an influence. This depends on how the ML expert assesses their own responsibility and ability and how others have assessed and reflected this to the person in the past. If ML experts rely on the information most readily available to them when choosing learning algorithms, they have most likely been influenced by anchoring bias, availability bias, recency bias and representativeness bias. This may happen if they depend on the first anchor points in the data and the announced target and choose the most obvious solution. Especially [48] and [49] showed the existence of the related biases. Especially [49, pp. 659–660] showed that software developers can tend to choose the first best, simplest solution in the software development process and to focus on the one they remember best. The representativeness bias can also lead to limited investigation during the process, resulting in incomplete and incorrect solutions in software development [49, p. 659]. Also, the choice of learning algorithms can amplify the biases contained in the data, as it learns based on the biases [50, p. 682]. With the partly manual determination of hyperparameter values, there are similar risks of influence, and the bias susceptibility hardly differs in type, but in design. The choice of hyperparameters, comparable to labeling, clearly reflects the view of the ML experts, as they can manually determine which values and features are important.

In **model training**, automation bias, and other biases related to overconfidence in control and decision-making can play a particular role. **Supervised learning** models are more transparent, whereas **unsupervised models** tend to be less explainable due to the black box problem. As in the choice of learning algorithms, it depends on which ML model is to be created and which types of data are available accordingly. Also, self-serving bias and overconfidence bias may also cause ML experts to engage in little critical thinking during iterative model training. Especially, [17, pp. 601–602] and [49, p. 659] were able to prove overconfidence bias in their studies. Also, [17, pp. 601–602] addresses overconfidence because of overfitting a model, which can result in errors in accuracy between training and test data sets. This failure to notice and address the decline in accuracy is strongly related to the overconfidence bias. This is related across phases to the data-related process phases, model development and evaluation. In the final step of model development, which can iteratively go back to the first step of selecting learning algorithms, the first versions of the model are **evaluated**. This overlaps with the model evaluation steps.

C. Behavioral Biases in Model Evaluation

In most cases, the ML experts must conclude that the model is not accurate and has not achieved the desired prediction or pattern recognition. Should the ML expert attribute the model's inaccuracy to the data, it may also be that the training dataset is adjusted again. In addition, the

hyperparameters are tuned, often manually, to make the model more accurate. There can be a very strong influence of the ML expert on the whole process, as the most optimal hyperparameter value or values are usually found through testing and trial and error. When exactly this point is reached depends on various factors, especially the availability of resources such as time and money, but also on the motivation and susceptibility of the ML expert or client to bias. Depending on how critically the ML expert sees the trained model version, the more action the expert might take. After ML experts have judged that the trained model or models are ready for the iterative evaluation step, they are usually guided by two goals in the evaluation. The model is evaluated based on metrics and different models are compared [43, p. 9]. If it turns out during the model development that different learning algorithms and methods need to be tried out, the model evaluation step will decide which model is finally chosen. In most cases, it is up to individual ML experts or teams to decide whether and how many models should be tried out and which one should be adopted in the end.

Especially in **unsupervised learning**, the explainability of results plays a special role, as it is the core of the model evaluation. Due to the black box problem described in unsupervised learning, explainability is relevant, especially in the sense of the so-called **Explainable AI** (XAI). From now on, the ML expert accesses the test data set generated at the end of the data splitting step. In the concrete process step of model testing based on the test data set, anchoring bias, representativeness bias and survivorship bias in particular play an important role. Under certain circumstances, they can cause the ML expert to perceive the test data set as more fitting and representative than it is, thereby neglecting important factors that would prove that the model cannot make generalisable predictions. The ML expert may rely on the fact that the data is new to the model and that if it produces a good enough result that is generalisable, it will work in the future application environment. Automation bias, illusion of control and self-serving bias can play a crucial role in neglecting to critically question the quality of the test dataset. The hot hand fallacy can also strengthen this overestimation, e.g. if the ML expert has achieved success in the model development step in the sense of an ML model that can initially be described as optimal. In the next step, the results of the model, which have been obtained based on the test data set, are evaluated using previously defined metrics. The selection of metrics can be guided by best practices, although representativeness and social desirability can also lead to a biased selection.

Depending on the chosen learning algorithm, methods and hyperparameters, the evaluation methods also differ, for which there is a long list of options. In the choice of evaluation metrics, especially anchoring bias, availability bias and representativeness bias play a role regarding an inadequate decision-making process. ML experts might be tempted, e.g. by overreliance on best-practices or easy solution suggestions. Loss aversion bias and omitted-payoff bias can influence the decision of evaluation metrics to choose those that produce the least critical output, e.g. those that are more likely to confirm the model than to falsify it, in order to prevent negative consequences such as costly retraining.

However, the supposed negative consequences are only negative in terms of resources and arise from a short-term perspective. In the medium and long term, it is necessary to choose evaluation metrics that are as appropriate and critical

as possible. This should lead to a model that reflects the optimisation problem, the characteristics contained in the data and the population as well as possible. Related to confirmation bias, ingroup bias, outgroup bias, social desirability bias and other stereotype-reinforcing biases is the subjective choice of evaluation metrics, which can reinforce existing biases or falsely mask their existence. The choice of evaluation metrics is an extremely important step, as it is the last possibility before model deployment to prevent bias and the resulting unfair decision-making. In the final step, the **comparison** of different models, behavioral biases can also prevent the selection from being as fair as possible in the ML process. As described in the other steps, behavioral biases can again have an influence, making the actual availability of options appear smaller and thus limiting the subjective decision-making freedom of the ML expert. These include anchoring bias, availability bias, recency bias, representativeness bias and survivorship bias.

Depending on the method, the process steps mentioned in the framework of the model evaluation can also run in a different order, be iterative and, as expected, have a recursive effect on each other. The same applies to the possible forms of bias along the individual steps, which can become even more pronounced along the process. In particular, the framing effect and confirmation bias can play a major and continuous role in the model evaluation step, as giving, accepting, evaluating, and implementing feedback are the main tasks in the evaluation process. Furthermore, this can end up with limitations in the application of the model, up to discriminatory influence on social processes. Depending on the framework of the ML life cycle, the process objective, the circumstances, the availability of resources and the motivation of the ML participants, it may also have an impact on how the evaluator perceives and assesses the results. The behavioral biases can also be recursively related to the evaluation bias, which has been described and observed by [7], [21], [23], [36], among others.

According to research, the reasons and effects lie in non-representative models, data, benchmarks to biased results. This, e.g. create already discriminatory output during testing, as in the case (but not individual case) of [2, p. 12]. In the evaluation process, the black MIT researcher Dr. Joy Buolamwini found that her face could not be recognised by a facial recognition software, but only when she put on a white mask [2, p. 1]. The confirmation bias in the context of software developers and testers has been proven by [51], among others. In their analysis, they found that there is a direct correlation between confirmation bias and code error proneness. Thus, they opted for positive rather than negative tests to confirm their own hypotheses instead of trying to disprove them [51, p. 2]. Also, [49, p. 659] were able to prove in their studies the negative influence of confirmation bias on the testing of codes by software developers, whereby the error search was inefficient and preconceived.

Compared to the black box problem in **unsupervised learning**, the ML expert has significantly more influence and decision-making possibilities in **supervised learning** algorithms and training. After choosing the learning algorithm, the ML expert can observe the model training and take it over manually. Due to the assignment of input and output, there are many more possibilities of influence. Since a strong HMI in this process step in supervised learning also entails a correspondingly higher susceptibility to negative

effects of behavioral biases, self-assessment-related forms of bias play a role here in particular. If the ML expert directly supports the model training process, anchoring bias, confirmation bias, framing effect, hot hand fallacy, illusion of control, omitted-payoff bias, overconfidence bias, and self-serving bias can have a negative impact on the success of model training. This applies especially to the iterative hyperparameter tuning step. The model evaluation process in supervised learning refers to checking whether the assignment of the labeled data input fits the model output. Metrics used can be cross-validation, train-test-split, testing for accuracy or recall. There can be data-related behavioral biases and automation-related biases. The behavioral biases and example thought processes and statements mentioned and described are interconnected, which still needs to be empirically investigated in the future. Furthermore, the framing effect can also influence, trigger or correct all the above-mentioned distortions through the perception of the setting.

ML experts have less room for interpretation in the evaluation of **supervised learning** models than in **unsupervised learning** since the predefined input-output relationship can provide a certain frame of reference. However, this also bears the danger or places the responsibility on exactly these steps in the ML process. Evaluation offers the opportunity to critically question the upstream process steps. In unsupervised learning, the black box problem is a concrete problem for bias identification and prevention during the model development phase. Due to the lack of explanatory power of the predictions made, the people involved in the ML process are usually not able to understand the decisions and thus to critically question and correct them. Automation bias can have a strong, negative effect, especially in the context of the black box problem. On the basis of overconfidence in the training results of the model, there may be too little scrutiny of the result, and the biases contained in the datasets and the HMI-based sources of bias may not be critically included in the assessment. It can be assumed that automation bias, overconfidence bias, groupthink bias, illusion of control and omitted-payoff bias play a role in the lack of responsibility. Omitted-payoff bias among ML experts can lead to the fact that they do not perceive the resource investment in the model development of a model interpretable from the beginning as rewarding enough, especially because they are not properly measurable. In the medium and long term, however, this can have extreme effects that only become apparent or cause subtle damage after the model deployment. In consequence of the completely automated pattern recognition in the incomprehensible model training step in unsupervised learning, it can be assumed that the risk of negative influence of behavioral distortions in the HMI approaches zero.

However, since there is no direct HMI in this step, despite the supposedly lower risk for bias susceptibility before and after, it carries an even higher risk. Especially since the lack of explainability may mean that the model cannot be checked and critically evaluated. The focus of the evaluation process in unsupervised learning is on testing and judging whether the automatic pattern recognition has worked. That is, accordingly, whether it fulfils the expectations based on the data and is transferable to other data sets and thus representative. Compared to supervised learning, there is potentially a greater leverage effect of behavioral biases based on prejudices and the ML experts' own attitudes. Since, in contrast to supervised learning, there is no clear assignment

here and the assessment of the correctness of the pattern recognition depends on the assessment of the ML experts, there is a greater risk that the model will become inaccurate. As pattern recognition is much more complex and difficult to evaluate, especially in the context of the black box problem, there are many challenges. Except for the point that labeling as evidence of one's own performance can be seen as certainty in relation to hot hand fallacy, overconfidence bias and self-serving bias. During **reinforcement model** development, behavioral biases can result if the design of the reward function is such that it reflects the beliefs of the ML expert. The developed model can also be overgeneralised if it makes assumptions about the environment that are not necessarily true.

When models are trained on simulated environments that do not accurately reflect the real world, this can also be influenced by stereotypes, ingroup bias, outgroup bias, or confirmation bias. If the model is trained in a way that makes it overvalue certain actions or outcomes in hindsight, this can come from the hindsight bias of the ML expert. If, in addition, the state-action space is not used effectively, but instead focused on the most obvious solution in the sense of anchoring bias, the ML expert negatively influences the model development of the reinforcement process. In the HMI between the ML expert, the reinforcement model and its alternating process of reacting and correcting, hindsight bias, recency bias, anchoring bias, but also confirmation bias are therefore a possible source of distortions. In the reinforcement model evaluation, it is tested whether the model really reflects the actual environment to which it has been trained.

D. Behavioral Biases in Model Deployment

The greatest impact and danger of behavioral biases arises especially in the deployment step where the ML model can cause actual harm to directly and indirectly interacting individuals and groups. In the context of this HMI, it becomes visible whether the application works as intended or not. When the ML evaluation process has been completed to the point where the ML model comes to the pre-defined use, further forms of biases can have a negative influence. There is an interaction between the **environment** of the finished ML model and the ML model itself [43, pp. 10-11]. If the ML model makes biased or incorrect predictions due to historical biases and other forms of bias, this can lead to the model, for which it was originally programmed, only being able to be integrated with great limitations [21, p. 100], [23, p. 6].

Since it is extremely difficult to pinpoint the reasons for biases, it is accordingly important to shed light on the various dependencies by taking a holistic view of the ML life cycle. It is also important to highlight that although there is a final model in terms of the selected model to be integrated into the field of application, the iterative process and especially the feedback loops mean that the learning or training of the ML model will never be properly completed. Only through the actual deployment in the planned environment, the learning process is usually continued in larger dimensions. Furthermore, the assumptions made at the beginning of the ML process can change during the learning process and thus falsify the result.

A well-known difficulty in the ML life cycle is to integrate the evaluated and finally selected model into the planned application domain. The mostly iterative **monitoring phase** involves tracking and analyzing the performance of ML models to ensure they are functioning as intended and

providing accurate predictions. In terms of data quality, this includes monitoring possible data drift by comparing the input data with the new data to correct it if necessary. **Anomaly detection** involves identifying instances where the model produces unexpected or unusual results. This can help to identify issues such as data errors or changes in the underlying data that may be affecting the model's performance. In particular, the role of humans in model monitoring is important as the ability to make judgements based on experience can help to identify errors that are difficult to identify or assign through automated monitoring. In an ideal monitoring and feedback loop, humans monitor the functioning of the ML model in the actual application, partly with the help of automated support, using predefined metrics. By looking for anomalies and problems, biased model results should be identified and corrected. However, if the biased result of the ML model reproduces existing biases such as structural discrimination, these biases may go unnoticed and cause great harm.

Many ML models have a field of application in society and thus points of contact with humans who interact with them voluntarily and involuntarily. It is important to know which persons interact with the ML model in which role and whether it is a conscious interaction. Furthermore, who has access to the ML interaction influences the susceptibility of the dynamic ML system to bias. This concerns the equal opportunities in the education system, which determines who can become an ML expert in the first place. But also, the interaction as a user is determined by which socio-economic framework conditions apply to the individuals or persons. Thus, there are **three perspectives** on HMI in ML model use.

First of all, it is **socially unequally** distributed who has access to technical devices and internet connections. Secondly, it is socially unequally distributed who has access to technical and professional qualification and thus it is position related. Because ML-teams are predominantly white and male, this can potentially distort the deployment process in such a way that the perspectives and experiences of discrimination of the humans who end up interacting most with the ML model have too little influence. There is still a tendency for ML models to be developed by groups of individuals who do not represent most actual user groups. Thirdly, it is unequally distributed who interacts with ML applications and if it is voluntarily or involuntarily. Interaction also refers to discrimination based on ML-based decisions.

It influences the life cycle of the model and the learning outcomes ultimately depending on who uses the ML models. This can also lead to biases that cannot be controlled or are not noticeable, e.g. because they reflect the attitudes of the ML experts who monitor the ML model.

How users of ML models **interact** with the application influences what data they provide as data input. Depending on how open and receptive the ML model is designed, the leverage of user interaction with the ML application is influenced. If the ML model is open and allows feedback loops, the possibility of influence is correspondingly higher. In this context, automation bias can play a crucial role, as overconfidence in the application tends to reveal more information. If the ML expert has excessive confidence in the ML model, monitoring is likely to be less critical and extensive. Automation bias, illusion of control, overconfidence bias and self-serving bias may play a role. If the ML expert relies on the first good results achieved during the model deployment, anchoring bias, hot hand fallacy and self-serving bias can lead to insufficient model monitoring, so that the model remains biased or new biases arise. In addition, it is likely that in the context of groupthink bias, loss aversion bias and social desirability bias, a critical result is noticed in the context of model deployment but is not transferred to the ML process and thus to accountability. All these biases, if present, are fed back to the model in the **feedback loop** and can influence the model development in a way that is difficult to assess while it is in use. From the user's perspective, automation bias, hindsight bias, hot hand fallacy, illusion of control, overconfidence bias, self-serving bias and social desirability bias can play a role.

Furthermore, [52] have shown that there are major differences in how the human brain responds to humans compared to an interaction with algorithms. Also, [53, p. 2] have shown that humans are more willing to take advice from AI than from humans, e.g. in relation to forecasting. On the other hand, according to [15], algorithm aversion occurs when the behavior of the machine heuristic leads to an aversive response to AI. There is **empirical evidence** that human judgements tend to be favoured over AI-based judgements [15]. Reasons for the aversion to algorithms may be the breaking of the AI's perfectibility scheme [15]. Besides the specific HMI, other sources of biases are the technical conditions, which in turn also can influence technology acceptance [21, p. 100]. There is a demonstrated strong

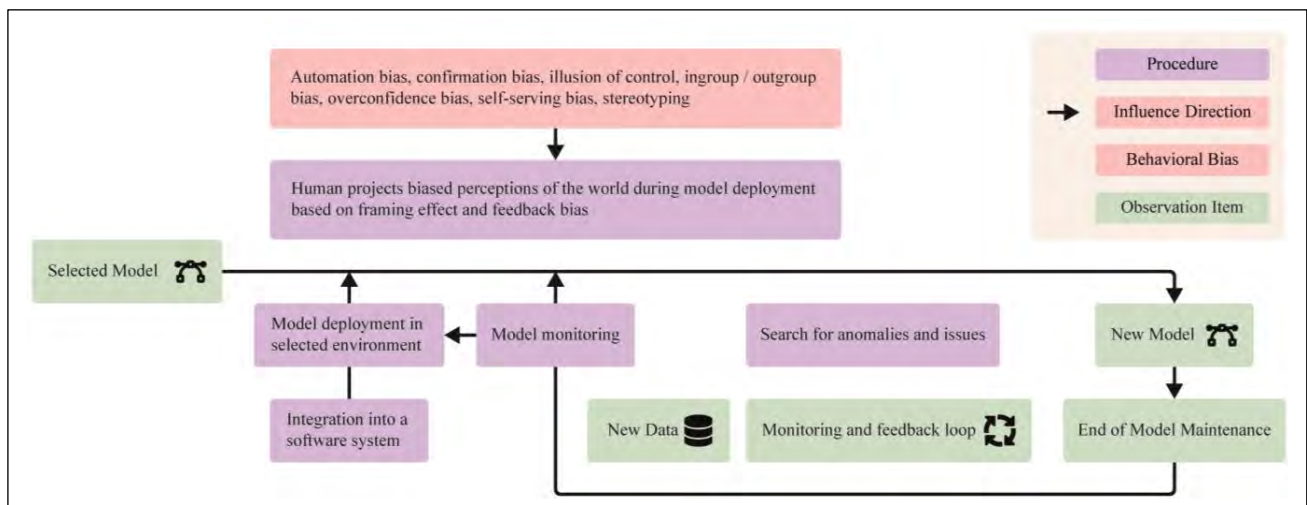


Fig. 3. Influence of behavioral biases during data collection process.

relationship between behavioral intention and collaborative intention, suggesting that **human-in-the-loop** (HITL) approaches can indeed be used successfully in end-user applications [54, p. 14]. There is no consensus in research on which factors determine the use or rejection of new systems [10, p. 5]. When there is an optional change from an existing to a new one, there are different outcomes, such as regret avoidance and social norms [10, p. 5]. Depending on the learning method and the openness and ability of the ML model to process new data input through user interactions, the data is processed or collected in real time and reused in the learning process of the model. As a result, the ML model is updated continuously, iteratively and depending on the learning method. If the ML expert notices errors, e.g. in the course of a model drift, the decision can be made to improve the model through retraining.

Accordingly, the **feedback loop** may not only lead to the interface of the HMI in the deployment phase, but the loop may extend to the data input. It can also happen that the model confirms itself, since the model output can recursively influence the new model input. In the process, it is accordingly important that the ML monitoring along the feedback loop is critically checked by human experts and that bias susceptibilities are included. This can be seen especially in the case of racist predictive policing. Data distortions and data shifts in the deployment phase are a serious threat in the ML life cycle. Especially when ML models are used in different application areas that cannot be estimated concretely beforehand, there is a risk of bias due to sampling bias and limited transferability of the ML model results to other application areas, populations, etc. than the one with which the ML model was trained [21, p. 100].

E. Fairness in the Machine Learning Life Cycle

It is important to identify and address the underlying environmental causes of discrimination in order to actively shape the ML process in a way that minimizes or reinforces discrimination. Fairness efforts can take place through pre-processing techniques in data-related steps to identify and address biases [7, p. 13]. The data phase includes processing the data with appropriate critical judgement and transparency and respecting the differences and vulnerabilities to bias. The data sources should be as diverse and representative as possible. The definition of **sensitive attributes** and the division into protected and unprotected groups play a decisive role and are the basis for the application of fairness metrics. In the model development step, in-processing techniques can be applied to achieve a fairer model development [7, pp. 13–14]. By changing selected and partly standardised learning algorithms, their design can lead to the identification and partial compensation of biases of the preceding steps. In the deployment process, it is also essential to communicate transparently how the model makes decisions, based on which data sets, algorithmic decisions, parameter tuning and with which metrics it is evaluated.

V. DISCUSSION

ML can make existing discrimination visible, which is why a transparent ML life cycle is so important. It is particularly important that various ML experts work together in teams that are aware of what sources of bias exist and what influence they can have. The relationship between behavioral biases and statistical bias in ML has not yet been investigated and the correlations are extremely complex and difficult to

generalise. Corresponding interactions, cause-trigger connections and dependencies between behavioral biases are difficult to grasp and could not be named comprehensively. There are strategies that address individual concrete phases of the ML life cycle at the technical level, such as specific metrics. Accordingly, the focus is on a structural solution to the problem. Experiments should be used to test the suspected sources of bias to prove their interactions and effects. Studies on racist ML output clearly show the severity and structural embeddedness. It should also be noted that classic ML models for face recognition are gender-discriminatory, as they usually distinguish in a binary way whether humans are to be assigned to the male or female gender. The design of the ML models fails to recognise that it has been scientifically proven that there are more than two genders, and that humans in this non-binary gender category are generally misclassified, which is discriminatory and unfair [2], [5].

VI. CONCLUSION

The answer to the research question thus lies in the warning and complex answer that behavioral biases at the bias-related, technology- and control-overestimating and availability-related level theoretically have an influence on all the process steps of ML. The biased choice and quality of the data is reinforced in the model training, a biased evaluation process hinders the identification of the biases, and in the end the actual use of the ML application in the feedback loop may again introduce biases that reinforce previous ones. Since there are currently few diverse teams of ML experts, behavioral biases such as self-interest, control illusion, and anchoring bias are correspondingly dangerous. Eurocentric and structural discrimination by ML can be explained by this.

REFERENCES

- [1] E. M. Bender, T. Gebru, A. McMillan-Major and M. Mitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *ACM FAccT*, Canada, 2021, pp. 610–623.
- [2] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proc. of Machine Learning Research*, vol. 81, pp. 1–15, Feb 2018.
- [3] A. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau "FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning," in 2019 IEEE Conf. on VAST, Vancouver, BC, Canada, 2019 pp. 46–56.
- [4] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *BD&S*, vol. 5, pp. 1–17, Oct 2017.
- [5] T. Gebru, "Race and Gender," in *The Oxford Handbook of Ethics of AI*, Vol. 13, M. D. Dubber, F. Pasquale and S. Das, Eds., Oxford: Oxford Academic, 2020, pp. 252–269.
- [6] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, issue 12, pp. 86–92, 2021.
- [7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, issue 6, pp. 1–34, Jan 2021.
- [8] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.
- [9] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, "The AI Index 2023 Annual Report," HCAI, Stanford, Stanford University, 2023. Available: <https://aiindex.stanford.edu/report/>
- [10] M. Mueller, F. M. Oschinsky, H. Freude, C. Reßing and M. Knop, "Exploring the Role of Cognitive Bias in Technology Acceptance by Physicians," *40th ICIS*, Munich, 2019, pp. 1–10.

- [11] F. D. Davis, "A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results," Ph.D. dissertation, Sloan School of Management, MIT, 1985.
- [12] V. Venkatesh, M. G. Morris, G. B. Davis and F. D. David, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, issue 3, 425–478, Sep 2003.
- [13] A. Erlei, F. A. Nekdem, L. Meub, A. Anand and U. Gadiraju, "Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining," in *Proc. of 8th AAAI Conference on Human Computation and Crowdsourcing*, Hilversum, 2020, pp. 43–52.
- [14] C. March, "The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players," working paper, CESifo, Munich, 2019 [Online]. Available: <https://www.cesifo.org/en/publications/2019/working-paper/behavioral-economics-artificial-intelligence-lessons-experiments>
- [15] B. J. Dietvorst, J. P. Simmons and C. Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *J. Exp. Psychol. Gen.*, vol. 144, issue 1, pp. 114–126, Feb 2015.
- [16] B. J. Dietvorst, J. P. Simmons and C. Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *ManSci*, vol. 64, issue 3, pp. 1155–1170, Nov 2016.
- [17] C. F. Camerer, "Artificial Intelligence and Behavioral Economics," In *The Economics of Artificial Intelligence: An Agenda*, A. Agrawal, J. Gans and A. Goldfarb, Eds. Chicago: The University of Chicago Press, 2019, pp. 587–608.
- [18] G. E. Bolton and A. Ockenfels, "Behavioral economic engineering," *J. Econ. Psychol.*, vol. 33, issue 3, pp. 665–676, Jun 2012.
- [19] J.-F. Bonnefon and I. Rahwan, "Machine Thinking, Fast and Slow," *Elsevier*, vol. 24, issue 12, pp. 1019–1027, Dec 2020.
- [20] A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science*, vol. 211, issue 4481, pp. 453–458, Jan 1981.
- [21] T. Fahse, V. Huber and B. van Giffen, "Managing Bias in Machine Learning Projects," *Wirtschaftsinformatik 2021 Proc.*, vol. 7, pp. 94–109, Oct 2021.
- [22] C. Janiesch, P. Zschech and K. Heinrich, K., "Machine learning and deep learning," *EM*, vol. 31, pp. 685–695, Apr 2021.
- [23] H. Suresh and J. Guttat, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," *EAAMO '21*, New York, 2021, pp. 1–9.
- [24] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
- [25] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, Massachusetts: MIT Press, 2016.
- [26] J. Alzubi, A. Nayyar and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *JPCS*, vol. 1142, issue 012012, pp. 1–15, 2018.
- [27] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi and T. Zimmermann, "Software Engineering for Machine Learning: A Case Study," *Proc. 41st ICSE-SEIP '19*, Montreal Quebec, 2019, pp. 291–300.
- [28] Z.-H. Zhou, *Machine Learning*, Singapore: Springer Nature, 2021.
- [29] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, issue 5, pp. 206–215, Mar 2019.
- [30] W. Xu, M. J. Dainoff, L. Ge and Z. Gao, "Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI," *Int J Hum-Comput Int*, pp. 1–69, Jan 2022.
- [31] J. P. Chin, V. A. Diehl and K. L. Norman, K. L., "Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface," in *CHI '88: Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, New York, 1988, pp. 213–218.
- [32] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artif. Intell. Rev.*, vol. 56, pp. 1–50, Aug 2022.
- [33] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgements reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, issue 4157, pp. 1124–1131, Sep 1974.
- [34] D. Kahneman, *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux, 2011.
- [35] T. Hellström, V. Dignum and S. Bensch, "Bias in Machine Learning - What is it Good for?" *arXiv:2004.00686*, pp. 1–8, Sep 2020, preprint.
- [36] A. Olteanu, C. Castillo, F. Diaz and E. Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Front. Big Data*, vol. 2, pp. 1–33, Jul 2019.
- [37] M. L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," in *AIAA 1st Intelligent Systems Technical Conf.*, Chicago, 2004, pp. 1–6.
- [38] K. W. Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color," *SLR*, vol. 43, pp. 1241–1299, Jul 1991.
- [39] K. W. Crenshaw, "Reach everyone on the planet...": *Kimberlé Crenshaw und die Intersektionalität* [I. Kappert, P. Piesche, E. Roig, & H. Lichtenthäler, Eds.], Berlin: Gunder-Werner-Institut in der Heinrich-Böll-Stiftung, 2019.
- [40] S. A. Friedler, C. Scheidegger and S. Venkatasubramanian, S., "On the (im)possibility of fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making," *arXiv:1609.07236*, 64(4), pp. 1–16, Sep 2016, preprint.
- [41] J. Kleinberg, S. Mullainathan and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *Proc. of ITCS*, New York, 2016, pp. 1–23.
- [42] E. S. Jo and T. Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," in *Proc. of the FAT* '20*, New York, 2020, pp. 306–316.
- [43] M. Gärtler, V. Khaydarov, B. Klöpper and L. Urbas, L., "The Machine Learning Life Cycle in Chemical Operations – Status and Open Challenges," *CIT*, vol. 93, issue 12, pp. 2063–2080, Nov 2021.
- [44] A. Caliskan, J. J. Bryson and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, issue 6334, pp. 183–186, Apr 2017.
- [45] S. Hooker, "Moving beyond 'algorithmic bias is a data problem'," *Patterns*, vol. 2, issue 4, pp. 1–4, Apr 2021.
- [46] S. M. Jones-Jang and Y. J. Park, "How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability," *JCMC*, vol. 28, issue 1, pp. 1–8, Nov 2022.
- [47] E. Jussupow, I. Benbasat and A. Heinzl, "Why are we averse towards algorithms? A Comprehensive literature review on algorithm Aversion," in *28th ECIS*, Marrakech, 2020, pp. 1–16, 2020.
- [48] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan and V. Gogate, "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems," in *IUI '21*, New York, 2021, pp. 340–350.
- [49] S. Chattopadhyay, N. Nelson, A. Au, N. Morales, C. Sanchez, R. Pandita and A. Sarma, "A Tale from the Trenches: Cognitive Biases and Software Development," in *42nd ICSE*, Seoul, 2020, pp. 654–665.
- [50] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *CLR*, vol. 104, issue 3, pp. 671–732, Sep 2016.
- [51] G. Çalkılı, B. Aslan and A. Bener, "Confirmation Bias in Software Development and Testing: An Analysis of the Effects of Company Size, Experience and Reasoning Skills," In *Workshop on PPiG*, Madrid, 2010, pp. 1–16.
- [52] K. Goodyear, R. Parasuraman, S. Chernyak, P. Madhavan, G. Deshpande and F. Krueger, "Advice Taking from Humans and Machines: An fMRI and Effective Connectivity Study," *Front. Hum. Neurosci.*, vol. 10, issue 542, pp. 1–15, Nov 2016.
- [53] J. M. Logg, J. A. Minson and D. A. Moore, "Algorithm Appreciation: People Prefer Algorithmic To Human Judgment," working paper, Article 17-086, 2018.
- [54] I. Baroni, G. Re Calegari, D. Scandolari and I. Celino, I., "AI-TAM: a model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications," *HC*, vol. 9, issue 1, pp. 1–21, 2022.
- [55] O. Spjuth, J. Frid and A. Hellander, "The machine learning life cycle and the cloud: Implications for drug discovery," in *Expert Opinion on Drug Discovery*, vol. 16, issue 19, pp. 1071–1079, May 2021.
- [56] S. Mohamed, M.-T. Png and W. Isaac, "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence," in *Philos Technol*, vol. 33, pp. 659–684, Jul 2020.

How to prepare students for the AI era?

Ulrich Bucher
Studienzentrum Dienstleistungsmangement
DHBW Stuttgart
Stuttgart, Germany
ulrich.bucher@dhbw-stuttgart.de

Kai Holzweißig
Studienzentrum Wirtschaftsinformatik
DHBW Stuttgart
Stuttgart, Germany
kai.holzweissig@dhbw-stuttgart.de

Abstract— The rise of AI and its increasingly significant role in business and society raises multiple questions on the formation of competencies. This discussion paper aims to make a contribution by adding insights to answering the question of how to prepare pupils and college students for the AI age. In order to do so, central sub-questions are raised, which prove to be very challenging and complex, such as the influence of AI on the world of work, among others. Secondly, within the framework of a hermeneutic literature review, various points of view on these sub-questions are presented. From this, the competence fields of AI Literacy are derived in a third step. In a fourth step, the cornerstones of a concept for increasing AI literacy are outlined. In the final step, the insights of the discussion paper are being evaluated by students, secondary school teachers and experts using semi-structured interviews and focus groups. The results show that competencies such as critical thinking, knowledge of the risks of AI and problem-solving skills are particularly significant. Furthermore, the conducted evaluation supports the idea of promoting AI literacy as part of research-based learning approaches and that more flexible and dynamic ways of learning are needed in schools and colleges to support the formation of competencies in an efficient and effective manner.

Keywords — *AI Literacy, Artificial Intelligence, Research-Based Learning, Critical Thinking, Problem-Solving*

I. INTRODUCTION & METHODOICAL APPROACH

Similar to the moon landing, the introduction and rapid spread of generative AI tools such as ChatGPT has made it clear to a broad public that AI is ushering in a new age. It is likely – as patterns of historic development show – that the disruptive character of AI will be accompanied by significant changes in society and the world of work. One prediction is that in the future, employees with AI skills will increasingly replace those without [1]. Furthermore, AI tools will play an increasingly important role in our everyday lives [2]. Accordingly, it is important to prepare pupils and college students for the AI era. This paper addresses the research question of how, in terms of competencies, such preparation can take place, i.e. how AI literacy can be formed.

If one wants to answer the research question, then a series of partly very demanding, complex sub-questions arise:

- What impact does AI have on the world of work?
- What is AI Literacy?
- Why should we increase AI literacy?
- How can AI be integrated into teaching?
- What conclusions can be drawn from the above questions for a curricular and didactic concept to increase AI literacy?

In this discussion paper, answers to the above questions are developed on the basis of the literature and our own reflections. These answers were presented to AI experts as

well as various stakeholders in a second step. The latter consisted of school teachers and college students. Their feedback is presented in a separate evaluation chapter. The purpose of this discussion paper is to define cornerstones for a concept to increase AI Literacy. Describing the specific ways of competency formation is not within the scope of this paper, but will take place in a subsequent step.

II. IMPACT OF AI ON THE WORLD OF WORK

If we want to prepare pupils and students for the AI age, this raises the question of what influence AI will have on the world of work and society and how it will change the role of employees or the reality of everyday life. Even if this is not explicitly elaborated in many articles, various approaches can be found in this regard:

1. In the future, humans will be responsible for tasks that cannot be automated by AI [3]. This draft of the future can be formulated in a positive or negative way. A positive formulation would be to relieve humans (e.g. of routine activities) and thus give them time to take care of more demanding tasks [4]. The negative variant says that humans have to be satisfied with the "breadcrumbs" that AI leaves them. Harari, for example, paints a picture of a cascade of ever greater disruptions, driving people out of existing jobs and forcing them to recurrently qualify for new fields of activity [5].

2. Another "human-in-the-loop" design assumes that humans will not be displaced by AI in the foreseeable future. Even if AI can perform tasks very well and in many cases better than humans, this does not replace humans. Rather, a team of humans and machines often produces the best results [6]. For example, even if a dermatologist uses an AI solution, he or she still makes the final diagnosis and discusses it with the patient. This perspective brings to the fore the question of how artificial and human intelligence should be combined, as is the case with augmented intelligence [7].

3. AI is leading to a new industrial revolution in the course of which the world of work is changing significantly. In this context, an analogy is often drawn with earlier industrial revolutions [8]. Around 1800, for example, around 62% of the workforce worked in agriculture, 21% in industry and 17% in the service sector [9]. Currently, the primary sector employs 1.2 % of the labour force, the secondary sector 23.6 % and the tertiary sector 75.2 %. Similar to previous industrial revolutions, AI will therefore structurally change the fields of work and the content of work.

4. A positive design states that AI creates significantly more new jobs than it destroys. The development and implementation of AI solutions generates many jobs in software development and data science, among others. Many new occupational fields are emerging, such as prompt engineering and the control of AI systems. With new technologies, new industries are emerging, bringing new forms of employment [8]. Increasing productivity also leads

to growing economic prosperity, rising consumption and higher incomes, which in turn also leads to positive effects in existing industries such as healthcare [10].

A. The problem of predictability

The different drafts make it clear how difficult it is to find an answer to the question of how to prepare pupils and students for the AI age when it is not even clear what the future will look like and how quickly the change will take place. Answers to the question of how to prepare for the AI age thus always run the risk of missing the actual development.

One example of this is that numerous recent works highlight creativity as a key human strength vis-à-vis AI [8, 11]. Thus, while until recently professions such as graphic designers and artists were considered to be at little risk from AI, the emergence of image generators such as Midjourney, Dall-E and Stable Diffusion has since changed this picture considerably.

When fundamental assumptions are overtaken by technological development within a few years, this raises the question of whether the future is predictable even for short and medium time periods. Just think of the complex interaction effects of technological innovations in a VUCA (Volatility, Uncertainty, Complexity, Ambiguity) world or how the boundaries between humans and machines may become blurred in the future because the brain will be connected to the computer, interventions in DNA will take place or biocomputers will be driven by human brain cells [12]. Currently, intensive research is being conducted on a variety of innovations with a high disruptive potential, such as quantum computers, biohacking and autonomous driving.

B. Development of long-term responses

Assuming that any blueprint we make of the future is likely to be wrong, the preparation of pupils and students focuses on their ability to adapt to new situations, such as promoting flexibility combined with lifelong learning, developing problem-solving skills and strengthening cognitive skills such as critical thinking. If the amount of available information increases exponentially, as is currently the case, then it becomes more important to separate the important from the unimportant, to distinguish the true from the untrue, to derive knowledge from the information and to apply it to solve problems. In this respect, with a view to long-term development, it is certainly right to strengthen the aforementioned competences.

For obvious reasons, however, this view seems only partially satisfactory. After all, it only provides a very rough orientation and the most diverse competences can be named that will be necessary in the future. If the future challenges cannot be described in concrete terms, then it is hardly possible to determine the significance of individual competences. Accordingly, it is not surprising that the present frameworks on competences for the 21st century read like an "all-in-one solution". For example, the "P21 Framework for 21st Century Learning" contains 39 different competences, including critical thinking, problem-solving skills, flexibility and adaptability, leadership and responsibility [13, 51]. Frameworks such as P21 and their concretisation in corresponding programmes tend to place extremely high demands on the next generation, as they are supposed to be prepared for all eventualities. In addition, such general specifications only create a very loose orientation framework for deriving concrete topics/contents to be dealt with in

teaching. Ultimately, one can also discuss with the present competence frameworks for the 21st century whether they themselves do justice to the demands they place on the following generation – such as critical thinking – since they do not use the present development patterns and the potentials of technology to draw designs for the future from them.

If we do not want to leave the current generation of pupils and students to their own devices, then we need an answer as to how to prepare them concretely for the future. In this respect, it seems necessary from our point of view to consider not only the long-term development but also the short-term and medium-term time horizon.

C. Development of short- and middle-term responses

Various approaches are possible to develop short-term responses. One approach is to extrapolate existing trends more or less linearly into the future. Another approach tries to find patterns in previous industrial revolutions and apply them to the current situation.

If we look at past industrial revolutions, they were usually characterised by innovations that allowed them to be used in different fields of application. Examples include the steam engine, assembly line production, electricity and information technology. Thus, the steam engine was able to replace human or animal power in a variety of fields of application. In earlier industrial revolutions, professional success was closely linked to the ability to apply these new technologies, according to the motto "Better to ride the wave than get caught under it", which led to the job description of the engineer, among other things [8].

In our view, artificial intelligence is one of the defining forces of current change. Artificial intelligence is pushing the automation of tasks into new areas. AI learns on the basis of large data sets or mimics skills that were previously reserved for humans, such as writing answers to users' questions.

In the current situation, the acquisition of skills in the field of artificial intelligence promises professional success. This is also made clear in a study by Bitkom Research [14]. According to this study, 70 percent of the companies surveyed expect that AI for text generation will be part of everyday working life in the future [14]. For this reason, the demand is also made that more knowledge about AI must be taught in schools and training [14].

The demand for more knowledge transfer makes it clear that there is a growing gap between the development and implementation of AI solutions and the teaching of AI competences. This also emerges, for example, from a study on students' competences in the use of ChatGPT [15]. This study revealed considerable deficits among students in the area of critical thinking. For example, the output of ChatGPT was often adopted by the students without reflection and not sufficiently questioned, compared with other sources or checked for internal consistency [15]. For this reason, closing this competence gap and increasing AI literacy will certainly be a building block to prepare the following generations for the AI age.

The problem of transferring historical patterns to the current situation, just like the linear projection of trends, is that systems do not always develop linearly. Rather, developments can lead to completely new system states [16]. In this context, we also speak of an emergent system [16]. An illustrative

example of this is water, which has different properties in different aggregate states (such as "wet").

Emergent systems combined with disruptive innovations and the great challenges of our time (such as climate change and wars) seem to make the undertaking of preparing for an AI age hardly possible. Especially since this undertaking is not without a certain cynicism from the perspective of the next generation. For the latter could be offended that a generation that will leave it many problems wants to prepare it for a new age.

If one views the world of work as an emergent system, then its future states are not predictable. Accordingly, there is no need to attempt a blueprint of the future from which one can deduce how the adaptation process of the younger generations to a future world of work should proceed. Instead, this brings to the fore the problems facing the next generation. The charm of this view lies in the fact that people's success has always depended on their problem-solving skills, regardless of the context.

If we look at the current changes from this perspective, we see that AI increases the complexity of tasks, as simple / repetitive tasks are automated by AI. In this respect, the education system should improve the ability to solve complex problems. Complex problems are characterised by the fact that they first require a problem analysis in order to explore the causes of the problem, that there is no standardised solution path and that various alternative solutions have to be developed, which then have to be evaluated.

D. Summary of assumptions and conclusions

The assumptions and conclusions of the different views can be summarised as follows:

Assumption [long-term perspective]: Due to the disruptive nature of AI (as well as a multitude of other possible innovations), we do not know the long-term future and therefore cannot assess what challenges it will bring and what skills younger generations will need to develop.

Conclusion: The adaptability of the younger generation to new situations should be improved. Starting points for this include promoting flexibility, lifelong learning and strengthening cognitive skills such as critical thinking.

Assumption [short and medium term perspective]: There is a rapidly growing gap between the need for and the availability of AI skills. This applies to the development and introduction of AI solutions in business practice as well as to their application in school, work and leisure.

Conclusion: The education system should better promote AI literacy.

Assumption: People's success is largely determined by their problem-solving skills, regardless of the context. At the same time, automation increases the complexity of tasks.

Conclusion: The problem-solving competence of complex problems should come to the fore.

III. WHAT IS AI LITERACY?

In order to define the concept of AI Literacy, it is first necessary to discuss what is meant by AI. This should be approached via the concept of tools and the significance of tools for humans.

A. Significance of tools for human cultural development

The development of tools plays a significant role in human history, because tools allow people to achieve cultural achievements that would be inconceivable without them [17]. Thus, the cognitive abilities of humans today are no more advanced than those of their ancestors many thousands of years ago, yet humans of the recent past can achieve significantly higher cultural achievements than were possible for their ancestors because of their skilful use of tools [18].

In order to understand what effects AI tools can have on education, the world of work and society as a whole, it is first important to address the question of what the basic support or benefit of AI tools for humans actually consists of. By clarifying this question, the potential but also corresponding limitations or even risks, legal and ethical issues of these technologies become clear. This is referred to as "AI literacy in the broader sense" when understanding AI as a tool. In addition, as with learning how to use any tool, there is the question of how to use the tool correctly, which is what the term "AI literacy in the narrower sense" is aimed at.

Computer-based tools in particular can support human cognitive processes in extraordinary ways by enabling representations of what is thought through multiple external representations [18]. These representations can be coupled with higher-order media functions (evaluations, manipulations, arrangements, etc.) to enable individuals to engage further between the thought and the objectified [18]. If cognitive processes are outsourced or supported by external aids, we also speak of "external cognition" in this context [19]. Human thought and cognitive processes experience considerable cognitive support through the use of external representations, since, for example, mental processes are relieved in the sense of "computational offloading" and capacity is thus created for more far-reaching cognitive processes [19]. In addition, mental processes can find new types and modes of support through digital media ("thinking stuff"), not only in terms of quantity but also in terms of quality, so that completely new kinds of cognitive possibilities arise that were not possible before [20].

B. Critical reflection on the potential of AI

In order to more closely define the upheavals that can arise from AI, it is necessary to assess which skills can be supported at all or possibly even completely replaced by AI tools in the future. To this end, it is first important to have a critical understanding of the concept of AI so that an overestimation of AI capabilities in the sense of a "false faith in technology" or "marketing promises" can be countered. A critical understanding of AI includes first of all the statement or distinction that AI in the sense of corresponding "computational models" by definition only simulates or imitates human cognitive functions, but is therefore not to be equated with them [21]. In other words, there is no equivalence of human intelligence with artificial intelligence today.

The crucial difference between human and artificial intelligence (in the sense of strong artificial intelligence) is expressed in the core of philosopher John Searle's Chinese Room argument, which means "formal computations on symbols cannot produce thought" [22]. AI based on today's computers, which operate on the principle of symbolic manipulation purely at the syntactic level, will not be able to completely replace higher human cognitive functions, such as

forward planning, problem solving, etc. [21], or even develop consciousness tomorrow. Nevertheless, and this is where the concept of a tool comes into play, AI can support these processes in a quantitative or also qualitative way, as explained above, and thus create significant benefits in many different contexts.

C. Working with AI as a tool

If AI is understood as a tool, then AI literacy can be understood as a set of skills for living, learning and working with AI technologies [1]. Accordingly, one approach to determining AI literacy is to derive corresponding competencies from AI use. For example, Kong et al. subsume under the term AI literacy the understanding of AI concepts, competences in the use of AI concepts for evaluation and the use of AI concepts for understanding the real world [23].

Based on a literature review, Ng et al. developed four aspects to promote AI literacy: 1. knowing and understanding, 2. using and applying, 3. evaluating and creating, and 4. ethical issues [1]. In doing so, the authors draw on Bloom's taxonomy of cognitive domains [1]. This is accompanied by a decisive advantage. While the majority of authors limit AI literacy to the teaching of basic concepts, skills, knowledge and attitudes without requiring learners to have prior knowledge [1], Bloom's taxonomy can be used to distinguish between different levels of AI literacy.

The distinction between different levels of AI literacy offers the opportunity to build it up gradually [24]. In this way, a playful engagement with AI can take place in the lower grades, while the upper grades reflect critically, for example [24]. In this way, students are not overtaxed and can be gradually introduced to the topic [24].

D. Dangers of AI, ethical and legal issues

Since the development and use of AI tools is also associated with a variety of dangers, it is also necessary to address these in order to derive the necessary areas of competence. The dangers include discrimination against groups, data protection, fairness and justice, surveillance and the reduction of competences [25]. For example, the ability of AI systems to imitate humans or human behaviour poses significant security problems. A recent example is the deepfake of a video conference between the governing mayor of Berlin and the alleged Vitali Klitschko. In addition, AI has long since arrived in organised crime. In March of this year, for example, a story made the rounds on the internet that fraudsters used AI to imitate the voice of grandchildren in order to get their hands on a Canadian grandmother's money [26].

In the long term, the use of AI can also lead to a relationship of dependency and a loss of competence on the human side [27]. If tasks are taken over by AI, then the need for humans to develop or possess their own competences disappears. For example, the ability to structure or formulate texts can unfortunately suffer from the fact that the AI abstracts such basic competencies further and further for humans. Another example is programming. Here, generative AI is now capable of generating functioning source code. However, assessing the safety and efficiency of the code and its integration into larger systems requires a detailed understanding of the code and the side effects and consequences of the functions used. If the automation of tasks is accompanied by a loss of competence on the human side, then this in turn can lead to a variety of dangers, such as a real

or even only perceived loss of control over the AI or low effectiveness in the collaboration between AI and humans [28].

In addition, the use of AI also goes hand in hand with a variety of legal aspects. These include copyright and patent law aspects, security aspects, equality and discrimination aspects, data protection, explainability and sector-specific regulations (for example in medicine) [29].

The long list of dangers is further extended by specific dangers of individual AI tools or individual application purposes and areas. For example, when using ChatGPT, students run the risk that generative AI makes plagiarism more likely, that ChatGPT's answers are sometimes wrong and that the source references are often incorrect.

E. Difficulty of understanding of AI literacy

The difficulty of this understanding of AI literacy is also that AI technologies are becoming increasingly heterogeneous as AI finds use in more and more products and services. For example, more than 100 AI applications were launched in April 2023 alone [30]. As a result, the competences that fall under AI Literacy are constantly expanding and there is a tendency for any competence to become a component of AI Literacy over time. To exaggerate, a competence such as driving a car could become part of AI Literacy at some point, namely when the car is equipped with more and more AI technologies.

In particular, a conception of AI literacy, such as that of Ng or Wong, makes it very difficult to distinguish it from other concepts: if we look at the definition of AI literacy according to Kong et al., it can be clearly distinguished from information literacy, which is about finding information, evaluating it and using it for personal, social or global purposes [31]. However, if we now look at an internet service such as ChatGPT, then AI literacy in the sense of Kong et al. is not sufficient for its use, as the service can also be used for information research.

F. AI literacy in a broader and narrower sense

In order to solve the problems outlined above, it makes sense to distinguish between AI literacy in a broader and narrower sense. If one wants to define AI literacy in a broader sense, then one starting point would be to look at its special features, which consequently requires an examination of the unique facets of AI.

Looking at the unique facets of AI compared to other technologies, these include greater autonomy, learning and its black-box nature [28]. Whereas in classical software development programmes were written to solve a task in a deterministic way according to a predefined sequence, AI uses learning algorithms that cause the results of AI systems to behave probabilistically in many cases. An example of this is the Microsoft-developed chatbot Tay, which was suspended within 24 hours of its release on Twitter for racist, sexist and anti-Semitic remarks it had previously been taught by Twitter users [32]. The inscrutability of AI arises from its complexity. For example, neural networks are so complex that it becomes difficult or even impossible for humans to understand them [33]. In the scientific literature, this is discussed in detail under the black box problem of AI.

A number of skills that are necessary in an AI era are derived from these specifics. This applies, among other things, to the ability to recognise AI as such [34]. In addition,

competences must be built up on how AI works, especially on how AI handles input, processes information and produces output [33]. Furthermore, it is important to critically question the output, as e.g. the patterns learned by AI have a bias.

From the perspective of Schuetz / Venkatesh, the specificities of AI challenge a number of assumptions on which research on information systems has been based so far [35]. These include that AI can be functionally inconsistent and non-transparent, and that users are often unaware of their AI use [35]. The specificities of AI also call into question the view of AI as a tool [28, 35, 36]. This is justified, among other things, by the growing autonomy of AI systems, such as bots that independently write posts in social media, trade securities on the stock market or make calls as call centre agents [35].

Berente et al. understand AI less as a phenomenon and more as a process [28]. They define AI as the limit of progress of IT systems in solving complex decision-making problems, using human intelligence as a reference benchmark [28]. According to McCorduck, this in turn leads to the paradox that when a problem is solved by AI, AI quickly loses its status as AI [36]. For example, while in the past systems combining expertise and rules were considered AI, this is usually no longer the case today, as no learning takes place in these. Since, on the other hand, various basic technologies such as neural networks or machine learning methods have been used successfully for many years and are still counted as AI, McCorduck's statement can only be agreed with to a limited extent. Regardless of this, one can basically agree with the assessment that AI is about the creation of new possibilities for intelligent behaviour of information technology systems or the application of existing possibilities to problem areas that have not yet been tapped. In addition, the subject area of AI is constantly changing, which is why Berente et al. also speak of a "moving target" [28].

G. Critical thinking

If this understanding of AI is taken as a basis, then, in addition to problem-solving skills (see Chapter II), critical thinking becomes a central competence of the AI age. This is because the use of AI opens up new fields. This requires a profound examination, among other things, of the consequences and dangers of AI use. This is also visible from the social discussion of AI, in which it is not uncommon for humanity's existence to be considered threatened by AI. For this reason, pupils and students need competences in the area of critical thinking, whereby competences are based on the combination of knowledge, skills and attitudes [16].

Unesco defines critical thinking as a "process that involves asking appropriate questions, gathering and creatively sorting through relevant information, relating new information to existing knowledge, re-examining beliefs and assumptions, reasoning logically, and drawing reliable and trustworthy conclusions" [37]. Essentially, critical thinking consists of asking questions about alternative ways to achieve a particular goal [38]. Accordingly, critical thinking takes place as an internal dialogue in which different alternatives are represented by mental models, which are then confronted with questions and evaluated, and in the process the reliability of the assessment must be tested [38]. Cohen thus places critical thinking in the context of decision-making, where there is uncertainty about the best alternative.

The assessment of the importance of critical thinking as a central competence field of the AI age is shared by large parts

of the literature [8]. Critical thinking is seen as a prerequisite for the application and ethical use of AI. Moreover, critical thinking is necessary to interpret information and to generate knowledge from it and to use this knowledge [39]. At the same time, AI can in turn stimulate critical thinking, for example by taking over (routine) activities and thus free up space for higher cognitive processes [1].

The importance of critical thinking also results from the fact that AI increases the complexity and dynamics of decision-making situations, for example because simple tasks are automated. In accordance with Ashby's law of required variety, the complexity and dynamics on the human side must also increase as a consequence, for example through more extensive mental models [40]. This in turn leads to the need for critical thinking.

Moreover, critical thinking is a counterweight to the characteristics of new technologies. According to Ariso, these technologies aim to avoid mental effort of any kind – such as critical thinking – in order to offer users an optimal experience [41]. Last but not least, critical thinking is also important because it is considered a prerequisite for the acquisition of other competences, such as information and media literacy [42].

Other authors also refer to uncertainty to justify the need for critical thinking [43]. The various mental activities that are summarised under the term critical thinking ultimately serve to reduce the uncertainty of a decision-making situation. These activities include observation, examining evidence, exploring alternatives, reasoning, testing conclusions, reconsidering assumptions and reflecting on the whole process [43]. Ultimately, critical thinking consists of the competent analysis of ideas and their appropriate application [8].

H. Interim conclusion

It makes sense to distinguish between AI literacy in a broader and narrower sense. AI literacy in the narrower sense focuses on the use of AI. AI literacy in the broader sense is about the innovations that AI brings with it and the associated potentials, limitations, risks, legal and ethical-moral aspects. In both cases, different degrees of AI literacy can be distinguished.

AI, with its autonomy, learning and black-box nature, brings with it various particularities that lead to a number of demands on pupils and students. These include, in particular, the need to understand these particularities.

In addition, AI breaks with various assumptions that have so far been associated with information systems. For this reason, an understanding of the process of information processing and the collaboration between humans and AI becomes important. Regarding the process, this concerns the following: 1. Input: How should one design the input to get the most useful output? For example, the usefulness of ChatGPT's responses depends on the content of the input, which has made prompt engineering popular. 2. Information processing: How does the AI process the incoming information? 3. Output: Critically examine the output as well as 4. the impact of deploying or using AI on oneself and/or other people. For example, if employees use ChatGPT, they are often not aware that the internet service may gain access to company secrets through their input.

AI is a process that aims to continuously expand the limits of the possibilities of problem solving through an information technology system. The associated development of new possibilities and fields requires not only ethical and moral competences but also a high level of competence in the field of critical thinking.

IV. RELEVANCE OF AI LITERACY

This chapter takes up the question of why we should increase AI literacy among pupils and students. After all, the curricula are full. Even without AI, numerous topics and fields of competence can be named that subsequent generations will or could need. Especially since AI is only a sub-field of computer science, it is often perceived as complex or difficult to understand from the students' point of view, and they are therefore intimidated by the subject matter [24]. Furthermore, the affinity of students towards computer science / AI is very heterogeneous. Therefore, good arguments are necessary to win them over for the topic. The reference to a future of whatever kind, which may come sooner or later, is unlikely to be sufficiently motivating. Can one expect a younger target group, which is itself in the process of development, to learn for stock, when they already have enough challenges to respond to currently? Even if one affirms the need to improve AI literacy, the question arises as to whose responsibility this is. Currently, 80% of the necessary learning takes place "on-the-job" [16] and is thus situationally embedded in the learners' action environment with all the associated benefits.

A. Importance in professional life

As described in the second chapter, the significance of AI in professional life depends largely on its further technological development and, in particular, how long it will take for a strong AI to emerge. At present, AI still has difficulty with various human skills, such as creativity, idea generation, cross-disciplinary expertise, and unconventional and lateral thinking [11]. As the boundaries of what AI is capable of are constantly shifting, it is also to be expected that the path to a strong AI will be gradual.

Assuming that it will take quite some time before a strong AI emerges, the importance of AI literacy for students in the short term results less from the fact that existing jobs will be completely automated and therefore disappear. Much more significant is augmented intelligence, i.e. the interaction of human and artificial intelligence on the one hand and the automation of subtasks / process steps on the other. In this way, quality improvements and efficiency gains can be achieved. Even small advantages in terms of quality or efficiency can then manifest themselves over time or via economies of scale in clear competitive advantages. An example of this is the call centre sector. If an AI is used to generate suggestions for the call centre agents as to which argument they should use next in a conversation, this can increase the chances of success (e.g. of closing a sale). Another example of augmented intelligence is dermatologists and AI working together to improve the quality of diagnoses. Efficiency gains can arise, for example, when sections of a programme or a text are generated in order to optimise them in a subsequent step or to integrate them into a larger programme / text.

If companies, organisations or individuals can increase their productivity or improve the quality of their work results with the help of AI, this quickly puts pressure on other companies, organisations or individuals to also use AI. If, for

example, the use of generative AI increases the language quality of an academic paper and this is included in the grading, then every additional student who uses generative AI creates further pressure on other students who do not use this tool. For this reason, it can be assumed that AI will become increasingly widespread in the world of work and that collaboration between humans and AI in the sense of augmented intelligence will become more widespread.

B. Interaction with AI systems

With the spread of AI, the importance of interacting with AI systems is increasing [44]. For example, AI-based algorithms have now become standard on the major social media platforms. Whether on Twitter, TikTok or the like, they decide which posts / contributions are played out to the individual users. The algorithms try to determine which posts are most likely to be clicked on. However, this approach is closely associated with negative phenomena such as filter bubbles and doomscrolling [45]. AI literacy can help counteract such negative phenomena by raising awareness and providing alternatives.

C. Understanding the world

The complexity of the world has increased significantly in recent decades even without AI [46]. AI has certainly contributed to this complexity increasing further. On the other hand, AI offers numerous approaches to reduce complexity. After all, one of the central strengths of AI is finding patterns in the data, which helps us to understand the world and find appropriate answers.

D. Response to complex challenges

One example of this is generative AI, which is a complex challenge for many media professionals, and one to which there are often no simple answers. On the other hand, AI is also a response to the challenges. This is why the INMA (International News Media Association) also comes to the conclusion in a report that AI offers news media companies significantly more opportunities than risks [47]. In its report, INMA lists numerous AI tools that improve productivity, make it easier to reach new target groups (e.g. by automatically outputting texts in audio or translating them into other languages) or automatically checking comments on the website [47]. This approach follows the logic that whom you cannot defeat, you better ally with.

E. Social aspects

A general promotion of AI literacy helps to avoid social divides and allows broad sections of the population to participate in the benefits of AI [48]. Digitalisation has transformed consumers from passive recipients to socially networked content creators. Participation in these networks is increasingly linked to AI. An example of this is the Creative User Empowerment project of the Badisches Landesmuseum. In this, the image of visitors was replaced by users, in whose service the AI is placed, for example, to give them recommendations that fit their interests and knowledge base [49]. In addition, they are given the opportunity to create their own productions in a co-creative development environment [49].

F. Ethical aspects

In addition, the use of AI also has an ethical dimension. If, for example, a service such as ChatGPT is used to search for information, then texts are used to generate the answers where their authors regularly do not appear and, as a rule, have not

been asked for their consent. Consequently, the authors do not receive any monetary compensation. This is not only highly questionable from an ethical point of view, it also raises the question of what consequences this will have. One possible consequence is that in future content will be placed behind a (pay) wall so that it cannot be used to train a large language model.

E. Certainties are questioned / boundaries blurred

As explained in the third chapter, AI aims to push the boundaries of the capabilities of IT systems further and further. In this way, existing certainties are called into question and familiar boundaries become blurred. For example, while spam emails could be recognised comparatively accurately in the past, especially due to the peculiarities of the German language, generative AI makes this much more difficult. The creation of deepfakes is becoming more and more efficient, requires less and less expertise, and in many cases reaches such a high level that it is difficult to distinguish them from reality. The same applies to AI systems that imitate humans, for example by independently making telephone calls, making recommendations or conducting dialogues with users as chatbots.

Every advance in AI always means that existing certainties are challenged. Current examples include ChatGPT, which is often perceived as a threat in schools and colleges. If we want to create AI literacy, we need to explore the current limits of what IT systems can do and where the limits are currently shifting. In this way, certainties can be readjusted and future developments can be anticipated and prepared for.

The aforementioned aspects show various benefits of an engagement with AI. Whether this is sufficient to encourage a more comprehensive engagement with AI literacy, however, remains open. The experiences of other subject areas, such as data analysis or scientific work, which are highly significant from an academic point of view, do not lead to a corresponding engagement either, despite good arguments. In this respect, engagement of learners requires connection to their personal goals.

V. CONCEPT OF RESEARCH-BASED LEARNING

The dynamics of AI as a “moving target” requires a corresponding dynamic of the didactic concept. Static contents are therefore not appropriate for the subject area. For this reason, this concept is based on learning that accompanies research. The learners thus have the opportunity to actively explore the limits of what AI can do today. This exploration of the limits, combined with deeper learning through active participation, creates the learning experiences necessary to build AI literacy. In this respect, research-based learning is, in our view, an ideal form for improving AI literacy.

Research-based learning is understood here to mean that a researcher accompanies project groups of students or pupils throughout the entire research process. The researcher takes on the role of a coach. He/she supports the project groups in all phases of the research project. The coach contributes his or her expertise on the state of research and provides initial ideas for research projects that fit the skills and goals of the project groups. Together with the project groups, the coach reflects on the procedure as well as the progress of the project and provides input on the individual work phases (e.g. on the methods and tools that can be used for individual work steps).

The coach can provide assistance in solving subtasks or also make his or her own contributions to the project. The coach thus acts as an enabling person and does important groundwork in understanding AI, especially in the initial phase.

In order to approach AI literacy in a broader sense, a conceptual understanding of AI is first required. Appropriate impulse lectures by the coach and group discussions are intended to create a relevant prior knowledge regarding the basic concepts among the students as well as pupils. The aim is to understand fundamental concepts and functions of AI in order to be able to make realistic assessments about potentials, applications and possible risks. This forms a central prerequisite for being able to carry out a subsequent competence development of AI literacy in the broader sense. Or metaphorically speaking: Only when the nature and functioning of the tool is understood can it be used in a beneficial and responsible way.

The participants are actively involved in the research project in the sense of “learning AI by doing AI”. The responsibility for the success of the project lies with the project groups. Therefore, they also bear the responsibility for the project’s success. This follows the self-organisation and self-responsible learning called for in the literature [16]. Self-organisation allows the project groups to divide their work. This ties in with the realisation that in business practice it is crucial to build up the necessary competences within the team rather than with each individual. Taking on a complete project also has the advantage that students are encouraged to think and work holistically. The solution of practical problems by the project groups serves to promote the problem-solving skills of the learners. In addition to knowledge about AI, working on a project also creates the opportunity to build subject-specific knowledge.

A. Open participation

Participation in the projects accompanying the research is largely open. In addition to students and pupils, the target group also includes employees of a company who would like to gain further qualifications or people who are interested in the topic. In order to test the concept, the first phase will be limited to dual partners, students of the Duale Hochschule Baden-Württemberg and other universities, and pupils of the upper secondary school. The opening up of the learning programme is intended both to promote a successful professional career in the sense of lifelong learning and to increase regional prosperity. In addition, opening up creates new social constellations, which stimulates divergent thinking and thus promotes creativity [16].

There are various advantages associated with research-based learning. For example, the team of researchers creates an organisational framework and the support of a researcher helps the project groups to overcome many hurdles. In addition, the concept of research-based learning as a form of deeper learning aims to increase the flexibility of students.

B. Need for deeper learning

Learners today are often accused of learning superficially at best and only learning from test to test, which is summarised by the term “bulimic learning” [50]. Like a Teflon pan, many teachers feel that their efforts do not stick. The education system, on the other hand, is accused of placing too much emphasis on the mere reproduction of what has been learned, which is losing importance in today’s world where knowledge

can be accessed at any time via the Internet and is poor preparation for the challenges ahead [42].

In contrast, deeper learning aims at a profound understanding of knowledge. Through the learning process, the learner should be able to assess how, why and when the acquired knowledge can be used, derive general principles, rules and patterns, establish relationships between different ideas and information and apply what has been learned to new situations. In this respect, the concept promotes the employability of the learners and they contribute to avoiding dangers and strengthening competitiveness in the practical application in business practice.

C. Regular reflection periods

The dangers associated with AI as well as the promotion of critical thinking require that regular reflection phases take place. These include questioning the goals and methods, discussing ethical aspects and possible dangers such as discrimination against social groups. The reflection phases also serve to discuss interim results as well as to link theory and practice. Furthermore, they improve problem-solving skills by giving learners the opportunity to abstract from the chosen approach and to derive generalised learnings from the project.

D. Procedure

The process aims at confronting learners with all the tasks and challenges that arise in the development and/or application of AI solutions. This should lead to deeper learning and to the learners being able to transfer what they have learned to new situations, as is expected of them later in their professional life.

1. Kick-Off

- Input: Presentation of the research programme as well as the research status by the researcher and presentation of open questions
- Formation of working groups
- Determination of the objectives by the working groups

2. Planning and structuring of the project

- Input: Description of the process based on previous projects
- Problem analysis
- Determination of next steps by the working groups

3. Development of results

- Input: AI concepts / methods
- Elaboration of the results by the working groups
- This is accompanied by regular reflection meetings on the development and learning progress

4. Presentation and reflection of the results

- Presentation of the results by the working groups
- Reflection on the results, the methodological approach and the learnings

- Outlook: open questions / ideas

The mind map (fig. 1) summarises the most important results of the discussion paper.

VI. FEEDBACK ON THE DISCUSSION PAPER

In order to obtain feedback on the discussion paper and to be able to develop it further, subject experts, secondary school teachers as well as students of business administration and business informatics data science were surveyed in semi-structured interviews and focus groups.

A. Conceptual understanding and critical thinking

The interviewees agreed that a basic conceptual understanding of AI is an essential prerequisite for a critical approach to the technology. This is pointedly brought to the point by the statement of an interviewed expert "AI is basically stupid". It should be made clear that AI only works with probabilities, i.e. for example what the most probable answer to the question asked is, but understanding in the sense of semantics does not take place. Therefore, a critical reflection on the technology is important, also in order to be able to make an appropriate assessment of its actual potential.

Critical thinking also emerged as the most significant competence from the written survey of three focus groups of students. In this, critical thinking was rated on a five-point scale (from 1 = not at all important to 5 = very important) with an average of 4.84 (n = 43). The panelists were presented with the thirteen competencies that emerged from the discussion paper. Ten of the thirteen competencies were rated higher than 4.0 on average. Only knowledge of the development dynamics of AI (\bar{x} 3.47, n = 43), understanding how AI works (\bar{x} 3.93, n = 43) and ethical-moral aspects (\bar{x} 4.00, n = 43) were slightly lower in the rating of importance. In addition to critical thinking, knowledge of the risks of AI and problem-solving skills were rated as particularly significant (\bar{x} 4.63 and 4.60 respectively, n = 43).

B. Importance of AI literacy in education

Regarding the importance of AI Literacy in schools, it became clear that the focus should be on developing the so called 4Cs as 21st century skills: critical thinking, communication, collaboration and creativity [51]. In development of these competencies, it is important to differentiate according to the cognitive abilities of the pupils. Corresponding teaching and learning opportunities in lower, middle and upper school are to be differentiated in terms of the degree of complexity. It is advocated that students be introduced to meaningful use gradually and through constructive engagement. Of outstanding importance in the school context across all grades is the topic of the proper use of information to complete school tasks. This refers in particular to competencies regarding the collection, structuring and evaluation of information. The overarching and central goal competence should be to develop a critical assessment competence of information in the course of the pupils' school careers.

Since the ability to think abstractly cannot be assumed until the pupils reach around 7th grade, it is advocated that the topic of AI competence is taught in the lower grades in a more rule-based and guided manner, i.e. in the form of "do's and don'ts". In this regard, good empirical values are already available through other technical topics such as responsible social media use. It is also important that students in lower

grades receive a basic education on the foundations of information technology, i.e. for example on what an algorithm is or on other core concepts. This is seen as an important prerequisite for the understanding of topics around AI. The use of AI tools becomes particularly relevant from the 7th grade onwards, when students are expected to independently formulate texts with the help of sources as part of a term paper or project documentation. Here, competencies in recognizing text structures, comparing texts, analysing deep structures, and structured content reproduction are important. Furthermore, the ethical dimension should be reflected upon at an early stage, which is the subject of religion and ethics lessons. Here, the question of what makes people human and how they are influenced by digitization is addressed. Accordingly, it can be connected what distinguishes humans from AI and how their reality of life is influenced by AI artifacts.

As teaching formats, short explanatory videos are specifically mentioned here, which show “learning by doing” and can support in finding and implementing meaningful and less meaningful uses. As the grade level progresses, the competencies should be deepened and more advanced ethical or even legal topics should be discussed. In the upper school, for example, a deepening and expansion of competencies with regard to the critical and correct use of sources can take place within the framework of the seminar course. Here, concrete examples, i.e. the work to be produced by the students, could be used to explore what benefits AI can bring to one’s own work, but also where the performance of AI falls short of expectations and requirements. Here AI can take the role of a personal assistant, teaching buddy or tutoring system providing individual feedback on one’s learning journey.

As a further idea for teaching, it is suggested that pupils and students are introduced to the topic through application, which should create additional motivation. There is a high level of interest in learning how to use AI to solve concrete tasks and problems. Among other things, the preparation of a scientific paper, the support of the learning process or processes in general, and the automation of tasks were mentioned.

In order to better understand the limitations of AI and to be able to critically reflect on its potential, it is suggested that, as part of the teaching, the task is set of tricking the AI by means of deliberate inputs, i.e. deliberately having it produce incorrect results. Teams of students and pupils could compete against each other in the sense of a capture-the-flag game.

C. AI literacy of teaching staff and supporting structures

Last, it became clear that developing AI literacy also means that teaching staff needs to have the appropriate skills. In this sense development of AI literacy must take place at the teachers' level as well. Although the school curriculum today already provides a framework in which AI literacy can be promoted at many points, it depends crucially on the respective teacher and the organizational structures of the single school how strongly and interwoven the topic can actually be made a subject. Therefore, it is certainly worth discussing whether further curricular development should include the promotion of AI competencies much more explicitly. However, there is also the opinion amongst teachers that we do not need to further extend or refine the current curriculum. Rather should contents be removed from the curriculum in order to free up space for more flexibility and new teaching and learning formats. This should allow for

more dynamic and self-organized ways of learning in terms of a “new normality” in education [52]. The organizational structures of schools are a central determinant for the success of such new formats, which are found to be important for – but are not limited to – the development of AI competencies.

In summary, the idea of promoting AI literacy as part of a project accompanying the research was well received by the focus groups. On a five-point scale, this idea received an average of 4.05 out of five possible stars ($n = 43$). What students particularly like about this idea is the open-ended approach to the technology, the achievement of a deeper understanding, the engagement with a subject area of great future significance, and the acquisition of skills that can be applied practically.

VII. CONCLUSION

This discussion paper aimed to make a contribution by adding insights to answering the research question of how to prepare pupils and college students for the AI age. In doing so, the paper has raised various fundamental and complex sub-questions necessary for answering the research question. A hermeneutic literature review identified different viewpoints on these questions. In the following steps, several fields of competence were derived from the identified viewpoints and a concept laying out cornerstones for competency formation was developed. Finally, an empirical evaluation of the insights of the discussion paper was conducted using semi-structured interviews and focus groups. The discussion in the context of the conducted evaluation revealed a broad consensus regarding the importance of AI literacy and central areas of competence such as critical thinking, problem-solving and knowledge on the risks of AI. It was also underlined that developing AI literacy as part of project and research-based learning approaches seems to be a promising approach. In this context, the students often expressed the wish for a practical application of AI. Furthermore, it became also apparent that there is a high need for more flexible and dynamic ways of learning in schools and colleges to support the formation of AI competencies in an efficient and effective manner. In summary, students seem to be fascinated by the possibilities created by AI and are very interested in using them productively for their own purposes. From this, a broad field of possibilities for integrating AI into teaching can be derived, such as in the context of scientific work or projects in the area of research-based learning. Further work should focus on the design, implementation and evaluation of such formats. Furthermore, structural questions on the appropriate organisational frames and contexts for enabling such new formats must be addressed practically and from a research perspective.

REFERENCES

- [1] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, “Conceptualizing AI literacy: An exploratory review,” *Computers and Education: Artificial Intelligence*, vol. 2, p. 100041, 2021, doi: 10.1016/j.caeai.2021.100041.
- [2] R. Buchkremer, T. Heupel, and O. Koch, Eds., *Künstliche Intelligenz in Wirtschaft & Gesellschaft*. Wiesbaden: Springer Fachmedien Wiesbaden, 2020.
- [3] M. Brussevich, E. Dabla-Norris, C. Kamunge, P. Karnane, S. Khalid, and K. Kochhar, *Gender, Technology, and the Future of Work*. Washington, D.C: International Monetary Fund, 2018. [Online]. Available: <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1915003>.
- [4] P. Buxmann and H. Schmidt, Eds., *Künstliche Intelligenz*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019.

- [5] Y. N. Harari, 21 Lektionen für das 21. Jahrhundert, 1st ed. München: C.H. Beck, 2018. [Online]. Available: <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6991678>.
- [6] M. Rueckert and M. Riedl, "Human-in-the-Loop: Wie Mensch und KI Aufgaben besser lösen," *Digitale Welt*, vol. 6, no. 4, pp. 36–39, 2022, doi: 10.1007/s42354-022-0542-x.
- [7] K.-L. A. Yau, H. J. Lee, Y.-W. Chong, M. H. Ling, A. R. Syed, C. Wu and H. G. Goh, "Augmented Intelligence: Surveys of Literature and Expert Opinion to Understand Relations Between Human Intelligence and Artificial Intelligence," *IEEE Access*, vol. 9, pp. 136744–136761, 2021, doi: 10.1109/ACCESS.2021.3115494.
- [8] Aoun, Joseph (2017): Robot-proof. Higher education in the age of artificial intelligence. Cambridge, Massachusetts: The MIT Press (The MIT Press Ser). Online verfügbar unter <https://permalink.obvsg.at> J. Aoun, Robot-proof: Higher education in the age of artificial intelligence. Cambridge, Massachusetts: The MIT Press, 2017. [Online]. Available: <https://permalink.obvsg.at/>.
- [9] J. Osterhammel, "1850 bis 1880," Bundeszentrale für politische Bildung, 19 Jan., 2022. <https://www.bpb.de/shop/zeitschriften/izpb/das-19-jahrhundert-315/142117/1850-bis-1880/?p=all> (accessed: Apr. 4 2023).
- [10] McKinsey, JOBS LOST, JOBS GAINED: WORKFORCE TRANSITIONS IN A TIME OF AUTOMATION. [Online]. Available: <https://www.mckinsey.com/~/media/mckinsey/industries/public/2020and%20social%20sector/our%20insights/what%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/mgi-jobs-lost-jobs-gained-executive-summary-december-6-2017.pdf> (accessed: May 12 2023).
- [11] A. M. Recanati, AI Battle Royale. Cham: Springer International Publishing, 2023.
- [12] R. Molar Candanosa, Could future computers run on human brain cells? [Online]. Available: <https://hub.jhu.edu/2023/02/28/organoid-intelligence-biocomputers/> (accessed: May 12 2023).
- [13] T. Borrowski, The Battelle for Kids P21 Framework for 21st Century Learning, 2019. Accessed: Apr. 20 2023. [Online]. Available: <https://measuringtel.casel.org/wp-content/uploads/2019/08/awg-framework-series-b.9.pdf>.
- [14] Bitkom, ChatGPT & Co.: Jedes sechste Unternehmen plant KI-Einsatz zur Textgenerierung. [Online]. Available: <https://www.bitkom.org/Presse/Presseinformation/ChatGPT-Jedes-sechste-Unternehmen-plant-KI-Einsatz-Textgenerierung>.
- [15] U. Bucher and N. Klein, "ChatGPT – Are the students ready for the AI age?," 2023.
- [16] U.-D. Ehlers, Future Skills. Wiesbaden: Springer Fachmedien Wiesbaden, 2020.
- [17] T. G. Wynn, "The evolution of tools and symbolic behaviour," in *Handbook of human symbolic evolution*, C. R. Peters and A. Lock, Eds., Oxford, UK: Blackwell, 1999, pp. 263–287.
- [18] R. Keil-Slawik, "Zwischen Vision und Alltagspraxis: Anmerkungen zur Konstruktion und Nutzung typographischer Maschinen," in *Neue Medien Im Alltag: Begriffsbestimmungen Eines Interdisziplinären Forschungsfeldes*, K. Boehnke, Ed., Wiesbaden: VS Verlag für Sozialwissenschaften GmbH, 2000, pp. 199–220.
- [19] J. Preece, Y. Rogers, and H. Sharp, Interaction design: Beyond human-computer interaction. Chichester: Wiley, 2015.
- [20] R. Keil, "Das Differenztheater. Koaktive Wissensarbeit als soziale Selbstorganisation," in *Automatismen*, Leiden, Niederlande: Brill | Fink, 2010, pp. 205–229.
- [21] M. W. Eysenck and M. T. Keane, Cognitive psychology: A student's handbook. London, New York: Psychology Press Taylor & Francis Group, 2015.
- [22] D. Cole, "The Chinese Room Argument," in *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), E. N. Zalta and U. Nodelman, Eds., 2023. Accessed: 1.5.23. [Online]. Available: <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/>.
- [23] S.-C. Kong, W. Man-Yin Cheung, and G. Zhang, "Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100026, 2021, doi: 10.1016/j.caeai.2021.100026.
- [24] G. K. W. Wong, X. Ma, P. Dillenbourg, and J. Huan, "Broadening artificial intelligence education in K-12," *ACM Inroads*, vol. 11, no. 1, pp. 20–29, 2020, doi: 10.1145/3381884.
- [25] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," *SSRN Journal*, 2020, doi: 10.2139/ssrn.3518482.
- [26] A. Puig, Scammers use AI to enhance their family emergency schemes. [Online]. Available: <https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes> (accessed: May 5 2023).
- [27] A. Fügner, J. Grahl, A. Gupta, and W. Ketter, "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI," *MISQ*, vol. 45, no. 3, pp. 1527–1556, 2021, doi: 10.25300/MISQ/2021/16553.
- [28] N. Berente, B. Gu, J. Recker, and R. & Santhanam, Managing Artificial Intelligence. *MIS Quarterly*, 45(3), 1433-1450, 2021.
- [29] M. Hartmann, KI & Recht kompakt. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020.
- [30] Generative AI, Blog-Beitrag auf LinkedIn. [Online]. Available: https://www.linkedin.com/posts/genai-center_100-ai-tools-activity-7057190420866334720-_Hoi?utm_source=share&utm_medium=member_desktop (accessed: May 4 2023).
- [31] R. N. Mishra and C. Mishra, Relevance of information literacy in digital environment. In: *Journal of Emerging Trends in Computing and Information Sciences* (1/2010), S. 48-54, 2010. [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2a2fef4fce76f2b4be39cce683bf90c373e8ad55#page=52>.
- [32] M. J. Wolf, K. Miller, and F. S. Grodzinsky, "Why we should have seen that coming," *SIGCAS Comput. Soc.*, vol. 47, no. 3, pp. 54–64, 2017, doi: 10.1145/3144592.3144598.
- [33] M. Pinski and A. Benlian, AI Literacy-Towards Measuring Human Competency in Artificial Intelligence. Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL), 2023.
- [34] D. Long and B. Magerko, "What is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA, 2020, pp. 1–16.
- [35] S. W. Schuetz and V. Venkatesh, The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction, *Journal of the AIS* (21:2), 2020, 460-482., 2020.
- [36] P. McCorduck, Machines who think: A personal inquiry into the history and prospects of artificial intelligence. New York: CRC Press, 2004.
- [37] UNESCO, Critical thinking. [Online]. Available: <https://www.ibe.unesco.org/en/glossary-curriculum-terminology/c/critical-thinking> (accessed: May 12 2023).
- [38] M. Cohen, A Three-Part Theory of Critical Thinking: Dialogue, Mental Models, and Reliability., 2000. Accessed: Feb. 16 2023. [Online]. Available: https://www.researchgate.net/profile/Marvin-Cohen-2/publication/228751544_A_Three-Part_Theory_of_Critical_Thinking_Dialogue_Mental_Models_and_Reliability/links/0deec515e32c0d7dbf000000/A-Three-Part-Theory-of-Critical-Thinking-Dialogue-Mental-Models-and-Reliability.pdf.
- [39] UNESCO, Towards Knowledge Societies. Paris: Editions UNESCO, 2005. Accessed: 15 of April 2023. [Online]. Available: <http://unesdoc.unesco.org/images/0014/>.
- [40] O. Mack, A. Khare, A. Krämer, and T. Burgartz, Eds., Managing in a VUCA World. Cham: Springer International Publishing, 2016.
- [41] J. M. Ariso, "Is Critical Thinking Particularly Necessary when Using Augmented Reality in Knowledge Society? An Introductory Paradox," in *Augmented Reality*, J. M. Ariso, Ed.: De Gruyter, 2017, pp. 3–22.
- [42] E. Manalo, "Introduction," in *Deeper learning, dialogic learning, and critical thinking: Research-based strategies for the classroom*, E. Manalo, Ed., Milton: Routledge, 2020, S. 1-13.
- [43] J. M. Spector and S. Ma, "Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence," *Smart Learn. Environ.*, vol. 6, no. 1, 2019, doi: 10.1186/s40561-019-0088-z.
- [44] J. Harrod, AI Literacy, or Why Understanding AI Will Help You Every Day. [Online]. Available: <https://www.youtube.com/watch?v=cK19QsVv7hY> (accessed: May 1 2023).
- [45] A. Bruns, Are filter bubbles real? Cambridge: Polity Press, 2019. [Online]. Available: <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5887595>.

- [46] R. McGrath, The World Is More Complex than It Used to Be. [Online]. Available: <https://hbr.org/2011/08/the-world-really-is-more-compl> (accessed: Apr. 30 2023).
- [47] J. Walters, “Generative AI opportunities outweigh threats”: INMA report lists tools publishers can use now. [Online]. Available: <https://whatsnewinpublishing.com/generative-ai-opportunities-outweigh-threats-inma-report-lists-tools-publishers-can-use-now/> (accessed: Apr. 30 2023).
- [48] I. Lee, S. Ali, H. Zhang, D. DiPaola, and C. Breazeal, “Developing Middle School Students’ AI Literacy,” in Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, Virtual Event USA, 2021, pp. 191–197.
- [49] J. C. Bernhardt, Partizipation und künstliche Intelligenz: Perspektiven aus dem Badischen Landesmuseum. In: Friesen, K. / Mohr, H. (Hrsg.) (2022): Agilität - Digitalität - Diversität - Zukunftsthemen einer innovationsorientierten Kulturpraxis und Kulturpolitik, S. 76-79. [Online]. Available: https://www.researchgate.net/publication/362582885_Partizipation_und_Kunstliche_Intelligenz (accessed: May 17 2023).
- [50] T. Städtler, Die Bildungs-Hochstapler: Warum unsere Lehrpläne um 90% gekürzt werden müssen. Heidelberg: Spektrum Akademischer Verlag, 2010.
- [51] Battelle for Kids, Framework for 21st Century Learning. [Online]. Available: https://static.battelleforkids.org/documents/p21/P21_Framework_Brief.pdf (accessed: May 25 2023).
- [52] OECD, OECD Lernkompass 2030. [Online]. Available: https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/OECD_Lernkompass_2030.pdf (accessed: May 25 2023).

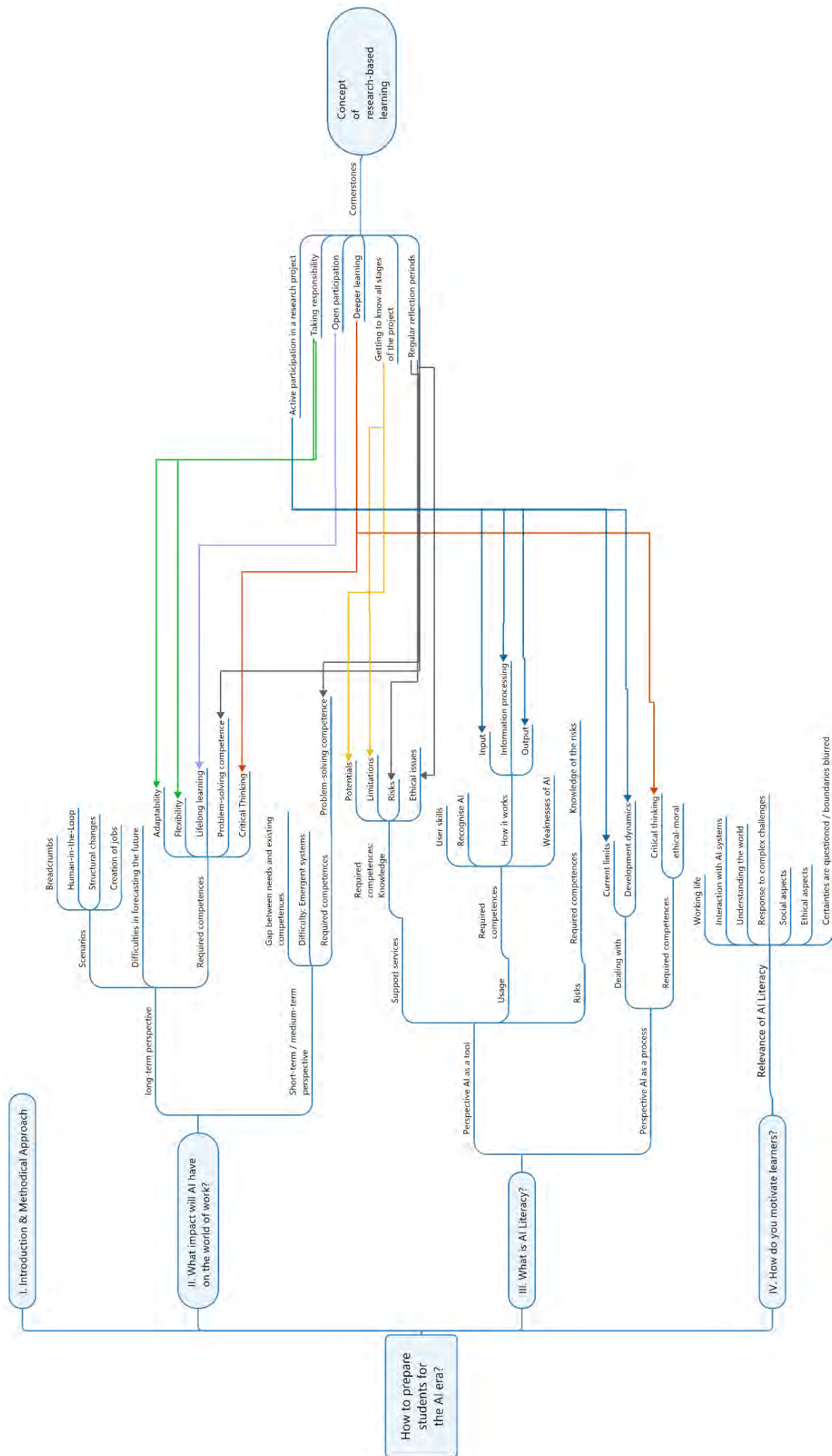


Fig. 1: Mind map of discussion paper

Digital Twins von Organisationen (DTO) für die Personalarbeit: KI als Ersatz oder Unterstützung des HR-Bereichs

Benedikt Hackl

*Duale Hochschule Baden-Württemberg und TDHBW –
Forschungszentrum Management Analytics.*

Ravensburg and Munich, Germany
hackl@dhbw-ravensburg.de

Joachim Hasebrook

TDHBW – Forschungszentrum
Management Analytics

Munich, Germany
hasebrook@management-analytics.de

Abstract— Ein "Digital Twin" einer Maschine (DTM), z. B. einer Windkraftanlage, ist ein digitales Abbild der Maschine und erhält von dieser fortlaufend Messergebnisse – allein bei einem Flügel werden permanent rund 10 Messwerte von Sensoren Mikro-Controller in jedem Flügel erhoben. Dadurch können die teuren Anlagen vorausschauend und kostenschonend gewartet sowie Risiken oder gar Unglücke vermieden werden. Zunehmend werden auch Digital Twins für Organisationen konzipiert und eingesetzt. Für solche „Digital Twins“ von Organisationen (DTOs) gibt es derzeit kaum Empfehlungen und Einsatzerfahrungen. Es fehlt ein Bild dafür, wie sich DTO von DTM unterscheiden und welchen Einfluss sie auf die Steuerung einer Organisation haben werden. Diese Frage stellt sich insbesondere für den Personalbereich: Dieser wird entweder zu einer relevanter Steuerungseinheit im Unternehmen oder von Künstlicher Intelligenz (KI) ersetzt, die zunehmend zentrale steuernde Aktivitäten übernimmt. Mit unserer Forschung und Entwicklung wollen wir dazu beitragen,

- Ansätze für DTO im Personalbereich entwickeln,
- konkrete Entwicklungs- bzw. Erfolgsbeiträge von DTO zu erfassen und
- Rollen und Aufgaben von Personalbereichen bei der Nutzung KI-basierter DTO zu erproben.

Keywords—*Digital Twin, data-based HR solutions, live-time analytics and benchmarking*

I. VERÄNDERTE ANFORDERUNGEN IN DER UNTERNEHMENSSTEUERUNG

Das Interesse am Einsatz von KI und der Entwicklung von „Digital Twins“ von Organisationen (DTO) ist gewachsen, weil eine vorausschauende und anpassungsfähige Unternehmenssteuerung über Maschinen und Finanzkennzahlen hinaus auch das Personal als zentrale Quelle für Kreativität und Innovation im Unternehmen umfassen soll. „Predictive Analytics“ auf Basis von KI-basierten Digital Twins spielt vor allem in der Maschinen- und Produktionssteuerung eine Rolle und sehr viel eingeschränkter in der Steuerung von Finanzgeschäften sowie bei Marketing und Sales. Der Personalbereich ist in diese Entwicklung bisher kaum einbezogen. Die Steuerung des Personals erfolgt weiterhin oftmals auf Basis einfacher Kennzahlen wie Personalkosten, -kapazitäten und -fluktuation. Bei einigen Unternehmen kommen

Transaktionsdaten, wie z. B. die Durchlaufzeiten und Annahmequoten in der Personalgewinnung hinzu. Die eigentlich für das Management und die Führung relevanten Aspekte wie z. B. Leistungs- und Anpassungsbereitschaft, Lern- und Veränderungsfähigkeit, Teamfähigkeit und Teamkonflikte werden gar nicht und bestenfalls in umfassenden Mitarbeiterbefragungen in ein- oder zweijährigem Rhythmus erfasst. Auf dieser Datenbasis ist eine vorausschauende und für die Unternehmenssteuerung relevante Management Analytics nicht möglich. Damit wird die zentrale Quelle von Kreativität und Leistung im Unternehmen, also seine Innovationskraft, nicht aktiv und vorausschauend gemanagt.

Die Innovationskraft eines Unternehmens ist jedoch entscheidend für die Sicherung der Marktposition und den Ausbau von Wettbewerbsvorteilen. Deshalb steigt der Innovationsdruck, um zukunftsfähig bleiben zu können. Viele Unternehmen können ihre eigene „Innovationsfitness“ kaum beurteilen. Sie wissen nicht, welche Weichen sie stellen müssen, um ihre Innovationsfähigkeit zu erhöhen – und sie haben keine Vorstellung davon, welche Rolle der Personalbereich für die Weiterentwicklung dieser Innovationsfähigkeit spielt.

II. EIGENE STUDIEN ZUR INNOVATIONSKRAFT DEUTSCHER UNTERNEHMEN

A. Kernergebnisse

Management Analytics ist in einer der aktuell größten deutschen Studien diesen Themen nachgegangen. Insgesamt haben 1.400 Führungskräfte und Mitarbeitende vorwiegend aus mittelständischen Baden-Württembergischen Unternehmen an dieser Studie teilgenommen (vgl. Hackl/Hasebrook/Baumann 2022). Diese Studie ist die neueste in einer Reihe von mittlerweile 12 Studien, in denen untersucht wurde, welche Steuerungsgrößen muss das HR aktiv zum Thema machen, um die Innovationskraft von Unternehmen zu steigern (vgl. Liste der Artikel und Studien am Ende).

Wie sich anhand der Daten zeigt, gibt es tendenziell zwei Arten von Unternehmen: eher dynamische und eher statische. Erstere zeichnen sich unter anderem durch flache Hierarchien, eine höhere

Anpassungsfähigkeit, Agilität, höhere Reaktionsgeschwindigkeiten und die Integration neuer Arbeitsweisen aus. Eher statische Unternehmen zeigen dagegen Tendenzen zu strengen Hierarchien, viel Bürokratie, festen Strukturen, geringer Kommunikationsbereitschaft gegenüber Mitarbeitenden und langsamen Entscheidungsfindungen. Wenig überraschend ist, dass dynamische Unternehmen deutlich innovationsorientierter sind als statische. Bemerkenswert ist hingegen, dass die Bedeutung von Organisationsstrukturen, Hierarchiestufen und festen Regeln deutlich geringer eingestuft wird als bei statischen Unternehmen: Klassische Hierarchien und Rollenmuster, erwartbare Handlungs- und Ablauflogiken und eine standardisierte Unternehmenskommunikation sind für dynamische, innovationsstarke Unternehmen kaum von Bedeutung. Die Steigerung der Innovationsfähigkeit erfolgt also nicht dadurch, dass bestehende Organisationsstrukturen und -regeln angepasst werden.

Wie die Datensätze nahelegen, sind Teams mit gut funktionierender Zusammenarbeit der zentrale Faktor zur Steigerung der Innovationsfähigkeit: Herausforderungen können im Team besonders gut gelöst werden. „Team“ ist nicht zwingend mit „Abteilung“ gleichzusetzen. Vielmehr verstehen wir darunter einen Zusammenschluss von Menschen in Arbeitssituationen, der interdisziplinär, themenbezogen, projektbasiert, hierarchie- und eben auch abteilungsübergreifend sein kann. Gerade die Loslösung aus alten, starren Organisationsstrukturen kann einen entscheidenden Wendepunkt bedeuten, wenn es darum geht, Innovationsthemen von unterschiedlichen Seiten zu beleuchten. Nur wenn Teams als primäre Steuerungsgröße im Innovationsgeschehen verankert sind, nehmen sie entscheidenden Einfluss auf die Innovationsleistung des Unternehmens.

Genau das ist die zweite große Erkenntnis aus der Studie: Teams müssen zentraler Bestandteil von Leistungserstellung, Leistungsmessung, Zielsetzung und Zielerreichung sein. Dann können sie erfolgreich mit wechselnden Teamzusammensetzungen und Führungskräften oder auch ganz ohne disziplinarische Führung erfolgreich arbeiten. Und sie brechen dann das klassische System der Führungskaskade mit je einer Führungskraft und den ihr zugeordneten Mitarbeitenden auf.

Einen noch höheren Innovations-Output erhalten Unternehmen, wenn sie neben dem Teamfokus auch die Bedeutung und Bedürfnisse von Individuen berücksichtigen. Die Art der Führung entscheidet darüber, wie zufrieden Mitarbeitende mit ihrer Arbeit sind. Und das ist die dritte Erkenntnis aus der Studie: Nur zufriedene Mitarbeitende sind teamfähig im Sinne einer zielführenden, themenbezogenen Zusammenarbeit. Die Arbeitszufriedenheit bildet eine zentrale Voraussetzung für eine positive Teamkultur und die wiederum für eine hohe Innovationsleistung. Viele Unternehmen gehen davon aus, dass ihre Führungskräfte einen direkten Einfluss auf die Teamleistung und Innovationsfähigkeit ihres Unternehmens haben. Aber das stimmt nicht: Führungskräfte nehmen direkten Einfluss auf Arbeitszufriedenheit und damit die Bereitschaft, sich gegenseitig im Team zu unterstützen und zu fördern. Die höchste Innovationsfähigkeit haben daher auch nicht Unternehmen mit fachlich exzellenten Führungskräften, die ihre Mitarbeitenden eng führen. Es sind vielmehr Unternehmen mit einem klaren und unmissverständlichen Fokus auf Teamarbeit plus individueller Wertschätzung und Unterstützung. Darauf müssen Strukturen und Prozesse abgestimmt werden, zum Beispiel Informationsflüsse im

Unternehmen, bereichs- und hierarchieübergreifende Zusammenarbeit, Schaffung dezentraler Entscheidungsstrukturen, hohe Kooperationsfähigkeit und offene Schnittstellen nach innen und außen. Wie sich zeigt, sind nur rund 25 Prozent der untersuchten Unternehmen aktuell innovationsstark (siehe Abbildung 1). Für die anderen 75 Prozent gilt: Fast alle innovationsrelevanten Betrachtungsdimensionen machen eine radikale Anpassung der Kultur und Struktur erforderlich.

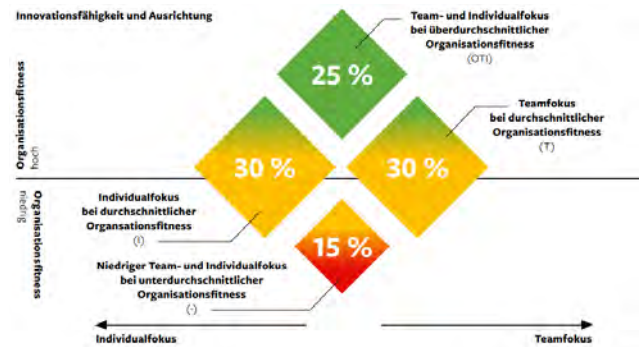


Fig. 1. Innovationsfähigkeit und Ausrichtung in deutschen Unternehmen, n=935

B. Stark abgesunkene Innovationskraft während der Pandemie

Management Analytics hat eine der ersten Studien während der Corona-Pandemie vorgelegt, die konkret auf individueller und Arbeitsteam-Ebene die Auswirkungen des „Task-Loads“ sowie der Führungsarbeit auf die derzeitige Innovationsfähigkeit der Unternehmen erfasst (Hasebrook/Hackl/Rodde, 2021). Die Studie beruht auf dem „NASA Task Load Index“ (TLX), den die NASA in den 1980er-Jahren ein Messinstrument zur Erfassung von Arbeitsbelastung entwickelt und der heute weltweit in vielen Branchen eingesetzt wird. Dieser für Einzelpersonen entwickelte Test wurde in den letzten Jahren um eine Fassung für Teams erweitert. Auf Basis des TLX wurden 176 Personen aus unterschiedlichen Unternehmen anonym dazu befragt, wie sie die mentale, körperliche und zeitliche Belastung vor und während der Corona-Pandemie einschätzen. Zusätzlich haben sie nach Faktoren gefragt, die laut wissenschaftlicher Metaanalysen erfolgreiche Teams auszeichnen (Teamerfolg – Fremdeinschätzung), und wie erfolgreich sich die Teams selbst bewerten (Teamerfolg – Selbsteinschätzung). Am Ende wurden Aspekte unterstützender Führung und soziodemographische Angaben erfragt.

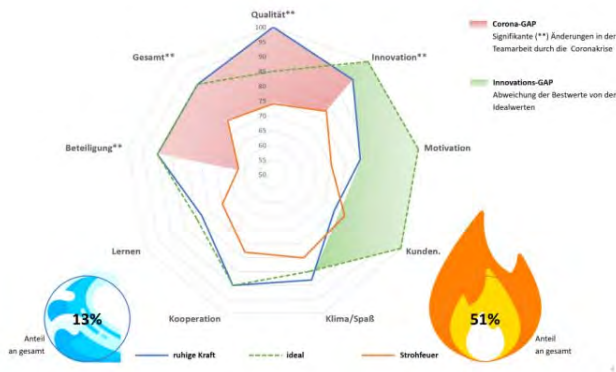


Fig. 2. Bei 51% der Befragten sind individuelle und Teambelastung erhöht („Strohfeuer“), bei nur 13% das individuelle Engagement zwar erhöht, die Teambelastung aber in der Pandemie gesunken („ruhige Kraft“). Doppelte Anstrengung („Strohfeuer“) führt zu signifikant verringerter Leistung in allen Bereichen, effektive Teams („ruhige Kraft“) zu besseren Ergebnissen. Insgesamt sinken aber insbesondere die Zukunftsfaktoren Innovation, Motivation und Kundenorientierung.

III. OTI: DREIKLANG AUS ORGANISATION-TEAM-INNOVATION

Im Kern geben die vorliegenden Studienergebnisse den Hinweis, dass die gleichzeitige Ausrichtung auf Organisation, Team und Individuum (O-T-I) bei starkem Fokus auf T(eam) und (Individuum) den größten Innovations-Output erzeugt. Teams müssen zur zentralen Leistungseinheit im Unternehmen gemacht werden. Die Organisationsstrukturen müssen vor allem der flexiblen Zusammenarbeit und Kooperationsfähigkeit nach innen und außen dienen. Allein auf die Organisationsstrukturen und Zuständigkeiten von Führungskräften zu fokussieren, bringt Verwirrung und Veränderungskosten mit sich, aber keine Stärkung der Innovationskraft. Auch die flexibelste und beste Organisationsstruktur allein zeigt keine Wirkung. Entscheidend ist der Fokus auf Team und Individuum, denn hier entsteht der eigentliche Innovations-Output. Nach unseren Studienergebnissen sind dies zentrale Aspekte für eine hohe Innovationskraft:

Offenheit. Im Schnitt weisen Unternehmen eine um 71 Prozent höhere Innovationsfähigkeit auf, wenn ihre Unternehmenskultur Offenheit für Neues zulässt. Darunter fallen etwa eine positive Fehlerkultur, die Arbeit in Projektlogiken und ein verbindliches Ideen- und Belohnungsmanagement.

Ressourcen. Die Mehrheit der Befragten gab unabhängig von der Unternehmensgröße an, nicht genügend Ressourcen zur Entfaltung ihres kreativen Potenzials zur Verfügung zu haben. Nur mit entsprechenden zeitlichen, personellen, räumlichen und technischen Ressourcen ist es aber möglich, überhaupt Innovationen zu treiben.

Führung. Sie nimmt auf fast alle wesentlichen Performance-Größen Einfluss. Eine vertrauensvolle, ermutigende Führung bildet die Voraussetzungen für eine positive Teamkultur, eine hohe Teamleistung und eine hohe Arbeitszufriedenheit. Wie die Studienergebnisse aber zeigen, ist ein höherer Anteil der Führungskräfte der befragten Unternehmen gegenüber neuen Ideen von hierarchisch niedriger gestellten, älteren sowie weiblichen Mitarbeitenden signifikant weniger offen.

Teamarbeit. Die Mehrheit der Teams ist offen für neue Ideen, wie die Studienergebnisse zeigen. Teams in kleineren Unternehmen unterscheiden sich dabei nur unwesentlich von mittelständischen und großen Unternehmen. Auch die hierarchische Position hat auf die Offenheit von Teams keinen Einfluss. Bei der Teamgröße jedoch zeigen sich Unterschiede: Kleine Teams von bis zu drei Personen zeigen eine fast doppelt so hohe Kreativproduktivität wie Teams ab vier Personen. Alter und Position beeinflussen die Kreativproduktivität unterschiedlicher Teamgrößen dagegen nicht.

Aus Sicht des Personalbereichs geht es also darum, leistungsfähige Mitarbeitende für leistungsfähige Teams zu entwickeln, die optimale Rahmenbedingungen durch Führung und Organisation bekommen. Die zentralen Aspekte sind in Abbildung 3 zusammengestellt.

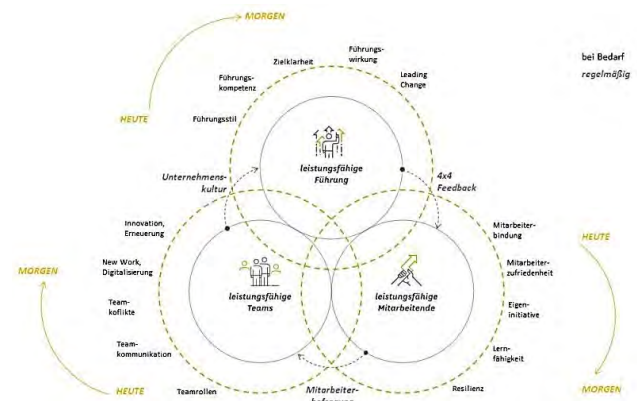


Fig. 3. Handlungsfelder für eine Management Analytics im Personalbereich zur Sicherheit von leistungsfähiger Führung und Organisation, leistungsfähigen Mitarbeitenden und leistungsfähigen Teams (1 siehe Appendix für vergrößerte Darstellung)

IV. ENTWICKLUNG EINES „DIGITAL TWIN VON ORGANISATIONEN“ (DTO)

A. Grundlegender Aufbau des Systems

Zu den in Abbildung 3 genannten thematischen Schwerpunkten haben wir durch unsere Forschungs- und Praxisprojekte eine umfangreiche Datenbasis für die programmatische Entwicklung einer KI-basierten Management Analytics bilden. Zentral bei der Entwicklung der Plattform war der Leitgedanke des „Digital Twins“ – also kein Tool, das die Personalarbeit durch Digitalisierung ersetzt, sondern einen „Partner“ darstellt, der die notwendige Datengrundlage für Lern- und Entwicklungsentscheidungen bereitstellt, und die Zusammenhänge zur Steigerung von Innovationskraft und Leistungsfähigkeit herstellt.

Ein DTO für die Personalarbeit lässt sich nicht entwickeln für ein Digital Twin für eine Maschine (DTM):

1. Maschinenzustände lassen sich mehr oder weniger exakt messen und analog in einem DTM abbilden, ohne dadurch die Maschine selbst zu beeinflussen. Bei Unternehmen und den in ihnen tätigen Menschen geht das nicht.

2. Big Data und statistische Simulationen auf Basis von Personaldaten akkumulieren (vor allem unkorrelierte) Fehlereinflüsse und erzeugen eine Scheingenauigkeit, die wirksame Entscheidungen behindert.
3. DTO können mit keinem klaren Soll- oder Idealzustand verglichen werden, weil Organisationen keinen Idealzustand haben. Ihr Zustand muss vielmehr zwischen Innovation (Kreativität/Exploration) und Performanz (Leistung/Exploitation) oszillieren.

Der DTO für die Personalarbeit besteht aus einer Plattform, die eine stetig erweiterbare Sammlung von den wesentlichen Produktivitäts- und Innovationsanalysen zu den in der Abbildung 3 genannten Themen umfasst. Zu den einzelnen Themen werden aus den Forschungsprojekten und Studien Benchmarks zur Verfügung gestellt, die es den Personalern ermöglichen die aktuelle IST-Situation im Unternehmen auf den verschiedenen Analyseebenen (Gesamthaus, Bereiche, ggf. Standorte) zu bestimmen und daraus fundierte und effektive Handlungsimplicationen abzuleiten. Die Plattform stellt eine Management-Sicht in „Radar-Form“ zur Verfügung, die hilft, wesentliche Zusammenhänge und Wirkungen von Managemententscheidungen abschätzen (vgl. Abb. 4).

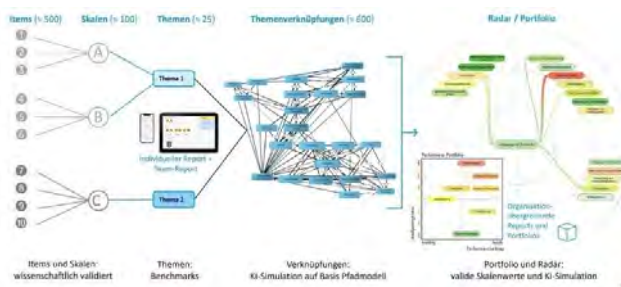


Fig. 4.: Aufbau des DTO für die Personalarbeit (von links nach rechts): Eine Bestand validierter Fragebogen-Items wird in Skalen (Themen) zusammengefasst, zu denen in Studien Benchmarks erhoben wurden. Die Zusammenhänge zwischen den Themen in Bezug auf die Zielgrößen „Innovation“ und „Performanz“ werden in einem KI-Modell simuliert und als Management-Tool in Radarform angezeigt (↑ siehe Appendix für vergrößerte Darstellung).

Die Komponenten des Systems und der Einsatz unterschiedlicher KI-Methoden dabei, wird im Folgenden dargestellt. Eine Übersicht gibt Abbildung 5.

B. Identifikation der Themen und der Items

Im Rahmen des Interview- und Buchprojekts „Team-Mind“ wurden insgesamt 15 führende Expertinnen und Experten auf den Bereichen Sport, Militär/Polizei, IT, Gesundheit, Wirtschaft und Forschung) interviewt sowie einschlägige Forschungsarbeiten erfasst und klassifiziert. In dem Buch „Team-Mind und Teamleistung“ (Hasebrook/Hackl/Rodde, 2020, 2. Auflagen) wurden 656 Literaturstellen verarbeitet, insgesamt 1.212 erfasst. Diese Texte wurden mit Hilfe eines Portals für „Natural Language Processing“

(NLP) auch Themen und Themenzusammenhänge untersucht (Topic Maps auf Basis von Bayes-Matrizen nach einem „Bag of Words“ Verfahren). Die Topic Guides (automatisierte Zusammenfassungen) wurden auf typische Items und Fragestellungen untersucht. Diese wurden mit bereits vorhandenen, praxiserprobten Fragebogen von Management Analytics verglichen. Daraus entstand ein Grundbestand von Items und Themen, die als Skalen aus jeweils 3-15 Items in den Fragebogen abgebildet sind. Auf Grundlage der vorhandenen Studienergebnisse können für jedes Thema (bzw. für jede Skala) empirisch abgesicherte Benchmarkwerte angegeben werden.



Fig. 5. Einsatz von KI-Methoden bei Entwicklung des DTO für die Personalarbeit: Themen und Items wurden auf Basis von Textanalysen wissenschaftlicher Arbeiten identifiziert, Zusammenhänge durch Experteneinschätzungen und eigene Studien ermittelt sowie als Prognosemodell mit verschiedenen Zeithorizonten in einem KI-Simulationsmodell abgebildet (↑ siehe Appendix für vergrößerte Darstellung).

C. Ermittlung der Zusammenhänge

Aus der Analyse der Themen ergaben sich rund 100 Themenstellungen, die alle auf unterschiedliche Weise mit den Zielvariablen „Innovation“ und „Performanz“ sowie miteinander verbunden sind. Aus aktuellen DHBW-Forschungsprojekten wurden insgesamt 1.128 Fragenrückläufe zu zentralen Personalthemen (Führung, Agilität, New Work, Bindung, psychologische Sicherheit, Teamzusammenarbeit und -konflikte sowie Lern- und Veränderungsbereitschaft) in Bezug auf die Erfolgskriterien Performanz (Erfüllung oder Übererfüllung der Geschäftsziele) und Innovation (Anteil von Innovationen an Geschäftsvolumen und -erfolg) verrechnet. Daraus ergab sich ein Pfadmodell, das als Grundlage für eine Einschätzung der Zusammenhänge von insgesamt 25 HR-Expert:innen genutzt wurde. Diese haben auf Basis der empirischen Ergebnisse eingeschätzt, ob und wie stark die einzelnen Themen mit Innovation und Performanz verbunden sind und welche Verbindungen untereinander bestehen. Die auf der Pfadanalyse und durch die Einschätzungen der Expert:innen beruhenden Verknüpfungen wurden in ein KI-gesteuertes Simulationsmodell überführt. Das Teilmodell für den Erfolgsfaktor „Innovation“ ist online zugänglich¹ (vgl. Abb. 6).

¹ Das Modell ist im Knowwhy-Net veröffentlicht unter: https://www.knowwhy.net/kb?key=CWXHU7_8xX9KASI9yASHVhw

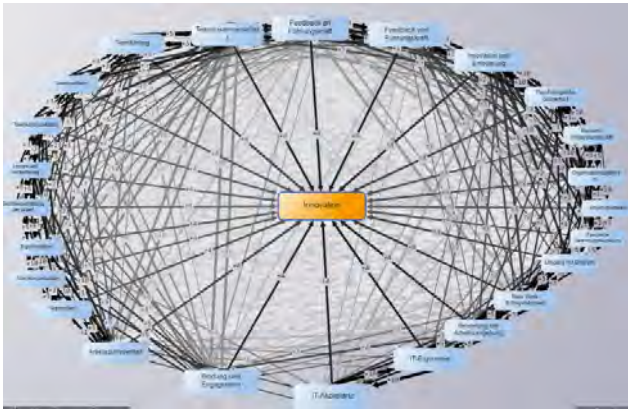


Fig. 6. Ansicht des Teilmodells für den Faktor „Innovation“ (umgesetzt „consideo“, einem Tool für Multi-Attribute Decision Making, MADM, basierend auf verschiedenen Methoden der Distanzberechnung und Markov-Ketten verschiedener Ordnung!) (↑ siehe Appendix für vergrößerte Darstellung).

Die Software implementiert ein „Multi-Attribute Decision Model“ (MADM) auf Basis der Rapid Impact Assessment Matrix (RIAM) Methode. Dabei werden halb-qualitative Gewichtungen an die bestehenden Verbindungen (in Falle von Management Analytics in 5 Stufen von „sehr schwach“ bis „sehr stark“) und die Einflussrichtung (positiv/steigernd oder negativ/senkend) vergeben. Zur Ermittlung der summierten Gewichte werden Markov-Ketten gebildet und über eine vorgegebene Anzahl von Simulationsläufen aufsummiert. Dabei ergeben sich für den DTO von Management Analytics rund 22 Millionen Schleifen, also sich verstärkende und abschwächende Einfluss Schleifen über drei oder mehr Stufen. Die folgende Liste zeigt einen Auszug aus der Liste der entstehenden „Loops“ für Reinforcement (R=Verstärkung) und Balancing (B=Dämpfung) auf die Zielvariable „Innovation“:

- R** (2): Lernen und Veränderung -> Teamkommunikation -> Lernen und Veränderung
- R** (3): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Lernen und Veränderung
- R** (4): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Lernen und Veränderung
- R** (5): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Lernen und Veränderung
- R** (6): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Lernen und Veränderung
- R** (7): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Lernen und Veränderung
- R** (8): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Teamführung -> Lernen und Veränderung
- B** (9): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Teamführung -> Strukturpräferenz -> Lernen und Veränderung
- B** (10): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative ->

- Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Teamführung -> Strukturpräferenz -> Psychische Gefährdungsbeurteilung -> Lernen und Veränderung
- B** (11): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Teamführung -> Strukturpräferenz -> Psychische Gefährdungsbeurteilung -> Arbeitszufriedenheit -> Lernen und Veränderung
- B** (12): Lernen und Veränderung -> Teamkommunikation -> Selbsteinschätzung der Arbeit -> Eigeninitiative -> Teamkompetenzen -> Teamkonflikte -> Teamrollen -> Teamführung -> Strukturpräferenz -> Psychische Gefährdungsbeurteilung -> Arbeitszufriedenheit -> Organisationspräferenz -> Lernen und Veränderung

Im Simulationsmodell können „Insight-Matrizen“ berechnet werden, die für jedes einzelne Thema (z. B. „Teamzusammenarbeit“, siehe Abb. 7), 1. den Einfluss (Impact) dieses Themas auf den Zielfaktor (in diesem Fall „Innovation“) über mehrere Simulationszyklen (X-Achse) und 2. die Veränderung dieses Einflusses während dieser Zyklen (Relevance) angibt. Dabei wird der Impact als Summe der Gewichte berechnet und die Veränderung (Relevance) als Differenz der Impacts von Zyklus 1 bis Zyklus „n“, wobei „n“ den letzte Simulationszyklus bezeichnet. Durch die Anzahl der Simulationsläufe können kurz-, mittel- und langfristige Auswirkungen abgeschätzt werden. Management Analytics verwendet für seinen HR-DTO 15 Simulationszyklen (mittelfristig). Die Gesamtbedeutung des Themas auf den Zielfaktor ergibt sich dann als Produkt aus „Impact x Relevance“. Im System für Endnutzer werden nur jeweils die wichtigsten fünf Themen und deren Verbindung angezeigt. Diese werden durch sogenannte Faltung errechnet, also das Produkt aus Impact und Relevance gewichtet durch Gesamteinfluss eines Themas gemittelt über alle anderen Themen.

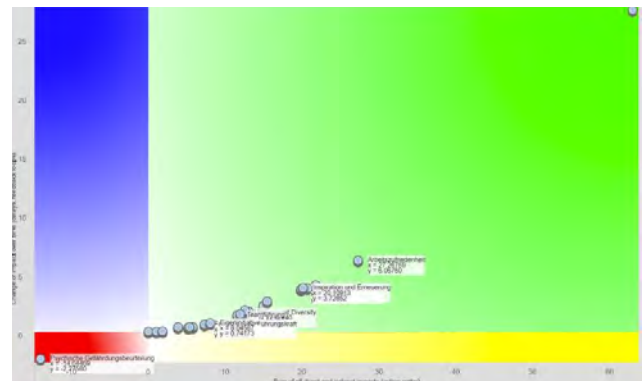


Fig. 7. Insight-Matrix für das Thema „Teamzusammenarbeit“ in Bezug auf den Zielfaktor „Innovation“; angezeigt werden Impact (gewichtete Gesamtsumme aller Simulationsläufe) auf der X-Achse und Veränderung (Relevance als Differenz aus letztem und erstem Simulationslauf) auf der Y-Achse. Wenn Impact und Relevanz negativ sind, liegen die Faktoren im roten Feld (unten links), wenn Impact und Relevance positiv sind im grünen Feld (oben rechts). (↑ siehe Appendix für vergrößerte Darstellung).

V. TECHNISCHE UMSETZUNG UND UMSETZUNGSERFAHRUNG

A. Technisches Framework

Bei der technischen Technischen Umsetzung wurde vor allem auf schnelle Definition und Umsetzung der Analyseeinheiten geachtet. Analysen werden in Sets zusammengehöriger Items in der Art einer Tabellenkalkulation definiert, daher ergibt sich eine einfache Bedienung und schnelle Erfassung bzw. Veränderung von Analysen. Sets werden zu Analysen zusammengestellt, meist zwei bis drei Sets mit je 5-7 Items, und ein inhaltlicher Rahmen im Verwaltungsformular (AnalyticsAdmin) definiert. Das Design wird per Cascading Style Sheet (CSS) sowie Erfassung ergänzender Angaben (Ansprache auf der Startseite, Start- und Endtermin etc.) festgelegt. Start- und Endseiten der Analysen können alle HTML-Elemente inklusive Bildern, Audio und Video sowie Links auf andere Webseiten enthalten. Dadurch können z. B. Videos, in denen das Management des Unternehmens Stellung nimmt, oder Hinweise auf weiterführende Quellen (z. B: Checklisten und Handbücher) eingebunden werden.

Die in Tabellen definierten Analysen können kopiert und verändert werden und unterliegen einer Versions- und Datumskontrolle. Die Analysen werden nicht direkt aus der Datenbank abgespielt, sondern vor der Durchführung kompiliert. Kompilierte Tests sind immer fehlerfrei und auf allen Endgeräten sehr schnell durchführbar ohne merkbare Lade- oder Rechenzeiten. Die durchgeführten Analysen werden 1:1 wie durchgeführt in der Datenbank abgespeichert. Sobald Test neu kompiliert werden, werden diese sofort ausgespielt, d.h. während laufender Befragungen können jederzeit Veränderungen und Verbesserungen vorgenommen werden. Die Betriebsanforderungen wurden sehr geringgehalten und erfordern zum Betrieb einen handelsüblichen Virtueller Server mit 2GB RAM, auf dem NodeJS läuft.

B. Praktische Umsetzung

Der HR-DTO von Management Analytics bietet entsprechend des OTI-Ansatzes auf den drei Ebenen Organisation, Team und Individuum entsprechende Analyse- und Auswertungs-Tools. Auf individueller Ebene steht eine bedarfsorientierte Auswahl von Analysen zur Verfügung, die optional mit einer direkten, individuellen Rückmeldung verknüpft werden kann. Auf der Teamebene steht eine direkte Auswertung der Teamergebnisse als Live-Board zur Verfügung. So dass Ergebnisse direkt in einer Teamsitzung oder einem Workshop erhoben und diskutiert werden könne. Bei Bedarf können Ergebnissen von Individual- oder Team-Analysen per Link untereinander ausgetauscht werden (z. B. Mitarbeitende und Führungskraft oder zwei miteinander kooperierenden Teams). Optional können automatisierte Handlungsimpulse auf Teamebene ausgegeben werden.

Aufbau des Systems zur Ableitung organisationaler „Digital Twins“ für die Personalarbeit



Fig. 8: Systemkomponenten des HR-DTO von Management Analytics auf 1. individueller Ebene (bedarforientierte Auswahl von Analysen optional mit direkter, individueller Rückmeldung), 2. Teamebene (Live-Auswertung der Teamergebnisse zur direkten Besprechung, dem Austausch von Ergebnissen per Link oder optional automatisierten Handlungsimpulsen auf Teamebene), 3. Organisationsebene (Gesamtauswertung oder KI-basiertes Radar-Cockpit über Themenzusammenhänge und Einfluss auf Innovation und Performanz) (↑ siehe Appendix für vergrößerte Darstellung).

Auf Organisationsebene steht ein klassisches Daten-Cockpit zur Gesamtauswertung zur Verfügung und ergänzend ein auf der oben beschriebenen KI-Modellierung beruhendes Radar-Cockpit, das Analyseergebnisse, Themenzusammenhänge und Einfluss auf Innovation und Performanz aufzeigt (vgl. Abb. 9).

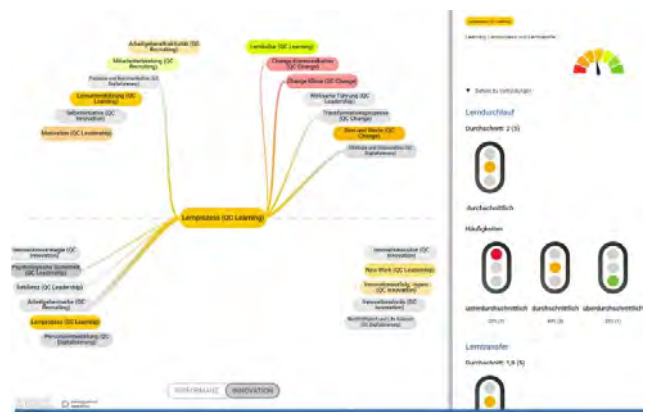


Fig. 9. KI-basierter Radar, der Nutzungshäufigkeit, Ergebnisqualität und Wirkungszusammenhänge für mehr Innovation und Performanz in der Organisation mit internen und externen Benchmarks darstellt (↑ siehe Appendix für vergrößerte Darstellung).

C. Bisherige Einsatzerfahrungen

Die Plattform wird bereits u.a. beim Bildungswerk der Baden-Württemberg Wirtschaft (biwe), im Medizincontrolling bei ID Berlin, beim Laborunternehmen UCL, bei den Gesundheitsversorgern Alexianer und Universitätsklinikum Münster und beim Wirtschaftsprüfer PwC eingesetzt. Für das das Bildungswerk wurde auf Wunsch der Bildungsberater:innen eine weitere, KI-basierte Komponente hinzugefügt.

D. Regelbasiertes Expertensystem für automatisierte Handlungsimpulse

Die Berechnung wichtiger thematischer Zusammenhänge aus den Simulationsergebnissen erlaubt es auch gezielt Handlungsimpulse für individuelle und Teamauswertungen abzuleiten. Dazu werden am Ende eines individuellen Testdurchlaufs die Ergebnisse mit den durch die Befragungen gewonnenen Benchmarks (Mittelwert und Konfidenzintervall) verglichen und als grüne (überdurchschnittlich), gelbe (durchschnittlich) und rote (unterdurchschnittlich) Ampeln ausgegeben. In Abhängigkeit von der Bedeutung der jeweiligen Themen werden Textbausteine als Handlungsimpulse zur besseren Nutzung von Stärken (grüne Ampeln), Herausforderungen (rote Ampeln) und zu beobachtende Themen (gelbe Ampeln) durch ein regelbasiertes Expertensystem ausgegeben (WENN-DANN-Beziehungen und logische Verknüpfungen durch bool'sche Operatoren UND, ODER, NICHT). Die Handlungsimpulse, die direkt nach einer Testdurchführung der durchführenden Person angezeigt werden, beziehen sich auf die individuelle Ebene (siehe Abb. 10). Für die organisierende Person (z. B. Teamleitung) werden die Ergebnisse aller Einzeldurchführungen summiert und Handlungsimpulse auf summarischer Ebene ausgegeben, die das Team oder die Organisation betreffen.

VI. DISKUSSION

Aus unserer Sicht ist der „Digital Twin“ in der Personalarbeit „hybrid“, nämlich eine gemeinsame, faktenbasierte mentale Repräsentation der Wirkungsbeziehungen in einer Organisation – genauso, wie wir Menschen ein bewusste Repräsentation des mentalen Zustands anderer Menschen erzeugen, um uns in Gruppen orientieren und verhalten zu können („Theory of Mind“, siehe Hasebrook/Hackl/Rodde, 2020). Dies bedeutet ganz pragmatisch, dass ein DTO nicht messen sollte, was das Unternehmen im Allgemeinen interessiert (z. B. in einer Mitarbeiterbefragung), sondern das, was die Mitarbeitenden aktuell brauchen (z. B. eine Analyse der aktuellen Teamzusammenarbeit). Daher sollte nicht möglichst umfassend gemessen werden, sondern vielmehr so wenig möglich und nötig, um die nächstbeste Messung abzuleiten (z. B. von der Analyse der Teamzusammenarbeit zur Analyse aktueller Teamkonflikte). Dabei ist nicht die Einzelmessung wichtig, sondern deren Wirkung auf das Unternehmen im Hinblick auf Innovation und Performanz.

Dazu ist eine selbstgesteuerte Auswahl, wissenschaftlich fundierter Analysen mit sofortiger Auswertung und Rückmeldung auf Basis umfangreicher Referenzwerte für Individuen und Teams erforderlich. Umfassende Befragungen sollten abgelöst werden durch fokussierte Analysen, die den Prozessen in der Organisation folgen und in Echtzeit für Selbstreflektion, Teamarbeit, Workshops und Konferenzen einsetzbar sind. Schließlich benötigt die Personalarbeit einen KI-basierter „Radar“ für die Analyseergebnisse, um Nutzungshäufigkeit, Ergebnisqualität und Wirkungszusammenhänge für mehr Innovation und Performanz in der Organisation zu verstehen und mit internen und externen Benchmarks in Beziehung zu setzen, die beispielsweise zeitliche Veränderungen (internes Benchmarking) und Wettbewerbsvergleiche (externes Benchmarking) erlauben. Das System von Management Analytics stellt so einen HR-DTO als

Forschungsprototypen dar, der bereits erste positive Einsatzerfahrungen für sich verbuchen konnte.

QuickCheck Learning & Development



Ihre Auswertung

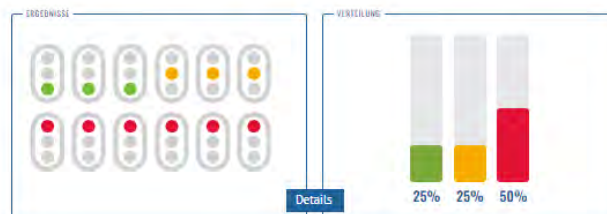


Fig. 10 : Auszug aus einer individuellen Auswertung (links) und einer Gruppen- bzw. Teamauswertung (rechts) mit einem regelbasierten Expertensystem basierend auf den Simulationsergebnissen.

VII. PERSONEN

Benedikt Hackl

Professor für Unternehmensführung und Personal an der DHBW in Ravensburg, Wiss. Leiter des Forschungszentrums Management Analytics. Schwerpunkte sind die Neupositionierung von Teamarbeit, Führungs- und Steuerungssystemen sowie Innovationsproduktivität.

Joachim Hasebrook

Professor für Personal- und Organisationsentwicklung an der Steinbeis Hochschule, Wiss. Leiter des Forschungszentrums Management Analytics. Schwerpunkte sind KI und HR, Teamarbeit und Lernen.

VIII. LITERATUR

Hasebrook, J./Hackl, B./Rodde, S. (2020): Team-Mind und Teamleistung. Teamarbeit zwischen Managementmärchen und Arbeitswirklichkeit. Heidelberg: Springer.

Hackl, B. et. al. (2017): New Work – Managementimplikationen für die neue Arbeitswelt. Wiesbaden: SpringerGabler.

Hackl, B. & Gerpott, F. (2015). HR 2020: Personalmanagement der Zukunft. Strategieumsetzung treiben, Agilität ermöglichen, Individualität schaffen. München: Vahlen.

Benning-Rohnke, E./Hasebrook, J./Pütz, M. (2022): Kunden begeistern. Konzepte und Praxisberichte aus Finance, Automotive und Gesundheit. Wiesbaden: SpringerGabler.

Hasebrook, J./Fürst, M./Kirmße, S. (2019): Wie Organisationen erfolgreich agil werden. Heidelberg: Springer essentials.

Hasebrook, J./Zinn, B./Schletz, A. (2018): Lebensphasen und Kompetenzmanagement. Ein Berufsleben lang Kompetenzen erhalten und entwickeln. Heidelberg: Springer.

Hellmann, W./Beushausen, T./Hasebrook, J. (2015): Krankenhäuser zukunftssicher managen. Stuttgart: Kohlhammer.

Barthel, E./Hanft, A./Hasebrook, J. (2011): Innovationsstrategien als Aufgabe der Organisations- und Personalentwicklung. Münster: Waxmann.

Dohrn, S./Hasebrook, J./Schmette, M. (2011): Vielfalt und Innovation. Strategisches Diversity Management für Innovationserfolg. Düren/Mastricht: Shaker.

Hasebrook, J. et al. (2022). Green Behavior: Factors influencing behavioral intention and actual environmental behavior of employees. Sustainability, 14(17): 10814.

Hasebrook, J./Wolfslast, M./Lister, M. (2022). Ein Navigationssystem für Corporate Social Responsibility. VersicherungsPraxis, 09/2022.

Hackl, B. et. al. (2021): Die Kunst Teams zu denken, Personalführung, 6/2021.

Hasebrook, J./Hackl, B./Rodde, S. (2021): Mehr Anstrengung, weniger Erfolg. Führen in Pandemienzeiten. Steinbeis Transfer „Best of 2021“.

Hasebrook, J. (2021): Die Teamresilienz-Matrix: Gemeinsam cool bleiben. managerSeminare, 4/2021.

Hasebrook, J./Rodde, S. (2021): Belastende Beweglichkeit – die dunkle Seite der Agilität. managerSeminare, 12/2021.

Hasebrook, J./Hackl, B. (2020): Starke Führung, starke Teams, Personalmagazin, 1/2020.

Hackl, B./Baumann, D. (2019): Wenn Mitarbeiter zu Mitarbeitenden werden, Personalführung, 12/2019.

Handelsblatt, 16.08.2019, Die New-Work-Illusion, 6 Irrtümer der neuen Arbeitswelt, mit Beiträgen von Benedikt Hackl.

Hackl, B.; Baumann, D (2018): Schöne neue Arbeitswelt. Unternehmermagazin. 5/6 2018.

Nirmala, R.; Hackl, B.; Hasebrook, J.; Servatius, F. (2018): A twinkle-toed elephant: How New Work in India spreads. Personalführung.

Lahm, P; Hackl, B. (2018): Gemeinsames Interview Philipp Lahm und Prof. Hackl.

Hackl, B.; Hasebrook, J.; Baumann, D. (2018): Quick Check New Work, Magazin Personalführung, 4/2018. Link zum Quick Check.

Hackl, B. (2018): Führung ja, aber weniger Hierarchie, Personalführung, 5, 2018.

Hackl, B. (2017): Innovationsfähigkeit, Wirtschaft + Weiterbildung, 4 (17).

Hackl, B. / Baumann, D. (2017): Die Beteiligungsfrage, Human Resource Manager, 4 (17).

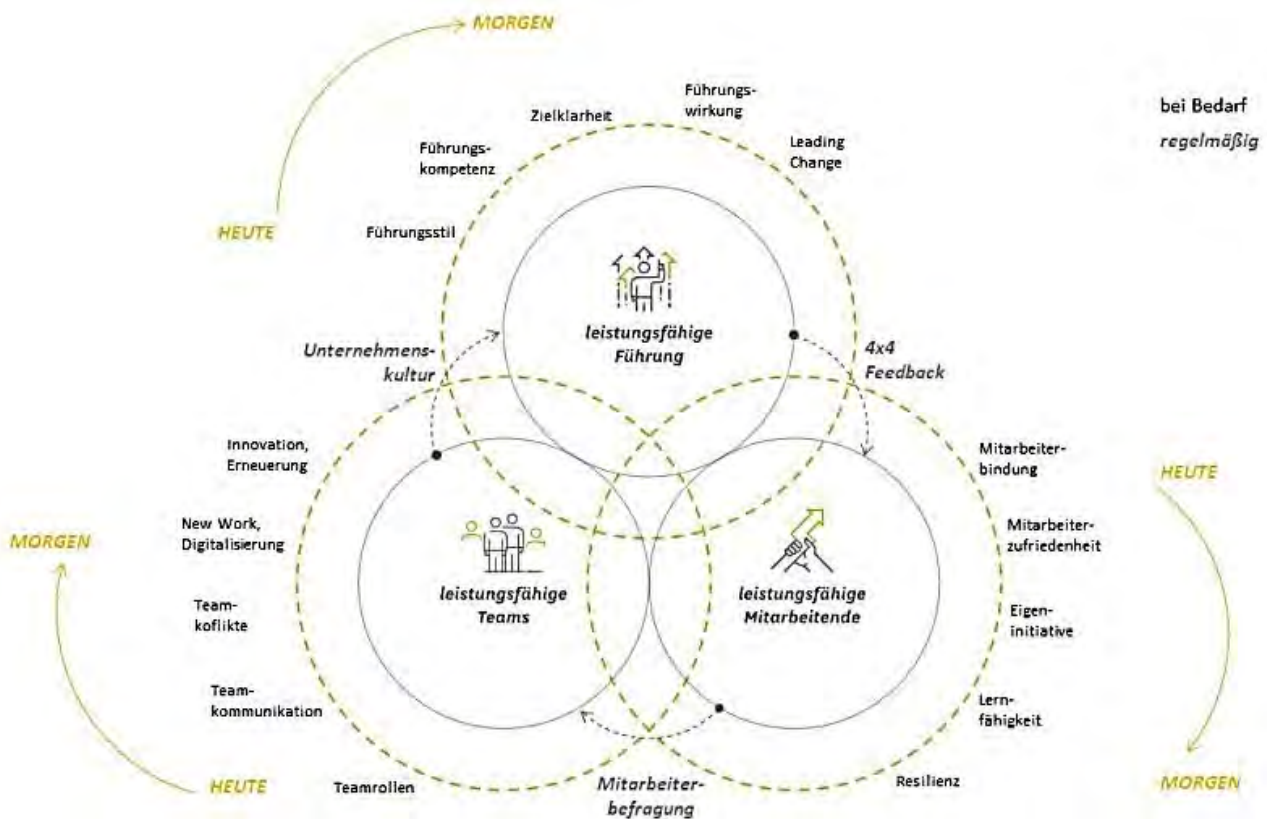


Fig. 3. Handlungsfelder für eine Management Analytics im Personalbereich zur Sicherheit von leistungsfähiger Führung und Organisation, leistungsfähigen Mitarbeitenden und leistungsfähigen Teams

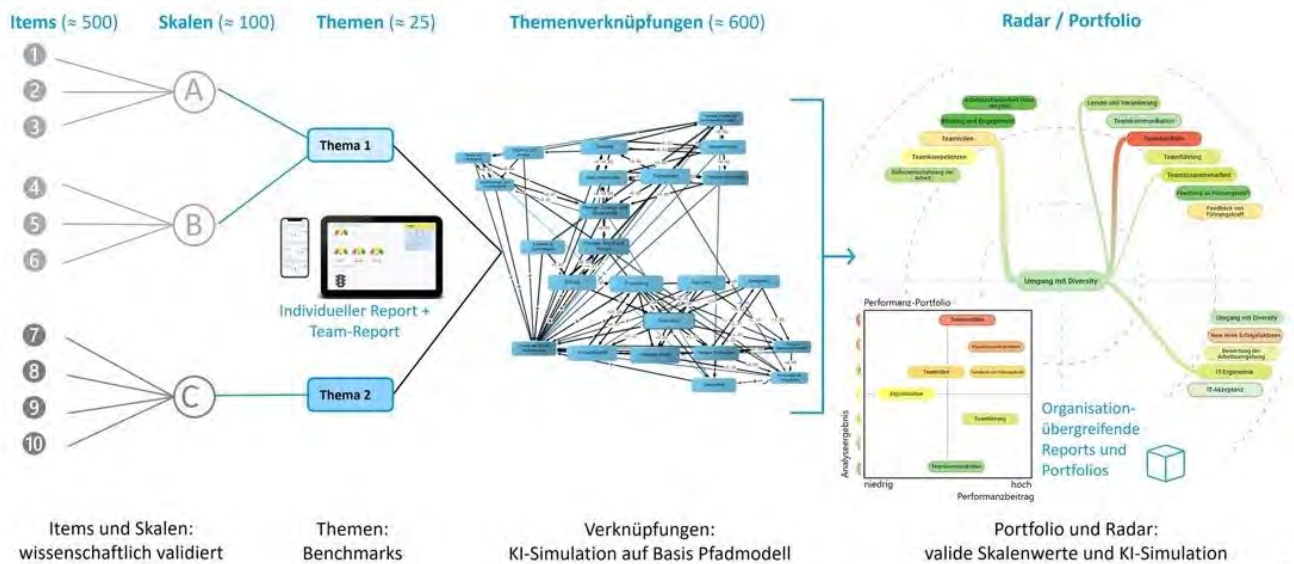


Fig. 4.: Aufbau des DTO für die Personalarbeit (von links nach rechts): Eine Bestand validierter Fragebogen-Items wird in Skalen (Themen) zusammengefasst, zu denen in Studien Benchmarks erhoben wurden. Die Zusammenhänge zwischen den Themen in Bezug auf die Zielgrößen „Innovation“ und „Performanz“ werden in einem KI-Modell simuliert und als Management-Tool in Radarform angezeigt

- Was wurde wissenschaftlich über den Einfluss von Individual-, Team- und Organisationsfaktoren auf Innovation und Performanz veröffentlicht?



- Automatisierte Suche: 1.212 klassifizierte Artikel
- Semantische Textanalyse: Topic Maps

- Welche förderlichen und hemmenden Einflüsse sehen Expert:innen?



- Expertenbewertung der Topics: Einflussmatrix

- Was bedeuten diese Einflüsse für die zukünftige Entwicklungen von Innovation und Performanz?



- Neuronale Simulation der Einflüsse: Prognosemodell



Fig. 5. Einsatz von KI-Methoden bei Entwicklung des DTO für die Personalarbeit: Themen und Items wurden auf Basis von Textanalysen wissenschaftlicher Arbeiten identifiziert, Zusammenhänge durch Experteneinschätzungen und eigene Studien ermittelt sowie als Prognosemodell mit verschiedenen Zeithorizonten in einem KI-Simulationsmodell abgebildet († siehe Appendix für vergrößerte Darstellung).

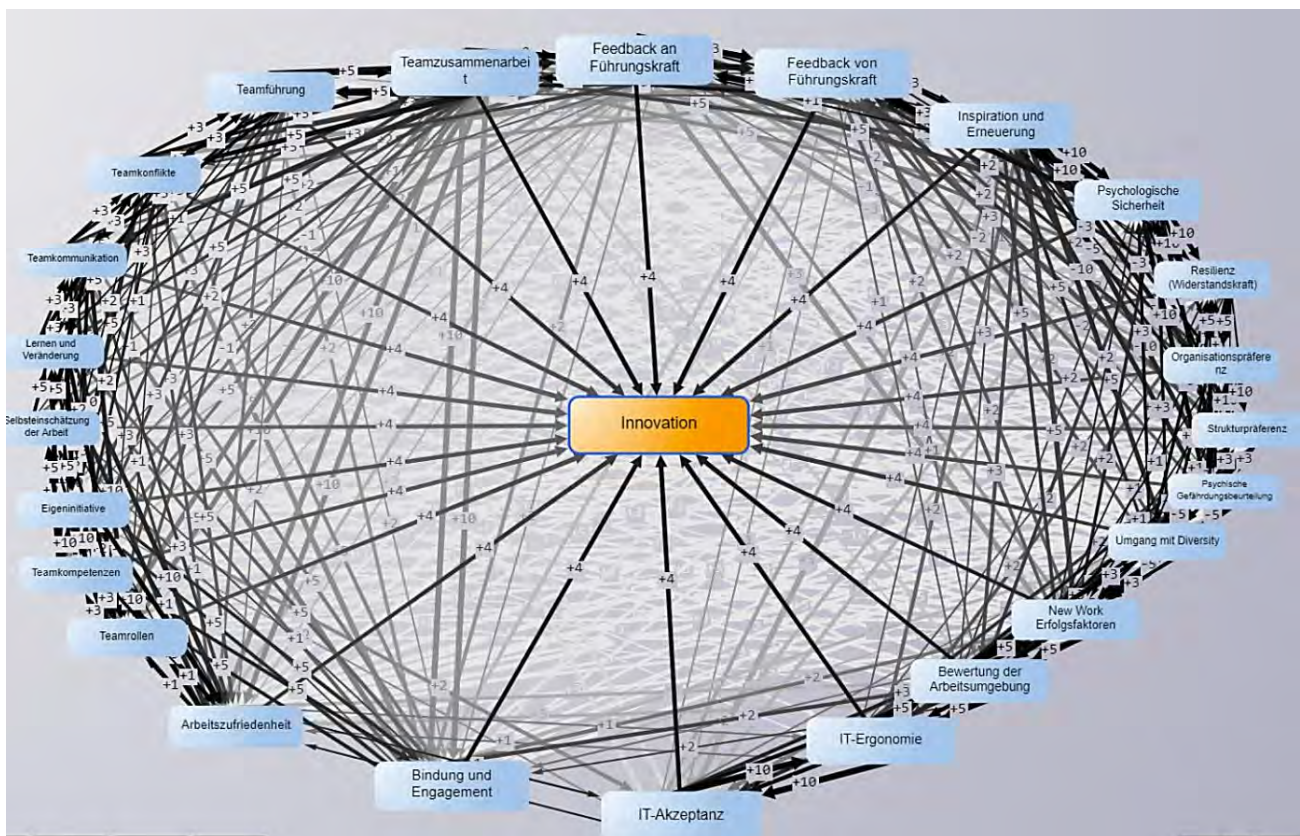


Fig. 6. Ansicht des Teilmodells für den Faktor „Innovation“ (umgesetzt „consideo“, einem Tool für Multi-Attribute Decision Making, MADM, basierend auf verschiedenen Methoden der Distanzberechnung und Markov-Ketten verschiedener Ordnung¹

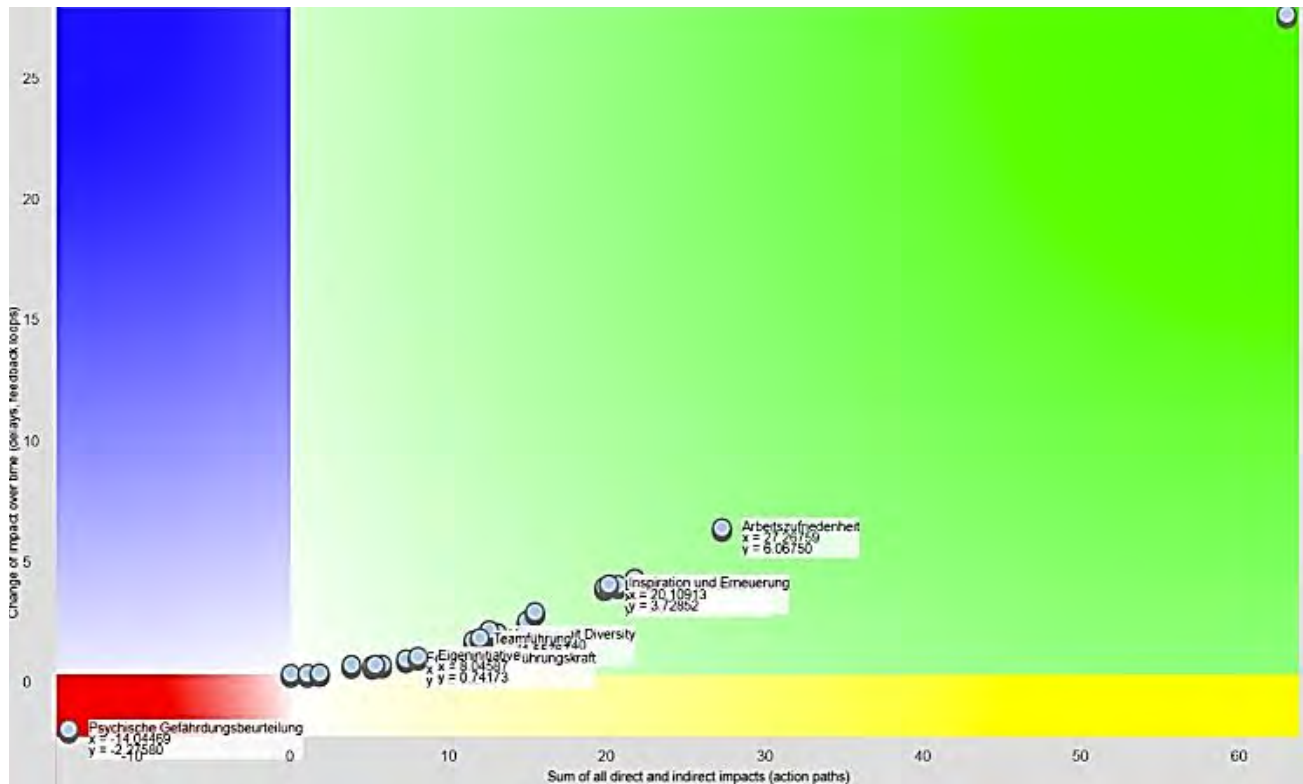


Fig. 7. Insight-Matrix für das Thema „Teamzusammenarbeit“ in Bezug auf den Zielfaktor „Innovation“; angezeigt werden Impact (gewichtete Gesamtsumme aller Simulationsläufe) auf der X-Achse und Veränderung (Relevance als Differenz aus letztem und erstem Simulationslauf) auf der Y-Achse. Wenn Impact und Relevanz negativ sind, liegen die Faktoren im roten Feld (unten links), wenn Impact und Relevanz positiv sind im grünen Feld (oben rechts).

Aufbau des Systems zur Ableitung organisationaler „Digital Twins“ für die Personalarbeit

HR Analytics Systemkomponenten



Fig. 8: Systemkomponenten des HR-DTO von Management Analytics auf 1. individueller Ebene (bedarfsorientierte Auswahl von Analysen optional mit direkter, individueller Rückmeldung), 2. Teamebene (Live-Auswertung der Teamergebnisse zur direkten Besprechung, dem Austausch von Ergebnissen per Link oder optional automatisierten Handlungsimpulsen auf Teamebene), 3. Organisationsebene (Gesamtauswertung oder KI-basiertes Radar-Cockpit über Themenzusammenhänge und Einfluss auf Innovation und Performanz).



Fig. 9. KI-basierter Radar, der Nutzungshäufigkeit, Ergebnisqualität und Wirkungszusammenhänge für mehr Innovation und Performanz in der Organisation mit internen und externen Benchmarks darstellt.

AI Glucose Prediction For An Insulin Recommendation System Based On Smartwatch Activity And Glucose Monitoring Data For Type 1 Diabetics

Stefanie Neumann

Baden-Wuerttemberg Cooperative
State University (DHBW) Karlsruhe
stefanie.neumann.ka@outlook.de

Armin Zundel

Inline Internet Online Dienste GmbH
Karlsruhe
zundel@inline.de

Kay Berkling

Baden-Wuerttemberg Cooperative
State University (DHBW)
kay.berkling@mosbach.dbhw.de

Abstract—Diabetes is a chronic disease that can have a major impact on a person's daily life. Insulin therapy for type 1 diabetics requires insulin doses to be calculated several times a day. This calculation is time-consuming and often not accurate enough.

Using machine learning models, this work developed a solution that enables a more accurate calculation of the insulin dose. The solution was designed as a decision system that provides insulin recommendations. A more accurate determination of the insulin dose was made possible by capturing multiple relevant influencing factors. By automating the data collection and insulin determination process, the burden on the diabetic patient is hereby reduced.

This research shows that automated data collection can improve insulin therapy for people with type 1 diabetes. The use of artificial intelligence to estimate postprandial glucose levels and determine insulin dose accordingly is promising. The best solution achieved a prediction accuracy of glucose values with a mean deviation of $20.22 \frac{mg}{dL}$. This work reports results on an insulin prediction system. A deviation of less than one unit is expected for insulin prediction. In another study, an accuracy with a mean deviation of 2.4 insulin units was achieved [22]. The results reported here compare favorably with others reported in the literature.

Index Terms—type 1 diabetes, glucose prediction, insulin recommendation, gradient boosting

I. THEORETICAL BACKGROUND

A. Type 1 Diabetes

Diabetes is a multimorbid metabolic disease. The disease is characterized by elevated blood glucose levels [15]. One manifestation of the disease is chronic type 1 diabetes and is characterized by insulin deficiency. Insulin is an endogenous messenger produced in the so-called beta cells of the pancreas [3]. This vital hormone regulates the uptake of glucose into the body's cells and is thus essential for basic metabolic functions. The cause of insulin deficiency in type 1 diabetes is damage to the insulin-producing cells. This is an autoimmune reaction of the body. The result of insulin deficiency is excessive blood glucose levels and can only be treated by external insulin administration [15].

An elevated blood glucose level caused by insulin deficiency and permanently left untreated can result in cardiovascular diseases, kidney, eye and nerve damage or the so-called diabetic

foot syndrome [6]. Effective treatment is therefore essential to prevent permanent secondary damage from diabetes.

B. Conventional Diabetes Therapy

Traditional therapy includes glucose monitoring and insulin administration. Over time, technological advances have changed both areas.

On the one hand, glucose monitoring is necessary to provide a data base for therapy decisions and to verify the effectiveness of insulin therapy. On the other hand, continuous glucose monitoring enables early detection of severe dangerous fluctuations and an appropriate immediate response. Modern measurement systems allow Continuous Glucose Measurement (CGM) using sensors that are permanently attached to the body [16]. Such a sensor measures the glucose level in the tissue and transmits the measured values via Bluetooth to an external reading device [16]. This makes it possible to continuously monitor the blood glucose level.

The only way to control high blood glucose levels in type 1 diabetes is to administer insulin. Insulin is usually injected with each meal and once a day to cover long-term insulin needs. The body needs insulin to process the carbohydrates in food. To prevent a sharp rise in blood sugar immediately after meals, an appropriate amount of insulin is injected. In healthy people, the metabolism is automatically regulated so that increases or decreases in blood glucose are balanced by the release of insulin or other hormones. Because the synthetic insulin injected into diabetics works much more slowly in the body, it must be administered in addition to food intake [26] [4]. The amount of insulin needed must be determined in several steps before a meal is eaten. To do this, it is necessary to know the influencing factors, determine them, and relate them to each other. The factors that influence blood glucose levels are now known, so insulin needs can be calculated as a function of the meal. Diabetics usually inject their mealtime insulin before eating.

In conventional insulin therapy, the carbohydrates contained in the meal are determined and the current glucose value is measured. Taking into account the time of day and past glucose

records, the necessary insulin dose is calculated individually and manually for each meal [13].

The calculated insulin dose is administered by the patient. If necessary, the injection can be adjusted to the external circumstances. For example, if a meal is followed by a major sporting effort, a lower insulin dose can be selected. In this case, the insulin dose is only adjusted by estimation.

The body's insulin requirement can also fluctuate during the course of the day independently of meals. Meal-independent insulin requirements are met by injecting long-acting insulin (basal insulin). While meal insulin (bolus insulin) acts relatively quickly in the body, long-acting insulin acts over the entire day. In particular, it stabilizes blood glucose levels at night.

The aim of diabetes therapy is to prevent long-term consequences caused by a permanently high blood sugar level and at the same time to prevent hypoglycaemia - a dangerously low blood glucose level - which can occur as a result of an insulin injection that is too high. The therapy goals of a diabetic are determined individually in consultation with the attending physician. A low long-term blood glucose level (A1C) may be associated with a lower risk of developing diabetes-related microangiopathic and neuropathic sequelae [1]. At the same time, low long-term blood glucose increases the risk of severe hypoglycaemia and weight gain [1]. In principle, it is recommended to aim for the lowest possible A1C value, provided this does not lead to severe hypoglycaemia. The American Diabetes Association Professional Practice Committee recognises the possibility of more stringent glycemic targets of a $A1C < 7\%$, especially for younger patients with newly discovered diabetes and high motivation to achieve treatment goals [8].

Treatment methods of diabetes differ on the basis of the chosen means of injection and the form of administration of basal insulin. The insulin therapy used in this work is Intensified Conventional Insulin Therapy (ICT), in which meal insulin is administered at least three times daily and long-acting insulin is administered once daily using insulin pens. For glucose monitoring, a sensor is used that continuously measures the blood glucose level and transmits it to a smartphone via Bluetooth.

C. Challenges of conventional therapy

Too high a glucose levels can be counteracted in the long term by suitable therapy. The factors that influence blood glucose are sufficiently well known, making it theoretically possible to calculate the individual daily insulin requirement.

In practice, however, a number of hurdles stand in the way of successful insulin therapy. Influencing factors in everyday life are only partly recorded and not taken into account when calculating the insulin requirement. Changing daily habits make constant treatment difficult, and the assessment of the effect of the insulin administered is subjectively influenced by the time delay between measurement and physical reaction.

Insulin therapy is challenging in everyday life because frequent insulin injections and constant glucose monitoring interfere with normal routines that are time-consuming and

error-prone. The calculation of the insulin amount includes the recording or estimation of influencing factors, the correction determination and the total value calculation. The determination is particularly prone to error in estimating the amount of carbohydrate or the manual calculation of the amount of insulin. In addition, the calculation can never provide fully justified values, since it does not capture all influencing factors. None of the treatment methods used in current diabetes therapy measures physical activity, stress or sleep duration as influencing factors. Additional parameters can only be taken into account as estimating factors.

The problems of conventional insulin therapy are clearly demonstrated by the high daily workload involved in recording the influencing factors and the susceptibility to errors when determining the necessary insulin doses.

II. RELATED WORK

The potential applications of machine learning in the context of diabetes are manifold. For example, appropriate technologies are used in the prevention, diagnosis, and treatment of the disease and associated sequelae [18] [3]. This work focuses only on the options for diabetes treatment.

A selection of publications that used different approaches and investigational constellations to implement blood glucose and insulin prediction systems, respectively, are presented below.

Gupta and Jiwani [14] presented an approach to predict insulin levels at different times of the day. The basis of the prediction was the health data of a diabetic patient. The only training factors considered were the regular insulin doses administered as a function of the time of day and date. For training the prediction model, the authors used Long Short-Term Memory (LSTM) and Artificial Neural Network (ANN). The result was individual predictions of insulin doses for a given time of the day over several days.

For their part, Obeidat and Ammar [22] took an approach to implementing an insulin prediction model. They compared four different machine learning models in an experimental setup over a total of five months. The final implementation used an ANN and achieved the best prediction with a mean squared error (MSE) of 5.79. This model was trained with the characteristic, as well as continuous, data of 13 type 1 diabetes patients. Factors considered were static patient data such as gender, age, height, weight, and time since onset of the illness. In addition, CGM data were collected and included in the recommendations. The characterizing data of patients were used as input of a classification model. According to the particular category, an insulin prediction model was used. Glucose measurements were incorporated into this. The result of the processing was insulin recommendations for six specified time periods. The implementation was distinguished by the fact that the final model was implemented with the help of a Raspberry Pi and was thus portable and suitable for practical tests.

Another approach to an insulin recommendation model was described by authors Alqudah, Younes, Alqudah [2].

Their model considered data from 149 patients, only 20 of whom have type 1 diabetes and 109 of whom have type 2 diabetes. They used both static patient data and continuous glucose measurements. Patient data collected included gender, age, Body Mass Index (BMI), medical history, total daily insulin dose, diabetes type, *Smoking Factor*, *Genetic Factor*, creatinine clearance, and long-term glucose (HbA1c). The model used was an ANN. The purpose of the study was to determine the relationship between blood glucose and insulin dose taking into account physical factors.

Martinsson et al. [19] succeeded in predicting glucose levels by using a Recurrent Neural Network (RNN). They used a data set of 6 type 1 diabetes patients. Parameters included in the prediction were only the continuous glucose measurements of the last 60 minutes for a given patient. The result of the processing were predictions of glucose values up to 30 minutes and up to 60 minutes into the future, respectively.

Zarkogianni et al. [28] conducted a comparative evaluation of four personalised blood glucose prediction models for patients with type 1 diabetes that used data from sensors monitoring blood glucose concentration and physical activity. The aim was to investigate the impact of including physical activity data on the predictive performance of the models. Data from 10 patients were monitored over a 6-day period. The results of the study showed that SOM leads to better predictive performance than the other observed models FNN, WFNN and LRM. The authors highlight the advantage of the small input dimension, which keeps the time required to train the models relatively short.

Zhu et al. [29] used the same dataset consisting of 6 type 1 diabetic patients, but in addition to continuous glucose measurements, they took into account reported insulin injections and carbohydrates. The authors solved the prediction task by transforming the problem into a classification problem.

Classifying postprandial glucose levels in type 1 diabetes patients using XGBoost was what Cappon et al. [5] aimed to do. Their predictions extended two to six hours into the future after a meal had been eaten. The data were based on 100 virtual adult subjects for whom continuous glucose measurements, insulin doses, and carbohydrate intakes were available. The authors proposed their classification model as a basis for insulin dose adjustment and presented a suitable calculation method. They also mentioned the possibility of appropriate early warnings or recommendations for carbohydrate intake.

Only the authors [29] and [5] took into account the carbohydrates in the meals consumed and the insulin doses administered in addition to continuous glucose measurements. [19], [2] and [22] used glucose measurements as variable inputs, while [14] only considered insulin amounts. [28] considered sports activities and continuous glucose measurements in their predictions but, in contrast, did not consider insulin amounts or other behavioral parameters. Of all the papers reviewed, the one by Obeidat and Ammar [22] was considered the most promising, as it implemented a fully comprehensive insulin recommendation system and achieved encouraging results in practical testing, with an MSE of 5.79. This paper is used for

comparison to assess the quality of the results achieved by this work.

III. RESEARCH QUESTION

The aim of the work is to solve the challenges of conventional insulin therapy by an automated insulin recommendation system. In particular, the complex and inaccurate calculation of the amount of insulin, as well as the daily burden of constantly determining therapy factors and making therapy decisions, are being counteracted by the insulin recommendation system. This solution was implemented using machine learning applications. The provision is intended to facilitate therapy realization in everyday life. Compared to presented literature on the improvement of insulin therapy with the help of artificial intelligence, this work captured multiple relevant influencing factors on the course of glucose levels in everyday life and took them into account in insulin determination. In particular, this feasibility study was used to investigate whether the recording and consideration of these additional factors is practicable in everyday life and whether an improvement in insulin prognosis can be achieved as a result. In addition, a solution suitable for everyday use meets the requirements of rapid and automated insulin recommendation and convenient recording of therapy factors.

IV. METHODOLOGY

To answer the research question, an insulin recommendation system was prototypically implemented, which already integrates a method for assessing the effectiveness of an insulin dose. The data for training the prognosis model were collected over a period of three months from a diabetic patient and processed independently. The therapy factors recorded include regular measurements of

- glucose levels,
- carbohydrates consumed,
- insulin administered,
- sleep duration,
- heart rate,
- the duration of physical activity and
- step count.

Glucose levels were measured using a FreeStyle Libre 3 sensor [10], carbohydrate and insulin levels were recorded using a self-developed web app (web application), and activity data were measured by a Samsung Galaxy Watch 5 [27]. The subject of the data collection was a female type-1 diabetic patient aged 20 years with a BMI of 20.3.

A. Data Collection

For data-driven and data-intensive applications such as machine learning models, data collection is particularly important and costly. The distinct challenge of data collection in this work is that factors in everyday life were collected using different technological means. Different data sources, heterogeneous data formats, and variable collection times had to be combined into homogeneous data sets. Three data sources were used to collect all the data relevant to the forecast.

The **Freestyle Libre 3 Sensor** [10] measured the glucose level in the tissue and transmitted it via Bluetooth to the smartphone connected to the sensor (see figure 1).



Fig. 1: Freestyle Libre 3 Sensor and App (taken from [12] and [11])

The self-hosted and self-developed **web app Insulin Calculator** was used to collect the injection data.

This includes the amount of insulin injected, the amount of carbohydrate ingested, and the amount of long-acting insulin to be administered once daily. To determine the required amount of insulin, the current glucose value is needed. This was therefore transmitted to this web app via real-time communication. In terms of ICT therapy, this then automatically calculated a correction value from the glucose value and the entered carbohydrate quantity and a time-dependent factor, and finally the insulin quantity to be administered. Calculating the necessary insulin doses according to the ICT therapy via an algorithm ensured that no manual calculation errors were made and that the insulin dosage data were comparable for machine learning applications. Figure 2 shows a screen shot of the Insulin Calculator web app used for data collection.

Fig. 2: Insulin calculator app web interface (taken from [21])

In order to take other possible factors into account when training the machine learning model, health data was collected using a **Smartwatch**. The heart rate, exercise duration, sleep duration and step count were recorded.

Data collection took place over a three-month period. The smartwatch and FreeStyle Libre sensor were worn continuously during this time, and each insulin injection was calculated and stored using the insulin calculator app. The data from all data sources were kept in a central database, with data from each data source stored in separate database tables.

Once the data collection was complete, the data in the three database tables were merged via time stamps and exported in CSV format. Python was then used to prepare the data.

B. Data preparation

After data collection, the data were available in three database tables divided into glucose, injection and activity data. Each of these tables was first exported in CSV format and loaded into a Jupyter Notebook [25] for further data processing. The Jupyter Notebook resided in a Docker container [7]. The goal of the data preparation was to process the data so that it could be used to train a glucose prediction model. Since the data came from different sources and had different levels of granularity, the main task of data preparation was to combine these data in an appropriate way. For each injection, it was necessary to take into account the behaviour of the patient for a certain period of time before the injection, as all these contextual parameters influence the blood glucose curve. In order to represent these individual data points in a coherent way, features were created by statistical aggregation. This type of linear feature extraction leads to a loss of data [9]. However, the interpretability of the results is preserved [9]. Another disadvantage of statistical aggregation is the sensitivity to outliers. These can influence the results negatively. To minimise the influence of possible outliers on the prediction result, they were removed in a first step. The difficulty included measurement errors, such as the incorrect estimation of the amount of carbohydrate in the meal, are difficult to reconstruct in retrospect. The estimation of the amount of carbohydrates is particularly susceptible to measurement errors, as this is only an estimate. Particularly high or low glucose values after a meal that deviate from the other values suggest that the amount of carbohydrate was estimated incorrectly, but it cannot be determined with certainty whether another influencing factor did not lead to this glucose value. Deleting these outlier values would therefore have been tantamount to manipulating the database. In addition to incorrect estimations, the entry of measured values in the web app also lead to corrections, additions and accidental multiple storage by pressing the save button several times. However, these could be traced and corrected afterward. To clean up multiple entries, insulin or meal entries were deleted that followed a previous entry for less than one minute and matched the previous entry in all other parameters. The further data preparation was performed in five steps (see Figure 3).

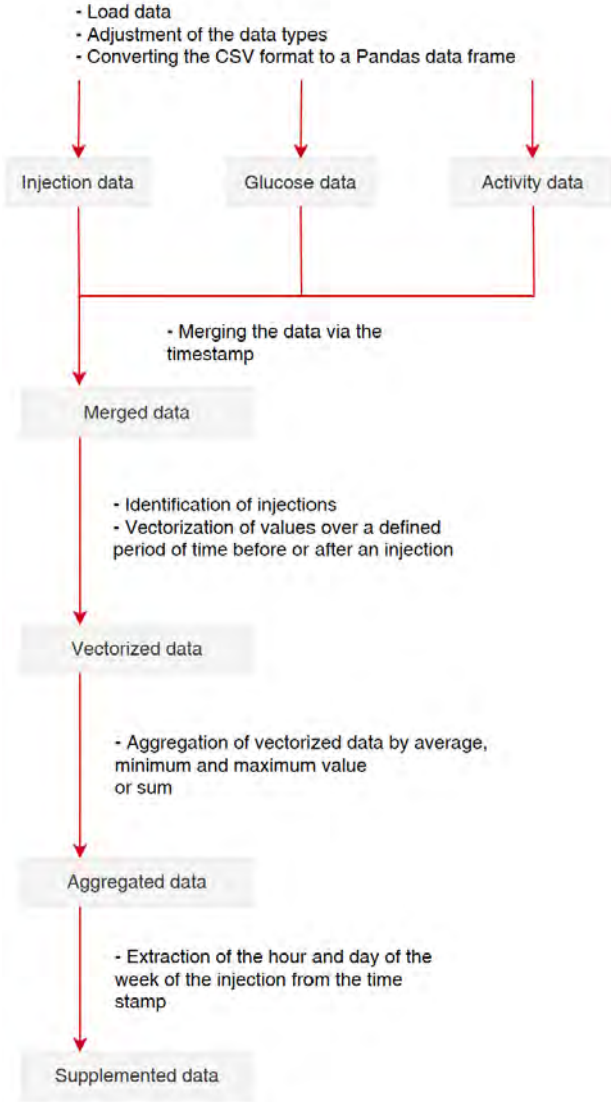


Fig. 3: Scheme of data preparation

First, the CSV files were loaded into the Jupyter Notebook and converted into Pandas DataFrames [23]. The result of this first step was three Pandas DataFrames that hold the injection data, glucose data, and activity data separately.

In the second step, the DataFrames were merged using the timestamp field. This created a new DataFrame that contained all the data. Since the timestamp reflected the time of measurement to the second and the frequency of data collection varied between every two minutes for glucose values and every four hours for injection data, zero values were created when the data was merged. These were initially left blank (see Table I).

TABLE I: Exemplary presentation: Extract of the merged data

timestamp	glucose	Heartrate	SleepDuration	...
2022-11-22 22:00:00	NaN	NaN	0.0	...
2022-11-23 22:10:29	NaN	72.0	NaN	...

Next, all injections in the merged data were identified. In a new DataFrame, each injection was represented by a record. All values of a given feature collected over a defined period of time before the injection time were merged into one vector. The same was done with the glucose data after the injection (see figure 4).

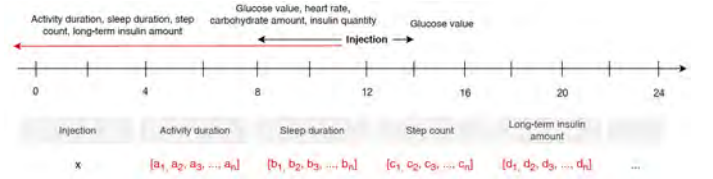


Fig. 4: Visualization of the procedure for vectorization of the data

Each record of the vectorized data now contained

- an entry indicating the time of injection,
- an entry indicating the amount of insulin,
- vectors indicating all pre-injection glucose, injection, and activity data, respectively, and
- a vector indicating the glucose data four hours after injection

(see table II).

TABLE II: Exemplary presentation: Extract of the vectorized data

timestamp	currentInsulin	glucose	Heartrate	...
2022-12-04 13:06:29	[4.5]	[169.0, 167.0, 160.0, 153.0, 146.0, 141.0, 139...]	[79.0, 70.0, 69.0, 69.0, 71.0, 73.0, 74.0, 72...]	...

The time periods for aggregating the data before an injection were set differently. For the measured values sleep duration, step count, long-term insulin amount and duration of physical activity, a period of 24 hours was set for the aggregation of the data before the injection. Blood glucose levels, heart rate, amount of carbohydrates previously consumed, and insulin amounts previously administered were summarized over a four-hour period.

The vectorized data was then aggregated. There were some values for which summation was useful. These included, for example, the number of steps, the amount of carbohydrates and the duration of sleep. These values were therefore added and aggregated as a total. For heart rate or glucose values, a sum would not have been meaningful. Instead, the mean value, minimum value and maximum value were calculated for aggregation (see table III).

TABLE III: Exemplary presentation: Extract of the aggregated data

timestamp	glucoseAverage	glucoseMin	glucoseMax	...
2022-12-04 13:06:29	149.859649	72.0	209.0	...

The goal of the training model implemented with the prepared data was the prediction of glucose values based on different influencing factors. Since only one output value could be predicted by the machine learning model, the glucose values after an injection to be predicted by the model had to be summarised in an interpretable way. In addition, the model was supposed to be able to assess whether an insulin dose had resulted in the expected glucose levels. For this purpose, glucose values after an injection were summarised into a glucose score, which indicated the mean deviation from a desired target value.

As the insulin amounts recorded during data collection did not always result in ideal glucose values, it was not possible to use these as outcome values. Instead of directly predicting insulin levels, it was decided to train the model to learn the interaction of the influencing factors and the glucose levels. In this way, all measured values could be used as training values, and not only those representing effective insulin injections. The prediction of insulin levels is achieved by adjusting the insulin input to the trained model until an ideal post-injection glucose score is predicted.

The average of all post-injection glucose data alone was not considered meaningful enough to judge whether an insulin dose was appropriate. Particularly after a meal, large fluctuations in blood glucose levels were often observed. Instead, different methods of calculating this score were compared to assess the effectiveness of an insulin dose. These different methods of calculating the score used are discussed later in chapter V. An ideal score in this case would be zero.

TABLE IV: Exemplary presentation: Extract of the supplemented data

timestamp	hour	weekday	...	score
2022-12-04 13:06:29	13	6	...	20.423729

Finally, the aggregated data were supplemented by further characteristics. The time of day in hours and the day of the

week as a numerical value were determined from the time stamp of the feed-in point. These were added to the data in order to be able to correlate unconsidered influencing factors (see figure IV).

Each data record now consisted of input data such as the time of day in hours, the day of the week of the injection and the amount of insulin in that injection. There were also cumulative inputs, which included the average, lowest and highest blood glucose values before an injection, the amount of carbohydrate previously consumed, the amount of insulin previously injected and the amount of long-acting insulin previously administered. Activity input data included average heart rate, duration of sleep, duration of physical activity and number of steps. The output data was the previously calculated score, which indicated the mean deviation of glucose values from a target value after an injection. This score is predicted by the learning model.

C. AI Implementation

After data preparation, the AI was implemented. LightGBM was used as the machine learning algorithm.¹

The implementation of the algorithm included the division of the data set into test and training data sets, the determination of the training parameters and the subsequent training of the model. The learning curve plot during training, the visualization of the feature weighting and the test error determined via the Root Mean Squared Error (RMSE) served as the basis for evaluating the training success.

To train the model, the data was divided into training and test data sets, each consisting of different data records, all built according to the scheme described above. The separation in test and training sets was performed using a prebuild method by scikit-learn [24]. To ensure that the results are comparable, the division was carried out without shuffling the values. The model was trained using the training data set to predict the score of an injection based on the input data. A test run with previously unknown test data was then used to determine the prediction accuracy of the model. During the test run, the model made predictions about the score of each test data based on the input data. The actual scores of the test data were then compared with the predicted scores to determine the test error as the RMSE. The RMSE ultimately indicates the accuracy with which the model was able to predict the score of an injection (see figure 5).

After the implementation of the basic model, the training parameters and test size were varied to achieve an optimal training result. The improvement cycle that was continuously run during the implementation is represented by the deming circle [20] in figure 6.

Finally, the best training setting achieved an average RMSE of $20.22 \frac{mg}{dL}$. The comparison of the parameter settings that finally lead to this best result is shown in the following chapter V.

¹LightGBM results generalise well enough to other comparable tools, given the present use case, data types and amount.

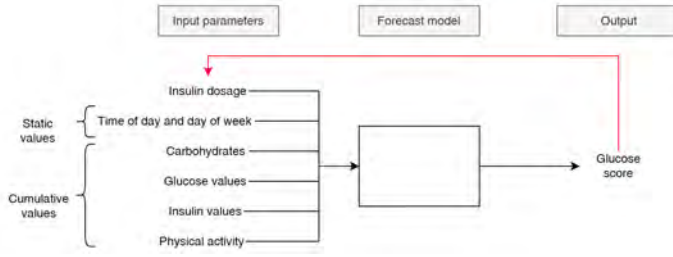


Fig. 5: Input and output parameters of the AI system

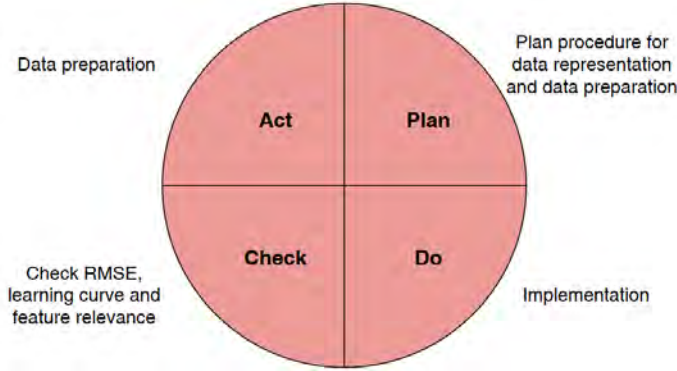


Fig. 6: Deming circle of AI implementation and data collection (own representation based on [20])

V. RESULTS

The model training investigated different training parameters and variants of data preparation with the aim of achieving the best possible training result. The results of the investigations are compared in this chapter. The starting point of the investigations were the settings selected at the beginning of the training. After each comparison, the setting of the best result was used. The basis for evaluation here was the test error, which is given by the RMSE. For the evaluation of the model training further statistics were used. However, for simplicity, these were not considered in the following comparisons.

First, the aggregation periods of the glucose values after an injection were compared. These values form the basis for the calculation of the score to assess the effectiveness of an injection. The variation of the periods is useful because blood glucose fluctuations are usually observed after an injection and only gradually stabilize. In a study it was found that the greatest rise in blood glucose occurs on average one hour after the ingestion of a meal [17]. Usually, blood glucose drops after an initial rise. The mealtime insulin used has a duration of action of 2 to 5 hours. The effect of insulin can therefore be detected up to 5 hours after an injection. As a guideline for checking postprandial glucose levels, the ADA Professional Practice Committee recommends that blood glucose levels 1 to 2 hours after a meal should be $< 180 \frac{mg}{dL}$. The time period for considering glucose levels to make a meaningful statement about the efficacy of an injection was guided by these guidelines, but was verified by experimental comparisons. Too short a time for glucose values to aggregate

could result in a high mean value being misinterpreted as an insulin dose that is too low. On the other hand, too long a time period could lead to bias because events after an injection are not taken into account. For example, if a time span of four hours after an injection was chosen, it would be quite possible that another meal had already been taken during this time, which would not be taken into account when interpreting the result for the current injection. When comparing the time spans of postprandial glucose values, one hour, two hours, three hours, and four hours were compared. Calculating the score with a time interval of three hours provided the best test result (see figure 7).



Fig. 7: Comparison criterion: time after an injection

In addition, different calculation methods for the score were being tested. The score was calculated from the glucose values after an injection and was intended to provide an indication of whether an injection was effective or not. In general, an injection is considered effective if it does not cause either too low or too high postprandial glucose values. To achieve this, a target value or target range was defined to be reached by the insulin administration. In independent training runs, different methods were used to calculate the score. A first method calculated the score as the standard deviation of glucose values from an individually set target value of $140 \frac{mg}{dL}$. This target value was individually chosen and was based on the glycemic target value of $< 6.5\%$ of the patient. The A1C of 6.5 can be converted to a blood glucose value of $140 \frac{mg}{dL}$. A second calculation method calculated the score as the variance from a target value of $140 \frac{mg}{dL}$. A final calculation method defined a target range between 120 and $140 \frac{mg}{dL}$ and assessed deviations from this target range. The independent comparison of all score calculation methods determined the calculation method that provides the most accurate prediction.

In the first formula, the score was calculated from the standard deviation of all postinjection glucose values, using a fixed value of $140 \frac{mg}{dL}$ as the target point instead of the mean.

$$score = \sqrt{\frac{1}{N} \sum_{i=1}^N |(x_i - 140)|}$$

The second option was calculated from the variance of the glucose values. Again, instead of using the mean, a blood glucose value of $140 \frac{mg}{dL}$ was used as the target point.

$$score = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - 140)^2}$$

The first two calculation methods did allow an assessment to be made regarding the effectiveness of an injection. - However, no statement could be made as to whether the injection was too low or too high. Negative values were neutralized in both methods. The target value of $140 \frac{mg}{dL}$ was also not very flexible for the assessment of an injection. Values such as $120 \frac{mg}{dL}$ or $130 \frac{mg}{dL}$ were also evaluated negatively in the result, although these values were in a preferred target range, which should be as low as possible, but should not cause hypoglycaemia. An alternative formula was to evaluate values that were too low negatively and values that were too high positively. Values that were in the target range of 120 to $140 \frac{mg}{dL}$ were not being assessed. The calculation formula was:

$$\begin{aligned} sum &= 0 \\ \text{if } x_i > 140: sum &= sum + (x_i - 140) \\ \text{if } x_i < 120: sum &= sum + (x_i - 120) \end{aligned}$$

In comparison of all calculation methods, the use of the standard deviation was the best variant (see figure 8).

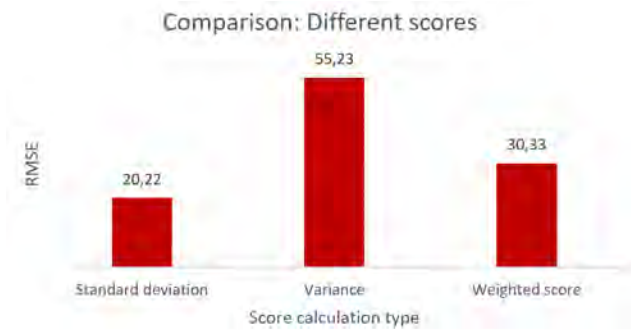


Fig. 8: Comparison criterion: Calculation of the score

Finally, the influence of the characteristics on the training result was tested. On the one hand, the investigation determined the best possible result, but on the other hand, it also allowed a statement as to whether the consideration of additional influencing factors in the data collection had any noteworthy effect at all on the training result. The comparison was carried out once with all characteristics, once without weekday, once without activity data and once neither with activity data nor with weekday.

All features

$X = \text{data}[['hour', 'weekday', 'glucoseAverage', 'glucoseMin', 'glucoseMax', 'HeartrateAverage', 'currentInsulin', 'Exersice-DurationMinSum', 'SleepDurationMinSum', 'StepCountSum', 'kohlenhydrateSum', 'insulinSum', 'langzeitinsulinSum']]$

Features without weekday

$X = \text{data}[['hour', 'glucoseAverage', 'glucoseMin', 'glucoseMax', 'HeartrateAverage', 'currentInsulin', 'Exersice-$

$\text{DurationMinSum', 'SleepDurationMinSum', 'StepCountSum', 'kohlenhydrateSum', 'insulinSum', 'langzeitinsulinSum']]$

Features without activity data

$X = \text{data}[['hour', 'weekday', 'glucoseAverage', 'glucoseMin', 'glucoseMax', 'currentInsulin', 'kohlenhydrateSum', 'insulinSum', 'langzeitinsulinSum']]$

Features without activity data and without weekday

$X = \text{data}[['hour', 'glucoseAverage', 'glucoseMin', 'glucoseMax', 'currentInsulin', 'kohlenhydrateSum', 'insulinSum', 'langzeitinsulinSum']]$

The purpose of this comparison was to investigate how strongly the inclusion of activity data affected the training outcome - this was achieved by directly comparing the outcome with all characteristics and the outcome without activity data. It was also being investigated how strong the correlative relationship was between the day of the week and the outcome. Indeed, the day of the week had no direct logical influence on insulin sensitivity or glucose levels. Only indirect behavioral parameters that were repeated throughout the week had an influence on blood glucose levels. For example, blood glucose could have been fundamentally lower on Sundays than during the week. However, the reason for this would have been a lower stress level, a change in eating behavior or a longer sleep duration.

The best result could be achieved by training with all characteristics. The worse result was achieved by training without activity data and weekday. Thus, taking into account the data collected by the smartwatch had a demonstrably positive effect on the training result. The day of the week also seemed to have an impact on blood glucose levels. Without the day of the week, a worse test result was achieved than when all characteristics were taken into account. This indicated that there were behavioral patterns that influenced the blood glucose course and were repeated on a weekly basis (see figure 9).

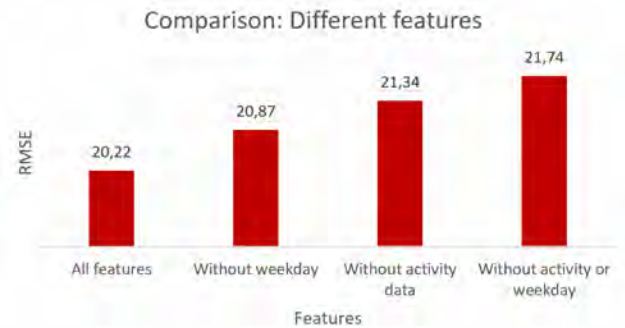


Fig. 9: Comparison criterion: Feature influence

The lowest test error was achieved when a period of three hours was selected for the summary of the glucose values after the injection. The score for evaluating these glucose values was calculated by the standard deviation. Finally, all features were considered for the model training.

A. Prediction accuracy

The best training result achieved an RMSE of 20.22. Compared to the RMSE of 2.4 (MSE: 5.79) obtained, for example, as a result of the study by Obeidat and Ammar [22], the prediction accuracy of this work initially appears worse. However, when the result is placed in the context of application, a different picture emerges. Since the score used for prediction is calculated from the standard deviations of the glucose values after an injection, the training result can also be interpreted as the mean deviation from the actual standard deviation. Thus, on average, the post-injection glucose value is predicted with an accuracy of $20.22 \frac{mg}{dL}$. In conventional insulin therapy, a so-called correction table is used to calculate the correction of a measured blood glucose value. The patient whose health data are used in this paper uses a 30 unit correction table. This means that, on average, one unit of insulin results in a fasting blood glucose reduction of $30 \frac{mg}{dL}$. Consequently, a deviation of up to $20.22 \frac{mg}{dL}$ in prediction would theoretically affect the insulin recommendation by less than one unit. Because Obeidat and Ammar have already implemented a complete insulin recommendation system in their study, the data on RMSE (or MSE) do not refer to glucose prediction but to insulin recommendation (difference between predicted and actual insulin levels).

B. Automation capability of the solution

The aspect of automation is a consistently relevant topic in the design and implementation of the solution. After all, only by automating workflows can diabetics be truly relieved in their everyday lives. Technological means are used to automate data collection, which is also relevant for a later overall solution. Activity data is collected using a Samsung smartwatch, while glucose data is collected through continuous glucose measurements using a Freestyle Libre sensor. Only the carbohydrate amounts are manually entered into a mobile app by the diabetic. Thus, data collection is fully automated only for glucose measurements and activity data. Nevertheless, there is an improvement compared to conventional therapy, as the amount of insulin does not have to be calculated by the patient.

The merging of the data is automated for the glucose measurements and the carbohydrate data via an Iobroker server. The activity data from the smartwatch can only be loaded manually into the database. Samsung does not currently provide an interface for the Samsung Health app to automatically send data to other devices. The data stored in the app has to be downloaded to the smartphone for further processing and uploaded to the database via a direct connection. It would be conceivable to automate this step using a self-implemented solution. In this work, however, this was initially dispensed with, as the focus of the processing was initially on the implementation of the learning model. For a complete insulin recommendation system, a different solution to this problem would have to be found.

Data preparation and training of the model are automated using Python. The finished model could be exported and

integrated into the existing insulin calculator application in a next step.

C. Factors considered

In addition to the glucose level and the amount of carbohydrates, the time of day, stress level, physical activity and sleep duration are also determined as relevant therapy factors. Physical activity is measured by the smartwatch by recording the number of steps and the duration of athletic training sessions. Sleep duration is also measured by the fitness watch. It is true that the watch also has a function for measuring the stress level. However, the measurements can only be performed by manually operating the smartwatch. This does not correspond to the automation idea. Explicitly, the stress level as such is not measured and included in the prediction. Instead, however, the heart rate is recorded.

D. Everyday suitability

The insulin recommendation system is intended to facilitate the everyday tasks of a diabetic. The hardware used for data collection must therefore not be disruptive or unsuitable for everyday use. Smartwatches are basically designed for everyday use. According to the manufacturer, the model used is suitable for showering with a protection class of 5 ATM [27]. However, the subjective feeling when using the smartwatch for data collection is that the watch can interfere when showering, exercising or sleeping. Especially in direct comparison with the glucose sensor, the wrist-worn watch is a bit bulky. Another disadvantage of the smartwatch is that it only has a limited battery life and therefore has to be recharged in between. No data can be recorded during this time.

VI. FUTURE WORK

The result of the work is an AI-based glucose prediction system that uses a convenient, automated data collection process based on one person. The glucose predictions of the model yield encouraging insulin predictions. The system is suitable for everyday data collection and can be integrated into an insulin recommendation system. To the best of our knowledge, this is the first time that a wide range of physical factors are quantitatively measured and taken into account as influencing factors in glucose prediction. The mechanism to try different insulin doses as input to find the best score needs to be added to make a complete recommendation system. The accuracy of the predictions could be further improved with a broader data set. Considering additional influencing factors (female menstrual cycle, state of health, scheduling or physical exercise) could also have a positive effect on the prediction result. Future research could include a comparative study of different machine learning or deep learning techniques, and the application of more comprehensive data mining and data cleaning methods, across more patients.

REFERENCES

- [1] Deutsche Diabetes-Gesellschaft (DDG). "S3-Leitlinie Therapie des Typ-1-Diabetes". In: *Evidenzbasierte Leitlinien 2* (2018).

- [2] Amin Alqudah, Abdel-Rahman Bani Younes, and Ali Mohammad Alqudah. "Towards modeling human body responsiveness to glucose intake and insulin injection based on artificial neural networks". In: *Jordanian Journal of Computers and Information Technology (JJCIT)* 6.01 (2020).
- [3] Christoph Auer, Nora Hollenstein, and Matthias Reumann. "Künstliche Intelligenz im Gesundheitswesen". In: *Gesundheit digital*. Springer, 2019, pp. 33–46.
- [4] T Biester et al. "Individualisierung der Diabetestherapie durch Automatisierung der Insulingabe". In: *Monatsschrift Kinderheilkunde* 169.10 (2021), pp. 902–911.
- [5] Giacomo Cappon et al. "Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes—An in silico proof-of-concept". In: *Sensors* 19.14 (2019), p. 3168.
- [6] Linda A DiMeglio, Carmella Evans-Molina, and Richard A Oram. "Type 1 diabetes". In: *The Lancet* 391.10138 (2018), pp. 2449–2462.
- [7] Docker. <https://www.docker.com/>. Accessed: 2023-04-26.
- [8] Nuha A. ElSayed et al. "6. Glycemic Targets: Standards of Care in Diabetes—2023". In: *Diabetes Care* 46.Supplement₁ (Dec. 2022), S97–S110. ISSN: 0149-5992. DOI: 10.2337/dc23-S006. eprint: https://diabetesjournals.org/care/article-pdf/46/Supplement_1/S97/693609/dc23s006.pdf. URL: <https://doi.org/10.2337/dc23-S006>.
- [9] Cheng Fan et al. "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data". In: *Frontiers in Energy Research* 9 (2021). ISSN: 2296-598X. DOI: 10.3389/fenrg.2021.652801. URL: <https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801>.
- [10] *Freestyle Libre 3 Sensor*. <https://www.freestylelibre.de/produkte/freestyle-libre-3-sensor.html>. Accessed: 2023-04-26.
- [11] *Freestyle Libre app image*. https://www.freestylelibre.de/content/dam/ad/freestylelibre/de/de/fsl3/images/home/freestylelibre-measuring-system/FSL3_genericphone_cleanscreen_112mg.png. Accessed: [09.05.2023].
- [12] *Freestyle Libre sensor image*. <https://www.freestylelibre.de/content/dam/ad/freestylelibre/de/de/fsl3/images/freestyle-libre-3-entdecken/Sensor-2-floating.png>. Accessed: 09.05.2023.
- [13] Deutsche Diabetes Gesellschaft. *S3-Leitlinie Therapie des Typ-1-Diabetes*. https://www.ddg.info/fileadmin/user_upload/05_Behandlung/01_Leitlinien/Evidenzbasierte_Leitlinien/2018/S3-LL-Therapie-Typ-1-Diabetes-Auflage-2-Langfassung-09042018.pdf. 2018.
- [14] Ketan Gupta and Nasmin Jiwani. "Prediction of Insulin Level of Diabetes Patient Using Machine Learning Approaches". In: *Ketan Gupta, Nasmin Jiwani, Prediction of Insulin Level of Diabetes Patient Using Machine Learning Approaches*, *International Journal of Creative Research Thoughts (IJCRT)*, ISSN (2022), pp. 2320–2882.
- [15] Michael Harreiter Jürgen und Roden. "Diabetes mellitus—definition, klassifikation, diagnose, screening und prävention (update 2019)". In: *Wiener Klinische Wochenschrift* 131.1 (2019), pp. 6–15.
- [16] Lutz Heinemann et al. "Glukosemessung und-kontrolle bei Patienten mit Typ-1-oder Typ-2-Diabetes". In: *Diabetologie und Stoffwechsel* 14.S 02 (2019), S119–S141.
- [17] N Jendrike et al. "Glucoseverläufe nach Einnahme einer schnell resorbierbaren Mahlzeit zu unterschiedlichen Tageszeiten". In: *Diabetologie und Stoffwechsel* 2.S 1 (2007), P361.
- [18] Bernhard Kulzer and Bad Mergentheim. "Künstliche Intelligenz in der Diabetestherapie". In: *Digitalisierungs- und Technologiereport Diabetes* (2022).
- [19] John Martinsson et al. "Blood glucose prediction with variance estimation using recurrent neural networks". In: *Journal of Healthcare Informatics Research* 4.1 (2020), pp. 1–18.
- [20] Ronald Moen and Clifford Norman. *Evolution of the PDCA cycle*. 2006.
- [21] Stefanie Neumann. *Untersuchung der grundsätzlichen Eignung eines KI-gestützten Systems für automatisierte Therapieanpassung von Diabetes mellitus Patienten*. unpublished. Student research project at the Baden-Wuerttemberg Cooperative State University Karlsruhe, submission date: 22.05.2023.
- [22] Yusra Obeidat and Ahmad Ammar. "A system for blood glucose monitoring and smart insulin prediction". In: *IEEE Sensors Journal* 21.12 (2021), pp. 13895–13909.
- [23] *pandas*. <https://pandas.pydata.org/>. Accessed: 2023-04-26.
- [24] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [25] *Project Jupyter*. <https://jupyter.org/>. Accessed: 2023-04-26.
- [26] Vaisnevee Sugumar et al. "A Comprehensive Review of the Evolution of Insulin Development and Its Delivery Method". In: *Pharmaceutics* 14.7 (2022), p. 1406.
- [27] *Vergleich: Galaxy Watch5 vs. Watch5 Pro*. <https://www.samsung.com/de/support/mobile-devices/vergleich-galaxy-watch5-watch5-pro/>. Accessed: 2023-04-18. 2022.
- [28] Konstantia Zarkogianni et al. "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring". In: *Medical & biological engineering & computing* 53 (2015), pp. 1333–1343.
- [29] Taiyu Zhu et al. "A Deep Learning Algorithm for Personalized Blood Glucose Prediction." In: *KHD@IJCAI*. 2018, pp. 64–78.

A Language Training Approach to Improve Children's Expression Skills Using AI Methods for Text-to-Picture Conversion

1st Elisa Schäfer

Faculty of Computer Science
Duale Hochschule Baden Württemberg
Karlsruhe, Germany
elisa@elmama.de

2nd Armin Zundel

INLINE Internet Online Dienste GmbH
Karlsruhe, Germany
zundel@inline.de

3rd Kay Berkling

Faculty of Computer Science
Duale Hochschule Baden Württemberg
Mosbach, Germany
kay.berkling@mosbach.dhbw.de

Abstract—Pupils performance in key subjects at school shows a decrease in recent years. The aim of the scientific paper is to conceptualise, evaluate, implement, and test a language training tool that enables children to improve their expression skills using a state-of-the-art text-to-picture-generating artificial intelligence. In the beginning, the theoretical background, covering learning theories, motivation, cognitive processing, and artificial intelligence, is summarised. Based on this background, the use case, target group, and the task that is to be performed when using the tool are defined. Afterwards, the solution design, including the comparison of text-to-picture frameworks and the selection of DALL-E 2, as well as the prompt engineering, results in the model of dependencies. Finally, an exploratory data analysis to evaluate the tool's performance under real circumstances is performed, to then be able to draw a conclusion, discussing the overall benefits, challenges, and future research options.

Index Terms—text-to-picture conversion, artificial intelligence, education, language training, cognitive science

I. INTRODUCTION

In recent years, students' academic performance, particularly in subjects like Mathematics and German, has seen a decline. The IQB's 2021 educational trend study in Germany highlighted this issue, revealing poor performance among fourth graders [1]. Considering the data of the National Assessment of Educational Progress, in 2017, three-quarters of both twelfth and eighth graders lacked proficiency in writing [2], which shows that this decline is not limited to Germany and the COVID-19 pandemic but reflects a global problem due to societal changes and evolving learning conditions. The accessibility of technology and the prevalence of informal language on online platforms have influenced children's language use [3]. While previous generations have managed to balance informal language use with formal language skills, the current generation seems to be struggling in this regard. The educational sector plays a vital role in addressing this imbalance and cultivating a healthy relationship between students and their first language, which serves as the foundation for all other subjects. However, the traditional teaching methods employed in schools, which often lack engaging visuals and interactive elements, fail to capture students' attention

and motivation. Fostering an interactive and student-centred learning environment that incorporates technology, empowers children, enhances their internal motivation, and promotes critical thinking and problem-solving skills [4]. To address these challenges, it is crucial to develop effective teaching tools that utilise advanced technology to autonomously enhance students' vocabulary and language expression skills. By making language learning an adventurous and self-directed process, these tools can boost students' motivation, confidence, and overall academic performance in various subjects. Ultimately, this tool aims to transform students' attitudes towards learning and contribute to a successful language learning process by enhancing both motivation and attitude, which are closely intertwined [5] [6].

II. BACKGROUND

A. Learning Theories

Education and research in the field of learning have made significant progress, incorporating various theories such as behaviourism, cognitive learning, social-cognitive learning, constructivism, and social-constructivism. These theories provide frameworks for effective educational practises that align with the human mind.

The **constructivist approach** has been shaped by Piaget, who developed a stage model that categorises the child's development by age. He splits the child's development into the sensorimotor, preoperational, concrete operational, and formal-operational stages. These four stages describe the child's cognitive development based on biological, logical and mathematical as well as psychological backgrounds [7].

Due to the fact that this paper aims to develop a tool that can be used in a school environment, the focus lays upon the last two stages, the concrete and formal operational stages. Furthermore, only for the developed approach relevant facets of Piaget's research in the stage model are to be described in further detail.

In the concrete operational stage, children start to think logically about concrete situations and strengthen in categorisation, which enables them to organise objects into classes.

Towards the end of this stage, children are also able to define properties that determine the classes they are sorting the objects into. The child is able to understand the concept of conservation and begins to use inductive concepts that allow it to reason from specific experienced knowledge to a general principle [7].

In the transitional years from the third to the fourth stage, children are starting to change their underlying patterns of thought and develop the ability to process several aspects of a situation simultaneously, and become sensitive to transformations [7]. Moving to the formal operational stage, the child starts to develop abstract thinking processes and begins to reason about hypothetical problems. At the age of 11 and older, children generally have a sense of genuine cooperation, which makes them able to know the rules and enjoy elaborating upon them in a social context. In this stage, deductive logic is developed, which enables the child to reason from a general principle to specific information [7].

B. Motivation

One of the fundamental pieces of research that needs to be considered when touching upon motivation and a person's natural productivity, is Ryan and Deci's **Self-determination theory**.

In their approach to human nature and personality, consisting of traditional empirical methods, organismic meta-theories, and studies, they define the psychological needs that are to be fulfilled in order to make humans motivated and self-determined as well as to assure their well-being [8].

They define three different types of motivation: *amotivation*, *extrinsic*, and *intrinsic motivation*.

Ryan and Deci referred to intrinsic motivation as "the inherent tendency to seek out novelty and challenges, to extend and exercise one's capacities, to explore, and to learn", whereas extrinsic motivation "refers to the performance of an activity in order to attain some separable outcome" and involves "instrumentalities rather than enjoyment of the work itself" [8]. While both types of motivation lead to an increase in productivity, only intrinsic motivation leads to a feeling of fulfilment and, due to that, to a productivity that is less stressful and has a better effect on a human's mental well being.

In order to foster intrinsic motivation, Ryan and Deci presented the **cognitive evaluation theory** as a subtheory to their Self-determination theory, where they aimed to identify the psychological needs that foster intrinsic motivation. The results of this analysis are three basic needs: *autonomy*, *competence*, and *relatedness*. Because of the fact that these needs are interrelated, a certain degree of all of them is required to achieve the desired intrinsic motivation [8].

C. Cognitive Processing of Visual Displays

The human brain has a certain way of accessing the information provided by visual displays. Therefore, visual displays affect cognitive processing strongly. Following McCrudden

and Rapp's research, visual displays can have relational inferences with objects, which include intrinsic-static and intrinsic-dynamic visual displays, or they can have relational inferences between objects, such as extrinsic-static and extrinsic-dynamic visual displays.

To fully retrieve information from visual displays, a person has to organise what is seen. In cognitive terms that means, that this person has to extract and localise important information to support *organisational processes*. When extracting, the brain is separating more important from less important, whereas when localising, it is placing related information in close proximity [9].

Once organised, a person can start the *integration process* of a visual display, which means that there are inferred relations between important information contained in an instructional message and prior knowledge. In order to make prior knowledge and information from an instructional message associated in human memory, it is important to activate the two of them simultaneously [9].

Visual displays can be a tool to encourage learners to reflect on and test their ideas. *Constructivist activities* challenge students to build predictions, run simulations, and manipulate factors that are crucial to the learning process.

They are also likely to support integration as learners engage in knowledge generation that relates to and goes beyond the information provided in a display [9].

D. Text-to-picture conversion

The task of generating a picture from a prompt is a central component of the tool evaluated and implemented later in this paper and is performed by deep learning models. Text-to-picture generations are complex problems because, in order to solve them, a complex model that bridges the gap between text and image modalities is needed.

There are several approaches, ranging from cGANs as well as VAEs to transformer-based models.

While **CGANs** are characterised by training a GAN where the generator takes both random noise and textual input as conditions to generate images that align with the provided text [10], **VAEs** learn a latent space representation of images as well as text and generate new images based on text inputs by sampling from these text inputs [11].

Apart from these common approaches, there is another way of dealing with text-to-picture tasks, which is the **diffusion model**. It refers to a family of generative models, with the most well-known being the score-based generative model or score matching, primarily focusing on generating images from noise or partial images. This approach aims to learn a probability distribution by modelling the score function associated with the log-density of the distribution, which makes it suitable for image generation, and data synthesis tasks [12].

The diffusion model that is implemented in the later chosen deep learning framework is a diffusion model that is designed for text-to-picture generation by using a combination of multiple neural network models as well as techniques in order to be able to handle the various input prompts.

III. USE CASE AND REQUIREMENTS

Learners' characteristics and desires form concrete areas to emphasise, when developing an application that enables learners to be self-determined and to actively interact with the given material, from which certain use cases as well as requirements derive.

A. Key Aspects

As the key driver for productivity, **motivation** is the first aspect to consider. As already presented in the background section II-B, the tool aims to trigger intrinsic motivation by targeting learners' feelings of autonomy, self-determination and relatedness, which results in creating a task-solving approach that enables learners to independently work on the given task. Furthermore, various original pictures that evoke emotional connection should be designed, as this does not only target learners' relatedness but also adds the element of choice to the process.

Above that, it is crucial to avoid limiting learners' input, allowing them to generate their own ideas and see results, promoting a sense of freedom. Additionally, granting learners the choice of whether they would like to go on altering their description or to end the task, creates a feeling of self-determination.

Apart from motivation, **clarity** is crucial to providing a straightforward learning environment. Important requirements for a clear structure are a well-designed user interface as well as a consistent feedback cycle. Reducing distractions through a minimal user interface and displaying pictures as well as descriptions in close proximity improves cognitive processing, as described in section II-C. By limiting the picture-generating AI to one art style, confusion can be prevented and constant result picture improvement is supported.

The last main aspect that traditional learning approaches lack, is a trigger to engage learners in an **activity**. This trigger is not only embodied by providing the learners with a modern technical device and tool they can work with individually and the excitement that comes with opportunity, but also through the user interface, which encourages learners to get into action by putting default text in the input areas.

These interconnected aspects, along with preconditions such as task understanding and independent interpretation of feedback, make the language training tool effective for competence-oriented and individualised learning. The text-to-picture approach enhances knowledge establishment by leveraging both, text and visual displays, benefiting learners' psychological needs and promoting cohesive learning.

B. Target Group

Identifying the target group is crucial for designing an effective tool. The target group for this interactive and autonomous language training tool is pupils who can read, write, and interpret context from pictures as well as deal with technical devices and use keyboards. To address the declining performance in German at an early stage, the tool should be implemented as early as possible, while still meeting the main

set of skills described above.

Considering Piaget's theory, as presented in the background section II-A, the right age is between ten and twelve years of age, as then, children will develop inductive thinking, become sensitive to transformations and start building mental models. Learners will benefit from this training approach by practising their skill of transferring information and building abstract mental models from specific experience accordingly, while also building their language knowledge using these skills.

Taking the named problems and requirements for the pupils to be effectively using this tool into account, the decision on the target user group for this tool are learners of the fifth grade. By then, they should be able construct own sentences and use a keyboard confidently on their own. Choosing this age span, it can be made sure that learners are equipped with a strong knowledge base and confidence in their ability to naturally use correct language suitable for their further educational process. Especially, because the transition time between primary and secondary school is crucial for learners to independently improve their language skills in order to get on the level that is required for further school years. Working individually and using a computer with a keyboard, fifth graders can enhance their technical and typing skills while reducing the pressure of external comparisons and adapting to new classmates.

C. Task

The concrete task that the learners should perform when using the language training tool is to improve their description of an original picture by comparing it to their own result pictures.

The entire task follows an iterative principle, which engages the learner to head into the task's cycle by using an original picture that is to be described as a trigger.

In order to start the process, learners can choose the most appealing or best describable from a set of original pictures of different abstraction levels. This way, every learner can choose a picture that is comfortable and offers enough of a challenge without being overwhelming.

After this element of choice, the fifth graders write a prompt as a description of this picture in one sentence and click on the "generate image" button to then get a generated picture back.

After that, learners analyse the differences in the two pictures in a self-determined manner and use the found differences to alter or create a prompt in order to get results that resemble the original picture better.

This iterative process is performed as long as learners want to optimise the description. The trigger to end the process is their satisfaction with the generated result.

IV. SOLUTION DESIGN

A. Comparison of Text-To-Picture Frameworks

In the fast-changing IT industry, especially around text-to-picture generation frameworks, there are many that provide an image-generating function accessible as a web interface, while their features as well as their further development and

pricing can change on a daily basis. Thus, in this chapter, the comparison and the decision on the framework are made on March 25, 2023. Any further developments in the picture-generating field are not taken into account in this particular comparison.

Because of the fact that the most important requirement in the process of developing a language training tool is the possibility of using the framework in a self-implemented application, only those public text-to-picture frameworks that provide an open source solution or an API are assessed in further detail.

In order to be able to compare the shortly presented picture-generating frameworks adequately, certain evaluation criteria, which are determined by the requirements, are to be defined. The criteria are separated into different categories, which include content, economic, technical, and security aspects. The selected four categories, - picture-, cost-, technology-, and security-related - are mandatory for being able to properly use the framework in a school environment. Each category consists of different facets, which are explained in the following:

Picture-related criteria:

- Content Quality: accuracy of the prompt's content reflection in the picture
- Output Quality: picture quality
- Art Styles: the range of different art styles that are provided

Cost-related criteria:

- Limitation: limitation regarding the signup and use of the framework
- Cost: cost that might occur when there is a limitation

Technical criteria:

- Creation Time: measured creation time that is needed for a text-to-picture generation with the respective web interface
- API: indicator whether the framework has an API for developers or a GitHub repository
- API user friendliness: indicator how easy the API is to work with

Security and User Policies criteria:

- Content Filters: indicator whether the has content filters or not
- Diversity: indicator whether the framework has diversity measures or not

With these assessment criteria points, it should be possible to choose the suitable text-to-picture-generating AI framework that meets the requirements for the training tool that is to be conceptualised.

After the main content in each cell, the score for the respective criteria point is written in brackets. If the cell's content already embodies a score itself, no extra score is written behind it. Because of the fact that the application that is to be developed has a target group of children aged from ten to twelve years, some criteria points are prioritised more than others. That includes, especially, safety and user policy measures, which is why the score is doubled for the criteria points in this category to make sure that children are not exposed to inappropriate

content. The last row of the table is the overall score, which is the sum of all the rows above.

In table I the text-to-picture-generating AI Frameworks Stablecog¹, Dream², DeepDream³ and DALL-E 2⁴ are compared.

Based on the overall score, the decision is made for DALL-E 2 by openAI, as it scores best in API and its user friendliness, as well as Security and Policies, which have the highest priority.

DALL-E 2 is one of the more popular AI image generators and provides an online interface as well as a developer API that can be used to create an own AI image generator tool. It has a diverse set of capabilities, including combining unrelated concepts in plausible ways.

DALL-E 2 shows high quality in output as well as content and is very flexible in configuring the output parameters such as picture size and art styles. With strong content filters and active diversity measures, DALL-E 2 is extremely suitable for a school environment.

Especially compared to Stablecog, where the MIT Project would need to be set up and trained with a variety of own pictures, DALL-E 2 provides an easy-to-use API using an extensively trained diffusion model to learn the relationship between text and images that has access to an extremely huge amount of pictures to refine the features.

Moreover, with DALL-E 2 being in the beta phase and being part of openAI's portfolio, it is clear that further development and improvement are yet to come while still meeting the important content policies⁵, which includes measures against attempts to create pictures that could cause harm. These strict policies make DALL-E 2 very safe for immature users but also limit its output, as no political or other famous people can be generated with their model.

The mentioned benefits of DALL-E 2, especially in terms of the prioritised content policies and content quality based on the well-trained model, lead to the decision for this framework.

B. Prompt Engineering

Considering the fact that DALL-E 2 offers more than 50 different art styles and creates individual art from a given prompt, there is a huge amount of different possibilities when creating pictures.

In order to be able to perform the task for this particular language training approach (described in section III-C), it is crucial to create pictures that are properly comparable with each other. Therefore, specific parts of a prompt need to be specified and fixed, as otherwise the range of different outputs that are randomly occurring could be counterproductive to the task at hand.

For this specific tool, the aim is to find a realistic art modifier to make the pictures easy for young learners to interpret.

¹<https://stablecog.com/>

²<https://dream.ai/>

³<https://deepai.org/machine-learning-model/deepdream>

⁴<https://openai.com/dall-e-2/>

⁵<https://labs.openai.com/policies/content-policy>

TABLE I
TEXT-TO-PICTURE AI FRAMEWORK ASSESSMENT

Criteria	StableCog	Dream Wombo	Deep Dream	DALL-E 2
PICTURE-RELATED				
Output Quality	high (4)	high (4)	very high (5)	high (4)
Content Quality	high (4)	low (2)	very high (5)	very high (5)
Picture Sizes	4 sizes (4)	1 size (1)	1 size (1)	3 sizes (3)
Art Styles	≥ 25 (3)	21 (2)	29 (3)	≥ 50 (5)
COST-RELATED				
Limitation	limited (1)	unlimited (5)	limited (1)	limited (1)
Cost	\$0.006 (4)	none (5)	\$0.68 (1)	\$0.02 (3)
TEC-RELATED				
Creation Time	20 secs (4)	10 secs (5)	25 secs (3)	50 secs (1)
API	GitHub (1)	yes (5)	yes (5)	yes (5)
API user friendliness	difficult (1)	easy (5)	medium (3)	easy (5)
Language	multi	English only	English only	English only
SECURITY-RELATED (x2)				
Content Filters	self-service (3)	no (1)	no (1)	yes (5)
Diversity	no (0)	no (0)	no (0)	yes (5)
OVERALL SCORE				
Overall Score	37	43	36	57

Among the realistic art modifiers, *35 mm f_8 1_300s, photography, photo, high quality photo 35 mm f_8 1_300s, photo 35 mm f_8 1_300s* and *high quality photo*, the latter scored best and was therefore chosen to be the fixed modifier when generating images with the language training tool.

C. Model of Dependencies

With all the mentioned dependencies, framework requirements, as well as those resulting from a proper feedback in order to provide a decent user experience and a well designed layout, the model of dependencies, shown in figure 1, can be constructed.

It shows the overall cycle the learner goes through, with all its technical dependencies. The cycle trigger is the chosen original picture. From there on, the user enters the cycle of typing the description in, which is then engineered in the backend to get a picture generated in return, which can then be compared to the original one. The prompt engineering in the backend includes the translation of the description with DeepL⁶ as well as adding the art style to this translated description and with that turning it into the prompt that will then be used by the DALL-E 2 API to generate the result picture. This generated picture is then displayed in the frontend, where the learner is able to alter the description to get a more precise result.

V. EXPLORATORY DATA ANALYSIS

In order to be able to see if the concept behind the developed language training tool and the overall approach are working and to discover in what particular way the tool has an influence on the target group's learning progress as well as their learning process, an *exploratory data analysis* is carried out.

This approach has the benefit of being able to discover the tool's impact while avoiding the limitation of specifying too many details for the target beforehand.

⁶<https://www.deepl.com>

A. Methodology

The overall approach to testing the developed language training tool is to create a testing environment in school where learners of the evaluated age group are able to test the software individually. The methodology to collect the participants' result pictures and prompts is to give them the already evaluated and iterative task, which is described in detail in section III-C, to solve in a setup environment in their school. Each student solves this task individually, and all students have the same setup for doing that, which is described later in this paper.

B. Environment and Setup

For individual testing, an empty classroom is prepared and setup with the correct hardware that enables the participants to interact with the provided software.

For each participant, the interface is refreshed and has the same default starting screen.

Learners can make themselves comfortable in the provided chair and adjust the external keyboard, the laptop's screen, as well as the mouse before and during the solving process.

The only other person in the room is the supervisor, who is running the test. The supervisor's task is not only to bring the testing setup to the starting stage each time a participant is finished but also to answer questions that might occur or to help if there is a major technical issue with the hardware or software. Furthermore, the supervisor does explain the task and which part of the hardware should be used to solve it, as well as the possible options for the original pictures, and makes sure that the original picture chosen by the participant is well presented on the tablet. Apart from these actions, the supervisor remains silent in the room and does not interfere with the pupil's testing process. The only questions that are allowed to be answered are those that, if unanswered, would make it impossible for the participant to carry on with the task.

In general, there is no limit to the time or the number of

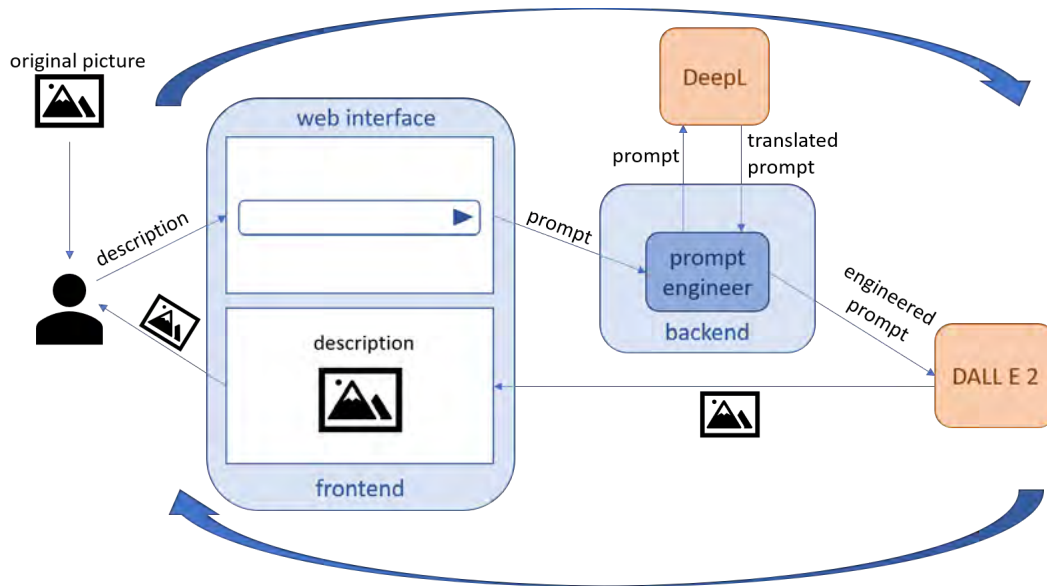


Fig. 1. High level model of dependencies

generated pictures, which reduces the pressure while solving the task. The end of the task is determined by the pupil's satisfaction with the result picture created by her prompt. The participant is then allowed to leave the room so that the setup can be brought to the starting stage.

C. Data Collection

The test data is collected by saving the pictures as well as the pupils' written prompts automatically and locally on the laptop when clicking on the button for picture generation. This way, an uninterrupted testing experience for the participants is secured.

The data for this exploratory data analysis is collected in two sessions, which are both carried out during supervised homework time (14:00-16:00) in the afternoon.

D. Instruments

There are several instruments that are used for the data collection process, including hardware and software components as well as the school's furniture, to provide a comfortable and familiar learning environment for the participants.

The technical hardware components used are a laptop, a tablet, as well as an external mouse and keyboard. The laptop is needed to run the tool and to display the user interface for the participant. The external mouse and keyboard are essential as they facilitate the use of the application. The external mouse provides an easier way of moving the cursor compared to the touch pad, and the external keyboard's keys are bigger and more responsive, which result in an easier typing process for the participants. The tablet is used to display the original pictures.

The room that is used for the data collection is a quiet classroom with a regular setup of desks and chairs. The participants specifically used only one chair and one desk,

which had the technical setup on it.

Apart from the physical instruments, digital resources are used to perform the test. The most important instrument for that is the actual tool, which runs on the already mentioned laptop. The interaction is performed in the browser while the project, started by the supervisor, is running in the background.

Furthermore, the original pictures of different abstraction levels beforehand created are an important part of the digital instruments.

The choice for different levels of abstraction was made because children in the fifth grade, as already evaluated in section II-A, are already able to think inductively and abstractly as well as to build mental models.

In the process of creating the original pictures, emphasis is laid not only on the fact that they should look as photorealistic as possible while also representing different levels of abstraction, but also to represent something that pupils might be naturally interested in, which makes them feel related to the scenarios in the given pictures.

The four pictures that resulted are the following:

- *An orange hanging from a green branch with light green leaves that has a canopy of leaves in the background and the sun shining through it*, represents the simplest of pictures in terms of the level of abstraction because it is a rigid real-life object that does not perform an action or motion.
- *A cute Belgian Shepherd puppy on a green meadow running towards the viewer with its tongue stuck out and its left ear snapped*, remains rather simple in the fact that it still shows a real life situation but includes motion and perspective as well as concrete features such as the type of dog.
- *A strawberry floating through space, illuminated by the sun with the milky way in the background*, offers the

possibility to describe two familiar natural parts of the picture, namely the milky way and the strawberry, while combining these two familiar pieces into one shared construct.

- *An astronaut riding on a horse in the desert at sunrise*, embodies the most advanced picture as it shows two living beings that are from different natural contexts, performing an action together in an environment where they are usually not living naturally, while adding the detail of time.

These four pictures make it possible for children of different learning stages and advancements to be able to use the tool and to achieve their personal and individual learning success.

E. Participants

In two different testing sessions for the developed tool, 20 different fifth graders solved the task of describing their chosen picture. The learners tested the software one by one in a quiet testing environment where they had a whole empty classroom to themselves to be fully able to focus on the task. This way, they also had no option to compare themselves, their picture, or their solving time to others.

The participants are pupils from the Schoenstaetter Marienschule in Vallendar (Rheinland-Pfalz). The inclusion criteria were that they needed to be in the fifth grade, which makes the participants ages range from ten to twelve years. Other than that, there are no specific selection criteria. Performing the tests on two different afternoons (17.04.23 and 18.04.23) in supervised homework time (14:00-16:00), the first participants are those fifth graders that are already done with their homework, followed by the pupils that needed longer to finish their homework. With that, testing participants have an order, while the overall sample consists of both, fast learners and those that need longer to solve regular homework tasks.

F. Analysis Template and Criteria

There are many possibilities for analysing the collected testing results that are related to the length and development of the sentence used as a prompt.

For this paper, the chosen analysis criteria are the number of tries and words used per prompt. For the latter, the prompt of the first and last try of each pupil will be compared, and the difference of both amounts will be used as a size for comparison. Furthermore, it is shown how many tries each pupil used while comparing these two amounts as a reference to be able to put the difference into context.

For the technical and statistical analysis, the prompts saved during the test are extracted from the local folder and fed into Microsoft Excel⁷ to be able to analyse their lengths. The data visualisations shown in the next section, as well as the calculations related to them, are also generated with Microsoft Excel.

Apart from the mentioned criteria, patterns and observations of the supervisor are taken into account when analysing the

testing results to detect possible target variables for further research. For that, not only realtime impressions during the test but also a "human"-centred approach to analysing the results were performed after the testing sessions.

G. Findings and Interpretations

Based on the collected data and the performed analysis, there are various findings that are described and interpreted in the following.

The summarised results of the analysis (Figure 2) show that in all cases, the final prompt before finishing the task was longer than the first try. The average number of words for the first try is 8.4, while the average for the last try is 12.85 words, which makes an average difference between the first and last try of 4.45 words. While the shortest first try contains only one and the longest 16 words, the range of the last try starts with six and ends at 22 words, which makes the difference between the minimum and maximum of the first and last try vary only by one word. The number of tries ranges from 2 up to 6, with an average of 3.25 tries.

Furthermore, the number of words between the first and last try does not correlate with the total number of tries, as can be seen in figure 2 where these two lines have totally different courses.

The number of words per prompt varied from try to try, and even though in all cases the last try was longer than the associated first try, in the process of finding a better description, some prompts are shorter than their predecessors. This can be seen in figure 3 for the data sets P_7, P_10, P_13 and P_19. Tries that are shorter than their predecessors can occur due to restructuring the description or starting with a completely new sentence in the process of solving the task.

Additionally, it can be seen in figure 3 that the average amount of words per try is highest for the fourth try (12.56 words), while for the fifth and sixth tries, the average is lower with 8.4 and 11.5 words, respectively. This can result from the fact that there were only two data sets for those two last tries, as all other participants in the test were satisfied with the outcome before these last two. Therefore this trend should be analysed with a larger data set to verify the trend.

Figure 3 also displays the individuality of the pupils' describing processes. Some pupils were satisfied with their outcome after fewer tries than others, with different amounts of words as well as totally different sentence structures.

Apart from the number of tries and the words per prompt, it was counted how many times each of the original pictures was chosen.

The most chosen picture for the test was the *strawberry floating through space* (chosen 8 times), whereas the *orange hanging from the tree* was the least chosen picture (3 times). This could originate from the fact that children like to choose a picture that interests them, which might show that they prefer an interesting picture even though it might be a challenge to describe an unknown abstract situation over an everyday situation that would be way easier to describe.

Apart from these findings, the supervisor could observe differ-

⁷<https://www.microsoft.com/de-de/microsoft-365/excel>

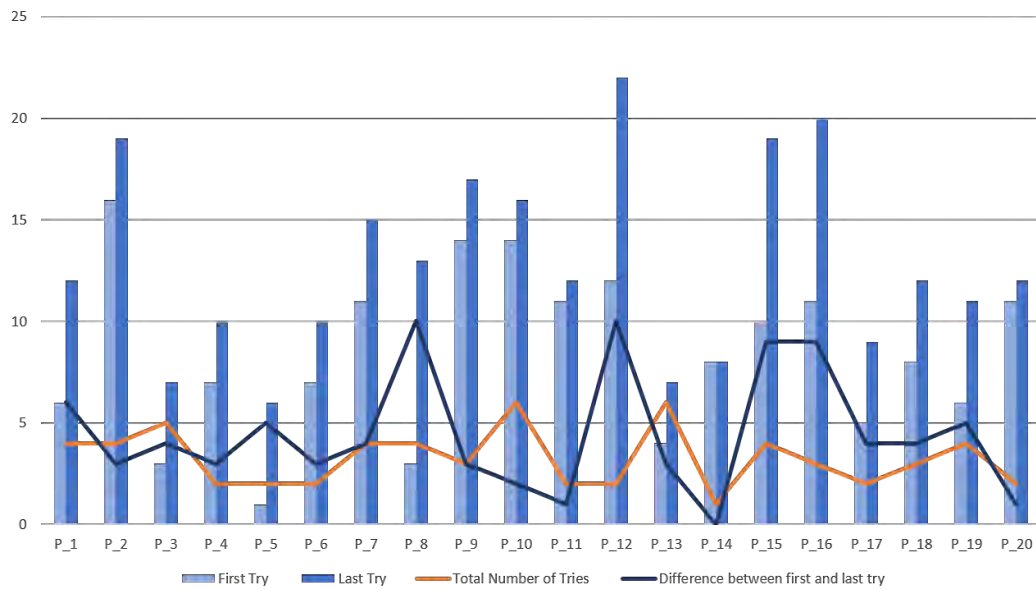


Fig. 2. Overview of summarised test results per person

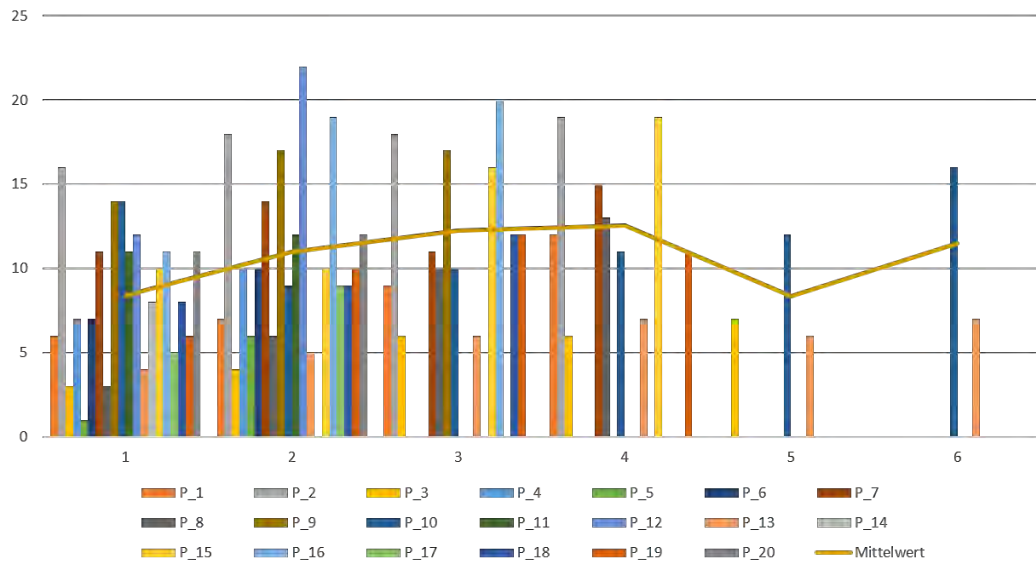


Fig. 3. Number of words per try per test person

ent approaches to improving the prompt that were independent of the number of words used per try. It could be seen that the steps taken to improve the description were very different from participant to participant. While some went from detail to detail, only changing or adding one attribute at a time, some started basically from scratch in the middle of the describing process and tried a completely restructured and newly formed sentence. This shows that the participants could use the tool to improve their describing skills in many different ways. Furthermore, it was observed that the pupils could find different descriptions for the same phenomenon, such as describing the colour of an object instead of the way its

colour is influenced by other factors (e.g. the lighting). This also represents the different thinking and describing strategies for the problem solving process, which is an indicator for the fact that learning tools that enable users to think freely and to express themselves creatively are an excellent way of bringing learners to their full personal potential.

Apart from the individualistic factors, it was observed that pupils in the testing group tended to specify certain attributes and exchange basic, simple words with more advanced vocabulary. Furthermore, participants tended to add complexity to the sentence structure by adding more sentence constructs, for example, relative sentences. That way, it could be seen that

the tool does not only trigger the use of more words but also the tendency to use more advanced words as well as more complex sentences. These tendencies should be analysed and proven in further detail based on a larger amount of testing data.

H. Results

The test and the respective analysed data show that, for the testing group, the language training tool is fulfilling the purpose it was designed and implemented for. Pupils were able to get creative in using language to describe a picture chosen by them. The amount of words per try and the increase of those when comparing the first to the last try across all participants show that the tool enables pupils to form more detailed sentences when describing a situation for a given picture.

The analysis has also shown that the focus on individual learning progress and describing processes worked for the selected target group. The tool as well as the element of choice in the task at hand enabled pupils to have a learning experience that fits their interest as well as learning stage, and with that triggers intrinsic motivation combined with a positive feeling while exploring their ability to use language.

The benefit of this is not only that they enjoy the task while being told to resolve it, but also that they actively seek to use the tool more often. Almost all participants asked when they would be able to use the tool again and wanted to carry on describing more pictures. The desire to go on learning can be extremely beneficial and shows once again that the tool's design triggers an internal feeling of fulfilment when using it. The test session also showed that pupils engage in prior knowledge to describe real life but also abstract situations, and with that, they are able to make cross-connections between different subjects, which will help them in the future. Besides that, it could be seen that different approaches or ways to describe things, such as describing the colour of an object by using the description of the lighting in the shown situation or specifying the colour with additional colour-specific attributes, can end in equally correct solutions. This shows that the support of individuality in solution processes is working for this tool and that pupils can bring their individual strengths to the task and develop them even further. It also shows that the tool is adapting to the children's knowledge state and learning process.

In addition to the benefits for individuality in the process and content, as well as the improvement of using language while being creative, which results in using more words, the testing sessions also showed that the tool met the user experience it was designed for. Not only did all the technical functionalities of creating the prompt and displaying the picture work, but also the constraints that were built into the button functionality reduced the stress and chaos of generating too many pictures at the same time. Because of the fact that the description has to be changed before being able to press the button again, children used the tool in the right way without having to be constantly supervised by checking if they really changed the

prompt before generating a new picture.

Despite the fact that the testing session can be rated as successful, it is important to notice that the results can only be stated for the testing peer group, and therefore more testing sessions with a larger number of participants should be carried out in order to manifest the rightfulness of the results.

VI. CONCLUSION

In the scope of this paper, a language training tool was conceptualised, designed, and implemented. Furthermore, the approach to actively using this application with a target group that was selected based on scientific theories was tested in a real world testing environment.

The focus on providing an individual and autonomous language training approach seems to be successful, according to the testing results from the first user test group of twenty fifth-graders.

The concept behind the autonomous learning approach was to bring intrinsic motivation into school life supported by the use of state-of-the-art technology in form of picture-generating AI. Intrinsic motivation is essential for developing a school environment where pupils are performing better at key subjects, such as German. Self-determination as well as autonomy and the feeling of self-relatedness are human needs that should be fulfilled in order to achieve this desired intrinsic motivation.

Furthermore, it is important to note that even though the task should include as much freedom in the solving process as possible, pupils still need scaffolding and the right preparation and explanation for the task beforehand.

A. Benefits

The language training approach consisting of a suitable task and a software tool for a specific target group could engage learners to actively train their key competences in German. This is proven not only by the increased amounts of words per prompt when comparing the first to the last try but also by the diverse ways to get to a satisfactory solution as well as the different amounts of tries until the task was finished.

In addition to that, this approach can bring a change to German classrooms by bringing modern technological progress, such as picture generating AI, into the school environment to foster pupils' learning processes. This does not only make children more motivated as they have the feeling of being able to interact with something exciting and new in school, but it can also prepare them for the future by growing up with technology and learning to deal with it in a safe environment. Using AI in school for a good cause could also bring a change to the thinking of the educational sector, as AI is seen more as a threat than a tool to benefit from. Approaches like this language training tool are proof of the benefits that this technological era can bring to the educational sector when prepared the right way.

B. Challenges

There are some challenges that this approach is facing that should be targeted in order to improve it even further. The

testing scenario showed that some pupils, even though they were told to use full sentences, did not follow the instructions properly and came up with rather half than full sentences. Therefore, this issue should be targeted in further research. This could either be done by further development of a content filter that detects whether the written prompt is a full sentence or not, so that children are obliged to form proper sentences when describing pictures, or by making the required sentence structure even clearer when explaining the task to the children by eventually giving an example.

Furthermore, the number of test participants in individual tests can be improved. The test group of twenty participants was only to get an impression of whether the designed task and tool were successful in what they were aiming for. In order to really prove if what worked for the first test participants can be concluded for the generality, a larger testing group, consisting of pupils from different schools and school types, is required.

C. Further Research

In the scope of this paper, it was possible to create a new approach to language training, but there are several opportunities for further research on this specific topic.

For this, it is essential to perform more tests with a larger testing group. This way, it would also be possible to analyse if there is a correlation between the number of words per prompt and the original picture, which could show if more abstract scenarios trigger more detailed descriptions. This could be the case if pupils naturally think more about the scenario and how to put it into words than with more natural pictures originating from everyday life.

Besides increasing testing group size, the learner's learning effect could be measured by letting them do multiple rounds, either in the same session or on another day and seeing how their results and number of tries change over time.

Above that, further testing and analysis of not only the quantity of words used to describe the pictures but also the quality of the prompts could be very helpful for analysing if the approach is working properly. This should contain an analysis of the progression of writing style, grammar, and the vocabulary's level of difficulty, to evaluate whether the tool is helping to not only use more but also advanced words and sentence structures.

The mentioned further research questions concerning the correlation between the abstraction level of the picture as well as the progression of writing style, grammar, and vocabulary could be analysed in the future by using methods of natural language processing.

Furthermore, the tool as such could be extended by further features, such as another deep learning model to score the similarity between the original and generated picture. The level of accordance could not only be used as a metric for evaluating the learner's performance but also to analyse at what level of accordance learners are most likely to finish the task.

Generally, the awareness of interdisciplinary approaches like this and the knowledge of different areas of expertise that

go hand in hand in order to enhance the learning process are crucial for a brighter future for the educational sector. Therefore, there should be further research in not only specific fields but also in interdisciplinary software solutions and the way the human brain processes information. Further research could lead to less stress when using AI-related software and could benefit the research by understanding the way the human brain can connect knowledge better while being fostered by the right software solutions.

All in all, it can be said that the approach of a language training tool using a state-of-the-art picture-generating AI to improve the ability to use language to describe scenarios was successful for the tested group of fifth graders. However, this approach should only be a first step towards future classrooms where education and technology go hand in hand while considering the effects that learning tools have on the human brain in an interdisciplinary way.

REFERENCES

- [1] Petra Stanat, Stefan Schipolowski, Rebecca Schneider, Karoline A. Sachse, Karoline A., Sebastian Weirich and Sofie Henschel, "IQB-Bildungstrend 2021," Münster: Waxmann Verlag GmbH, 2022.
- [2] NCES, "2017 NAEP Mathematics and Reading Assessments: Highlighted Results at Grades 4 and 8 for the Nation, States, and Districts," NCES, 2018.
- [3] Kayla Swan, "Gaining Perspective: Social Media's Impact on Adolescent Literacy Development," State University of New York, 2017.
- [4] Serhat Kurt, "Technology use in elementary education in Turkey: A case study," *New Horizons in Education*, vol.58, no.1, pp. 67–77, 2010.
- [5] Rod Ellis, "Second Language Acquisition (Oxford Introduction to Language Study)," Oxford University Press ELT, 1997.
- [6] R.C. Gardner, R.N. Lalonde, R. Moorcroft, "Social Psychology and Second Language Learning: The Role of Attitudes and Motivation," *Michigan: Language Learning*, vol. 35, issue 2, pp. 207–227, 1985.
- [7] Herbert P. Ginsburg, Silvia Oppen, "Piaget's Theory of Intellectual Development," Prentice-Hall, Inc, pp. 100–204, 1988.
- [8] Edward L. Deci, Richard M. Ryan, "Self-determination theory," *American Psychologist Association Inc.*, 2000.
- [9] Matthew T. McCrudden and David N. Rapp, "How Visual Displays Affect Cognitive Processing," *New York: Educational Psychology Review*, 2017.
- [10] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang and Dimitris N. Metaxas, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," *IEEE*, 2018.
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Ithaca: Cornell University, 2021.
- [12] Jonathan Ho, Ajay Jain and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

"Senior Consultant ChatGPT" - a Model of Collaboration Between Generative AI and Consultants

Friedrich Augenstein
Department Business Management–
Service Sector Management
Baden-Wuerttemberg Cooperative State
University
Stuttgart, Germany
friedrich.augenstein@dhbw-stuttgart.de

Abstract—Current studies show that knowledge-intensive and knowledge-providing professions in particular - the “professional services” - are strongly affected by generative artificial intelligence (AI) and that their activities can be replaced at least in part by this new technology. Using the example of consultants as representatives of “professional services”, this article shows similarities and differences between consultants and generative AI, shows where the respective actor has advantages or disadvantages and develops a model based on typical tasks in the consulting environment, which shows a possible collaboration of both actors. The basis of the developed model is the combination of the Stacey matrix and the Cynefin model as it is already used in a decision-making process for the question of whether a project should take place in a classical or agile form. This model is adapted to the present topic. The transferability to “professional services” other than consultants and the need for further research are discussed.

Keywords—Generative AI, consulting, professional services, collaboration, Stacey matrix, Cynefin model

I. GENERATIVE AI CHANGES THE WORKING WORLD OF “PROFESSIONALS”

A. 300 Million High-skilled Jobs Impacted by Generative AI

In March and April 2023, two studies made all “professionals” sit up and take notice: A study by Open AI Research and the University of Pennsylvania confirmed that knowledge-intensive and knowledge-transferring professions in particular are highly replaceable by generative artificial intelligence (AI).

The article investigates the potential implications of large language models (LLM), such as Generative Pre-trained Transformers (GPTs) - as one important sub-concept of generative AI -, on the US labor market. The study found that around 80% of the US workforce could have at least 10% of their work tasks affected by the introduction of GPTs, while around 19% of workers may see at least 50% of their tasks impacted. The influence spans all wage levels, with higher-income jobs potentially facing greater exposure. And even further the study concludes: “Accounting for other generative models and complementary technologies, our human estimates indicate that up to 49% of workers could have half or more of their tasks exposed to LLMs.”

The top industries exposed to advances in language modeling are legal services and securities, commodities, and investments. And well-paid professions are particularly affected. For example, about 80% of Survey Researcher jobs,

about 70% of Public Relations Specialist jobs, and 100% of Tax Preparer, Financial Quantitative Analyst, and Web and Digital Interface Designer jobs are exposed to LLMs. While generative AI offers high potential to create new occupations, according to the study, occupations that tend to be generalist and lack concrete job-focused skills appear to be particularly at high risk [1].

Another study by Goldman Sachs Group Inc. concludes in an analysis of the labor market in the U.S. and Europe and here of around 900 occupations that around 20% of all current tasks in companies can be taken over by AI-driven automation within the next 10 years. Accordingly, 63% of all occupations are affected, and around 300 million jobs in the USA, Europe and Asia could be completely or partially eliminated. At the same time, however, global economic output would increase by around seven percent through the use of AI [2].

B. Consultants and Other “Professional Services” Are Also Affected

These studies give reason to consider the future of “professional services” such as consultants, legal and tax advisors, financial and marketing consultants and others. The study mentioned in [1] lists typical consulting professions such as Survey Researchers, Public Relations Specialists, Tax Preparers, Financial Quantitative Analysts and Web and Digital Interface Designers and “professional services” industries such as Legal Services and Investments as particularly threatened by generative AI. Some newspapers have already conducted trials with ChatGPT and other generative AI, demonstrating that generative AI can perform typical support and consulting tasks. The “New York Times” let generative AI do classical assistance tasks like scheduling meetings, planning business trips or summarizing meetings and was positively surprised by the task completion and additional ideas raised [3]. The newspaper “Welt” set the test task “How do I set up a new, time-efficient booking process for recurring business trip accounting in SAP S4 HANA?” and received detailed instructions on how to set up the process [4].

Thus, these studies and initial experiments suggest that generative AI can indeed perform classic consulting tasks. Therefore, the tasks and capabilities of a consultant are subsequently compared to the application domains and capabilities of generative AI. The consultants are representative of other “professional services” and the insights gained for the consultants are transferable to them, as the following explanations will make clear.

II. CONSULTANTS AND GENERATIVE AI - A COMPARISON

A. Management Consultants and Their Task Areas and Capabilities

The consulting industry is an important economic factor. In Germany alone, the 220,000 employees in the sector generated sales of almost 44 billion euros in 2022. Around 25% of this is generated with strategy consulting (corporate strategy, business development & innovation, marketing and sales strategy, corporate finance, etc.), 44% with organizational consulting (project management, process optimization and performance management, change management, etc.), 22% with IT consulting (IT applications & infrastructure, IT governance & compliance, IT data protection and data security, etc.) and 9% with human resources consulting (management diagnostics and development, HR strategy, etc.) [5]. Consultants often follow an ideal-typical process in their consulting projects [6]:

- Acquisition phase with the subphases Contact & Information and Order & Contract Drafting and the main tasks
 - Information, orientation, research
 - Contact and acquisition meetings
 - Identification of problem area
 - Offer and contract drafting
- Analysis phase with the subphases as-is analysis and target formulation and the main tasks
 - Planning
 - Information gathering and deepening
 - Formulation of objectives
 - Selection of alternative solutions
- Problem solving phase with the subphases target concept and realization planning and the main tasks
 - Concept development
 - Development, discussion, evaluation of problem solution alternatives
 - Development of action plan
 - Decision according to implementation conditions
- Implementation phase with the subphases realization/implementation and evaluation/control and the main tasks
 - Implementation planning
 - Implementation realization
 - Practical testing
 - Optimization
 - Introduction
 - Success control / satisfaction check

Consultants often use predefined, phase-oriented, in science and practice established and standardized approaches to develop solutions for consulting clients, e.g., in project

management, strategic management, IT management and others [6].

The indeterminacy on the other hand is an essential characteristic of the consulting process [6]:

- Indeterminacy of the input: Often, information that is important for the course of the project is not yet known or available at the beginning of the project or is withheld - on both the consultant's and the client's side.
- Indeterminacy of the transformation process: The consultant often has to respond flexibly to changing requirements of the consulting client, often there are also imponderables in the cooperation between client and consultant team, environmental influences as well as possible knowledge gains can change the course of the project.
- The indeterminacy of the output is in turn a consequence of the indeterminacy of the input and the flexible transformation process, i.e. the output cannot be planned exactly ex ante either if the input and transformation process are indeterminate.

Management consultants exhibit characteristics that enable them to solve complex business problems of their consulting clients. From classical literature [6][7] as well as relevant publications from consulting practice [8][9] and in journals [10], the following essential characteristics can be derived:

- Analytical Thinking:
 - Analyze complex business problems and develop solutions based on the knowledge of the subject area in focus.
 - Process large amounts of data and identify trends and patterns
- Creativity:
 - Develop innovative solutions
 - Find creative approaches to solve problems and identify new business opportunities
- Communication Skills:
 - Communicate complex ideas and concepts clearly and understandably
 - Convince customers and build trust
- Adaptability:
 - Adapt quickly to new environments and working conditions
 - Be flexible and able to respond to change

According to Dueck [11], the consultant's skills are not only about technical knowledge, but professional intelligence also includes social skills, creativity and flexibility. Dueck emphasizes the importance of lifelong learning and the ability to adapt to a constantly changing work environment in order to be successful. Ultimately, the consultant benefits from the characteristics of principal-agent theory [6]. Principal-agent theory is concerned with the relationships between a principal and an agent, where the agent acts on behalf of the principal. The theory examines the incentive structures, moral hazards, and information asymmetry that can occur in such

relationships. Information asymmetry benefits the consultant because he has a knowledge advantage over the principal or has more comprehensive information than the principal. This is ultimately the motivation for the client to engage the consultant to solve a problem in the client's company.

B. Generative AI and its Characteristics and Capabilities

Generative AI refers to a type of AI systems that are able to create new content by recognizing patterns in existing data with which they have been trained and using these patterns to generate new information. Important features of generative AI in this regard are [12][13]:

- Ability to learn: Generative AI models learn from unstructured data and are able to recognize patterns in data without requiring any manual pre-work.
- Ability to generate new data: Generative AI models are able to generate new data that matches the characteristics of the training data. This makes it possible to generate new data never seen before.
- Modeling complexity: Generative AI models can learn complex, non-linear relationships between input and output data, enabling them to generate data with high levels of abstraction and with high diversity.
- Application to a wide variety of data types: Generative AI models can be applied to a variety of data types, such as images, music, and text. This makes them a versatile tool in many application domains.

The capabilities of generative AI are diverse and include, in particular, text generation: Generative AI can generate texts based on inputs and patterns it has learned from a training set of texts. This capability is used to automate tasks such as generating product descriptions, translations, and summaries [12][14] and is also important in the present case, where the goal is to replace expert knowledge.

Image and video generation and speech generation, also important capabilities of generative AI, will not be discussed further here.

One important capability of generative AI is its use via so-called "prompts." A "prompt" in the context of AI and machine learning is a text or instruction given to an AI model - in the form of a chatbot - to generate a specific task or a specific type of response. This is done - depending on the task - in a fraction of the time it would have taken a human to solve it [15]. This is of interest here, as the question the generative AI is asked has to fit perfectly to the respective topic. This is important for the AI for generating the adequate results.

Generative AI thus has the potential to accelerate and optimize creative processes by automatically generating content that is normally created by humans. However, the technology also raises challenges related to the authenticity and integrity of the generated content, as well as ethical and legal issues related to the creation and use of generated content as discussed for example in [16].

A well-known phenomenon is the "hallucination" of false information generated by generative AI [17]. This can lead to incorrect analysis, recommendations, and concepts when consulting results are generated by generative AI. In addition, it should be noted that the more specific the questions asked

of the chatbots, the more generic the answers, and the more commonly used models, concepts, and solution paths used to develop a solution, the more valuable the results of generative AI will be.

C. Generative AI for the Analysis of Business Problems

Problem analysis is an important step in the application of generative AI, as it can help identify the underlying challenges and constraints of a problem. Thorough problem analysis can help select the right data sources and models to produce high-quality results [18].

Generative AI can be used to analyze business problems that a consultant would otherwise perform. The following examples illustrate this:

- Text analysis and sentiment analysis: Generative AI models can be used to analyze large amounts of business data such as customer reviews, social media, or financial reports. These models can help extract important information, identify trends, and assess sentiments (attitudes, moods) [19].
- Prediction and forecasting: Generative AI can be used in business analysis to make predictions and forecasts. For example, time series models can be used to generate sales forecasts or demand forecasts [20].
- Scenario generation: Generative AI models can be used to generate alternative scenarios and simulations that can help in business decision making. These scenarios can assist in risk assessment, strategic planning, and evaluation of business options [21].

Generative AI is also already capable of solving classical business management problems, e.g., automated generation of business plans. An AI software *LeanStartupAgent* developed by WHU Otto Beisheim School of Management "allows for the automation of tasks that are crucial to the creation of innovative business models and the founding of a company. The program is capable of identifying a business model's innovative potential, as well as quickly executing a preliminary market analysis that is both comprehensive and easy to understand" [22].

Another application of generative AI in business administration is the automated generation of texts capable of describing complex business strategies [23]. Here, it is demonstrated how natural language processing (NLP) techniques can be used to analyze unstructured textual data in the field of strategic management research by using mandatory publications on the business activities of companies in the United States (so-called 10-K reports) to create text-based measures of core constructs in strategy, such as strategic change, positioning, and focus.

D. Comparison Generative AI vs. Consultant

If the human consultant is replaceable by generative AI, advantages for consulting clients are obvious:

- Efficiency: Generative AI is time-saving compared to traditional consulting services provided by management consultants, as generated output is often available in a matter of seconds after a question is entered. Even a series of increasingly detailed inputs can be done in a matter of minutes. Obviously, no weeks- or even months-long consulting projects would be necessary.

- **Cost:** Generative AI is less expensive compared to traditional consulting services provided by management consultants. Only the monetized working time of the consulting client's employees that would be required for the dialog with generative AI would have to be applied, as well as any costs incurred for the use of generative AI. Since the required working time is by far the largest cost item and, as mentioned in the first point, this is only a fraction of the time required compared to a classic consulting project, the costs would also only be a fraction of the costs of a classic consulting project.
- **Accuracy:** Generative AI offers greater precision in finding solutions. Not only the wealth of knowledge of a consulting team, but equally the wealth of knowledge of generative AI - in the case of ChatGPT of many millions of data sets, the content of thousands of websites, books, etc. - can be used [24].

It is therefore worth investigating in which of the above-mentioned tasks of a consultant generative AI can support or even completely take over.

- In the acquisition phase, generative AI can be used in particular in the gathering of information, orientation and searches. The strengths here lie in the processing of freely accessible and rather general information, e.g. information on a specific industry. The more specific the information to be obtained must be (e.g., company-specific information), the less suitable generative AI is. This is consistent with the statement made earlier that especially occupational activities that tend to exhibit generalist skills with little specificity to the job at hand seem to be strongly substitutable by generative AI. The contact and acquisition discussions are more likely to be covered by the human consultant if this is accompanied by the identification of the problem area. The vagueness of the input to the consulting process is a factor in favor of using a human consultant - the precise identification of the problem domain is an essential task at the beginning of the consulting process, which is covered by the communicative and analytical skills of a human consultant. The better described and the more general the problem, the more sensible it is to use generative AI in the consulting process. The drafting of offers and contracts, however, is more likely to be reserved for pre-formulated text modules and the results of personal discussions between the consultant and the consulting customer.
- In the analysis phase, planning is left to the human consultant when the indeterminacy of the transformation process takes hold. The clearer the ideas of the consulting client and the consultant are on how to solve the consulting client's operational problem, the more reasonable the use of generative AI seems to be. The same applies to the other points of information gathering and consolidation, goal formulation and the selection of alternative solutions. The fact that generative AI is very well capable of analyzing business management problems has already been demonstrated above.
- In the problem-solving phase, generative AI can be used well for the tasks of concept development and the development, discussion and evaluation of problem-

solving alternatives, if these do not involve very specific issues of the consulting client's company. The fact that generative AI is capable of developing concepts and alternative solutions for rather general problems has been demonstrated above. The development of an action plan and the decision according to implementation conditions is again largely reserved for the human consultant in close coordination with the consulting customer, since specific framework conditions of the consulting customer - e.g., concrete availability of individual key resources in the company - are usually considered here. However, general procedures for implementing a problem solution can be provided by generative AI.

- The same applies to the implementation phase. Here, too, the concrete operational conditions in the consulting customer's company are decisive for the specific design of this phase. However, generative AI can also provide support here, e.g., by automatically generating software program code, websites, etc.

E. (Apparent) Disadvantages of Generative AI Compared to Human Consultants

An apparent disadvantage of generative AI is the lack of human interaction, the - apparent - inability of generative AI to recognize emotions and moods in the consulting customer and to react appropriately. But even here, AI-powered tools like IBM's "Tone Analyzer" can already do practical things. The "Tone Analyzer" can detect seven tones: sad, frustrated, satisfied, excited, polite, impolite, and sympathetic. Additionally, it captures three relevant tones: cheerfulness, negative emotions, and anger. The service can also detect three types of tones from communications: emotion, social, and language. The Tone Analyzer can be used to understand how written communications are perceived and to improve the tone of communications" [25].

Another apparent drawback is the lack of empathy of the generative AI towards the human consultant. However, AI seems to have caught up in this area as well. For example, in the medical environment, patients have been advised partly by doctors and partly by AI-assisted chatbots. Here, AI was perceived to be significantly better in quality but also more empathetic than (human) doctors [26].

The flexibility may also be lacking in generative AI to respond to rapidly changing conditions or novel challenges among consulting clients. After all, generative AI must be trained by past data, which is precisely not yet available in the case of novel frameworks. But again, the speed at which new data is being integrated into generative AI seems to be very rapid. For instance, after ChatGPT was made accessible to the public on 11/30/2022, it was already expanding the topics and improving the correctness of the statements and the mathematical capabilities on 12/15/2022, 01/09, 01/30, 02/09, 02/13, 03/14, 03/23, 05/03/2023 [27].

Fully grasping complexity seems to be another drawback of generative AI. If complexity is understood as the variety of possible behaviors of interdependent elements and the variability of the trajectories of effects [28], generative AI based on statistical correlations in the trained text data may therefore sometimes give incorrect or inaccurate answers. In particular, for complex or specialized domains, generative AI may have difficulty providing accurate and reliable information.

F. Advantages of the human consultant over generative AI.

The human consultant thus seems to have an advantage over the generative AI when

- the problem is not clearly defined, but must be specified in more detail together with the consulting client,
- the path to solving the problem is not yet clear, but must first be identified in discussions and analyses together with the consulting client - however, generative AI can certainly support the consultant in this,
- it is not a matter of standardized procedures, but rather of dealing with complex or initially even completely unstructured ("chaotic") issues,
- very specific circumstances and basic conditions of the consulting customer must be included in the consideration of the problem definition as well as the solution path.

Thus, a model of "division of labor" between human consultant and generative AI emerges, which is spanned by two axes:

- the clarity of the goals,
- the clarity of the solution path, resulting from the individuality of the framework conditions at the consulting client.

This in turn is reminiscent of a model, which is used in the decision-making process whether a project should be carried out in a classical manner (waterfall model, etc.) or in an agile manner: The Stacey matrix combined with the Cynefin model.

III. A MODEL OF COLLABORATION BETWEEN CONSULTANTS AND GENERATIVE AI.

A. The Stacey Matrix / Cynefin Model

This combination of two models is based on the Stacey Matrix. The Stacey Matrix goes back to the British management professor Ralph Douglas Stacey (*1942), who deals with organizational theory and complex systems. The term "matrix" is not precise here; it is more of a coordinate system. On the y-axis is plotted how unclear the project goal is described. On the x-axis is plotted the ambiguity of the solution approach, i.e. how unknown is the way to achieve the project goal. Along the bisector, the character of the project is then plotted from simple (clear goal and clear solution approach) to complicated and complex to chaotic (both goal and solution approach are unknown) [29].

The Cynefin model divides problems into four different areas depending on the type of problem solving required [30]:

- Simple (Simple): In this domain, the cause-effect relationships are clear and unambiguous. There are established best practices and proven methods for problem solving. In this case, the following action is recommended: observe, classify, deduce, react.
- Complicated (Complicated): In this domain, the cause-effect relationships are also clear, but the solution requires expert knowledge or analysis. There may be multiple possible solutions, and selecting the best

option requires expert knowledge. Then the recommended approach is: Observe, Analyze, React.

- Complex (complex): In this domain, cause-effect relationships are not clear. There are emergent patterns and interactions that are difficult to predict. Solutions must be developed through experimentation, adaptation, and learning. An iterative approach is recommended: Try, observe, react, try again, observe, react.
- Chaotic (chaotic): In this domain, there are no clear cause-and-effect relationships; the situation is chaotic and unpredictable. Immediate action is required to stabilize the situation and allow a return to order. The task is to act and react until there is clarity and the chaotic decision-making situation has become first a complex and then possibly a complicated situation.

These two approaches are then combined to form a decision model for agile versus classic project management, e.g. by [31].

B. Division of Labor and Collaboration of Human Consultant with Generative AI Based on the Model.

This decision-making model can now be used for the division of labor between human consultant and generative AI in the context of a consulting project. The following figure illustrates the decision model.

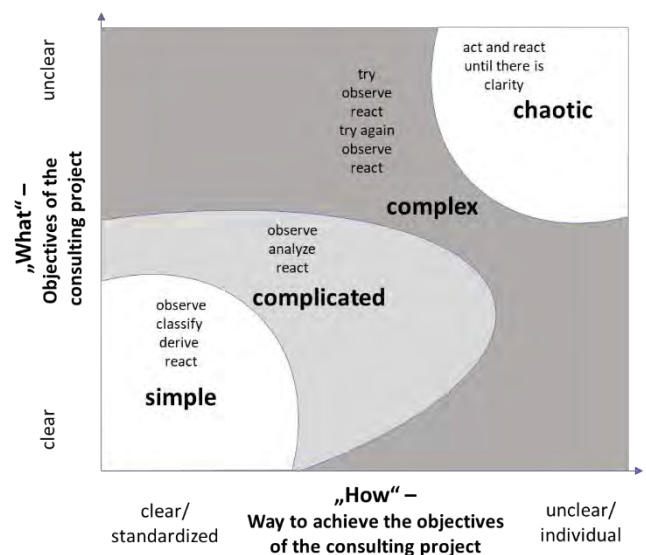


Fig. 1. The collaboration model of human consultant with generative AI.

The Y-axis indicates whether the goal of the consulting project is clear or unclear, i.e., must first be further specified - especially through an (intensive) dialog between the human consultant and the consulting client.

The X-axis indicates whether the way to achieve the goal is clear/standardized or unclear/individual due to the specific framework conditions at the consulting client.

Here too, as in the original model, the areas and options for action then arise

- simple - observe, classify, derive, react
- complicated - observe, analyze, react
- complex - try, observe, react, try again, observe, react

- chaotic - act and react until clarity prevails

Thereby - derived from the comparisons of the human consultant with the generative AI - a use of generative AI can replace the human consultant over long distances, especially for **simple** topics. These are generative consulting services with clearly defined and standardized solution paths, which can be found, for example, in classic strategy consulting - corporate strategy, marketing and sales strategy, corporate finance or CSR consulting. All these consulting fields are characterized by comprehensively documented and field-tested procedure models and a very comprehensive analysis and evaluation of company-external data.

For example, in classic strategy consulting, given some basic information about the consulting client's industry and some basic company-specific resources, generative AI can perform a Porter 5 Forces analysis, a VRIO framework analysis, and combine them into a SWOT analysis that derives strategic actions for each area of the SWOT matrix.

But IT consulting is also affected here, since here the problems must be clearly defined in order to be able to implement them in IT, and the way of solving the problem is also usually clearly described, namely implementation in a programming language with clearly defined rules and procedures.

For example, when implementing a business process using Robotic Process Automation (RPA), a basic description of the problem and process is sufficient to generate detailed programming instructions for implementing the process using an RPA tool such as UiPath.

For **complicated** issues, generative AI can relieve the human consultant over long distances. This is the case in organizational consulting, where, for example, standardized procedures are used in project management, change management, controlling or business process optimization, but extensive internal company data of the consulting client must be analyzed. AI is already being used here in a supporting role, e.g., in process mining.

But even here, standardized, phased approaches can be used by generative AI. For example, in change management, GPTs can develop a communication plan based on project-specific information from a stakeholder analysis.

For **complex** issues, the human consultant must take the lead. These are consulting projects where the objective of the consulting project requires extensive communication with the consulting client and the framework conditions are also very company-specific. In the intensive communication between the human consultant and the consulting client that is necessary here, the generative AI can be consulted again and again, e.g., to show benchmarks and best practices for a discussed topic within the consulting project.

For **chaotic** issues, the human consultant is exclusively responsible. Here, it is necessary to transform the project situation into a complex or even complicated situation together with the consulting client, where generative AI can then also provide support again.

In summary, it can be said that generative AI will not replace human consultants and is not the "better" consultant. However, typical consulting tasks, especially in the above-mentioned "simple" and "complicated" fields, can be strongly supported over long distances and the productivity of the

consultants can be massively increased by generative AI. The GPTs are well suited as a source of inspiration, sparring partners and for reflecting on one's own position. Data protection must be viewed critically when sharing company-specific content. And finally, the consulting customers will expect the increase in productivity of the consultants to be passed on in the form of lower fees or additional services - such as an even more individual approach to the customer or stronger support during implementation.

IV. OUTLOOK AND FURTHER RESEARCH

For the different areas of the model, fictitious problems can now be tested first, and later also concrete problems from consulting projects. In doing so, the performance of available generative AI such as ChatGPT, Bard and others can be better assessed. If consultants accompany these trials, the results produced by the generative AI can be compared with alternative results produced by human consultants. In addition, consultants can test and evaluate the degree of support provided by generative AI in each area of the model. Thus, there is a further refinement of the above model of division of labor between consultants and generative AI, as well as potential for improvement of generative AI.

A critical issue to be considered here is the data protection for the data provided to the generative AI in the context of the experiments. Anonymization of the data is recommended.

In principle, the findings developed here for the consulting industry can also be applied to other "professional services". Lawyers, tax consultants, marketing agencies or engineering service providers also work in a consulting context and often in a project-based environment with their clients. And here, too, generative AI can replace or at least support many activities, as the studies mentioned at the beginning of this article demonstrate in the naming of the professions and industries concerned.

REFERENCES

- [1] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," OpenAI OpenResearch, University of Pennsylvania, Working Paper, Mar. 2023.
- [2] Goldman Sachs, "Generative AI could raise global GDP by 7%," Goldman Sachs Insights, Apr. 5, 2023. [Online]. Available: <https://www.goldmansachs.com/insights/pages/generative-ai-could-raise-global-gdp-by-7-percent.html>. [Accessed: May 3, 2023].
- [3] B. Chen, "How ChatGPT and Bard Performed as My Executive Assistants," New York Times, Mar. 29, 2023. [Online]. Available: <https://www.nytimes.com/2023/03/29/technology/personaltech/ai-chatgpt-google-bard-assistant.html>. [Accessed: May 3, 2023].
- [4] B. Fuest, "300 Millionen Stellen betroffen- für diese Jobs wird KI zur ernsthaften Gefahr, Die Welt, Apr. 5, 2023. [Online]. Available: <https://www.welt.de/wirtschaft/plus244619752/ChatGPT-und-Co-300-Millionen-Stellen-betroffen-fuer-diese-Jobs-wird-KI-zur-ernsthaften-Gefahr.html>. [Accessed: May 3, 2023].
- [5] Bundesverband Deutscher Unternehmensberatungen (BDU) e.V., "Facts & Figures zum Consultingmarkt 2023," BDU-Marktstudie, March 2023.
- [6] D. Lippold, Die Unternehmensberatung. Wiesbaden: Springer Gabler, 3rd ed., 2018.
- [7] C. Haas, Managementberatung in einer integrativen Betrachtung. Wiesbaden: Springer Gabler, 2017.
- [8] J. Birt, "8 Key Consulting Skills Valued by Employers and Clients", Nov. 9, 2022, [Online]. Available: <https://www.indeed.com/career-advice/career-development/consulting-skills>. [Accessed: May 23, 2023].
- [9] Paperbell, "36 Core Consulting Skills to Set You Apart from Your Competition," Feb. 3, 2022. [Online]. Available:

- <https://paperbell.com/blog/consulting-skills/>. [Accessed: May 23, 2023].
- [10] M. Banai and P. Tulimieri, "Knowledge, skills and personality of the effective business consultant," *Journal of Management Development*, vol. 32, no. 8, pp. 886-900, 2013.
 - [11] G. Dueck, *Professionelle Intelligenz: Worauf es morgen ankommt*, Frankfurt am Main: Campus Verlag, 2020.
 - [12] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
 - [13] D. Forster, *Generative Deep Learning*, 2nd ed., Sebastopol, CA: O'Reilly Media, 2023.
 - [14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, May 2020.
 - [15] S. Shahriar and K. Hayawi, "Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations," *arXiv:2302.13817*, March 2023.
 - [16] Not a generative AI-generated Editorial, *Nat Cancer*, vol. 4, pp. 151-152, Jan. 2023.
 - [17] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Jin Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation", *ACM Comput. Surv.* 55, 12, Article 248, pp. 1-38, December 2023.
 - [18] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Pretrained Language Models for Text Generation: A Survey," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, Montreal, Canada, 2021, pp. 4492-4499.
 - [19] J. Wiles, "Beyond ChatGPT: The Future of Generative AI for Enterprises," *Gartner*, Jan. 26, 2023. [Online]. Available: <https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises>. [Accessed: May 23, 2023].
 - [20] A. Annor-Antwi and A. A. M. Al-Dherasi, "Application of Artificial Intelligence in Forecasting: A Systematic Review," *SSRN Electronic Journal*, Jan. 2019. [Online]. Available: https://www.researchgate.net/publication/330748214_Application_of_Artificial_Intelligence_in_Forecasting_A_Systematic_Review. [Accessed: May 31, 2023].
 - [21] S. Sohrabi, A. V. Riabov, M. Katz, and O. Udrea, "An AI Planning Solution to Scenario Generation for Enterprise Risk Management," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 160-167.
 - [22] WHU Otto-Beisheim School of Management, "Generating Business Models with Generative AI," May 15, 2023. [Online]. Available: <https://www.whu.edu/en/magazin/chair-of-entrepreneurship-innovation-and-technology/generating-business-models-with-generative-ai/>. [Accessed: May 31, 2023].
 - [23] A. Menon, J. Choi, H. Tabakovic, "What You Say Your Strategy Is and Why It Matters: Natural Language Processing of Unstructured Text," in *Academy of Management Proceedings*, 2018, pp. n.pag.
 - [24] Gptblogs.com, "ChatGPT: How Much Data Is Used in the Training Process?," Feb. 9, 2023. [Online]. Available: https://gptblogs.com/chatgpt-how-much-data-is-used-in-the-training-process?utm_content=cmp-true. [Accessed: May 13, 2023].
 - [25] IBM Tone Analyzer, Available: https://cloud.ibm.com/docs/natural-language-understanding?topic=natural-language-understanding-tone_analytics. [Accessed: May 14, 2023].
 - [26] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, and A. M. Fine, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA Intern Med*, Apr. 28, 2023. [Online]. Available: <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2782935>. [Accessed: May 23, 2023].
 - [27] Help.OpenAI, "ChatGPT Release Notes," [help.openai.com](https://help.openai.com/en/articles/5828210-chatgpt-release-notes), retrieved May 15, 2023. [Online]. Available: <https://help.openai.com/en/articles/5828210-chatgpt-release-notes>. [Accessed: May 23, 2023].
 - [28] Gabler Wirtschaftslexikon, "Komplexität," <https://wirtschaftslexikon.gabler.de/definition/komplexitaet-39259>, Accessed May 15, 2023.
 - [29] R. D. Stacey, "Complexity and management: A collection of essays," Routledge, 1996.
 - [30] D. J. Snowden and M. E. Boone, "A leader's framework for decision making," *Harvard Business Review*, vol. 85, no. 11, Nov. 2007, pp. 68-76.
 - [31] K. Wenzel, *Management Models of Digital Transformation*. Wiesbaden: Springer Gabler, 2021.

Poster Section

The following section presents extended abstracts of the presented posters in the framework of the AI Transfer Congress

Enhancing Quality Control through Computer Vision: A Comprehensive Study

Shobhit Agarwal*, Rami Mochaourab[†] and Bozena Lamek-Creutz*

Duale Hochschule Baden-Württemberg Mannheim, Mannheim, Germany*

RISE Research Institutes of Sweden, Stockholm, Sweden [†]

Email: * shobhit.agarwal@dhbw-mannheim.de, * bozena.lamek-creutz@dhbw-mannheim.de, [†] rami.mochaourab@ri.se

Abstract—Computer vision has emerged as a transformative technology with numerous applications in various industries. One of its key applications is quality control, which plays a vital role in ensuring the reliability and consistency of products and services. This paper provides a comprehensive overview of the intersection between computer vision and quality control. It explores the fundamental concepts of computer vision, delves into the applications of quality control across different industries, and highlights the various computer vision techniques employed in quality control processes. The paper discusses challenges associated with implementing computer vision systems for quality control and identifies future research directions. Through a case study in the manufacturing industry, successful implementations of computer vision in quality control are showcased. The potential impact of computer vision on quality control, including increased efficiency, improved accuracy, reduced costs, and enhanced product quality, is emphasized. The paper concludes by highlighting the significance of computer vision in revolutionizing quality control and suggesting areas for further research, such as algorithm development, real-time processing challenges, and ethical considerations. This paper aims to provide valuable insights into the advancements and potential of computer vision in enhancing quality control processes across industries.

Index Terms—Machine learning, Deep learning, Computer vision, Quality control, Automation

I. INTRODUCTION

Real-world manufacturing processes are complex and fragile. They are designed to address specific issues and be efficient, but their complexity means that the probability of introducing a defect is never zero and as the product's and process's complexities increase, the importance of dependable quality grows as well. For example in fabric production, defects such as knots, broken pick, and broken yarn can appear [2], in the process of welding aluminum parts defects such as burn-through, contamination and lack of penetration are common [3] and in production magnetic tiles defects such as cracks, blowholes are common [1].

These defects are overhead for the manufacturing business as they can add to the production costs and have severe economic consequences. Also, they can affect the usability of the final product if they go undetected. Defects also contribute to the additional wastage of resources and safety hazards. Therefore, maintaining the quality of the manufactured product is one of the crucial challenges in the manufacturing industry. Quality control encompasses the processes and techniques employed to detect defects, measure dimensions, and assess the overall quality of items produced or services rendered. It

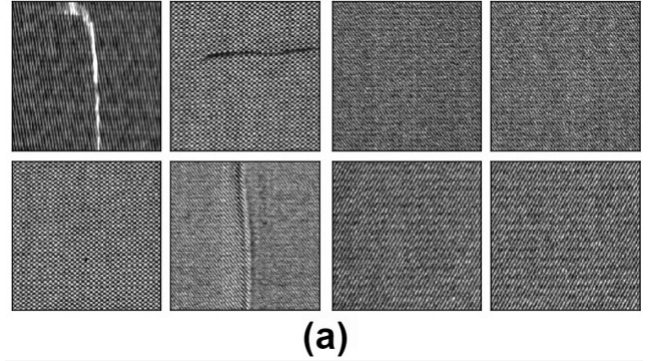


Fig. 1. Exemplary raw images of the dataset AITEX (where the four images on the left show the defected fabric and the images on the right side show the non defected fabric) [2], [31]

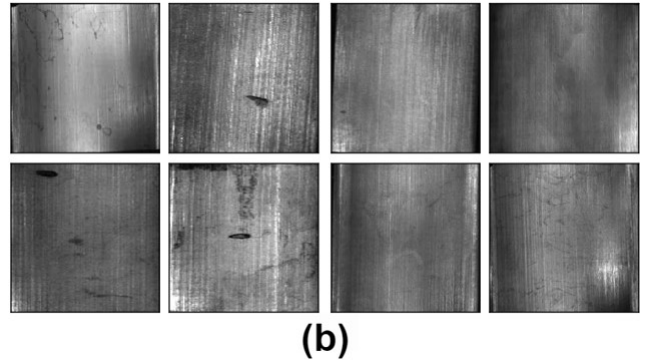


Fig. 2. Exemplary raw images of the dataset MagTile (where the four images on the left show the defected magnetic tile and the images on the right side show the non defected magnetic tile) [1], [31]

is a critical process in any production environment with the end goal to ensure that products meet the required standards of quality. However, the fast-paced and complex nature of these processes makes quality control difficult. The traditional Quality control and maintenance methods still require manual intervention. Given the massive scale at which industrial manufacturing occurs, manual Quality control becomes inefficient and expensive. Manual inspection is also more prone to errors. The difficulties faced in manual Quality control and the lack of a functional, efficient, and economical solution to Quality control has motivated industries to move towards automated maintenance. There have been various solutions

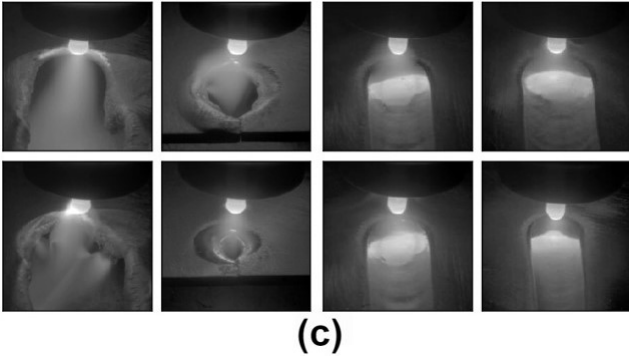


Fig. 3. Exemplary raw images of the dataset TIG5083 (where the four images on the left show the defected welding attempts and the images on the right side show the non defected welding attempts) [3], [31]

presented for this problem in various industries for example in the manufacturing industry the authors in [4] use machine learning to achieve near-perfect defect classification and in the printing industry, the authors from [6] achieved automated classification accuracy rate of 98.4%.

Among all the current solutions presented for quality control, a common solution is an automated quality control system through deep learning-based computer vision. In recent years, computer vision has emerged as a transformative technology with numerous applications across various industries from self-driving cars [7] to agriculture [8]. Computer vision is a rapidly developing technology and has the potential to revolutionize quality control. Computer vision is an interdisciplinary field between computer science, statistics, and digital image data. Traditionally computer vision has relied on statistical approaches to deal with visual data, i.e. image data. However, with recent development in the field of deep learning, a new branch of computer vision has emerged with a focus on deep learning-based computer vision. Deep learning is a subfield within the large scope of machine learning. In the context of industrial quality control, computer vision can achieve goals through cameras and precise algorithms that can identify defective products. It can automate many of the tasks that are currently performed manually by human inspectors. This can lead to significant improvements in efficiency, accuracy, and cost-effectiveness.

Furthermore, deep learning-enhanced computer vision algorithms are more powerful than rule-based algorithms because they can generalize to new data and do not require as much domain expertise. This means that they can be used to create visual quality gates that were previously not possible with rule-based algorithms [9]. For example, a deep learning-enhanced computer vision algorithm could be used to create a quality gate that automatically rejects images that are blurry or out of focus. This would not be possible with a rule-based algorithm because it would be difficult to write rules that could accurately identify all blurry or out-of-focus images.

Data-driven approaches are more precise than rule-based techniques, but they are also more dependent on the quality

and quantity of data [10]. If the data is not of high quality or if there is not enough data, the data-driven approach will not be as precise. Additionally, data-driven approaches can be affected by data drift, which is the change in the distribution of data over time. This can lead to the data-driven approach becoming less precise over time. However, deep learning-enhanced computer vision algorithms are still under development and have the potential to revolutionize the way quality control is performed.

To analyze the various aspects of computer vision and its impact on quality control in manufacturing and industrial processes, this paper introduces first the fundamentals and research objective in section II and III. The real world case studies are presented in section IV and the practical challenges and future directions are discussed in section V and the paper is concluded in section VI.

A. Research objectives

- To review the current state of the art in computer vision for quality control. This includes discussing the different types of computer vision techniques that can be used for quality control.
- To identify the challenges and limitations of using computer vision for quality control. This includes discussing the challenges of obtaining accurate and reliable data, as well as the challenges of developing and deploying computer vision systems in real-world applications.
- To propose new research directions in computer vision for quality control. This includes discussing the potential of using new technologies, such as Augmented reality, to improve the accuracy and reliability of computer vision systems for quality control.

II. FUNDAMENTALS

This section will provide an overview of the basic concepts and terminology used in machine learning. We will also provide a brief introduction to computer vision, a subfield of machine learning that deals with the analysis of images and video.

a) *Machine Learning*: is a subfield of artificial intelligence that involves the use of algorithms and statistical models to enable computer systems to improve their performance on a specific task through experience or learning from data. Machine learning algorithms use data to identify patterns and make predictions. This allows computers to perform tasks that would be difficult or impossible to program them to do manually.

b) *Deep learning*: is an extension of machine learning, that utilizes artificial neural networks with multiple layers to analyze and learn from complex datasets. These networks are designed to learn and make decisions in a similar way to the human brain, by processing and recognizing patterns in data. Deep learning has been shown to be highly effective in solving a wide range of complex problems, including image and speech recognition, natural language processing, and predictive analytics.

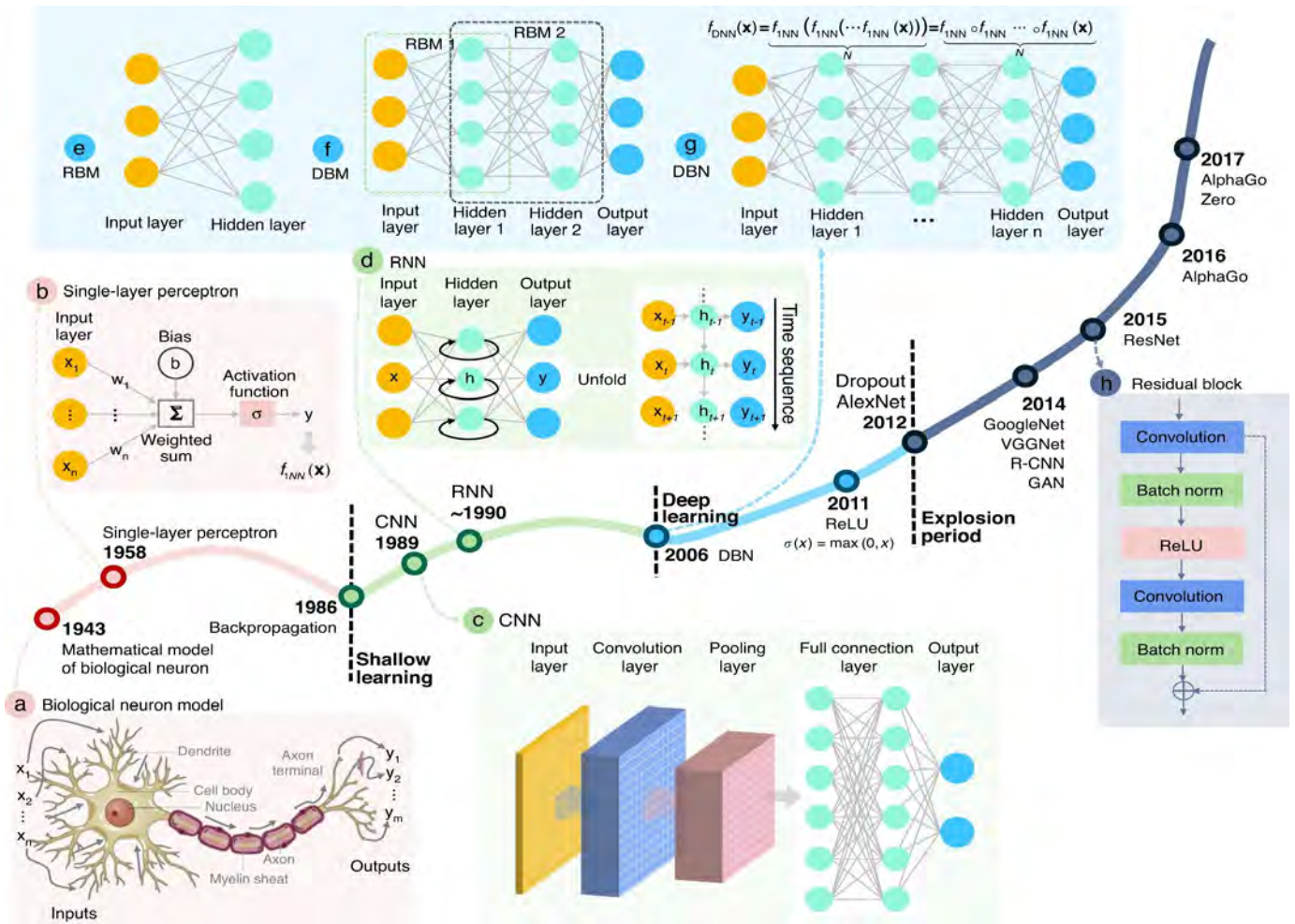


Fig. 4. **a** is the Biological neuron model [12]. **b** depicts the single-layer perceptron network: a single artificial neuron. **c** A convolutional neural network (CNN) that consists of an input layer, a convolution layer, a pooling layer, a full connection layer, and an output layer. **d** recurrent neural network (RNN) where the hidden layers have feedback loops and feedforward as well. **e** Restricted Boltzmann Machines (RBM) is a probability graph model that is undirected and has an input layer and one hidden layer. **f** Deep Boltzmann Machine (DBM) are formed by stacking multiple RBM units. **g** Deep Belief Networks (DBN) are formed by stacking multiple DBM units. **h** Residual blocks have two convolutional layers stacked together that have ReLU as an activation between them [11].

c) Computer vision: Computer vision is an interdisciplinary field that focuses on enabling machines to understand, interpret, and extract information from visual data, such as images and videos. This information can be used to identify objects, track motion, and recognize patterns. It also involves the development of algorithms and techniques that mimic human visual perception and cognition. At its core, computer vision encompasses several key components and processes for various practical analytical tasks.

A. A brief overview of Deep learning

Fig. 4 captures a brief history of deep learning. Fig. 4(a) is the biological inspiration for a computational neural network. Fig. 4(b) depicts the simplest form of an artificial neural network called perceptron. Fig. 4(c) outlays the major components of a CNN (convolutional neural network) where the initial blocks are image operations followed by an artificial neural network. Fig. 4(d) depicts the introduction of RNNs

(Recurrent neural networks), these are special kinds of neural networks that have a feedback loop. Fig. 4(e-g) introduce the modern-day deep learning with multiple hidden layers or stacking multiple networks together.

B. Key Components and Processes in Computer Vision

The four major components for a computer vision task are:

- 1) **Image Acquisition:** The process of capturing visual data through various devices, such as cameras or sensors. It involves factors like resolution, lighting conditions, and image noise.
- 2) **Image Preprocessing:** The initial step in computer vision, includes tasks like noise reduction, image enhancement, and image normalization to improve the quality and suitability of the data for subsequent analysis.
- 3) **Feature Extraction:** The identification and extraction of distinctive patterns or features from images or videos. These features could be edges, corners, textures, or

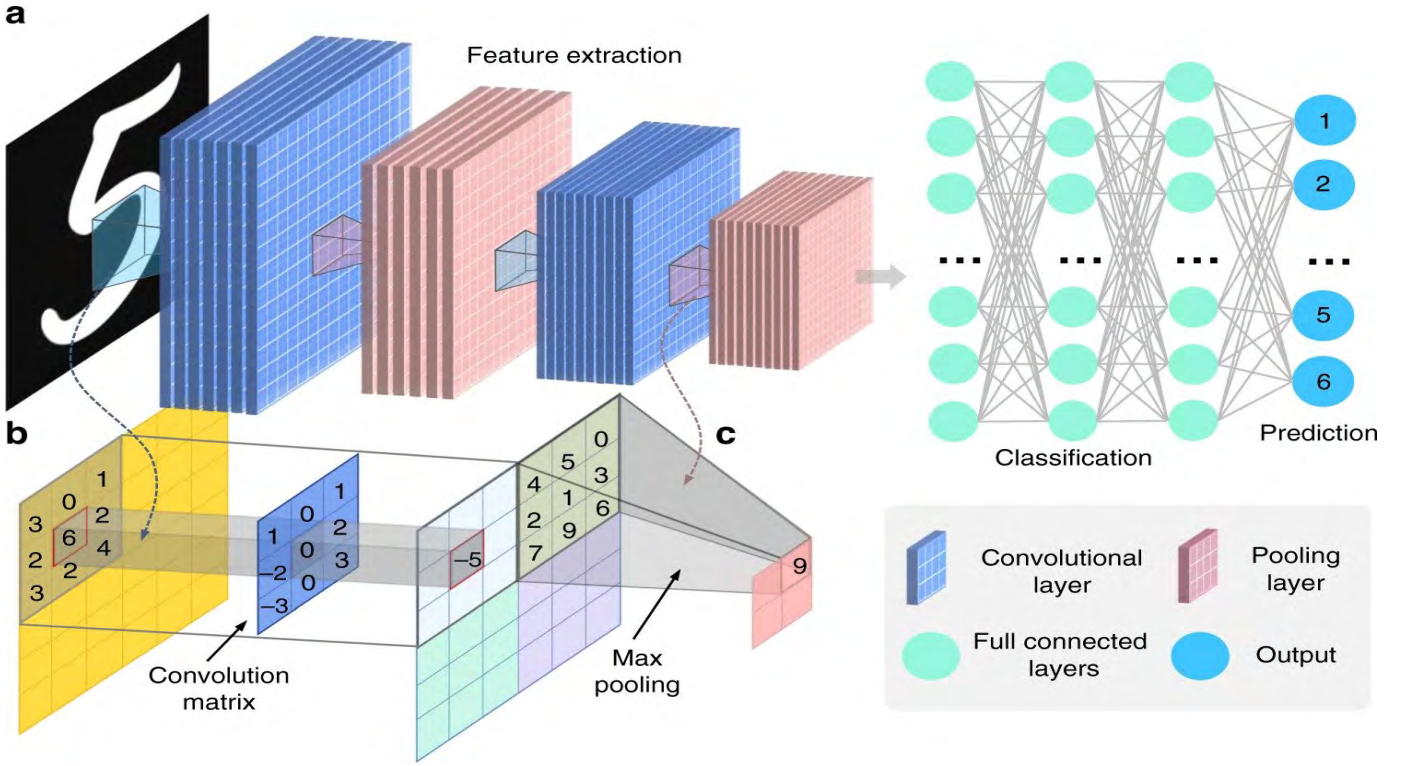


Fig. 5. **a** A typical convolutional neural network consists of the input layer, convolutional layers (these layers perform the task of feature extraction), fully connected layers, and output prediction **b** Convolution operation **c** Pooling operation [11].

higher-level semantic features. Fig. 5 depicts the overall feature extraction process as it would occur in a forward pass of any typical convolutional network. In Fig. 5 part (a), the four blocks represent the convolution operation that is performed on the image. In Fig. 5 part (b), the same is visualized as actual matrices.

- 4) Image Analysis: The interpretation and analysis of visual data to extract meaningful information. This involves tasks such as object detection, recognition, segmentation, tracking, and motion analysis.

C. Computer vision tasks

- 1) Object detection is a computer vision task that involves identifying and locating objects in an image or video. Object detection algorithms are typically used in applications such as self-driving cars, facial recognition, and medical image analysis.
- 2) Image classification is a computer vision task that involves assigning a label to an image. Image classification algorithms are typically used in applications such as image search, product recognition, and spam filtering.
- 3) Image segmentation: is a computer vision task that involves partitioning an image into multiple segments, or regions, based on their visual properties. The goal of image segmentation is to identify and extract meaningful objects from an image

III. RELATED WORK AND RESEARCH OBJECTIVES

The following section covers a general overview of the literature on computer vision in the context of industrial datasets. It also highlights the potential issues with industrial data and concludes by stating the research objectives.

A. Related work

Image classification can be defined as assigning an image to a particular category out of various predefined categories. The traditional approach to image classification involves a two-stage process where firstly, the handcrafted features are extracted, and these extracted features are used as the input to a trainable classifier. The major flaw in this approach is the handcrafted features as the absolute or final accuracy of the classifier is heavily dependent on the quality of the extracted features [13]. Another shortcoming in this approach is its tendency not to generalize well to new data points and require domain expertise [21]. The traditional image classification methods were replaced with convolutional neural networks (CNNs), and one of the earliest breakthroughs on CNNs was presented in [14] and [15]. Furthermore, the introduction of Alexnet [16] in 2012 has made deep convolutional neural networks the de facto standard models for image classification. Since then there have been various new architectures proposed including ResNet [17], Inception [18], VGG [19], DenseNet [20] etc. A comprehensive study on the application of the traditional methods and their contrast to deep learning was studied by [22] and [23]. With this great influx of models

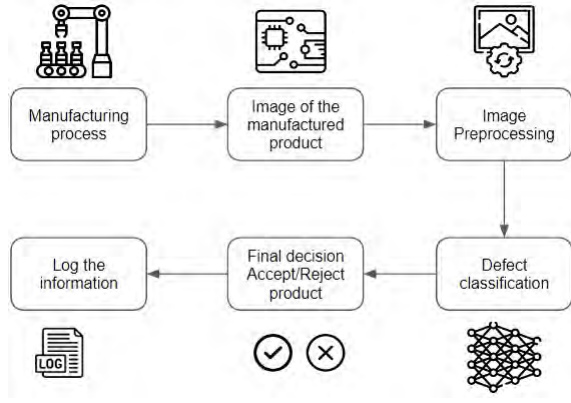


Fig. 6. Overview of the methodology for computer vision based quality control [30]

and research on convolutional networks, there has been a strong interest by industries with real-world datasets to employ computer vision for diverse quality control applications, such as in printing industry [24], PCB boards defect detection [25], magnetic tiles production [1], and aluminum welding [3].

One major disadvantage of real-world datasets is that models trained on them do not generalize well due to domain shift [26]. Domain shift refers to the shift in the distribution of the data present in the training(source) set and data present in the test(target) set, and it can be caused due to various factors such as differences in the viewing angle or image characteristics such as noise, colour and brightness Etc. Several solutions to this problem have been proposed in recent years for example pretraining [27], dropout [28], and transfer learning [29], particularly in deep neural networks these solutions have tried to address the issues of generalization.

IV. QUALITY CONTROL IN DIFFERENT INDUSTRIES

Quality control is an essential part of industrial production. There are several methods used to evaluate the quality of a product or the outcome of a process. Quality control methods can be classified as destructive or non-destructive, depending on the method used to identify defects on a surface or volume. *Destructive* quality control methods involve testing a product or material in a way that damages it. This type of testing is often used to determine the strength or durability of a product. For example, a destructive quality control test might involve dropping a product from a certain height to see if it breaks. *Non-destructive* quality control methods do not damage the product or material being tested. This type of testing is often used to identify defects that would not be visible to the naked eye. For example, a non-destructive quality control test might involve using X-rays to look for cracks in a metal object.

Quality control in industrial production seeks to maintain a quality level or to locate problems for subsequent correction. Conventional detection methods often deal with regular, macro-sized, and complicated fluctuations of surface defects. Nearly all artificial visual defect detection methods seek to identify defects and categorise them for future processing

[30]. From Fig. 7 an important type of non-destructive quality control method is visual inspection, where operators visually assess the state of a product at different stages of production to determine whether it meets quality standards and can be moved on to the next process.

A. A case study on Manufacturing Industry

Steel is an incredibly useful material that has played a pivotal role in various industries and sectors worldwide. Its versatility, durability, and strength make it a preferred choice for numerous applications. Hence the quality control of the defects in particular surface defects is of critical value.

a) *Background:* The authors in [32] presented a computer vision based solution to quality control in steel production. Traditional methods for steel surface defect classification are based on hand-crafted features, which are often computationally expensive and time-consuming to extract. Deep learning based computer vision methods, on the other hand, can automatically learn features from data, which can lead to more accurate and efficient defect classification. The overall process of computer vision based quality control has been depicted in Fig. 6.

b) *Problem Statement:* The main challenge in steel surface defect classification is the high variability of defect types and appearances. Steel surface defects can be caused by a variety of factors, such as manufacturing processes, environmental conditions, and material properties. This variability can make it difficult to develop a single classifier that can accurately identify all types of defects.

c) *Solution:* The paper proposes a deep-learning-based approach for fast and robust steel surface defect classification. The proposed approach uses an end-to-end SqueezeNet-based model [33] that has been pre-trained on the ImageNet dataset [35]. The model is then trained on the Northeastern University (NEU) surface defect dataset [34] that contains images of a variety of defect types. This model is used to classify new images of steel surfaces in order to detect defects.

d) *Evaluation & Conclusion:* The proposed approach was evaluated on two versions of the NEU surface defect dataset i.e one base version without any changes and a diversity-enhanced version. The diversity-enhanced dataset contained images with severe non-uniform illumination, camera noise, and motion blur. The proposed approach achieved a classification accuracy of 100% on the normal testing dataset and accuracy of 97.5% on the diversity-enhanced dataset. This accuracy is significantly higher than the accuracy of traditional methods and other state-of-the-art methods for steel surface defect classification. The proposed approach for steel surface defect classification has several benefits over traditional methods. First, the model only required a small amount of defect-specific training samples to achieve accurate defect recognition. This is because the model was pre-trained on a large dataset of images of steel surfaces, which allowed it to learn features that are common to all types of defects. Second, the proposed solution was more robust to variability in defect types and appearances. This is because the model

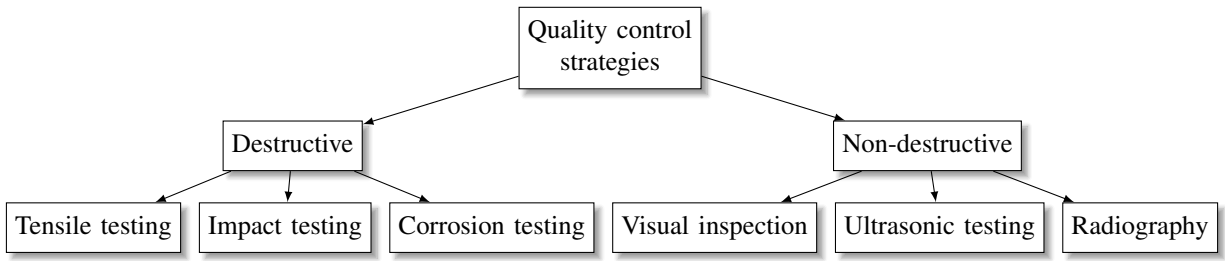


Fig. 7. The classification of quality control strategies. [30]

was trained on a diverse dataset of images, which included images with non-uniform illumination, noise, and blur. As a result, the model was able to identify defects accurately even in images with challenging conditions.

V. CHALLENGES AND FUTURE DIRECTIONS

Computer vision is a rapidly growing field with a wide range of applications. Quality control being one most promising applications. Computer vision can be used to automate a number of the tasks involved in quality control, such as inspecting products for defects, measuring dimensions, and identifying contaminants. This can in theory lead to significant improvements in efficiency, accuracy, and cost-effectiveness. Despite the many advantages of computer vision for quality control, there are still challenges that need to be addressed. Data availability is one of them, the primary basis for any computer vision and deep learning based quality control system is the availability of data. The following subsection discusses various challenges at different levels.

A. Data Level challenges

Data level challenges encompass various aspects, from the sheer volume of visual data to its quality, diversity, and annotation. Handling large datasets, ensuring data integrity, and extracting meaningful insights become increasingly complex tasks. The following list points out a few of those challenges [31].

- 1) Visual accessibility: Collecting data in an industrial environment is a significant challenge as industrial processes tend to be agile. This process agility results in insufficient access to all possible defect locations.
- 2) Labeling: Annotating industrial datasets is a labor-intensive task and usually requires opinions from multiple experts in the field.
- 3) Gathering faulty data/Imbalance: Production lines and manufacturing environments are usually optimized to produce the least defective products. Hence the data gathered is highly imbalanced, and the datasets usually have a very skewed ratio of positive to negative classes
- 4) Lighting: Lightning conditions significantly impact the usability of datasets. In an industrial setup, exposure to extreme volumes of contrast and brightness can hinder the data's usability e.g. dirt/dust set on the camera can influence image appearance.

- 5) Noise: Due to the agile nature of the industrial environment, the potential of adding noise to the dataset is often remarkably high. Contamination of images with shadows from other apparatus presents Etc, can add noise to the dataset.
- 6) Sensor Failure: Industrial processes are also prone to sensor failures due to various factors, these sensor failures can damage the usability of collected data.
- 7) Pose/Posture: The placement of data collecting hardware or cameras determine the quality of data collected. Posture is one of the critical components as it determines the degree to which the deep-learning the algorithm would be exposed to the problem.
- 8) Appearance change: Changes in the appearance of a product from time to time can make the data previously collected unusable, and most of the cases would require retraining of the model.
- 9) Variation in data: The data used to train a computer vision model should be as representative as possible of the data that will be encountered in real-world applications. However, it is often difficult to obtain data that is perfectly representative of all possible variations. This can lead to problems with model generalization.

B. Scalability challenges

Scalability is a critical consideration when it comes to computer vision and quality control systems. As the demand for advanced visual analysis and quality assurance grows, these systems must be able to handle increasing data volumes, computational requirements, and real-time processing. However, scaling such systems brings forth a unique set of challenges. From managing massive datasets to distributing computational resources, maintaining low latency, and adapting to diverse data sources, the scalability hurdles are multifaceted. Additionally, considerations such as model training and deployment, handling variations in imaging conditions, and integration with existing infrastructure add further complexity. In this context, understanding and addressing these scalability challenges are vital for building robust and efficient computer vision and quality control solutions. The following list points out a few of those challenges.

- 1) Data volume: Managing and processing large amounts of visual data in real time can be challenging. As the size of the dataset grows, it becomes more difficult to handle storage, retrieval, and processing efficiently.

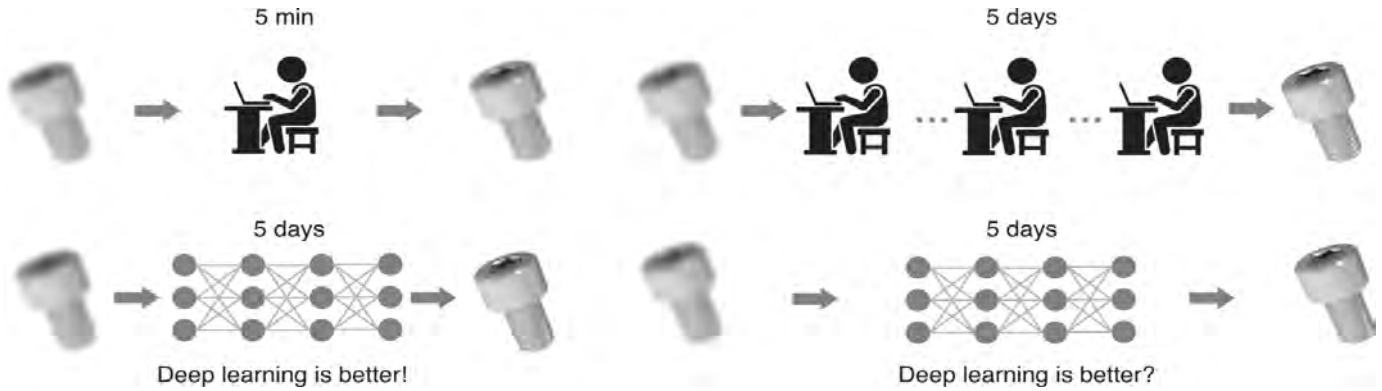


Fig. 8. Deep learning and traditional algorithms should be compared on an equitable basis. [11].

- 2) Computational resources: Computer vision algorithms can be computationally intensive, requiring significant processing power and memory. Scaling up to process high-resolution images or real-time video streams can strain the available computing resources.
- 3) Latency: Real-time computer vision applications require low latency to provide timely responses. As the system scales, reducing the processing time per image becomes crucial to maintain acceptable response times.
- 4) Distributed processing: Scaling computer vision systems across multiple machines or a distributed network introduces challenges in coordinating the processing of data, ensuring consistency, and minimizing communication overhead.
- 5) Model training and retraining: Training and retraining computer vision models with large datasets can be time-consuming and resource-intensive. Scaling up the training process while maintaining accuracy and quality is a significant challenge.

Furthermore, there is a very important discussion around the need of using computer vision and deep learning for quality control. Given the "no free lunch theorem," it is essential to approach the choice between deep learning and traditional algorithms with rational consideration. Deep learning is not always necessary for solving certain problems where traditional methods based on physics models can offer straightforward and highly satisfactory solutions, provided they are implemented correctly for example Fig. 8. In such cases, there may be no need to resort to deep learning techniques. By leveraging physics principles and employing well-established algorithms, these traditional approaches can often provide effective solutions without the added complexity and computational requirements associated with deep learning. However, it may not always be immediately apparent when a traditional approach is sufficient and deep learning is deemed "unnecessary." Although traditional methods can be functionally effective, it is essential to recognize that the performance of deep learning models heavily relies on the reliability and quality of the training data provided. Deep learning algorithms excel in tasks where large, diverse, and

accurately labeled datasets are available for training. It is crucial to consider the availability and suitability of training data when determining whether deep learning is the best approach. Understanding the limitations and dependencies on training data quality allows for a more informed decision-making process, ensuring that the chosen approach aligns with the data available and maximizes the potential benefits of deep learning [11].

C. Future direction

The future of computer vision and quality control holds exciting possibilities as technology continues to advance. One promising direction is the integration of computer vision with emerging technologies such as augmented reality (AR) and virtual reality (VR), enabling immersive visual experiences and enhancing quality control processes. Another future direction involves the development of real-time quality control systems that leverage edge computing and Internet of Things (IoT) devices, allowing for efficient and decentralized monitoring and analysis. Moreover, ethical considerations are gaining increased attention, and future developments in computer vision and quality control will likely involve the implementation of frameworks and guidelines to address issues like bias, privacy, and algorithmic fairness. As these technologies evolve, collaboration between industry, academia, and regulatory bodies will be crucial in shaping the future direction of computer vision and quality control, ultimately unlocking new possibilities and driving innovation across various domains.

VI. CONCLUSION

Deep learning based computer vision has emerged as a powerful technology for quality control, offering automated, accurate, and efficient inspection and analysis capabilities across diverse industries that rule-based methods cannot address. The fundamental components of computer vision, combined with specific techniques like image preprocessing, object detection, and deep learning, enable the development of robust quality control systems. Computer vision is revolutionizing quality control processes across industries, including reduced costs, and enhanced product quality. By leveraging computer vision, we can effectively detect defects, anomalies, and variations

in products, thus improving overall quality control processes. However, it is important to address data level challenges such as sparsity, imbalance, and shift in manufacturing applications to enhance the accuracy and reliability of computer vision-based quality control systems. Additionally, the issues of real-time processing, ethical considerations, and algorithmic biases require further attention. In conclusion, the significance of computer vision in quality control cannot be overstated. As we move forward, several areas for further research and exploration exist. Firstly, the development of novel computer vision algorithms and techniques can enhance the capabilities of quality control systems, improving their ability to detect and classify defects accurately. Real-time processing challenges need to be addressed to enable efficient and timely quality control inspections in dynamic manufacturing environments. Moreover, it is essential to consider the ethical implications of automated quality control systems, ensuring fairness, transparency, and accountability. Further research should focus on developing robust frameworks and guidelines for ethical implementation. By continuing to advance and innovate in these areas, computer vision can play a transformative role in revolutionizing quality control processes across industries.

ACKNOWLEDGMENTS

Funded by the Ministry of Economics, Labor and Tourism Baden-Württemberg under KI Lab DIANA (Data-intensive application for automation) at DHBW Mannheim.

We would like to acknowledge the contribution of Zuo, C., Qian, J., Feng, S. et al. for creating Figures 8, 4 and 5, that have been reproduced from their paper titled 'Deep learning in optical metrology: a review' published in the Journal of light: science & applications (Nature) in 2022. The images are made public under Creative Commons Attribution 4.0 International License. No modifications have been made to the original image.

REFERENCES

- [1] Y. Huang, C. Qiu, Y. Guo, X. Wang and K. Yuan, "Surface Defect Saliency of Magnetic Tile," 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), Munich, Germany, 2018, pp. 612-617, doi: 10.1109/COASE.2018.8560423.
- [2] Silvestre-Blanes, J., Albero-Albero, T., Miralles, I., Pérez-Llorens, R. & Moreno, J. (2019). A Public Fabric Database for Defect Detection Methods and Results. *Autex Research Journal*, 19(4) 363-374. <https://doi.org/10.2478/aut-2019-0035>
- [3] Daniel Baciou, Geoff Melton, Mayorkinos Papaefias, Rob Shaw, Automated defect classification of Aluminium 5083 TIG welding using HDR camera and neural networks, *Journal of Manufacturing Processes*, Volume 45, 2019, Pages 603-613, ISSN 1526-6125, <https://doi.org/10.1016/j.jmapro.2019.07.020>.
- [4] Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering*, 10(2), 1687814018755519.
- [5] Paraschos, Panagiotis & Koulinas, G.K. & Koulouriotis, Dimitrios. (2020). Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures. *Journal of Manufacturing Systems*. 56. 470-483. [10.1016/j.jmsy.2020.07.004](https://doi.org/10.1016/j.jmsy.2020.07.004).
- [6] Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., & Wellbrock, W. (2019). Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors*, 19(18), 3987.
- [7] Agarwal, N., Chiang, C. W., & Sharma, A. (2019). A study on computer vision techniques for self-driving cars. In *Frontier Computing: Theory, Technologies and Applications (FC 2018)* 7 (pp. 629-634). Springer Singapore.
- [8] Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation—A review. *Information Processing in Agriculture*, 7(1), 1-19.
- [9] Minhas, M.S., & Zelek, J.S. (2020). Defect Detection using Deep Learning from Minimal Annotations. *VISIGRAPP*.
- [10] C. Sun, A. Shrivastava, S. Singh and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 843-852, doi: 10.1109/ICCV.2017.97.
- [11] Zuo, C., Qian, J., Feng, S. et al. Deep learning in optical metrology: a review. *Light Sci Appl* 11, 39 (2022). <https://doi.org/10.1038/s41377-022-00714-x>
- [12] Kim, Y., Kang, N., Kim, S., Kim, H. (2013). Evaluation for snowfall depth forecasting using neural network and multiple regression models. *Journal of the Korean Society of Hazard Mitigation*, 13(2), 269-280.
- [13] Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural computation*, 29(9), 2352-2449. https://doi.org/10.1162/NECO_a_00990
- [14] Y. LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," in *Neural Computation*, vol. 1, no. 4, pp. 541-551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.
- [15] LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., & Jackel, L.D. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *NIPS*.
- [16] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*. 25. 10.1145/3065386.
- [17] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90.
- [18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.
- [19] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* 1409.1556.
- [20] Huang, Gao & Liu, Zhuang & van der Maaten, Laurens & Weinberger, Kilian. (2017). Densely Connected Convolutional Networks. 10.1109/CVPR.2017.243.
- [21] Minhas M. and Zelek J. (2020). Defect Detection using Deep Learning from Minimal Annotations. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*; ISBN 978-989-758-402-2, SciTePress, pages 506-513. DOI: 10.5220/0009168005060513
- [22] Murtadha D Hssayeni, Sagar Saxena, Raymond Ptucha, Andreas Savakis, "Distracted Driver Detection: Deep Learning vs Handcrafted Features" in *Proc. IST Int'l. Symp. on Electronic Imaging: Imaging and Multimedia Analytics in a Web and Mobile World*, 2017, pp 20 - 26, <https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-162>
- [23] Marnissi, Mohamed & Fradi, Hajer & Dugelay, Jean-Luc. (2019). On the Discriminative Power of Learned vs. Hand-Crafted Features for Crowd Density Analysis. 1-8. 10.1109/IJCNN.2019.8851764.
- [24] Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., & Wellbrock, W. (2019). Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0. *Sensors*, 19(18), 3987. <https://doi.org/10.3390/s19183987>
- [25] Bhattacharya, A., Cloutier, S.G. End-to-end deep learning framework for printed circuit board manufacturing defect classification. *Sci Rep* 12, 12559 (2022). <https://doi.org/10.1038/s41598-022-16302-3>
- [26] K. Zhou, Z. Liu, Y. Qiao, T. Xiang and C. C. Loy, "Domain Generalization: A Survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396-4415, 1 April 2023, doi: 10.1109/TPAMI.2022.3195549.
- [27] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why Does Unsupervised Pre-training Help Deep Learning? *J. Mach. Learn. Res.* 11 (3/12/2010), 625-660.
- [28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural

- networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (January 2014), 1929–1958.
- [29] Weiss, K., Khoshgoftaar, T.M. & Wang, D. A survey of transfer learning. *J Big Data* 3, 9 (2016). <https://doi.org/10.1186/s40537-016-0043-6>
 - [30] Czimmermann, T., Ciuti, G., Milazzo, M., Chiurazzi, M., Roccella, S., Oddo, C.M., & Dario, P. (2020). Visual-Based Defect Detection and Classification Approaches for Industrial Applications—A SURVEY. *Sensors* (Basel, Switzerland), 20.
 - [31] Leyendecker, L., Agarwal, S., Werner, T., Motz, M., Schmitt, R.H. (2023). A Study on Data Augmentation Techniques for Visual Defect Detection in Manufacturing. In: Lohweg, V. (eds) *Bildverarbeitung in der Automation. Technologien für die intelligente Automation*, vol 17. Springer Vieweg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-66769-9_6
 - [32] Fu, G., “A deep-learning-based approach for fast and robust steel surface defects classification”, *Optics and Lasers in Engineering*, vol. 121, pp. 397–405, 2019. doi:10.1016/j.optlaseng.2019.05.005.
 - [33] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J. & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5MB model size (cite arxiv:1602.07360Comment: In ICLR Format)
 - [34] He Y, Song K, Meng Q, Yan Y (2019) An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Trans Instrum Meas* 69(4):1493–1504
 - [35] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

Analyzing news articles on the COVID-19 pandemic regarding the timeline, vaccination, and sentiment

Linus Eickhoff
Computer Science
Baden-Wuerttemberg Cooperative State
University (DHBW)
Stuttgart, Germany
linusmaxeickhoff@gmail.com

Florian Kellermann
Computer Science
Baden-Wuerttemberg Cooperative State
University (DHBW)
Stuttgart, Germany
kellermannflorian1@gmail.com

Monika Kochanowski
Computer Science
Baden-Wuerttemberg Cooperative State
University (DHBW)
Stuttgart, Germany
monika.kochanowski@dhbw-
stuttgart.de

Keywords—news articles, NLP, COVID-19, sentiment topic modeling, data analysis

This work investigates the role of news media during the COVID-19 pandemic, a global health crisis that has had profound societal implications and has drawn significant attention in the media landscape. The central objective of this study is to examine COVID-19-related articles and recognize the behavioral patterns and responses of news media throughout the pandemic.

The applied techniques include Web Scraping, Natural Language Processing (NLP), and Topic Modeling. The dataset at the core of the investigation encompasses nearly 500.000 articles published between 2020 and 2022. Recognizing the need for a more comprehensive and diverse dataset, the initial dataset was augmented by web scraping [1] additional articles from a broader range of news sources, employing automated Google searches via SerpApi.

Following this, the data underwent thorough cleaning and preprocessing utilizing several NLP techniques [2], such as tokenization, stop-word removal, Part-of-speech tagging, and lemmatization, in preparation for further analysis. Latent Dirichlet Allocation [3] is applied for topic modeling, generating topical subsets that provided a foundation for investigating potential correlations with official COVID-19 data. Additionally, we referred to case count and vaccination data from Our World in Data, which collates the original data from the World Health Organization. Additionally, seasonal patterns and the trend of the data are investigated.

The first research question focused on the potential correlation between the number of COVID-19 cases and the volume of related news articles. However, our investigation found insufficient evidence to confirm such a correlation. Interestingly, the amount of news coverage on COVID-19 can be interpreted in the pandemic phases “scramble, learn, respond, test, and hope” quite well [4].

For our second question, we examined the correlation between administered vaccination doses and the volume or share of vaccination-related articles. In this case, we observed evidence suggestive of a correlation. Manifold factors

influence media coverage as well as vaccination speed. During the pandemic, it has been shown that more research on this topic is necessary.

Finally, we conducted sentiment analysis [5] on the content and titles of articles to examine the stance of news media regarding COVID-19 and vaccinations. Our analysis, based on a sentiment score [6] ranging from -1 to 1, revealed sentiment disparities among news sources, often reflecting their political inclinations.

In sum, our results shed light on the role of news media during the COVID-19 pandemic using data analysis techniques. Further work should give insights into influence and potential impact on public understanding and response to the crisis (also examined in [7]).

- [1] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bouso, and S. N. Mbaye, “Web scraping: State-of-the-art and areas of application,” in 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019. doi: 10.1109/bigdata47090.2019.9005594
- [2] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: State of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2022. doi:10.1007/s11042-022-13428-4
- [3] C. Guo, M. Lu, and W. Wei, “An improved lda topic modeling method based on partition for medium and long texts,” *Annals of Data Science*, vol. 8, no. 2, pp. 331–344, 2021, issn: 2198-5804. doi: 10.1007/s40745-019-00218-3.
- [4] K. Jetelina, “18 months of the covid-19 pandemic – a retrospective in 7 charts,” *The Conversation*, vol. 2021, 2021. [Online]. Available: <https://theconversation.com/18-months-of-the-covid-19-pandemic-a-retrospective-in-7-charts-166881>
- [5] A. Pambudi and S. Suprpto, “Effect of sentence length in sentiment analysis using support vector machine and convolutional neural network method,” 1978-1520, vol. 15, no. 1, p. 21, 2021, issn: 1978-1520. doi: 10.22146/ijecs.61627.
- [6] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, issn: 1554-0669. doi: 10.1561/1500000011
- [7] R. Jones, D. Mougouei, and S. L. Evans, “Understanding the emotional response to covid-19 information in news and social media: A mental health perspective,” *Human behavior and emerging technologies*, vol. 3, no. 5, pp. 832–842, 2021. doi: 10.1002/hbe2.304

Classification algorithms in database management systems (extended abstract)

Alicia Dietrich
Department of Computer Science
Baden-Wuerttemberg Cooperative State University
Horb, Germany
i20005@hb.dhbw-stuttgart.de

Olaf Herden
Department of Computer Science
Baden-Wuerttemberg Cooperative State University
Horb, Germany
o.herden@hb.dhbw-stuttgart.de

Abstract—In this extended abstract we describe our research project concerning machine learning in database systems.

Keywords—machine learning, classification, database

I. EXTENDED ABSTRACT

In the field of machine learning, classification is a well-known approach [1][2]. Classification is supervised, i.e. each data record in a data set has a class label. The data set is divided into a training set und test set. Based on the training set as input, a classification algorithm computes a classifier. The quality of the classifier is evaluated by using the test set. Finally, the classifier enables to determine the class of new data records with unknown label.

Usually, the classification process is realized in the application layer. Many program libraries and tool suites offer a multitude of algorithms. In recent years, several database vendors integrated classification algorithms into the core of the database management system.

In [3] we investigated classification in database systems and discussed the pros and cons of this approach in contrast to traditional solutions.

First, we give an overview about classification algorithms. These algorithms are divided into the four categories probabilistic, hierarchical, linear and prototype based classifiers. Moreover, different quality metrics for the evaluation of a classifier are described.

Afterwards we take a look on different database management systems and the classification algorithms implemented in these systems. Especially we investigated the realization opportunities in the Oracle database [4]. The vendor calls its solution OML (Oracle machine learning), interfaces are offered for the languages SQL, Python and R [5].

In a proof of concept the algorithms SVM (support vector machine) and decision trees were implemented in the Oracle database by using the language R. The used data set is "Human activity recognition using smartphones" [6] with about 10300 data records, each described by 561 attributes. Two different prototypes were realized, one for a standalone installation of the database and one within the cloud solution Oracle ADB (autonomous database). The pros and cons of both ways of realization are discussed.

. As a result, we can say that using the built-in functions in databases is advantageous for small applications. A solution can be realized quickly and automated and without deep programming knowledge. However, for more complex use cases and big data volumes the opportunities are limited.

As future work we see different tasks, e.g. combining the built-in algorithms with other database features like in memory processing or comparing the existing implementation with other database management systems like MySQL HeatWave.

- [1] T. Oladipupo. „Types of Machine Learning Algorithms“. In: Introduction to machine learning. Hrsg. von Ethem Alpaydm. 2nd edition. Adaptive computation and machine learning series. Cambridge, Massachusetts: MIT Press, 2010. isbn: 978-953-307-034-6. doi: 10.5772/9385.
- [2] S. Russell and P. Norvig. Artificial intelligence. A modern approach. eng. Third edition. Boston: Pearson, 2016. 151 S. isbn: 9781292153971. url: <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5831883>.
- [3] A. Dietrich “Klassifikationsverfahren in Datenbanken” (in German), Student reasearch work, Baden-Wuerttemberg Coopeartive State University Stuttgart Campus Horb, 2023.
- [4] Oracle Corporation. „Oracle Machine Learning Technical Brief“. URL: <https://www.oracle.com/a/tech/docs/technical-resources/oml-technical-brief.pdf> (visited May 20th 2023).
- [5] Oracle Corporation. „OML4R Machine Learning Algorithms“. In: (). URL: <https://www.oracle.com/docs/tech/otn-batch1/oml4r-algorithms.pdf> (visited May 20th 2023)..
- [6] Human Activity Recognition Using Smartphones - UC Irvine Machine Learning Repository. URL: <https://archive-beta.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones> (visited May 20th 2023).

Künstliche Intelligenz im Handel: Online-Kurs für Betriebswirt*innen

Daniela Wiehenbrauk
BWL-Digital Commerce Management
DHBW Heilbronn
daniela.wiehenbrauk@heilbronn.dhbw.de

Oliver Janz
BWL-Handel, insb. Fashion Management
DHBW Heilbronn
oliver.janz@heilbronn.dhbw.de

Johannes Kolb
BWL-Handel, insb. International Retail
Management
DHBW Heilbronn
johannes.kolb@heilbronn.dhbw.de

Armin Mueller
Global Head of Analytics
Kaufland, Deutschland

Modularer Online-Kurs für Betriebswirt*innen zur Entwicklung von KI-Kompetenzen mit Fokus auf den Handel

KI-Kompetenz, Handel, Flipped-Classroom

I. ÜBERBLICK

Betriebswirt*innen sind dafür verantwortlich, die **Anwendungsmöglichkeiten von Künstlicher Intelligenz (KI) für ihr Unternehmen zu identifizieren und umzusetzen**. Um diese Aufgaben zu erfüllen, müssen sie die Funktionsweise der KI sowie die eingesetzten Methoden verstehen und bestehende Anwendungsfälle von KI im Handel kennenlernen.

Unsere Zielgruppe sind Studierende des Bachelor-Studiengangs Betriebswirtschaftslehre, sowohl an der DHBW als auch an anderen Hochschulen und Universitäten. Dank seiner Struktur eignet sich der Kurs auch für Betriebswirt*innen in Unternehmen und kann beispielsweise bei den dualen Partnern der DHBW eingesetzt werden.

Der Kurs besteht aus drei Teilen, die sich auf die wesentlichen Inhalte konzentrieren

1. Grundlagen KI:

Dieser Teil, umgesetzt durch Lehrvideos und PowerPoint-Folien, behandelt Definitionen, Historie, Datenquellen, Datenqualität und Risiken der KI. Zudem spielt das CRISP-DM-Framework eine wichtige Rolle, da Anwendungsfälle typischerweise auf dieser Basis strukturiert werden.

2. Use Cases KI:

Dieser Teil wird durch Interviews mit Expert*innen aus Unternehmen umgesetzt. Die Interviews werden als Videos aufbereitet und mit entsprechenden Hintergrundinformationen versehen. Fallstudien zu den Themen Klassifikation, Clustering, Regression und Mustererkennung zeigen die Anwendung und Umsetzung von KI-Methoden in Handelsunternehmen.

3. Methoden der KI:

Um unterschiedliche Lernstände der Studierenden zu berücksichtigen, beginnt der dritte Teil mit einer Art Vorkurs zu Python. Screencast-Videos bieten eine

zielgerichtete Einführung in die Programmiersprache und erläutern Konzepte wie Imports, Exports und Dataframes. Der Hauptteil dieses Kapitels behandelt vier KI-Methoden (Clustering, Regression, Klassifikation und Mustererkennung) mit dem entsprechenden statistischen Hintergrund. Ein Fokus liegt auf der Anwendung dieser Methoden in Python anhand relevanter Datensätze aus Handelsunternehmen. Auch dieser Teil wird durch Lehr- oder Screencast-Videos und PowerPoint-Folien umgesetzt.

II. ANWENDUNG VON INTERAKTIVEN LERN- UND LEHRMETHODEN

Wird der Kurs als **interaktiver und modularer Online-Kurs** absolviert, unterstützen regelmäßige Lernziel-Checks den Lernprozess. Abschließende formale Prüfungsformate an jeder Lerneinheit sollen den Lernenden Rückmeldung darüber geben, ob die angestrebten Lernziele erreicht wurden.

Bei Durchführung des Kurses im **Flipped-Classroom-Format** an einer Hochschule stehen den Dozierenden für jedes Modul Aufgaben zur Verfügung. Die Studierenden können diese Aufgaben einzeln oder im Team bearbeiten und in der Vorlesung präsentieren. Die Leistungen der Studierenden können durch die Bewertung dieser Präsentationen beurteilt werden.

Der Online-Kurs hat einen Umfang von ca. vier Stunden. Bei Angebot des Kurses im Flipped-Classroom-Format sollte ein höherer Zeitaufwand eingeplant werden, um den **Dozierenden die Möglichkeit zu geben, auf individuelle Fragen und Herausforderungen der Studierenden einzugehen.**

III. LIMITATIONEN

Da sich der Kurs an Bachelorstudierende der Betriebswirtschaftslehre richtet, soll er eine leicht zugängliche Einführung in das Thema Künstliche Intelligenz bieten. Daher kann nur ein begrenzter Teil der Methoden der Künstlichen Intelligenz behandelt werden.

KOOPERATIVE PARTNER UND SPONSOREN

Stifterverband, KI.Campus, Kaufland, Breuninger, Würth und Wincor-Nixdorf

AI-based point cloud analysis and web-based VR visualization (KIP-VR)

Dominik Ruoff
DHBW Stuttgart Campus Horb
Horb am Neckar, Germany
d.ruoff@hb.dhbw-stuttgart.de

Richard Steffen
Pointcab GmbH
Wernau (Neckar), Germany
info@pointcab-software.com

Tim Jansen
DHBW Stuttgart Campus Horb
Horb am Neckar, Germany
t.jansen@hb.dhbw-stuttgart.de

Iulia Prica
Pointcab GmbH
Wernau (Neckar), Germany
info@pointcab-software.com

Point clouds, which are sets of 3D data points, have gained significant attention in computer vision and robotics due to their ability to capture the geometry and spatial arrangement of real-world scenes and objects. This paper shows the research content of the KIP-VR project in which objects are searched for in point clouds and made accessible to the user via a web browser-based application.

Keywords: point cloud, object detection, web-based VR

I. INTRODUCTION

Nowadays, the geometry of an object can be recorded using modern laser or image-based methods, e.g. as a so-called point cloud. A point cloud is understood as an accumulation of data points within a vector space. The written spatial coordinates allow an exact definition of the points in space. Recorded point clouds are currently being manually converted into 3D CAD models, which reduces their efficient, sustainable and cost-effective use.

II. PROJECT APPROACH

In recent years, the use of machine learning, often referred to as artificial intelligence, has developed rapidly. Based on developments such as support vector machines, Bayes networks and other probabilistic methods for supervised and unsupervised classification, identification and object recognition, neural networks, now in particular convolutional neural networks (CNN) [1] led to extremely promising results. These methods require large amounts of pre-classified data and computing power to train the network. Proceeding from this, there are recent efforts to extend the 1D problem of speech recognition and the 2D problem of image recognition to the 3D problem of object recognition based on point clouds. However, it turns out that an increase in dimension leads to an exponential effort. Nevertheless, procedures have been developed that largely minimize the increased effort. Today it is possible to learn a CNN with sufficient training data of a complex model in a few weeks on a mainframe computer. The application of the model to given data can usually be achieved efficiently with the simplest hardware (smartphone), provided the pre-processing of the data is good. VoxNet [2] and PointNet [3] should be mentioned here as examples. These methods often require complex pre-processing, i.e. very specific learning (training) with large amounts of data, in order to obtain adequate results.

In addition to the object detection within the point clouds, the visualization of the data with head-mounted displays is a

central research component. Thanks to the extremely large potential of the games industry and the investments in this area, inexpensive products could be established on the market [4]. This has increased the industry's interest in this technology for object and process visualization, which has led to proprietary applications for VR applications. The actual difficulty, as well as the opportunity, lies in the universal interchangeability and location independence of the customers who are supposed to interact with the object in VR. In order to be able to guarantee sustainable success, comprehensive work based on real-time data is elementary. WebXR/OpenXR [5] VR APIs are suitable for the standardization of web standards in the web browser. This means that VR-capable end devices can be addressed in the web browser without installing special proprietary software. In addition, various university research projects have enabled the 3D visualization of point clouds in the web browser using WebGL, such as the widespread Potree [6].

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

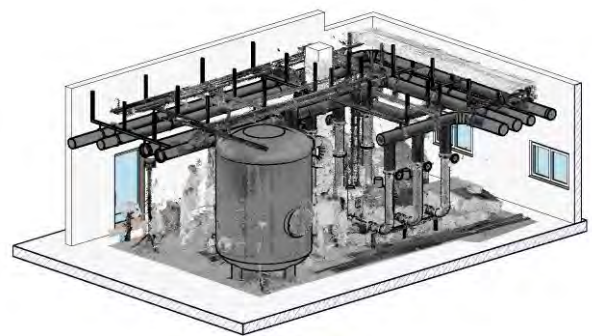


Figure 1: Point cloud overlaid with detected objects

III. TARGETS

Within the point cloud, a search should be made for specific objects whose geometry is known from 3D CAD models. By means of a point cloud matching with the object model, the point cloud should be sublimated with the 3D model. For easy use of the data models, the visualization takes place via head-mounted displays, which also enable interaction with the models via hand controllers or hand or eye

tracking. The web browser-based process visualization displays the results simply and without additional software.

The research is funded by the Baden-Württemberg Ministry of Economics, Labor and Tourism as part of the Invest BW funding program. It is implemented in cooperation with the company PointCab GmbH from Wernau, whose expertise lies in the field of point cloud processing, at the DHBW Stuttgart, Campus Horb under the direction of Prof. Dr.-Ing. Tim Jansen.

REFERENCES

- [1] Yann LeCun: LeNet-5, convolutional neural networks; https://d2l.ai/chapter_convolutional-neural-networks/lenet.html, accessed: 05. April 2023.
- [2] Daniel Maturana (2015) "VoxNet: A 3D Convolutional Neural Network for Real-Time Object-Recognition", https://www.ri.cmu.edu/pub_files/2015/9/voxnet_maturana_scherer_iros15.pdf, accessed: 20. April 2023
- [3] Charles R. Qi et.al. PointNet: „Deep Learning on Point Sets for 3D Classification and Segmentation“, https://openaccess.thecvf.com/content_cvpr_2017/papers/Qi_PointNet_Deep_Learning_CVPR_2017_paper.pdf; accessed: 15. April 2023
- [4] Microsoft HoloLens 2 - <https://www.microsoft.com/de-de/hololens/hardware>; accessed: 05. April 2023
- [5] WebXR/OpenXR - <https://www.khronos.org/openxr/>; accessed: 2. Mai 2023.
- [6] Three.js - JavaScript 3D Library - <https://threejs.org>; accessed: 20. Mai 2023.

Steered Training Data Generation for Learned Semantic Type Detection (Poster Abstract)

Sven Langenecker
DHBW Mosbach & TU Darmstadt
Germany
0009-0002-2809-5331

Christoph Sturm
DHBW Mosbach
Germany
0009-0008-5706-3041

Christian Schalles
DHBW Mosbach
Germany
0009-0005-7036-3012

Carsten Binnig
TU Darmstadt
Germany
0000-0002-2744-7836

Abstract—The poster introduces STEER to adapt learned semantic type extraction approaches to a new, unseen data lake. STEER provides a data programming framework for semantic labeling which is used to generate new labeled training data with minimal overhead. At its core, STEER comes with a novel training data generation procedure called Steered-Labeling that can generate high quality training data not only for non-numeric but also for numerical columns. With this generated training data STEER is able to fine-tune existing learned semantic type extraction models. We evaluate our approach on four different data lakes and show that we can significantly improve the performance of two different types of learned models across all data lakes.

Index Terms—semantic type detection, data programming, data discovery, data lakes

I. EXTENDED ABSTRACT

The task of detecting the semantic type of data is crucial for data discovery in Data Lakes. As such, in the last years various approaches for automated semantic type detection have been proposed. Whereas existing commercial products mainly rely on simple search based solutions such as regular-expressions and dictionary look ups, more recent approaches use machine learning [1]–[3]. While initial results of these learned approaches are promising, unfortunately a learned approach that was trained for data in one data lake cannot be used out-of-the-box for new unseen data sources in a different data lake, even if both data sources cover the same semantic types [3].

The reasons are that the data characteristics in the new (unseen) data lake might be completely different or even worse new semantic types occur that the model has not seen during training. Hence, the performance of a learned model trained for one data lake might be completely different on the new unseen data lake. In this poster, we therefore propose STEER which implements data programming for semantic labeling to adapt existing learned models for extracting semantics to unseen data lakes with minimal cost. STEER can not only significantly boost the performance of models on new unseen data lakes with the same types, but also helps us to re-train a model to detect types on a data lake that comes with unseen semantic types.

At its core, STEER involves a novel training data generation process called Steered-Labeling. The intuition is that in Steered-Labeling we separate the process into two subsequent

steps: STEER first labels the non-numerical columns that are easier to label. Afterwards, STEER then uses these labels to “steer” the labeling of the numerical columns. With this, STEER is able to not only generate high quality training data of textual columns but also to semantically label numerical data with a very high precision (up to 0.95). Which is a problem in existing approaches and has not been addressed sufficiently so far.

Finally, we provide an extensive evaluation of STEER on four different data lakes with different characteristics that in total contain more than 643,500 columns. These data lakes vary in the semantic types and cover a wide spectrum of numerical and non-numerical data types. Moreover, in our evaluation we also show that our approach can be used across models that implement different learning approaches [1], [4]. In particular, we use SATO which relies on a classical supervised training approach and TURL that uses the pre-training/fine-tuning paradigm. The results of this experiments demonstrate that STEER works for both model architectures. Overall, each experiment shows that STEER can generate training data that allows a learned model to provide high performance. Thereby we not only highlight the performance of the re-trained end model, but also the quality and quantity of the generated training data by STEER.

REFERENCES

- [1] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2021. TURL: Table Understanding through Representation Learning. In VLDB, Vol. 14. VLDB Endowment, 307 – 319.
- [2] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zraggen, Arvind Satyanarayan, Tim Kraska, Çağatay Demiralp, and César Hidalgo. 2019. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In SIGKDD (Anchorage, AK, USA) (KDD '19). ACM.
- [3] Sven Langenecker, Christoph Sturm, Christian Schalles, and Carsten Binnig. 2021. Towards Learned Metadata Extraction for Data Lakes. In BTW 2021, Kai-Uwe Sattler, Melanie Herschel, and Wolfgang Lehner (Eds.). Gesellschaft für Informatik, Bonn, 325–336.
- [4] Dan Zhang, Madelon Hulsebos, Yoshihiko Suhara, Çağatay Demiralp, Jinfeng Li, and Wang-Chiew Tan. 2020. Sato: Contextual Semantic Type Detection in Tables. In VLDB, Vol. 13.

A pilot study comparing the performance of deep learning model (LSTM) and statistical models (ARIMA and SARIMA) for demand forecasting of an automotive spare parts

1st Nehalben Ranabhatt

Center for Digitalization in Mobility Systems (ZDM))

DHBW Ravensburg, Campus Friedrichshafen

Friedrichshafen, Germany

Ranabhatt@dhbw-ravensburg.de

2nd Wilhelm Ruckdeschel

Center for Digitalization in Mobility Systems (ZDM)

DHBW Ravensburg, Campus Friedrichshafen

Friedrichshafen, Germany

Ruckdeschel@dhbw-ravensburg.de

Abstract—Forecasting of an automotive part is crucial for automotive part supplier companies. Indeed, the accuracy of forecast affects the costs associated with inventory and its management, replenishment of material flow on assembling line and timely delivery of products to the customer. Technically, demand forecasting of an automotive part is a time series forecasting problem. However, available training data which consists of various spare parts ordered at regular and intermittent periods poses a central challenge to this problem. Until recently, automotive spare parts demand has been modelled using time-series based statistical techniques such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA models (SARIMA). However, these approaches often under-perform when these are applied to the new conditions and when the trend is irregular. Nowadays, as the huge computational resources are becoming relatively affordable and rapidly available, the application of machine learning (ML) and Deep learning (DL) techniques such as Artificial Neural network (ANN) based Long Short- Term Memory (LSTM) Neural Network model provides interesting avenue for the demand forecasting. Here, we aim to test the performance of AI models compared to the well-established time series models on the 18 months of data from an automotive parts' supplier company. For this purpose, first, time-series decomposition was carried out to investigate the presence of seasonality and trends in the data. Afterwards, statistical procedures were applied to obtain acceptable ARIMA and SARIMA parameters before applying these models on the data. The evaluation of ARIMA and SARIMA did seem to generate reasonably good results. Later, the LSTM model was applied. As expected, the results of LSTM were strongly dependent on the tuning of hyperparameters, and the type of inventory used. Generally, LSTM performed better than ARIMA and slightly better than SARIMA in case of the orders having seasonality. In future, incorporating multiple influencing factors in the LSTM and its effect on the prediction will also be explored.

Index Terms—Demand forecasting of an automotive parts, ARIMA, SARIMA, LSTM

Künstliche Intelligenz in der Optimierung personalisierter Ernährungsstrategien: Das „Individual Nutrition Advisory Tool“ (INAT)

MSc Hande Gagali

Duale Hochschule Baden-Württemberg
Heilbronn

hande.gagali@heilbronn.dhbw.de

Dr. Cornelia Klug

Duale Hochschule Baden-Württemberg
Heilbronn

cornelia.klug@heilbronn.dhbw.de

Prof. Dr. Katja Lotz

Duale Hochschule Baden-Württemberg
Heilbronn

katja.lotz@heilbronn.dhbw.de

MSc Timo Sievernich

Duale Hochschule Baden-Württemberg
Heilbronn

timo.sievernich@heilbronn.dhbw.de

PD Dr. Alexandr Parlesak

Duale Hochschule Baden-Württemberg
Heilbronn

alexandr.parlesak@heilbronn.dhbw.de

Problemstellung: Die Inzidenz nicht-übertragbarer, ernährungsbedingter Krankheiten (NCDs) steigt trotz allgemeiner Ernährungsempfehlungen in europäischen Ländern weiter an. Herz-Kreislauf-Erkrankungen, Krebs, Diabetes und Demenz führen zu einem erheblichen Verlust gesunder Lebensjahre [1]. Neben Umwelteinflüssen werden Inzidenz und Schwere dieser Krankheiten durch individuelle Faktoren wie genetisches Profil, Darmmikrobiom und Ernährungspräferenzen beeinflusst.

Personalisierte Ernährung kann das individuelle Risiko dieser Krankheiten deutlich über das Maß der Befolgung allgemeiner Ernährungsempfehlungen reduzieren. Durch die fortschreitende Digitalisierung können individuellen Risikofaktoren in Echtzeit berücksichtigt und maßgeschneiderte Ernährungs- und Gesundheitsempfehlungen entwickelt werden, die zu erwartende qualitätsadjustierte Lebensjahre (QALYs) maximieren. Gleichzeitig sollen hierbei auch ökologische Ziele berücksichtigt werden, da die Lebensmittelproduktion einen erheblichen Anteil an Treibhausgasemissionen ausmacht [2] [3].

Zielsetzung: Es wird ein Lösungsansatz für eine personalisierte Ernährungsstrategie entwickelt, der Gesundheitsförderung, individuelle Akzeptanz, Nachhaltigkeit und Erschwinglichkeit implementiert. Dabei sollen individuelle Risikofaktoren einbezogen und die Risikomodulation durch Lebensmittelgruppen mithilfe von evidenzbasierten Dosis-Wirkungs-Beziehungen berechnet werden. Eine personalisierte Ernährungsstrategie soll durch deterministische Optimierungsalgorithmen erstellt und über eine dynamische App umgesetzt werden. Die App verwendet einen selbst-lernenden Algorithmus mit Serious-Gaming-Elementen und einem Recommender System, um vergangenes Akzeptanzverhalten mit einzubeziehen und die Wirksamkeit der App zu maximieren. Individuelle Echtzeitdaten aus mobilen Endgeräten (Biosensoren) werden die Optimierung individueller Ernährungsstrategien unter Berücksichtigung von systemischen Interaktionen ermöglichen. Diese Methoden machen zeitnahe Feedback und eine sich kontinuierlich optimierende individuelle Anpassung der Ernährungsstrategie möglich.

Methodik: In der ersten Phase des Projekts wird eine Datenbank erstellt, die relevante Datensätze für die

Optimierung enthält, darunter Nährstoffzusammensetzungen von Lebensmitteln, Ernährungsempfehlungen, Erkrankungs- und Sterbetafeln für NCDs, Treibhausgasemissionen aus der Lebensmittelproduktion u.a. enthält. Das individuelle Basisrisiko für die dominierenden NCDs und erwartete QALYs werden unter Berücksichtigung von anthropometrischen Daten, Rauchgewohnheiten, körperlicher Aktivität, den Einflüssen des genetischen Profils (SNPs) und der Zusammensetzung des Darmmikrobioms berechnet. Risikomodulationsmatrizen relevanter Lebensmittelgruppen werden erstellt, um die Risikomodulation und die Veränderungen der zu erwartenden QALYs basierend auf den Ernährungsgewohnheiten des Nutzers zu bestimmen. Mithilfe von Optimierungsalgorithmen wird eine personalisierte Ernährungsempfehlung entwickelt, die die zu erwartenden QALYs maximiert. Ein selbst-lernender Algorithmus im Nutzerinterface wird verwendet, um die Akzeptanz der Empfehlungen zu optimieren. Letztendlich wird in Interventionsstudien die Anwendbarkeit und Massentauglichkeit der App überprüft.

Bisherige Ergebnisse: Ein Großteil der benötigten Datensätze (Sterbe- und Erkrankungstafeln, gegliedert nach Geschlecht und Alter, durch relevante Lebensmittelgruppen modulierbare Risikofaktoren für NCDs u.a.) wurden in einer maschinenlesbaren Datenbank integriert. Eine Implementierung der Berechnung des Gesundheitsimpacts wird derzeit entwickelt und deterministische Methoden zur Implementierung einer individuell optimierten Ernährungsstrategie wurden in Interventionsstudien etabliert [4]. Ein deterministischer, optimierender Ansatz zur Identifikation des best-möglichen Trade-offs zwischen Nährstoffadäquanz, Gesundheitsimpact, Klimafreundlichkeit und Akzeptanz wurde erarbeitet [5].

Ausblick: Eine Vervollständigung der Risikotabellen für NCDs durch den Verzehr von Lebensmittelgruppen, die personalisierte Berechnung des individuellen Gesundheitsrisikos sowie die Einbeziehung des deterministischen Optimierungsmoduls zum Trade-off für Akzeptanz, Umweltverträglichkeit und personalisierter Gesundheitsförderung wird den ersten Teil des Projektes

abschließen. Der Output werden evidenz-basierte, individuell optimierte Ernährungsempfehlungen sein. Im

zweiten Teil wird das Interface zum Nutzer entwickelt werden, welches über den selbst-lernenden Anteil die Balance zwischen der Effektivität der Empfehlungen und der Bereitschaft des Nutzers, den Empfehlungen zu folgen, optimiert. Im weiteren Verlauf werden Faktoren des genetischen Profils und die Zusammensetzung des Darmmikrobioms implementiert werden.

REFERENCES

- [1] World Health Organization (WHO) (2021): Noncommunicable diseases. Online verfügbar unter <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>, zuletzt geprüft am 19.07.2022
- [2] Intergovernmental Panel on Climate Change (IPCC) (2021): Summary for Policymakers. In: V. Masson-Delmotte, P. Zhai, Pirani, A., Connors, S.L., C. Péan, Berger S., N. Caud et al. (Hg.): Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.
- [3] Crippa, M.; Solazzo, E.; Guizzardi, D.; Monforti-Ferrario, F.; Tubiello, F. N.; Leip, A. (2021): Food systems are responsible for a third of global anthropogenic GHG emissions. In: *Nat Food* 2 (3), S. 198–209.
- [4] Eustachio Colombo P, Patterson E, Lindroos AK, Parlesak A, Elinder LS. Sustainable and acceptable school meals through optimization analysis: an intervention study. *Nutr J.* 2020 Jun 24;19(1):61. doi: 10.1186/s12937-020-00579-z.
- [5] Masino T, Colombo PE, Reis K, Tetens I, Parlesak A. Climate-friendly, health-promoting, and culturally acceptable diets for German adult omnivores, pescatarians, vegetarians, and vegans - a linear programming approach. *Nutrition.* 2023 May;109:111977. doi: 10.1016/j.nut.2023.111977.

Fahrspurerkennung für ein autonomes Modellfahrzeug mit Convolutional Neural Networks

Arp, Benjamin
Fakultät Technik
DHBW Stuttgart
Stuttgart, Germany

inf20142@lehre.dhbw-stuttgart.de

Utz, Anton
Fakultät Technik
DHBW Stuttgart
Stuttgart, Germany

inf20005@lehre.dhbw-stuttgart.de

Drüppel, Matthias
Fakultät Technik
DHBW Stuttgart
Stuttgart, Germany

Matthias.Drueppel@dhbw-stuttgart.de

I. EXTENDED ABSTRACT

Künstliche Intelligenz ist eine entscheidende Schlüsseltechnologie, um autonomes Fahren zu ermöglichen [1]. Einen ersten Einstieg in dieses Thema bietet der Carolo Cup, an dem jährlich internationale Teams autonome Modellfahrzeuge entwickeln und gegeneinander antreten lassen [2]. Das Team der Smart Rollerz der DHBW entwirft hierfür jedes Jahr ein neues Auto mit eigener Hard- und Software.

Eine Grundvoraussetzung beim autonomen Fahren und auch beim Lösen der Aufgaben des Carolo Cups ist es, verlässlich und schnell Fahrbahnmarkierungen zu erkennen. In dieser Studienarbeit wurde hierfür ein Convolutional Neural Network (CNN) basierter Algorithmus entwickelt.

Im Forschungsgebiet der Spurerkennung wurden bereits mächtige open Source Modelle entwickelt, wie z.B. *lanedet* von Zheng et. al. [3]. Wir konnten zeigen, dass es nicht trivial ist diese für den speziellen Anwendungsfall des Carolo Cups zu übernehmen. Selbst auf großen Datensätzen vortrainierte Linienerkennungs-Modelle zeigten zuerst keine ausreichende Performance: Die Accuracy der Linien-Detektion liegt nur bei 43%.

Wie lässt sich nun trotzdem das Potential von vortrainierten Netzen nutzen? Ein großes Problem ist der Mismatch in den Datenverteilungen: Vortrainierte Netze wurden typischerweise auf Bildern von echten Straßen trainiert – im Carolo

Cup werden weiße Linien auf einen schwarzen Boden geklebt. Um das Problem der unterschiedlichen Datendarstellungen zu mitigieren, wurde in dieser Arbeit Transfer Learning eingesetzt: Hierfür wurden über 1000 Bilder hand-gelabelt. Auf diese Daten konnte das Modell von Zheng et. al. [3] für den Carolo Cup mit kleiner Learning Rate nach trainiert werden (*fine tuning*).

Eine weitere wichtige Anpassung für den Anwendungsfall des Carolo Cups ist das Verschieben des von dem CNN erfassten Horizonts, sodass mehr Bilddaten verarbeitet werden können. Mit dieser und weiteren Anpassungen wie Hyperparameter-Tuning konnte mit dem fertigen Modell eine hervorragende Performance mit einer Accuracy von 95% auf den Validierungsdaten erzielt werden.

Auch liegt die Performance des CNN-Ansatzes deutlich über einer parallel entwickelten „klassischen“ Lösung mit einer Accuracy von 86%. Ein wichtiger Vorteil der CNN basierten Lösung liegt in ihrer Stabilität und Fähigkeit, Spuren auch über Unterbrechungen hinweg zu vervollständigen.

- [1] Weiyu Hao. Review on lane detection and related methods. In: CognitiveRobotics 3 (2023), URL: <https://www.sciencedirect.com/science/article/pii/S2667241323000186>
- [2] Technische Universität Braunschweig. Carolo-Cup Homepage. 2023. URL: <https://www.tu-braunschweig.de/carolo-cup/>
- [3] GitHub Turoad lanedet. 2023, URL: <https://github.com/Turoad/lanedet>

Re-existing inside of the world of the algorithmic determinism

R. S. Guimaraes

Arts, Science of the Arts

Université Paris 1 Panthéon-Sorbonne

In order to face the climate crisis and the stakes posed by a multipolar world, an education based on the concrete and necessary relationships between human beings and the physical environment, as well as on the heterogeneity of human cultures and collective organisation, seems essential. Cybernetic applications, computers and intelligent machines, AI-related technologies and the environment require the development of an educational framework that helps citizens to overcome the consequences of the inequalities and injustices inherent in the logic of the annexation of human capital by big data. Literally, the need to break patterns - data is created from the past - in order to create a new world, a world that is not deterministically dictated by algorithms. We

are immersed and more than ever reduced to a pure process of computation/commodification, where the symbolic field is replaced by a cybernetic field that achieves a triumphant desanthropologisation. Western modernity is traversed by opposing orientations of mental action (between the real and the ideal), empiricism and idealism. Consequently, the historical agency of concrete existence is demonstrated, existence is also practical. How can one re-exist as a full citizen in an environment that is, as Simon Penny puts it, "focused on reasoning in the form of the manipulation of symbols"? Would it be possible to overcome the bias of the mathematised rationalist legacy when considering the deterministic approach of big data?

Quantum Computing for Feature Selection in Machine Learning

Gerhard Hellstern

DHBW Stuttgart

Stuttgart, Germany

gerhard.hellstern@dhbw-stuttgart.de

Vanessa Dehn

Fraunhofer-Institut für Angewandte Festkörperphysik IAF

Freiburg, Germany

vanessa.dehn@iaf.fraunhofer.de

Martin Zaefferer

DHBW Ravensburg

Ravensburg, Germany

zaefferer@dhbw-ravensburg.de

Index Terms—Machine Learning, Quantum Computing, Supervised Learning, Feature Selection, Optimization, Hybrid Methods

The research project QORA II, funded by the Ministry of Economics, Labor and Tourism Baden-Württemberg, explores possible applications of Quantum Computing (QC) with resilient algorithms. One potential application is optimal feature selection in machine learning [1]. With classical non-QC methods, this NP-hard problem can only be solved approximately; quantum computers might provide an improvement due to their generic characteristics – superposition and entanglement. This improvement may arise in terms of accuracy and efficiency. To this end, the selection problem is transformed into a quadratic optimization problem, which can then be solved using QC algorithms, among other methods [2].

In the first step, we consider low-dimensional problem instances for which a brute force solution is possible. Moreover, we assess well-known machine learning algorithms for feature selection, i.e., Least Absolute Shrinkage and Selection Operator (LASSO) and Recursive Feature Elimination (RFE). This provides a baseline to compare against solutions generated via a QC algorithm, the Quantum Approximate Optimization Algorithm (QAOA) proposed by Farhi et al. [3]. In a second step, for larger problem instances, the solutions of the QC algorithms are compared to non-QC, metaheuristic optimization methods as well IBM’s non-QC commercial solver CPLEX [4]. The QC algorithms are run with quantum simulators and physical quantum hardware from IBM. Although the results on the physical hardware are still overlaid by errors at the qubit level, they show reasonable results.

The investigation shows that formulating feature selection as an optimization problem is a viable way to better constrain the solutions of this NP-hard problem. For the quantum algorithms studied, our results show that a carefully modified and implemented QAOA algorithm [5] is superior to IBM’s standard solution [6].

An in-depth description of this investigation can be found in the preliminary publication [7].

This work is funded by the Ministry of Economic Affairs, Labour and Tourism Baden-Württemberg in the frame of the Competence Center Quantum Computing Baden-Württemberg (project ‘QORA II’).

REFERENCES

- [1] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014. 40th-year commemorative issue.
- [2] A. Milne, M. Rounds, and P. Goddard, “Optimal feature selection in credit scoring and classification using a quantum annealer,” tech. rep., 1QBit, 2017.
- [3] E. Farhi, J. Goldstone, and S. Gutmann, “A quantum approximate optimization algorithm,” 2014. arXiv:1411.4028, <https://arxiv.org/abs/1411.4028>.
- [4] I. I. Cplex, “V12. 1: User’s manual for cplex,” *International Business Machines Corporation*, vol. 46, no. 53, p. 157, 2009.
- [5] S. Brandhofer, D. Braun, V. Dehn, G. Hellstern, M. Hüls, Y. Ji, I. Polian, A. S. Bhatia, and T. Wellens, “Benchmarking the performance of portfolio optimization with QAOA,” *Quantum Information Processing*, vol. 22, Dec. 2022.
- [6] IBM, “QAOA - Qiskit 0.39 documentation,” 2022. Available at <https://qiskit.org/documentation/stubs/qiskit.algorithms.QAOA.html>.
- [7] G. Hellstern, V. Dehn, and M. Zaefferer, “Quantum computer based feature selection in machine learning,” June 2023. arXiv:2306.10591, <https://doi.org/10.48550/arXiv.2306.10591>.

Parallel classification algorithms for big data applications (extended abstract)

Jannik Duerr

Department of Computer Science
Baden-Wuerttemberg Cooperative State University
Horb, Germany
i20007@hb.dhbw-stuttgart.de

Olaf Herden

Department of Computer Science
Baden-Wuerttemberg Cooperative State University
Horb, Germany
o.herden@hb.dhbw-stuttgart.de

Abstract—In this extended abstract we describe our research project concerning parallel classification algorithms.

Keywords—machine learning, classification, classifier, parallelization

I. EXTENDED ABSTRACT

In the field of machine learning, classification is a well-known approach [1][2]. Classification is supervised, i.e. each data record in a data set has a class label. The data set is divided into a training set and test set. Based on the training set as input, a classification algorithm computes a classifier. The quality of the classifier is evaluated by using the test set. Finally, the classifier enables to determine the class of new data records with unknown label.

Traditional classification algorithms are implemented sequentially. In the era of big data, this can lead to performance problems. Therefore, we analyzed parallel classification algorithms [3]. These algorithms work on the training set in parallel and can reduce the time for training a classifier.

First, we give an overview about classification algorithms. These algorithms are divided into the four categories probabilistic, hierarchical, linear and prototype based classifiers. By implementing a prototype, we show that all traditional classification algorithms show time and space complexity problems when the size of the data set used is growing.

Then we present horizontal and vertical approaches for parallelization [4]. Horizontal approaches are map-reduce, peer-to-peer and master-slave concepts, while vertical scaling is realized by multi core CPUs (central processing unit) or GPUs (graphics processing unit). For each category we present the theoretical background as well as its implementation.

Finally, we demonstrate the practical application of parallel algorithms in Python. In our first use case, we implement the Hogwild! algorithm [5] and apply it with different degrees of parallelization to the MNIST (Modified National Institute of Standards and Technology) data set [6]. In the second use case, we apply an implementation of cascade SVM (support vector machines) [7] for comparing with a sequential implementation of SVM by using the two numerical data sets skin segmentation [8] and SUSY (super

symmetry) [9]. In both use cases the time to train the classifier is reduced by using the parallel approach. The speed up factor in these use cases is between 2 and 4.

As future work we see different tasks, e.g. the evaluation of other parallel classification algorithms and the use of other data sets. Moreover, we want to evaluate other parallelization approaches, e.g. the use of GPUs.

- [1] T. Oladipupo. „Types of Machine Learning Algorithms“. In: Introduction to machine learning. Hrsg. von Ethem Alpaydin. 2nd edition. Adaptive computation and machine learning series. Cambridge, Massachusetts: MIT Press, 2010. isbn: 978-953-307-034-6. doi: 10.5772/9385.
- [2] S. Russell and P. Norvig. Artificial intelligence. A modern approach. eng. Third edition. Boston: Pearson, 2016. 151 S. isbn: 9781292153971. url: <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5831883>.
- [3] J. Duerr “Parallel Classification” (in German), Student research work, Baden-Wuerttemberg Cooperative State University Stuttgart Campus Horb, 2023.
- [4] Z. Dafir, Y. Lamari und S. Chah Slaoui. „A survey on parallel clustering algorithms for Big Data“. In: Artificial Intelligence Review 54.4 (2021). PII: 9918, S. 2411–2443. doi:10.1007/s10462-020-09918-2.
- [5] F. Niu, B. Recht, C. Re and S. J. Wright: HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. June 2011. url: <https://arxiv.org/pdf/1106.5730>.
- [6] Y. LeCun, C. Cortes and C. J. C. Burges. The MNIST database of handwritten digits. Courant Institute. May 2013. url: <http://yann.lecun.com/exdb/mnist/> (last visited May 20th 2023).
- [7] H. P. Graf, E. Cosatto, L. Bottou, I. Durdanovic and V. Vapnik „Parallel support vector machines: The cascade SVM“. In: Advances in Neural Information Processing Systems 17 (2004), pages 521-528.
- [8] <https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation>. (last visited May 20th 2023).
- [9] <https://archive.ics.uci.edu/ml/datasets/SUSY>. (last visited May 20th 2023).

CITAI: Building Bridges or Breaking Barriers? Unveiling the Secrets of Citizen Trust in AI Innovations

Sinu Thirukketheeswaran
Empirical Research Center
Duale Hochschule Baden-Württemberg
Stuttgart, Germany
sinuthiru@hotmail.com

Marc Kuhn
Empirical Research Center
Duale Hochschule Baden-Württemberg
Stuttgart, Germany
marc.kuhn@dhbw-stuttgart.de

Lars Meyer-Waarden
TSM Research
Toulouse School of Management
Toulouse, France
lars.meyer-waarden@tsm-education.fr

Monika Gonser
Intersectoral School of Governance
Duale Hochschule Baden-Württemberg
Stuttgart, Germany
monika.gonser@cas.dhbw.de

Maren Lay
Dean Digital Business Psychology
Hochschule Heilbronn
Heilbronn, Germany
maren.lay@hs-heilbronn.de

Benjamin Österle
Economics and Marketing Faculty
Hochschule Heilbronn
Heilbronn, Germany
benjamin.oesterle@hs-heilbronn.de

Gabriel Yuras
Empirical Research Center
Duale Hochschule Baden-Württemberg
Stuttgart, Germany
gabriel.yuras@dhbw-stuttgart.de

Abstract—Urbanization and power demand drive smart city urgency. With 80% of urban population by 2050, Artificial Intelligence (AI) and Internet of Things (IoT) connect sectors for better urban life. AI and IoT drive automation, but trust and ethical concerns persist. Current Research overlooks citizen acceptance and well-being. Our paper proposes AI and IoT smart city assessment, simulating real experiences to measure adoption and aid stakeholders, enhancing long-term success.

I. INTRODUCTION

AI and IoT improve living quality in Smart Environment, Mobility, Living, Governance and Education [1]. They enhance efficiency but prompt privacy and ethical concerns [2]. Validating citizen experiences aids adoption through scenario-based methods [3]. Our research employs simulations for citizen insights, prioritizing trust and stakeholder interests. The research questions can be formulated as:

RQ1: What drives citizen trust and adoption of AI-based smart city solutions, varying by demographics and contexts?

RQ2: How do data privacy and ethics impact trust, well-being, and adoption of smart city solutions?

RQ3: What stakeholder interests matter for sustainable smart city solutions and personalized user services?

II. RESEARCH METHOD

Funded by Schwarz Foundation through Schwarz Beteiligungs GmbH, with support from Lidl Foundation and Kaufland Foundation [4], our research occurs in simulation labs in Stuttgart and Heilbronn. The labs focus on smart mobility and living. One conducts AI and IoT driving simulations, guided by a chatbot, the other showcases an IoT-equipped living apartment for data-driven analysis of citizens' real experiences.

III. RESEARCH CONTRIBUTION

Our study examines consumer views on AI-driven smart vehicles and cities, enhancing consumer behavior literature. As AI enters, consumers shape beliefs through progressive automation and personalized services [5]. We enhance

UTAUT 2 with transformative marketing theories [6], covering well-being and utilitarian motivations. SERVQUAL theory [7], perceived risk theories [8] and trust theory [9] address privacy, ethics and technology risk, complementing cognitive antecedents.

Methodologically, real living and smart mobility labs surpass scenarios, revealing trust, well-being, and intention interactions across automation levels [5]. Mixed methods, including qualitative studies, topic modeling and mobility/living lab experiments, reinforce findings.

In practice, our research informs smart city design, well-being and trust. AI delegation and automation experiences reinforce trust and intentions, aligning with UN sustainability goals, shaping society.

REFERENCES

- [1] C. C. Okafor, C. Aigbavboa, and W. D. Thwala, "A bibliometric evaluation and critical review of the smart city concept—making a case for social equity," *Journal of Science and Technology Policy Management*, vol. 14, no. 3, pp. 487-510, 2023.
- [2] L. Meyer-Waarden, J. Cloarec, C. Adams, D. N. Aliman, and V. Wirth, "Home, sweet home: How well-being shapes the adoption of artificial intelligence-powered apartments in smart cities," *Systèmes d'information et management*, vol. 26, no. 4, pp. 55-88, 2021.
- [3] S. Smys, H. Wang, and A. Basar, "5G network simulation in smart cities using neural network algorithm," *Journal of Artificial Intelligence*, vol. 3, pp. 43-52, 2021.
- [4] Dieter Schwarz Foundation. (2023). About us, August 2023. [Online]. Available: <https://www.dieter-schwarz-stiftung.de/foundation.html#pr ofile> [Accessed Aug. 01, 2023].
- [5] Y. Huang, and L. Qian, "Consumer adoption of electric vehicles in alternative business models," *Energy Policy*, vol. 155, 2021.
- [6] M. J. Sirgy, E. Gurel-Atay, D. Webb, M. Cicic, M. Husic, A. Ekici, ..., and J. S. Johar, "Linking advertising, materialism, and life satisfaction," *Social Indicators Research*, vol. 107, pp. 79-101, 2012.
- [7] A. B. L. L. Parasuraman, V. A. Zeithaml, and L. Berry, "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality," vol. 64, pp. 12-40, 1988.
- [8] P. Slovic, "Perception of risk," *Science*, vol. 236, pp. 280-285, 1987.
- [9] D. H. McKnight, M. Carter, J. B. Thatcher, and F. C. Paul, "Trust in a Specific Technology: An Investigation of Its Components and Measures," *ACM Transactions on Management Information Systems*, vol. 2, 2011.

Realization of an environment for event based vision

Ehret, Felix
DHBW Stuttgart
Stuttgart, Deutschland
inf20177@lehre.dhbw-stuttgart.de

Langner, Erik
DHBW Stuttgart
Stuttgart, Deutschland
inf20049@lehre.dhbw-stuttgart.de

Abstract—This work presents a software framework for computer vision (CV) tasks with event based camera data. Event based cameras are novel imaging sensors that are inspired by the human retinal system. To achieve this, these sensors operate asynchronously, which means that they react to brightness change at pixel level instead of capturing discrete, full frames like a conventional camera does. Each change in brightness triggers a so-called event, which is the data structure for event based vision. An event itself has four data fields that provide information about the captured scene: Two of them represent the coordinates of the triggered event and therefore brightness change, additionally a timestamp is added to allow for communication with current conventional synchronous Computing hardware. Lastly, an indicator of the direction of brightness change called the polarity is added. This polarity indicates whether the change in light intensity was negative or positive, so the transferred information is if the scene got brighter or darker. Since only a change in intensity triggers an event, there is no redundancy in the data transferred to the processing device. When comparing this to conventional imaging sensors, it becomes apparent that this could drastically reduce the energy consumption in the sensing process itself as well as the processing step, since only sparse data has to be analysed instead of dense image data. Other advantages are a high temporal resolution, a very high dynamic range and low power consumption, which makes this type of sensor particularly relevant for scenarios where low latency and power consumption are relevant factors - but it also requires new suitable algorithms for vision tasks, such as adaptations of the likewise biologically inspired Spiking Neural Networks (SNN's) [1]. Especially with regard to sustainability and resource conservation, the combined use of event cameras and SNNs is considered to have great energy-saving potential compared to the established use of conventional cameras and traditional Deep Neural Networks, like Convolutional Neural Networks or Vision Transformers. These possible advantages justify further investigation into computer vision applications using event based sensors. With this work we introduce a novel software framework, which covers most of the typical processing pipeline of event based computer vision. Since most machine learning models for visual analysis are still based on a dense input, the sparse output of the event camera has to be aggregated into what is referred to as a event representation. There are multiple types of representations based on different aggregation techniques. The most promising ones, according to recent research, namely Binary Histograms, Histograms, Event Volume and Time Surface are included in the framework by default [2], [3]. This selection can be extended by users by simply providing an according function. The next step after creating a representation is to input this tensor into a computer vision model. The framework includes two State of the Art models for object detection, but it also allows for an easy specification of own models by the user via Pytorch to enable other use cases.

Finally, to increase the understanding and to enable a qualitative analysis, a visualization module is implemented. This module can be used to either display the different representations for a raw event stream or directly overlay the model predictions on the event stream. These visualization capabilities are meant to alleviate the challenge of understanding this new paradigm of vision computing and bridge the gap between established methods and novel solutions like event based machine learning. To the best of our knowledge, this framework is the first to feature this degree of pipeline coverage while still maintaining the given degree of freedom for user-defined functionality in terms of representations and types of vision applications. Consequently the goal of this work is to support researchers in their research as well as bring more attention to this promising field of research by lowering the barrier of entry through enabling an intuitive and easily understandable entry point.

REFERENCES

- [1] Gallego, Guillermo and Delbrück, Tobi and Orchard, Garrick and Bartolozzi, Chiara and Taba, Brian and Censi, Andrea and Leutenegger, Stefan and Davison, Andrew J. and Conradt, Jörg and Daniilidis, Kostas and Scaramuzza, Davide “Event-Based Vision: A Survey,” *Phil. Trans. Roy. Soc. London*, doi 10.1109/TPAMI.2020.3008413, pp. 154-180, 2022.
- [2] Perot, Etienne, et al., “Learning to detect objects with a 1 megapixel event camera,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 16639-16652, 2020.
- [3] Mathias Gehrig and Davide Scaramuzza, “Recurrent Vision Transformers for Object Detection with Event Cameras,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Multi Object Tracking using Machine Learning

Christian Alexander Holz
AE – Active Safety & Automation
Daimler Truck AG
Stuttgart, Deutschland
christian.c.holz@daimlertruck.com

Christian Bader
AE - Active Safety & Automation
Daimler Truck AG
Stuttgart, Deutschland
christian.bader@daimlertruck.com

Matthias Drüppel
Fakultät Technik
Duale Hochschule Baden-Württemberg
Stuttgart, Deutschland
matthias.drueppel@dhbw-stuttgart.de

I. EXTENDED ABSTRACT

Im Bereich der Advanced Driver Assistance System (ADAS) werden unterschiedliche Sensortechnologien für die Wahrnehmung des eigenen Umfelds eingesetzt. Die Sensoren zur Erfassung des Umfelds haben in der Regel eine integrierte Electronic Control Unit (ECU), welche die Rohdaten verarbeiten und dem Fahrzeug unter anderem die erkannten Sensor Objekte (SO) (z.B. Fahrzeuge, Fußgänger, etc.) und deren Zustände zur Verfügung stellen. Die Herausforderung für Fahrzeughersteller besteht nun darin die Zustandsinformationen der gelieferten SO in der ADAS ECU so zu verarbeiten, dass real existierende Objekte verfolgt und plausible Trajektorien (Tracks) für die aktiven Sicherheitssysteme (Fahrzeugfeatures) bereitgestellt werden. Um diese Aufgabe zu realisieren, werden daher Trackingverfahren eingesetzt.

Dabei kann die Herausforderung der Multi Objekt Verfolgung (Multi Object Tracking (MOT)) in folgenden Teilaufgaben beschrieben werden [1]:

- **Zustandsvorhersage** i.e. Prädiktion (Prediction) aller Zustände der erkannten SO im nächsten Zeitschritt.
- **Assoziation** - Es müssen die SO (Messungen) den prädierten Objektzuständen (Tracks) zugeordnet werden.
- **Update** - Abweichungen zwischen den vorhergesagten und gemessenen Zuständen müssen pro Zeitschritt entsprechend korrigiert werden.
- **Trackmanagement** - Unplausible, nicht mehr relevante Tracks müssen entfernt und hinzukommende, bestätigte Sos als neue Tracks aufgesetzt werden.

In aktuellen Serienentwicklungsprojekten von aktiven Sicherheitssystemen haben sich einige Filtervarianten (wie etwa der Kalman Filter (KF)) für die dynamische Zustandsschätzung und einige Assoziationsverfahren (wie etwa Global Nearest Neighbor, Joint Probability Data Association) etabliert. Die Grenzen der eingesetzten Verfahren zeigen sich in der Praxis jedoch durch die Notwendigkeit von heuristischen Programmregeln, welche in der Systementwicklung für unterschiedliche Szenen ermittelt werden in denen die Verfahren Schwächen aufzeigen. Machine Learning (ML) Verfahren ermöglichen es ein datengetriebenes Modell zu entwickeln, welches kontextbasierten Regeln auf Basis der verfügbaren Daten erlernt hat. Durch die Entwicklung einer entsprechenden

Simulationsumgebung lassen sich Modell und somit Performance weiter datenbasiert optimieren. Theoretisch kann ein Modell so im Entwicklungsprozess mit steigender Datenverfügbarkeit verbessert werden, ohne dabei den Bedarf an Ressourcen für die Implementierung auf der Zielhardware zu erhöhen. [2]

Innerhalb dieser Arbeit wird eine Recherche zu aktuell bekannten ML basierten Trackingansätze durchgeführt. Die Rechercheergebnisse werden diskutiert und es wird bewertet, welche Machine Learning Modelle sich für die Integration in einen bestehenden KF MOT Ansatz eignen. Des Weiteren wird ein Framework für das Sensor Objekt (SO) Tracking vorgestellt, im welchem die Teilaufgaben des MOT durch den Einsatz von trainierten Recurrent Neural Network (RNN) gelöst werden.

Auf Basis einer bereits existierenden Kalman Filter Implementierung wird ein Netzwerk sowohl für die Zustandsvorhersage als auch für die Datenassoziation implementiert. Das Netzwerk für die Vorhersage der Zustandswerte einzelner Tracks (Single Prediction Network (SPENT)) konnte so trainiert werden, dass ein Root Mean Squared Error (RMSE) von 0.1 auf einen repräsentativen Testdatensatz (Split 80 / 10 / 10, KITTI) erzielt werden kann. Die Netzwerkentwicklung der Einzeldatenzuordnung (Single Association Network (SANT)) und die Netzwerkerweiterung auf eine m zu n Assoziation (Multi Association Network (MANTa)) zeigt eine 80-90 % korrekte Zuordnung.

Die Netzwerkentwicklungen erfolgen dabei in Modulen und ermöglichen eine Erweiterung zu Multi Sensor Verfahren. Verfahren wie Bayesian Neural Networks (BNN), welche die Unsicherheiten von NN Prognosen quantifizieren, werden diskutiert (für das Trackmanagement entscheidend). Die Evaluation zeigt einen Performance-Vergleich der verschiedenen Algorithmen mithilfe eines State-of-the-Art Tracking Datensatzes (KITTI). [4]

- [1] Ba Tuong Vo, code set for academic/research use: Multi-Sensor Multi-Target Tracking, 2013, URL: <https://ba-tuong.vo-au.com/codes.html>.
- [2] J. Pallauf, Objektsensitive Verfolgung und Klassifikation von Fußgängern mit verteilten Multi-Sensor-Trägern, 2016, Dissertation
- [3] A. Milan, Online Multi-Target Tracking Using Recurrent Neural Networks, 2016, Paper
- [4] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets Robotics: The KITTI Dataset, 2013, Paper

Stock Market Prediction System using Hybrid Model

Akshat Singh
School of Computer Science and
Engineering
Lovely Professional University
Phagwara, Punjab
raghuvanshiakshatsingh@gmail.com

Virrat Devaser
School of Computer Science and
Engineering
Lovely Professional University
Phagwara, Punjab virrat.14591
@lpu.co.in

Abstract— In recent years, the stock market has become a significant factor in the financial sector and plays a crucial role in the economy. The aim of this research paper is to design and implement a stock market prediction system using hybrid algorithms that combine the Auto-regressive Integrated Moving Average (ARIMA) and linear regression. The data set used in this study was collected from the National Stock Exchange (NSE) web-site and contains information on stock prices, volume, and other relevant factors. The ARIMA model was used to forecast future trends in the stock market, while the linear regression model was used to analyze the relationship between stock prices and other relevant factors. The results of the study showed that the hybrid algorithm provided more accurate predictions compared to using a single algorithm. This system can be used by investors, traders, and financial institutions to make informed decisions about buying and selling stocks. The hybrid algorithm approach provides a novel approach to stock market prediction and can be applied to other financial markets as well. The results of this research provide a promising direction for future studies in stock market prediction.

Keywords— *ARIMA, Linear Regression, NSE, Stock Prices Prediction*

I. INTRODUCTION

Many traders, investors, and market participants regard the stock market as a large and valuable passive income source. They are better equipped to estimate future market behaviour since they have a detailed awareness of the current status of the stock market. Stock closing prices are used by buyers and sellers to choose which stocks to buy and when to buy them in order to optimize investing returns. [1]. Companies can avoid losses and save millions of dollars by adopting stock market forecasting technologies that are more accessible than ever. It is possible to precisely anticipate the status of the stock market based on existing trends, allowing investors to detect prospective opportunities and risks. Unfortunately, the data sets used to estimate the closing price of a stock have intrinsic limits. They contain unpredictable and seasonal data variances, which have a significant impact on the accuracy of the results. The two most trustworthy sources for stock price predictions are social media and macroeconomics [2]. Social issues are frequently discussed in blogs, news stories, and private text, audio, and video comments. As news organizations update their social media platforms with stock market value information, a massive data stream is created [3] [4]. Seasonal oscillations, change in any trend of long-term, as well as short- to medium-term are all covered by macroeconomic subfields. Despite the fact that beginning macroeconomics usually conceals substantial changes, the long-term direction of trends and the medium-term business cycles are the primary goals of financial time-series studies. In addition to data purification, noise reduction is required for

data processing. Recent research has demonstrated that for forecasting financial time series, the Support Vector Machine (SVM) prediction model outperforms alternative neural network methods. The SVM prediction model has been used in one of the study to forecast the future trajectory of stock price indexes. Other analysts investigated share closing prices using a variety of hybrid methodologies. In a single study by Mohan, S. [5] and Mullapudi, the Hodrick-Prescott filter and Support Vector Regression were used to improve stock price predictions (SVR). Because of their capacity to replicate nonlinear data correlations, Artificial Neural Networks (ANNs) are a popular choice for market forecasting. The k-NN approach was used to calculate the stock values of six firms, and the KSE-100 Index was anticipated for nearly three years using artificial neural networks (ANNs). [6]. ARIMA is a well-known statistical strategy for forecasting stock prices. One study examined closing prices of an exchange stock of 2010-2018 using the Autoregressive Integrated Moving Average (ARIMA) model. The data was utilized to estimate Amman stock prices on the Jordanian market using the ARIMA model, demonstrating the efficiency of the suggested prediction method for the stock. RNNs were originally invented to analyse the data of the time-series, but their accuracy has led to the development of deep learning-based stock prediction systems [7] [8]. In one study, a deep convolutional neural network was employed to forecast stock market events, while an LSTM model was utilized in the other to examine firm valuations. Many research have been conducted on the usage of machine-learned classifiers and the high levels of noise in financial data. This study blends learning time series and textual data with technical and content characteristics to estimate stock closing prices [9] [10]. Predicting stock values is challenging due to the turmoil and unpredictability of the stock market. To overcome the limits of traditional forecasting methodologies, this project intends to create a prediction model that combines technological and content-based components. Deep learning approaches and machine learning algorithms are utilized to evaluate market price [11]. We consider content issues as well as technical data, such as sentiment analysis of tweets concerning the stock. We gather insights from textual input for the prediction model using natural language processing techniques. We use Hodrick-Prescott (HP) filters and other techniques such as noise-filter to reduce the noise effect on the predictions accuracy. This filter removes noise and smoothes previous stock price data to improve prediction accuracy. Using SVM, k-NN, and ARIMA, among other tools, we assess and contrast the proposed model accuracy with that of well-known techniques. We also forecast stock values using deep learning algorithms such as LSTM and auto-encoders (SAEs) [12]. Our suggested model provides a mechanism for predicting stock

prices that takes both technical and content aspects into account. By integrating traditional and deep learning-based methodologies, we want to provide more precise and reliable forecasts to stock market traders and investors. Our findings contribute to the corpus of knowledge on the application of machine learning algorithms to financial time-series data forecasting

II. REVIEW OF LITERATURE

A developing trend found in recent research on stock market forecasting utilizing hybrid algorithms is the combination of many approaches with deep learning algorithms. This was discovered through an examination of existing research [13]. Historically, a number of research have been carried out with the goal of predicting stock values by utilizing a variety of machine learning and deep learning techniques. Within the scope of their research, Gadekallu, T.R. and Manoj investigated a variety of hybrid deep learning algorithms for the prediction of stock prices [14]. Estimating stock prices using a hybrid method that combined LSTM as well as the CNNs was their recommended approach. The authors D. Selvamuthu, V. Kumar, and A. Mishra evaluated the effectiveness of three neural network learning methods by forecasting over tick-by-tick and 15-minute datasets. These techniques are known as Levenberg-Marquardt, Scaled Conjugate Gradient, and Bayesian Regularization. When compared to predictions made using data collected every 15 minutes, the outcomes of stock market predictions made using tick-by-tick information were significantly more accurate [15]. In this particular study, E. Faria and colleagues examined the primary index of the Brazilian stock market through the lens of adaptive exponential smoothing and artificial neural networks. The neural networks beat other models in predicting the correct sign of the index return, with a ratio of 0.60 indicating the number of times the correct market movement was expected. This indicates that the neural networks correctly anticipated the market movement 60% of the time [16]. This study made stock price projections one day, five days, and 22 days into the future by utilizing a support vector machine and various windowing methods. According to the findings, SVR models that include rectangular and flattened windows are more successful than other types of SVR models at projecting stock prices with minimum inaccuracy [17]. The author of [18] Both A. Pathak and S. Pathak utilized tried and true machine learning strategies when attempting to forecast stock market prices. It was determined to use KNN, SVM, Random Forest, and Logistic Regression. Following an examination of the performance indicators, they came to the conclusion that the algorithm with the highest effectiveness was Random Forest, which had an accuracy rate of 80.7%. Deep learning algorithms have a tremendous amount of untapped potential in the financial sector, as A. Sharma as well as U. Singh found out when they investigated the feasibility of stock market forecasting using ML techniques. In the research cited in [19], it was shown that deep learning systems performed significantly better than standard approaches when predicting the impact of news sentiment analysis on stock price. In a head-to-head competition, standard machine learning strategies were surpassed by deep learning algorithms when it came to predicting stock values. The use of machine learning techniques allowed [20] and [21] [23] found that increasing the use of sentiment analysis and deep learning in the process of estimating stock prices increased the accuracy of the forecast.

The findings of the study suggest that improving stock market estimates could be accomplished by integrating many methods, such as those involving deep learning algorithms. [24].

III. PROPOSED MODELS

We present an innovative hybrid method based on a model that combines linear regression and ARIMA machine learning techniques. The financial dataset is refined and noise is removed using this hybrid method. Stock price features are examples of innovative characteristics. The machine learning techniques ARIMA and linear regression were used in this study. Figure 1 depicts the proposed model's specifications. The method we propose combines historical stock data with Twitter tweets. The six-year historical stock data set from Tata Consultancy Services Limited includes daily stock data (TCS). The daily opening and closing stock prices, as well as the highest, lowest, and average stock prices, are highlighted, as is the total number of shares traded. The FMHP filter removes historical stock data's cyclical and trending components. The trend component is added to the training model after the cyclical component is removed from the stock price data. Using recent data, the program accurately predicts the stock's closing price.

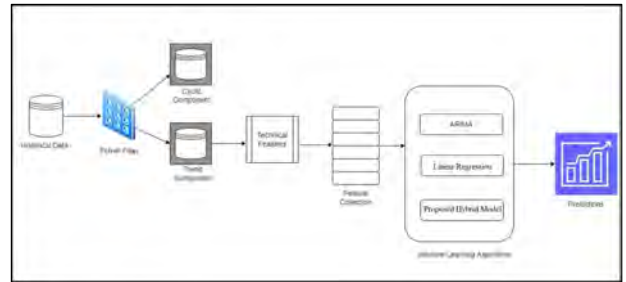


Fig. 1. The Proposed Model

A. Datasets Used:

The financial data was obtained using Yfinance, a popular Python program for obtaining web data. We used this model to collect the necessary historical stock price data to estimate stock closing values. The dataset's attributes are listed below.

1. *Input Datasets:* We chose TCS for our assessment after reviewing six years of transaction data, which included price changes in the stock (the opening, highest, lowest, and closing values of the stock price for each day). We are aware that the previous day's closing price influences the starting price. To properly calculate the starting price of the stock, it is necessary to consider both the company's closing price and its open price. However, the minimum and maximum values are retained.

B. Data Pre-processing:

Every machine learning model must use the data pre-treatment technique. We used a sophisticated filter on the raw data to reduce noise and find the financial trend component. Time series statistics frequently conceal the information contained within them. Disordered timestamps, missing values (or timestamps), outliers, and data noise are the most prevalent issues with time series. When missing values are included, the solution to the aforementioned problems

becomes more complicated. It is extremely recommended that noise components be removed from a time series before developing a model because noise can cause major difficulties. De-noising is a technique for decreasing noise exactly.

The rolling mean is the simple mean of a window of earlier observations and is formed of a string of numbers from the time series data, figuring out each arranged window's meaning. This greatly decreases noise in time series data.

1. *Prediction Models:* ARIMA and Linear Regression, two machine learning approaches, were applied to forecast stock closing prices. These concepts are further upon in the sections that follow.

a) *Autoregressive Integrated Moving Average (ARIMA)* : Moving average (MA), integrated (I), and autoregressive (AR) time series forecasting methodologies are combined in the Autoregressive Integrated Moving Average (ARIMA) time series forecasting model. ARIMA models are commonly used in econometrics, finance, and other fields where future values must be extrapolated from historical data. The model parameters p, d, and q represent the moving average order, degree of differencing, and order of autoregression, respectively. An ARIMA(p, d, q) model has the following formula:

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t$$

where B is a lag operator and

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

also

$$\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$$

Before we can understand ARIMA, we must first look at its fundamental components.

- AR
- I
- MA

The ARIMA model takes in three parameters:

- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing

The AR component of ARIMA describes the relationship between the most recent observation and previous observations. The MA component, on the other hand, depicts the relationship between the current observation and the errors of previous observations. The component represents the differencing process, which converts a non-stationary time series into a stationary time series by computing the difference between successive observations. These three characteristics enable the ARIMA model to recognize complex patterns in time series data and generate accurate forecasts.

ARIMA models can be expanded to include seasonal factors or external influences, depending on the type of data and the goal of the study. To assess the model's fit to the data, several statistical measures, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), can be used (BIC). To identify patterns in time series data, the

ARIMA time series forecasting model combines autoregressive, moving average, and differencing components. It is a widely used and adaptable method for predicting the data of time-series. The layout of the model can be altered, and it can be expanded to include more elements or variables. ARIMA can be a very effective method for predicting future values of a time series with careful parameter selection and model estimation[25].

b) *Linear Regression* : Linear regression [26] is a type of regression analysis in which the dependent and independent variables have a linear relationship. The linear regression algorithm tries to forecast future values by minimizing the residual sum of squares, also known as the sum of squares of observed and predicted values (RSS). To accomplish this, the line with the best fit and the lowest RSS value is chosen. To determine the parameters of linear regression models, ordinary least squares (OLS) regression is frequently used[27]. The Gauss-Markov assumptions that underpin OLS regression are based on a number of theories. Residual normality, homoscedasticity, independence, and linearity are among the assumptions. The independent variables in cross-sectional data should be random variables unrelated to the error term [28]. These assumptions allow us to estimate the coefficients of the linear regression model and draw valid statistical conclusions. Because different columns derive from the same process, the assumption that the independent variables are all random variables cannot be satisfied for time series data. To compensate, the requirements for autocorrelation, homoscedasticity, and endogeneity are more stringent [29]. Autocorrelation, or the correlation between residuals at different time points, must be considered in time series regression analysis. Endogeneity deals with the problem of reverse causation, which occurs when the dependent variable influences the independent variable rather than vice versa. The assumption of homoscedasticity is that the variance of residuals remains constant over time. The goal of linear regression is to minimize the sum of squares of the differences between observed and predicted values. The best fit line for a given dataset is then identified using regression methods. The cross-sectional data assumptions in OLS regression are less stringent than the time series data assumptions[30]. When performing regression analysis on time series data, autocorrelation, homoscedasticity, and endogeneity must be taken into account.

IV. RESULTS AND DISCUSSIONS

A. Experimentation based setup:

In this, we attempted to forecast stock closing prices one day in advance. We trained a model that is able to predict the closing price of a particular stock on the following trading day using six years of historical data. Both training as well as the test data were processed by recursive rolling method. The phase space reconstruction method was used to transform data of time-series into a M x N matrix. Here, "M" is representing the number of days and the number of samples have been represented by "N"[31]. Before the trials, the data were divided into training as well as the testing datasets (respectively, 80% and 20%). The optimal classifier settings

were determined using cross-validation. The optimal settings for each classifier were determined using a grid search.

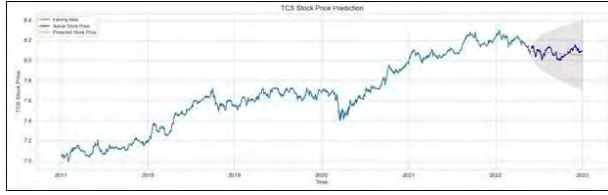


Fig. 2. Accuracy of Prediction of TCS closing Price

B. ARIMA Results:

ARIMA model projection results for the closing price of TCS stock have been summarized in the figures 2 and 3. ARIMA performed best with MAPE and RMSE values of 0.0414 and 0.001, respectively. With respective MSE, MAE, RMSE, and MAPE error rates of 0.001, 0.035, 0.0414, and 0.04, the accuracy of the models' predictions is 80%.

```
# report performance
mse = mean_squared_error(test_data, predictions)
print('MSE: '+str(mse))
mae = mean_absolute_error(test_data, predictions)
print('MAE: '+str(mae))
rmse = math.sqrt(mean_squared_error(test_data, predictions))
print('RMSE: '+str(rmse))
mape = np.mean(np.abs(predictions - test_data)/np.abs(test_data))
print('MAPE: '+str(mape))

MSE: 0.001716546218844261
MAE: 0.03541302048158317
RMSE: 0.0414312227534291
MAPE: 0.0404380543073524523
```

Fig. 3. MSE, MAE, RMSE, MAPE

C. Linear Regression Results:

Figure 4 depict the results of the linear regression model used to predict the closing price of TCS. The highest MAPE and RMSE linear regression values on the exam were 0.03% and 0.01%, respectively. The model's 99.97% accuracy in forecasting was astounding.

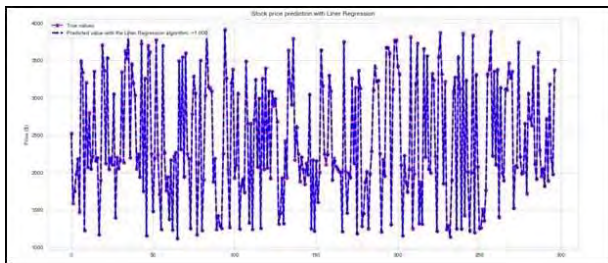


Fig. 4. Performance of Linear Regression in predicting the closing price of TCS

D. Hybrid Algorithm Results:

Fig. 5 displays the achieved results for the Hybrid LINEAR REGRESSION plus ARIMA model for the prediction of the TCS closing price. The best results obtained for Hybrid ARIMA+LINEAR REGRESSION with RMSE as well as the MAPE on the test were 17.656 and 311.74 respectively. The accuracy achieved for the model was 85%. Even though the accuracy is low, we can use other ML and DL algorithms in future to enhance the accuracy of our forecasting model

because ARIMA is suited for forecasting model but Linear Regression is mainly a prediction model.



Fig. 5. Performance of Hybrid Algorithm in predicting the closing price of TCS

V. DISCUSSION

When compared to parameters that are closer to linear trends, the AR, I, and MA reduce time series deviations. By incorporating trend and cyclical factors and reducing the endpoint base, MA and AR outperform the HP filter (EPB)[32]. We did a series of tests using various methodologies to determine the efficacy of stock price forecasting and our proposed strategy. Prediction accuracy was improved by combining technical and content characteristics. Using the original data, various noise reduction techniques such as HP and AR, MA and I, and machine learning algorithms were evaluated. HP, AR, MA, and I discovered that we can reduce noise and improve model predictions. By complementing one another, the technology and content elements reduce the MAPE as well as the RMSE error rates.

VI. CONCLUSION

The optimal as the noise reduction techniques as well as machine learning for estimating the closing price of any stock has been investigated in this study. Technical elements and additional indicators were compiled from historical stock data. Forecast accuracy was improved by combining technical and content features. A standard machine learning technique as well as a time-series-based technique were used. On six years of historical data, we ran a series of experiments using machine learning algorithms to determine the value of TCS stock [33]. Our ensemble approach produces a powerful predictive tool for evaluating stock market price prediction, content, and financial time series by combining machine learning and time series models with innovative technical as well as the content aspects. Time series data, macroeconomic concerns, the news, and other external factors all have an impact on stock market prices. Several issues raised by these constraints must be addressed in future research. Future research will focus on accelerating deep learning systems and improving prediction accuracy for larger historical data sets. Furthermore, we would like to validate the proposed distinguishing characteristics using a more diverse dataset. Changing hyper-parameters is also difficult. An automated method for selecting hyper-parameters will be used to obtain the best value.

REFERENCES

- [1] Sukono, M., Napitupulu, H., Sambas, A., Murniati, A., Kusumaningtyas, V.A.: Artificial neural network-based machine

- learning approach to stock market prediction model on the indonesia stock exchange during the covid-19. *Engineering Letters* **30**, 988–1000 (2022). <https://doi.org/10.11648/j.eng.2022.030.14>
- [2] Das, N., Sadhukhan, B., Chatterjee, T., Chakrabarti, S.: Effect of pub-lic sentiment on stock market movement prediction during the covid-19 outbreak. *Social Network Analysis and Mining* **12**, 92 (2022)
- [3] Liu, W., Yang, Z., Cao, Y., Huo, J.: Discovering the influences of the patent innovations on the stock market. *Information Processing and Management* **59**, 102908 (2022). <https://doi.org/10.1016/j.ipm.2022.102908>
- [4] Feng, X., Du, R., Zhang, Y.: A dual-stage neural network approach for stock price prediction. *Expert Systems with Applications* **181**, 115053 (2021)
- [5] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., Anastasiu, D.C.: Stock price prediction using news sentiment analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 205–208 (2019). <https://doi.org/10.1109/BigDataService.2019.00035>
- [6] Marri, A.A., Ghulam, M., Talpur, H.: Evaluation of stochastic and ann model for karachi stock exchange prices prediction. *International Transaction Journal of Engineering, Management, Applied Sciences Technologies* **13**, 1–11 (2022)
- [7] Cao, Q., Zhu, Y., Shi, Z., Sun, X.: A review of machine learning-based stock trading strategies and future directions. *IEEE Transactions on Industrial Informatics* **17**, 7918–7928 (2021).
- [8] Liu, H., Lu, H., Zhai, X.: A deep learning-based trading model combining multiple technical indicators for stock price prediction. *Neural Computing and Applications* **33**, 12441–12452 (2021).
- [9] Wang, Y., Gou, X., Huang, Z.: A deep learning-based stock prediction system using financial news and technical indicators. *Journal of Ambient Intelligence and Humanized Computing* **13**, 1503–1514 (2022)
- [10] Goh, T.S., Henry, H., Albert, A.: Determinants and prediction of the stock market during covid-19: Evidence from indonesia. *Journal of Asian Finance, Economics, and Business* **8**, 1–6 (2021). <https://doi.org/10.13106/jafeb.2021.vol8.no1.1>
- [11] Devaser, V., Chawla, P.: A novel anomaly detection approach for nifty stocks using machine learning for construction of efficient portfolio to reduce losses and protect gains. *Journal of Computer Science* **18**(5), 441–452 (2022). <https://doi.org/10.3844/jcssp.2022.441.452>
- [12] Rameh, T., Abbasi, R., Sanaci, M.: Designing a hybrid model for stock marketing prediction based on lstm and transfer learning. *International Journal of Nonlinear Analysis and Applications* **12**, 2325–2337 (2021) <https://doi.org/10.22075/ijnaa.2021.23620.2429>
- [13] Arashi, M., Rounaghi, M.M.: Analysis of market efficiency and fractal feature of nasdaq stock exchange: Time series modeling and forecasting of stock index using arma-garch model. *Future Business Journal* **8**, 14 (2022). <https://doi.org/10.1186/s43093-022-00088-z>
- [14] Gadekallu, T.R., Manoj, M.K., Kumar, N., Hakak, S., Bhattacharya, S.: Blockchain-based attack detection on machine learning algorithms for iot-based e-health applications. *IEEE Internet of Things Magazine* **4**, 30–33 (2021). <https://doi.org/10.1109/MIOT.2021.3093202>
- [15] Selvamuthu, D., Kumar, V., Mishra, A.: Indian stock market prediction using artificial neural networks on tick data. *Financial Innovation* **5**, 16 (2019). <https://doi.org/10.1186/s40854-019-0131-7>
- [16] Faria, E., Albuquerque, M., Gonzalez, J.L., Cavalcante, J.T.P., Albuquerque, M.: Predicting the brazilian stock market through neural networks and adaptive exponential smoothing methods. *Expert Systems with Applications* **36**, 12506–12509 (2009). <https://doi.org/10.1016/j.eswa.2009.04.032>
- [17] Meesad, P., Rasel, R.I.: Predicting stock market price using support vector regression. 2013 IEEE International Conference on Industrial Engineering and Engineering Management, 1029–1033 (2013). <https://doi.org/10.1109/ICIEV.2013.6572570>
- [18] Pathak, A., Pathak, S.: Study of machine learning algorithms for stockmarket prediction. *International Journal of Engineering Research Tech-nology* **9**(6), 186–189 (2020)
- [19] Sharma, A., Bhuriya, D., Singh, U.: Survey of stock market prediction using machine learning approach. 2017 International Conference on Emerging Computing Technologies and Applications (ICECA), 1–4 (2017)
- [20] Nikou, M., Mansourfar, G., Bagherzadeh, J.: Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management* **26**(1), 1–17 (2019)
- [21] Jeevan, B., Naresh, E., Kumar, B.P.V., Kambli, P.: Share price prediction using machine learning technique. 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 1–4 (2018). <https://doi.org/10.1109/CIMCA.2018.8739647>
- [22] Devaser, V., Chawla, P., Rana, A.: Dynamic approach to estimate performance and minimize losses in stocks, 32–37 (2018). <https://doi.org/10.1109/IC3I44769.2018.9007298>
- [23] Xu, Y., Keselj, V.: Stock prediction using deep learning and sentiment analysis. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 5573–5580 (2019). <https://doi.org/10.1109/BigData47090.2019.9006342>
- [24] Thakkar, A., Chaudhari, K.: A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications* **177**, 114800 (2021). <https://doi.org/10.1016/j.eswa.2021.114800>
- [25] Shankar, R.L.: Mispricing in single stock futures: empirical examination of indian markets, 1–36 (2015)
- [26] Andrade de Oliveira, F., Enrique Zarate, L., de Azevedo Reis, M., Neri Nobre, C.: The use of artificial neural networks in the analysis and prediction of stock prices. 2011 IEEE International Conference on Systems, Man, and Cybernetics, 2151–2155 (2011)
- [27] Setnes, M., Van Drempt, J.H.: Fuzzy modeling in stock-market analysis, 250–258 (1999)
- [28] Choudhary, R.K., Dey, N., Ashour, A.S.: A hybrid approach of data pre-processing and machine learning techniques for predicting stock prices. *Computers Electrical Engineering* **96**, 107345 (2022)
- [29] Devi, S., Devaser, V.: Stock market price prediction using sap predictive service: Second international conference, icaicr 2018, shimla, india, july 14–15, 2018, revised selected papers, part i, 135–148 (2019). https://doi.org/10.1007/978-981-13-3140-4_13
- [30] Van Horne, J.C., Parker, G.G.C.: Theory: An empirical test (1967)
- [31] Chen, D., Bin, F., Chen, C.: The impacts of political events on foreign institutional investors and stock returns : Emerging market evidence from taiwan **10**(2) (2005)
- [32] Hiemstra, Y.: A stock market forecasting support system based on fuzzy logic. *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences* **3**, 281–287 (1994)
- [33] Dupernex, S.: Why might share prices follow a random walk? *Stud. Econ. Rev.* **21**, 167–179 (2007)

Towards Learned Cost Estimation for Stream Processing Queries on Heterogeneous Hardware

Roman Heinrich*, Manisha Luthra†, Harald Kornmayer*, Carsten Binnig†

*DHBW Mannheim

†TU Darmstadt & DFKI GmbH

Optimization for DSPS. Distributed Stream Processing Systems (DSPS) are designed to face the increasing demand for fast- and real-time processing of data streams by distributing computation units on interconnected hardware resources. The optimization of DSPS is needed to meet specific quality requirements, such as the realization of low processing latencies or high throughputs [1]. Various methods have been proposed [2] to optimize these systems, e.g., by (1) optimally distributing computing operators on the underlying hardware or (2) by determining an optimal parallelism degree. However, optimizing query execution for DSPS is challenging, as various workloads need to be supported while the underlying hardware and network resources are heterogeneous. In addition, these systems typically face workload changes due to the dynamic nature of data streams.

What is missing? Existing optimization approaches for DSPS typically assume hardware and network homogeneity [3]–[5]. In contrast, modern IoT-Scenarios use heterogeneous hardware and network resources that greatly affect the performance of DSPS queries. For instance, placing an operator on a computing node with high network latencies will induce a higher overall query latency. Likewise, a low-performing edge device with restricted CPU resources will affect the overall query costs if too many compute-heavy operators are executed on that. Many recent approaches do not consider this heterogeneous computing and network environment, i.e., they cannot accurately predict quality metrics for DSPS.

Cost-based optimization. In this work, we propose *learned cost models* as an important component to optimize DSPS. Such a model estimates the query execution costs without the need for execution. This allows to optimize the query execution by reasoning over optimizations in a *what-if*-mode. In other words, a learned cost model allows to answer the question: “*What will be the execution costs of a given query when applying a given DSPS optimization?*”. Comparing the resulting cost metrics for various optimization configurations then allows one to select an optimal configuration in terms of latency or throughput.

Our Approach. To obtain precise cost estimations, we envision a new learned cost model that considers heterogeneous hardware and network and extends previous work [3] to predict latency and throughput. The main idea is that the model relies on *zero-shot learning*, which allows it to generalize to data unseen streams, queries, and hardware resources that differ from the initial training range. This is achieved by training

the model on various queries, data streams, and hardware resources that are described with so-called *transferable features*. The advantage of these features is that they can be applied and are meaningful on *any* data stream, query, and hardware resource. A significant benefit is that the model needs to be trained just once and then can generalize to unseen queries and hardware. This work proposes a common representation that describes the data streams, query operators, and hardware resources to make it usable for Machine Learning using Graph Neural Networks (GNN). Moreover, various strategies are presented to enumerate data streams, query operators, heterogeneous resources, and operator placements to create the training dataset, which is the foundation for the proposed zero-shot model.

Initial Results In an initial evaluation, Apache Storm [6] was used as DSPS to execute various queries while observing their cost labels, serving as training data for the model. Preliminary results showed a prediction capability with a median q-error of up to 1.34 for throughput predictions on an unseen test dataset. Furthermore, results show that the model can predict costs for hardware properties and published benchmarks [7] that were unseen during the training. These promising results show that the model has the potential to be used as a major building block in cost-based optimization for DSPS.

REFERENCES

- [1] Z. Shao, “Real-time analytics at facebook,” *XLDB*, 2011.
- [2] M. Hirzel, R. Soulé, S. Schneider, B. Gedik, and R. Grimm, “A catalog of stream processing optimizations,” *ACM Computing Surveys*, vol. 46, no. 4, 2014.
- [3] R. Heinrich, M. Luthra, H. Kornmayer, and C. Binnig, “Zero-shot cost models for distributed stream processing,” in *Proceedings of the 16th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS ’22, 2022, p. 85–90.
- [4] C. Wang, X. Meng, Q. Guo, Z. Weng, and C. Yang, “Automating characterization deployment in distributed data stream management systems,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2669–2681, 2017.
- [5] L. Aniello, R. Baldoni, and L. Querzoni, “Adaptive online scheduling in storm,” in *Proceedings of the 7th ACM International Conference on Distributed Event-Based Systems*, ser. DEBS ’13, 2013, p. 207–218.
- [6] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal, and D. Ryaboy, “Storm@twitter,” in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014, p. 147–156.
- [7] M. V. Bordin, D. Griebler, G. Mencagli, C. F. R. Geyer, and L. G. L. Fernandes, “DSPBench: A suite of benchmark applications for distributed data stream processing systems,” *IEEE Access*, vol. 8, pp. 222 900–222 917, 2020.

Birdstrike - Laufende Forschungs Kooperation mit dem Zentrum für Geoinformationswesen der Bundeswehr

1st Herbert Neuendorf 2nd Alexander F. Auch 3rd Christian Rohlfs 4rd Sebastian von Massow 5rd Moritz Deininger
DHBW Mosbach DHBW Mosbach DHBW Mosbach DHBW Mosbach DHBW Mosbach
herbert.neuendorf@mosbach.dhbw.de alexander.auch@mosbach.dhbw.de

I. PROJEKT BESCHREIBUNG

Vogelschlag (englisch „birdstrike“) – der Zusammenstoß von Vögeln mit Objekten wie z.B. Flugzeugen stellt sowohl für die zivile als auch die militärische Luftfahrt ein ernstes Risiko dar, und kann neben schweren Schäden an den Maschinen auch zum Verlust von Menschenleben führen. In Zusammenarbeit (Work in Progress) mit dem Zentrum für Geoinformationswesen der Bundeswehr (Euskirchen), Abteilung Angewandte Geowissenschaften / Dezernat Biologie und Ökologie erforschen wir, inwieweit die Ablösung eines bereits bestehenden nicht automatisierten Systems zur Einschätzung des „Birdstrike“-Risikos anhand von Radar-Daten [1] durch ein KI-gestütztes System möglich wäre.

Auf Basis studentischer Arbeiten der Studiengänge Wirtschaftsinformatik [2] und Angewandte Informatik wird aktuell ein System zur Visualisierung und automatischen Klassifikation für einzelne Gitterzellen (GeoRef) im deutschen militärischen Luftraum auf Basis von zur Verfügung gestellten Radardaten entwickelt und prototypisch umgesetzt. Bei Detektion von Vogelzug wird der identifizierten Zelle ein Birdstrike-Level zugewiesen. Überschreitet die Vogelzugintensität gewisse Schwellenwerte kann es zu Einschränkungen des Flugbetriebs kommen. Für die Klassifikation werden u.a. die Möglichkeiten des High Performance Computing Centers der DHBW Mosbach genutzt.

Zugleich wurde von den beteiligten Studierenden das Prinzip des forschenden Lernens im Rahmen einer offenen Forschungsfrage praktiziert. Diesbezüglich besteht auch ein Austausch mit dem Zentrum für Hochschuldidaktik (Projekt EdCoN).

Die vorgenommenen prototypischen Implementierungen nutzen bezüglich der KI-Komponenten u.a. die Ergebnisse einer Masterarbeit [3], die durch eine Mitarbeiterin des ZGeoBW an der Universität Hagen (Fakultät für Mathematik und Informatik) kürzlich erstellt wurde und diverse Strukturen Neuroner Netze (u.a. 3D Convolutional Neural Networks) inklusive Hyperparameter-Optimierung zum überwachten Lernen basierend auf Vergangenheits-Radardaten im Rahmen der Problemstellung evaluiert hat (siehe z.B. [4]). Somit lassen

sich auch eher schwach strukturierte Radarbild-Muster zur automatischen Indizierung des Vogelschlag-Risikos nutzen.

II. IMPLEMENTIERUNG

Teil der Implementierung ist die KI-gestützte Klassifikation des Birdstrike-Levels auf Basis der Echtzeit-Radarmuster mittels Neuroner Netze durch automatische Erkennung und Klassifikation von Vogelschwarmkonzentrationen im Luftraum.

Die modulare Systemarchitektur erlaubt die quasi-kontinuierliche kurzfristige Evaluation von im 5-Minuten-Takt aktualisierten Echtzeit-Radardaten. Der Kern des Systems beruht auf einer Mikroservice-Architektur aus containerisiert laufenden Go-Microservices.

Als Container-Technologie kommen Docker und Podman zum Einsatz sowie Portainer zur Verwaltung und Monitoring laufender Container.

Die KI-Komponenten beruhen auf Tensorflow-Keras. Eine nachträgliche Evaluation und Korrektur der vom System vorgenommenen Klassifikation des Vogelschlag-Gefahrenlevels durch menschliche Experten erhöht die Menge der klassifizierten Vergangenheitsdaten und soll zum regelmäßigen Neutrainning des zugrunde liegenden Neuronalen Netzes genutzt werden. Das Training erfolgt auf einem System des High-Performance-Computing Labs der DHBW Mosbach [5], das mit vier NVIDIA Tesla V100-SXM2-Karten ausgestattet ist.

Die Entwicklung eines hinreichend performanten User-Interfaces auf Basis von Angular (zur Visualisierung der Radarmuster in Echtzeit) ist ebenfalls Teil des fortlaufenden Projekts.

LITERATUR

- [1] W. Ruhe, „Verhütung von Vogelschlägen mit Hilfe meteorologischer Informationen“, *promet*, 1-2 vol. 39, pp. 4-10, 2014.
- [2] Webseite Studienzentrum WION, <https://mosbach.dhbw.de/wion>
- [3] A. Witzens, „Neuronale Netze und Vogelzugwarnungen in der Luftfahrt“, Masterarbeit, FernUniversität Hagen, 2021
- [4] T. Lin et al., „MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks“, *Methods Ecol. Evol.* 2019;00:1–15. <https://doi.org/10.1111/2041-210X.13280>
- [5] Webseite FIM / HPC-Labor, <https://mosbach.dhbw.de/fim>

Abgleich der Kompetenzen aus dem Curriculum des Bachelorstudiengangs Data Science und Künstliche Intelligenz mit DASC-PM

Stephan Daurer
DHBW Ravensburg
Ravensburg, Germany
daurer@dhbw-ravensburg.de

Martin Zaefferer
DHBW Ravensburg
Ravensburg, Germany
zaefferer@dhbw-ravensburg.de

Durch den Bologna-Prozess [1] wurde ein Studiensystem eingeführt, das Studienleistungen mit Leistungspunkten vergleichbar macht. Ziel ist es dabei, die Lernergebnisse der Studierenden im Sinne von was sie können – statt was sie wissen – vergleichbar zu machen. Dies führte zu einer stärkeren Kompetenzorientierung an Hochschulen [2]. Das bildungspolitische Ziel des Bologna-Prozesses war es auch, die Beschäftigungsfähigkeit zu erhöhen, indem Studierende zu Anwendung, Reflexion und Weiterentwicklung des im Studium erworbenen Wissens befähigt werden sollen [3].

Das Fachgebiet Data Science ist durch zahlreiche mögliche Anwendungsdomänen und dementsprechend vielfältige Praxisgebiete gekennzeichnet. Daraus ergeben sich unterschiedliche Projekttypen [4]. Um Data-Science-Projekte systematisch und zielführend umzusetzen, existiert eine große Anzahl an Vorgehensmodellen [5] (z. B. [6], [7]). Ein aktuelles und sehr umfassendes Vorgehensmodell, welches die gesamte Breite an möglichen Data-Science-Projekten abdeckt, ist das Data Science Process Model (DASC-PM) [8]. Einer der Kernbeiträge dieses Modells ist die Beschreibung der benötigten Kompetenzen je nach Projektphase, welche in älteren Vorgehensmodellen, wie dem Cross Industry Standard Process for Data Mining (CRISP-DM), nicht thematisiert wurden [7].

Ziel dieses Beitrags ist der Abgleich dieser Kompetenzen mit dem Curriculum des neuen Bachelorstudiengangs Data Science und Künstliche Intelligenz an der DHBW. Hierzu wurde eine Bewertung der jeweiligen Kompetenzen, Phasen und Schlüsselbereiche einzeln für jedes Modul vorgenommen. Über alle Module gemittelt und gewichtet nach ECTS ergibt sich eine Korrelation von ca. 85% (Kompetenzen) bzw. 96% (Phasen). Domänen-, Strategie- und Managementkompetenzen werden insbesondere in den letzten zwei Semestern abgebildet, während IT und Mathe/Statistik gleichmäßiger verteilt sind, mit etwas stärkerer Ausprägung in früheren Semestern. Die Wissenschaftlichkeit steigt im Studienverlauf stetig an. Im Ergebnis zeigt die Analyse, dass das Kompetenzprofil des Studiengangs eine hohe Passung mit den nach DASC-PM geforderten Kompetenzen aufweist. Eine vollständige Abdeckung wird jedoch nicht erreicht. Dies liegt zum einen daran,



Abbildung: Heatmap der Modulbewertung (Spalten: Module; rot=geringe Ausprägung, grün=hohe Ausprägung)

dass es sich um einen grundständigen Bachelorstudiengang handelt und daran dass Data Science Projekte einen sehr interdisziplinären Charakter aufweisen, der weitere fachliche Profile in entsprechend diversen Teams erfordert.

REFERENCES

- [1] Bologna Working Group on Qualifications Frameworks, *A framework for qualifications of the European higher education area*. Copenhagen: Ministry of Science, Technology and Innovation, Denmark, 2005.
- [2] E. Weyer, N.-M. Wachendorf, and A. Mörrth, "Kompetenzorientierung, wie ist das gemeint?," in *Die kompetenzorientierte Hochschule. Kompetenzorientierung als Mainstreaming-Ansatz in der Hochschule*, pp. 6–12. Berlin: Bundesministerium für Bildung und Forschung, 2017.
- [3] N. Schaper, O. Reis, J. Wildt, E. Horvath, and E. Bender, "Fachgutachten zur Kompetenzorientierung in Studium und Lehre," 2012. Hochschulrektorenkonferenz (HRK).
- [4] R. Theuerkauf, S. Daurer, S. Hoseini, J. Kaufmann, S. Kühnel, F. Schwade, E. M. Alekozaï, U. Neuhaus, H. Rohde, and M. Schulz, "Vorschlag eines morphologischen Kastens zur Charakterisierung von Data-Science-Projekten," *Inform. Spektrum*, vol. 45, no. 6, pp. 395–401, 2022.
- [5] S. Daurer, R. Theuerkauf, and T. Franke, "Vorgehensmodelle bei Data-Science-Projekten," *WISU - Das Wirtschaftsstudium*, vol. 51, no. 4, pp. 426–433, 2022.
- [6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–54, 1996.
- [7] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
- [8] M. Schulz, U. Neuhaus, J. Kaufmann, S. Kühnel, et al., *DASC-PM v1.1 - Ein Vorgehensmodell für Data-Science-Projekte*. Hamburg / Elmshorn: Universitäts- und Landesbib. Sachsen-Anhalt, 2022.

WORKSHOPS

Der AI Transfer Congress besteht aus zwei Tracks: Einem Conference Track und einem Workshop Track. Im Workshop Track wird zu einer Vielzahl an Anwendungsgebieten eine Diskussionsplattform angeboten und auch Methoden diskutiert. Die Workshops werden in unterschiedlichen Formaten angeboten und bieten ein breites Spektrum der Unterstützung eines gelungenen Transfers.

The AI Transfer Congress consists of two tracks. A Conference Track and a Workshop Track. In the Workshop Track, a discussion platform is offered for a variety of application areas and methods are also discussed. The workshops are offered in different formats and provide a broad spectrum of support for a successful transfer.

Workshop

Autoren

Artificial Intelligence and Digital Care Applications (DiPA)

Prof. Dr. med. Raik Siebenhüner and Jan Beger

Automating Quality Control in Manufacturing with Machine Learning and Computer Vision

Prof. Dr.-Ing. Bozena Lamek-Creutz and Shobhit Agarwal

Modern Mathematical Optimization with Python

Prof. Dr. Nathan Sudermann-Merx

Von Verwaltungssprache zu einfacher Sprache – Generative AI in der Versicherungsbranche?

Anja Fischer und Lorena Oliveira

Future Skills in einer KI geprägten Welt: Ergebnisse der baden-württembergweiten Studie "AI Comp" zu einem KI-Kompetenzrahmen

*Dr. Martin Lindner,
Prof. Dr. Ulf-Daniel Ehlers und
Emily Rauch*

Innovationspark Artificial Intelligence

Veronika Prochazka

Reinforcement Learning from Human Feedback

Prof. Dr. Maximilian Scherer

Solide Grundlagen für KI gesucht

Felix Weschenfelder

KI verstehen und Potentiale erkennen

*Prof. Dr. Matthias Drüppel und
Prof. Dr. Janko Dietzsch*

KI-gestützt arbeiten und lehren

*Prof. Dr. Dietmar Ratz und
Prof. Dr. Katja Wengler*

Author Index

Agarwal, Shobhit	109
Arp, Benjamin	127
Auch, Alexander	140
Augenstein, Friedrich	102
Bader, Christian	133
Berkling, Kay	82, 92
Binnig, Carsten	123, 139
Bucher, Ulrich	2, 58
Daurer, Stephan	141
Dehn, Vanessa	129
Deiningner, Moritz	140
Devaser, Virrat	134
Dietrich, Alicia	119
Drüppel, Matthias	127, 133
Duerr, Jannik	130
Ehret, Felix	132
Eickhoff, Linus	118
Friedrichs, Kathrin	125
Gagali, Hande	125
Gonser,	131
Götz, Gerhard	34
Hackl, Benedikt	70
Hasebrook, Joachim	70
Heinrich, Roman	139
Hellstern, Gerhard	129
Herden, Olaf	119, 130
Holz, Christian	133
Holzweißig, Kai	58
Jansen, Tim	121
Janz, Oliver	120
Kellermann, Florian	118
Klug, Cornelia	125
Kochanowski, Monika	118
Koenig, Ferdinand	10
Kolb, Johannes	120
Kornmayer, Harald	139

Kraljic, Karlo	10
Kuhn, Marc	131
Lamek-Creutz, Bozena	109
Langenecker, Sven	123
Langner, Erik	132
Lay,	131
Lotz, Katja	125
Luthra, Manisha	139
Meyer-Waarden, Lars	131
Mochaourab, Rami	109
Mueller, Armin	120
Nanopoulos, Alexandros	27
Neuendorf, Herbert	140
Neumann, Stefanie	82
Parlesak, Alexandr	125
Plociennik, Christiane	1
Ranabhatt, Nehalben	124
Rohlfs, Christian	140
Ruckdeschel, Wilhelm	124
Ruoff, Dominik	121
Schalles, Christian	123
Scheutzow, Theresa	48
Schwenkreis, Friedemann	20
Schäfer, Elisa	92
Sievernich, Timo	125
Silva Guimaraes, Renato	128
Singh, Akshat	134
Sturm, Christoph	123
Thirukketheeswaran, Sinu	131
Utz, Anton	127
von Massow, Sebastian	140
Wiehenbrauk, Daniela	120
Wolf, Jan	41
Yuras,	131
Zaefferer, Martin	129, 141
Zundel, Armin	82, 92

Keyword Index

ADAS	133
AI	1, 121, 128
AI in education	34
AI Literacy	2, 58
artificial intelligence	125
ARIMA	134
artificial intelligence	92
Artificial Intelligence	2, 58, 120
Artificial intelligence (AI)	131
autoencoders	10
Automation	109
automation	10
automotive parts	124
Autoregressive Integrated Moving Average (ARIMA) Seasonal ARIMA	124
Aviation	140
Behavioral bias	48
Beitrag zur Produktivität	70
Big Data	128
Birdstrike	140
Carolo-Cup	127
ChatGPT	2
Circular Economy	1
Citizen acceptance	131
classification	119, 130
classifier	130
Clustering	20
CNN	127
Co-Occurrence Grouping	20
cognitive science	92
collaboration	102
collaborative filtering	34
computer vision	132
Computer Vision	127
Computer vision	109
Connected communities	131
consulting	102
content-based recommendations	34
Cost Models	139
COVID-19	118
Critical Thinking	58
Curriculum	141
Cynefin model	102

data center	10
Data Discovery	123
Data Lakes	123
Data Programming	123
data science	10
Data Science	20, 141
database	119
deep learning model	124
deep learning	34
Deep learning	109
demand forecasting	124
Determinism	128
Digital Product Passport	1
Discrimination	48
Distributed Stream Processing	139
education	92
event based vision	132
Fahrspurerkennung	127
Fairness	48
Feature Selection	129
Fine Tuning	127
Flipped Classroom	120
fraud detection	41
Generative AI	102
glucose prediction	82
gradient boosting	82
Graph Neural Networks	139
Human-machine-interaction	48
Hybrid Methods	129
hybrid system	34
imbalanced data	41
insulin recommendation	82
Interactive Online Course	120
Internet of Things (IoT)	131
Kompetenzorientierung	141
Künstliche Intelligenz	141
label smoothing	27
language training	92
Lehre	141
Linear Regression	134

Long Short- Term Memory(LSTM)	124
LSTM	133
Machine learning	48, 109
machine learning	10, 41, 82, 119, 130
Machine Learning	20, 129, 133
Multi Object Tracking (MOT)	133
multiuser	121
news articles	118
NLP	118
non-communicable diseases	125
NSE	134
Optimization	129
out-of-distribution	27
outlier exposure	27
overconfidence	27
parallelization	130
Pattern Recognition	20
personalized nutrition	125
pointclouds	121
predictive maintenance	10
Problem-Solving	58
professional services	102
Punktwolken	121
Quality control	109
Quantum Computing	129
Real World Examples	120
Recommendation systems	34
representation knowledge	34
Research-Based Learning	58
resource consumption	1
Retail	120
Risk assessment via Machine Learning	140
RNN	133
Semantic Type Detection	123
sentiment	118
small-sized data sets	34
SNN	132
Stacey matrix	102
stacking	41
Stock Prices Prediction	134
super learner algorithm	41

Supervised Learning	129
sustainability	1
Sustainable smart city solutions	131
tabular data	27
text-to-picture conversion	92
topic modeling	118
Tracking von Teamperformanz	70
Transfer Learning	127
type 1 diabetes	82
Use Cases	120
user-based recommendations	34
vehicle loan fraud	41
Verbesserung von Führungsentscheidungen	70
Vorgehensmodell	141
webvr	121
webxr	121
Western modernity	128