



**Upper-Rhine Artificial Intelligence Symposium 2024**

Applied AI in the age of Generative AI

14 November 2024  
Offenburg





Proceedings of the  
Upper Rhine Artificial Intelligence Symposium  
2024

Editors: Klaus Dorer and Janis Keuper

November 2024



Oberrhein | Rhin Supérieur

**License**

Copyright (c) 2024 by the authors



Creative Commons License

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

See <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>

**Publisher**

Offenburg University

**Editors**

Prof. Dr. Klaus Dorer Prof. Dr.-Ing. Janis Keuper

Contact: [janis.keuper@hs-offenburg.de](mailto:janis.keuper@hs-offenburg.de)

# Contents

<b>1</b>	<b>URAI 2024</b>	<b>7</b>
<b>2</b>	<b>Peer-Reviewed Contributions</b>	<b>9</b>
2.1	Content Creation with GenAI: Use Case on ClimaClic’s Instagram Marketing . . . . .	11
2.2	Can Machine Learning and Explainable Artificial Intelligence Help to Improve an Expert Model for Predicting Thermomechanical Fatigue? . . . . .	21
2.3	Semi-supervised mold differentiation using typical laboratory results as label data . . . . .	31
2.4	Leveraging Generative AI for Enhancing Domain-Driven Software Design . . . . .	41
2.5	Prediction of PV Power Production with Neural ODEs on the base of Weather data . . . . .	51
2.6	A Commissioning-Oriented Fault Detection Framework for Building Heating Systems Using SARIMAX Models . . . . .	59
2.7	Artificial Intelligence for Quality Assurance and Troubleshooting in Industry . . . . .	69
2.8	Exploration of Neural Network Architectures for Inertia Parameter Identification of a Robotic Arm . . . . .	77
2.9	AI-Guided Noise Reduction for Urban Geothermal Drilling . . . . .	85
2.10	Supervised Transfer Learning Framework for Fault Diagnosis in Wind Turbines . . . . .	95
2.11	The Potential of Generative AI for Systematic Engineering Innovation . . . . .	105
2.12	Evaluating AI-Generated Solution Ideas: A Comparative Study of AI and Human Assessments for Sustainable Process Design . . . . .	113
2.13	Specification and Identification of Relationships between Products in the Food Segment . . . . .	123
2.14	AI Chatbots in Design Thinking . . . . .	131
2.15	AI Adaption within SMEs: Analysis of Impedances and Suggested Approaches . . . . .	143
2.16	Data augmentation for pathogen segmentation in vinewood fluorescence microscopy images . . . . .	153
2.17	Enhancing Phishing Email Detection with Context-Augmented Open Large Language Models . . . . .	159



# 1 URAI 2024

The Upper Rhine Artificial Intelligence Symposium is the annual scientific conference and networking event of the tri-national AI research community of TriRhenaTech.

URAI 2024 was hosted by the Institute for Machine Learning and Analytics at Offenburg University on November 13th and 14th.

**General chairs:** Klaus Dorer (HS Offenburg) and Janis Keuper (HS Offenburg)

**Program Committee:** Klaus Dorer (HS Offenburg), Uwe Haneke (HS Karlsruhe), Janis Keuper (HS Offenburg), Thomas Lampert (Université de Strasbourg), Tobias Lauer (HS Offenburg), Jean-Philippe Lauffenburger (Université de Haute-Alsace), Stefan Naumann (HS Trier), Daniela Oelke (HS Offenburg), Maja Temerinac-Ott (HS Furtwangen), Christoph Reich (HS Furtwangen), Susanne Suter (FH Nordwestschweiz), Dieter Wallach (HS Kaiserslautern)

## URAI 24 in Numbers

- Conference Attendees: 108
- Tutorial Attendies: 30+30
- Paper submissions: 22
- Accepted Orals: 8
- Accepted Poster: 10

**Website:** <http://urai24.org>

**Online Proceedings:** <https://journals.hs-offenburg.de/index.php/urai/>





## **2 Peer-Reviewed Contributions**

All submissions received at least two independent reviews from members of the PC in a single blind setting. The following papers were accepted for oral or poster presentation.



# Content Creation with GenAI: Use Case on ClimaClic's Instagram Marketing

Diana Kolbe<sup>1,2</sup>, Andrea Mueller<sup>1</sup>, Barbara Woerz<sup>2</sup>

<sup>1</sup> Hochschule Offenburg – University of Applied Sciences, Badstrasse 24, 77652 Offenburg, Germany

[diana.kolbe@hs-offenburg.de](mailto:diana.kolbe@hs-offenburg.de)

[andrea.mueller@hs-offenburg.de](mailto:andrea.mueller@hs-offenburg.de)

<sup>2</sup> Burda Direct GmbH, Hubert-Burda Platz 2, 77652 Offenburg, Germany

[barbara.woerz@burda.com](mailto:barbara.woerz@burda.com)

**Abstract.** The availability of generative artificial intelligence (GenAI) tools has substantially increased, resulting in numerous positive impacts on the marketing sector. However, issues related to misinformation and deepfakes, biases and fairness, privacy, and ethical concerns, among others, have been highlighted. This research aims to examine the effects of utilizing GenAI for text, image, and audio creation in Instagram marketing. Employing the Customer Experience Tracking method, the study evaluated the differences between traditionally created and AI-generated Instagram Reels. The findings indicated that AI-generated content can garner higher levels of user attention, thereby enhancing brand interest. Negative effects such as mistrust or ethical concerns associated with AI were not substantiated in this study. These results suggest that companies can enhance their social media campaigns by integrating AI tools for content creation.

**Keywords:** Content Creation; CXT; GenAI; Instagram Marketing

## 1 Introduction

The recent widespread of ChatGPT and other GenAI tools has highlighted its vast array of use cases [1]. The emergence of large language models (LLMs) has made AI-powered user interfaces accessible to ordinary users, thereby increasing interest in the use of artificial intelligence [2]. In the marketing context, GenAI offers the potential to overcome important productivity limits by increasing the quantity of content without sacrificing the quality and vice versa. GenAI can assist in generating high-quality content across various modalities, including text, images, and certain types of videos [3].

Nevertheless, the use of GenAI agents has also raised concerns about the challenge of differentiating between human and AI authorship and has renewed discussions about the significance of traditional human endeavors [4]. Also, companies are concerned about the brand reputation and the trust of consumers in the technology and brand. Recent research in content marketing found that questions on how GenAI can be applied in the realm of visual advertising need to be addressed [3], especially focusing on the reaction of consumers on AI-driven manipulated advertising [5]. With this background, the present study aims to explore the impact of AI-generated Instagram content on customers.

We contribute to the debate on the use of GenAI in content marketing in several ways. We first test GenAI-tools for Instagram content creation. Based on our findings, then we assess and discuss the acceptance of this content for applied Instagram marketing. Another contribution lies

in the setting of the research; extant studies are of conceptual nature while this paper provides empirical evidence. The present use case is part of a postdoctoral industry-company collaboration between Offenburg University and BurdaDirect, which is a division of Hubert Burda Media, one of Germany's largest publishing houses. This use case was developed for ClimaClic, a social lottery with focus on climate projects under BurdaDirect, as part of the applied marketing project within the Marketing Management course of the Dialogue Marketing and E-Commerce Master's Program at Offenburg University.

The remaining part of the paper is organized as follows. In the second part, we explain the concept of content creation with GenAI and describe the situation of the German social lottery market building the theoretical framework of this paper. Afterwards, in the methodology part, we discuss the experimental setup of the study and provide information on the design of the study and participants. After that, we present the results of the study followed by a discussion of these. The paper closes with a conclusion providing main findings, practical implications, future avenues as well as limitation of the study.

## 2 Theoretical Framework

### 2.1 Content Creation with GenAI

According to their systematic literature review, AI is playing an increasingly significant role in the content creation domain of advertising [5]. Studies in this domain emphasize GenAI's potential to improve marketing communications, especially content marketing and content design [6], [7]. What sets GenAI apart as groundbreaking is its capacity to enhance the value system of marketing content management. Recent developments have shown that GenAI can overcome the quality-quantity trade-off through the generation high-quality content across different content types.

GenAI allows to generate new, original content by learning data distribution patterns [8]. Large Language Models (LLMs) enable the understanding and production of text the processing of inputs and the prediction of the next word or token (part of the next word) in a sequence. In November 2022, OpenAI made the generative pre-trained transformer (GPT) ChatGPT available to the public; GPTs utilize large databases of text as input, train the model using a self-supervised language modeling objective, and employ reinforcement learning from human feedback [9]. This methodology enables LLMs to emulate human response behavior [10], [11].

GenAI's specific applications in advertising content creation primarily involve three modules: image/video creation, copywriting, and content planning [12]. AI can generate custom images and certain kinds of videos in real-time based on individual user data and preferences, enhancing personalization and engagement [8].

While AI possesses the potential to transform the advertising industry, it also introduces challenges and ethical concerns. A primary challenge is the potential for biases embedded within training data, which could result in unfair or discriminatory targeting and dissemination of information. Another concern pertains to privacy; advertisers must ensure transparency and ethicality in the gathering and utilization of consumer data, respect privacy rights, and provide

clear opt-out options. Moreover, AI technologies raise concerns related to accountability and trust, posing regulatory and legal challenges [5].

## 2.2 Social Lottery Market in Germany

The German social lottery market is characterized by a variety of lotteries aimed at supporting social causes. These lotteries not only provide entertainment but also contribute to societal benefits such as environmental protection, education, and community development. In 2024, there are 8 permitted social lottery companies listed on the whitelist of permitted gambling providers in Germany: Aktion Mensch, BildungChancen, Chariety (Locasa), ClimaClic, Deutsches Hilfswerk Stiftung des Bürgerlichen Rechts (Deutsche Fernsehlotterie), DoGood (Onestly), Dreamify (Traumhausverlosung), Gewinnspareverein, Postcode Lotterie [13].

According to the 2022 annual report of the German Gambling Authority, social lotteries have a market share of 5.5% in the permitted market. The share of products in gross gaming revenue is distributed as follows: 59% Aktion Mensch, 21% Postcode Lotteries, 19% Deutsche Fernsehlotterie, and 2% Others [14]. The most prominent social lottery is the Aktion Mensch, established in 1964, which focuses on improving the lives of people with disabilities and promoting social inclusion [15]. The German Postcode Lottery is a social lottery that raises funds for charitable causes by allowing participants to win prizes based on their postal codes [16]. Another example is the Deutsche Fernsehlotterie, which has been operating since 1956 and supports various social projects [17].

## 2.3 Use-Case ClimaClic

In recent years, there has been a growing interest in climate-focused lotteries. ClimaClic, was founded in 2022 by the non-profit ClimaClic gGmbH, is a social purpose lottery dedicated to funding climate projects. To maintain non-profit status, ClimaClic must prioritize social impact over financial gain, in accordance with regulations from the Joint Gambling Authority. Consequently, organizers must be non-profit entities, as is the case with ClimaClic gGmbH. Burda Direct GmbH, on behalf of the organizer, handles the execution of the social lottery [18].

ClimaClic operates exclusively online and offers three different ticket types, each corresponding to separate prize categories and supporting projects in areas such as Nature & Landscape, Environment & Resources, and Education & Research [18].

Despite the emergence of new social lotteries, the market share of climate-focused lotteries remains relatively low, at less than 1%. Therefore, there is a pressing need to reinforce brand communication through technology-driven campaigns. This is central not only for customer acquisition but also for long-term customer retention amidst competition within the media and entertainment industry, where Burda Direct operates.

Social media platforms provide affordable options to reach a wider audience and are therefore especially important for ClimaClic as the brand is relatively new in the market and has a low level of brand awareness among lottery players. Thus, in the present use case the specific example for Instagram was chosen.

### 3 Methodology

To answer the above defined research question, a qualitative study using the at Offenburg University developed Customer Experience Tracking (CXT) method has been chosen. CXT integrates advanced technologies and novel research methods to pinpoint and assess consumer needs, enhancing product and service performance. Its innovative approach centers on merging established user experience measurement techniques with valid emotional response metrics during the purchasing journey. The CXT process included the following steps: (1) Kick-off meeting where a group of Offenburg University's professors and researchers, the students of the Marketing Management course and Burda Direct's ClimaClic team initiated the project, (2) Expert evaluation of previously selected best-performer Instagram content through the project team, (3) Research design and stimuli including the selection of GenAI-tools, reel generation with GenAI-tools, preparation of the questionnaire, in-depth interviews and testing scenario (4) Data gathering, (5) Data analysis.

#### 3.1 Expert evaluation

As part of the expert evaluation, top-performing reels and the presence of ClimaClic on Instagram were analyzed by the project team. Regarding the top performing reels, the following areas of opportunity were identified: (1) The reels contain a vast amount of content including various text bubbles which are displayed in quick succession which make it difficult to understand the message, (2) the reels lack in using auditive stimuli such as music or voice-overs (3) The focus of the reels lies on winning in combination with traveling - less on the lottery's USP which is climate protection, (4) The call-to-action (CTA) is always "Buy Now" which tend to be hardly achieved in early stages of the digital funnel.

About their Instagram page, ClimaClic has had few interactions (e.g., likes or comments) by the time of the study. The few comments under the posts highlight the principal pain point of the users: lack of trust in the company or product.

#### 3.2 Research design and stimuli

To address the presented research question, a qualitative research design was chosen. The aim was to analyze an existing Instagram Reel, create and evaluate a new Instagram Reel using GenAI tools. The study comprised two main parts: (1) content creation with GenAI tools (stimulus) and (2) the research design permitting the evaluation and comparison of the two reels.

For content creation, existing GenAI agents for text-, image-, and voice-generation were evaluated. Video generation was not pursued at this time due to limited alternatives. For the final reel, ChatGPT-4o was employed to generate ideas for the storyboard, captions and hashtags of the reel. All the prompt results were verified and manually improved before implementation to ensure a better match with the brand's identity. Midjourney v6 was used to produce a series of images. As AI video production was no option to this point due to the limited alternatives, the idea was to generate a series of images for the reel. According to the ideas generated in the storyboard, these images should show a couple that matches with the characteristics of the target audience. To show the same person in different settings, character reference (cref-function in Midjourney) was applied (Figure 1). The initial prompt used for the character reference function

was the following: “a realistic full body portrait of two people, Tom and Lucy, a happy married couple in their mid-forties. Lucy with brown shoulder long hair, Tom with short black hair with grey strands. Tom is taller than Lucy and is wearing a watch. Create multiple poses and expressions, high resolution, white background, white floor--ar 9:16--style raw--v 6--stylize 50.”



**Fig. 1.** Midjourney (v6) character references in different settings [19]

However, the outputs were not free of errors. Figure 2 illustrates outputs from different prompts that exhibited suboptimal results, such as the appearance of additional personas that were not requested, unrealistic and deformed body parts and objects (chairs, benches), disproportionate body parts like giant hands (with the inclusion of a key in the prompt), and outputs that were not photo-like.



**Fig. 2.** Midjourney (v6) character references fails for different prompts [19]

Elevenlabs was utilized for audio generation to create the voice-over for the reel. Eleven Multilingual v2 offers a selection of voices for text-to-speech conversion. Not all voices performed equally well. For the final reel, the voice of Serena performed best and was the most suitable for the brand's requirements.

After the content creation with the AI tools, the research design was defined. The research design comprised three different methods: (i) neurophysiological measures of visual attention through eye-tracking technology recorded during the visualization of the stimuli, (ii) an online survey administered before and in between the different stimuli, and (iii) an in-depth interview conducted after the testing. Participants' visual attention was recorded using fixed eye-tracking equipment available at the lab. Furthermore, the lab disposes two additional cameras and the screen recording that recorded the participants and their reactions in different angles. A total of three short online surveys were administered to the participants before and during the experiment, immediately following exposure to each stimulus. The first questionnaire aimed to gather sociodemographic information and behavioral data on gambling and Instagram use. The second and third questionnaires focused on attitudinal and behavioral data. All surveys were presented using Lamapoll. The experiment concluded with a semi-structured in-depth interview. The interview guide comprised a total of 14 questions on attitudinal and behavioral data, including the recall and level of understanding of the displayed message, the comparison of the videos addressing the perception of differences (e.g., the use of voices, embedded text in the video), the use of AI, the purchase intention related to AI-generated content, and concrete proposals for improvement. To minimize distractions, the whole procedure was carried out in the Customer Experience Tracking Laboratory at Offenburg University.

### 3.3 Data gathering and analysis

For the sample, participants were selected based on the criteria of the buyer personas provided by ClimaClic. These criteria included an age range of 34 to 65 years, residence in Germany, a gross income of over €3,500 per month, an affinity for lotteries, and active engagement with social media. The final sample included 14 participants including 4 participants from 34 to 45 years, 4 participants from 46 to 55 years and 6 participants from 56 to 63 years old. Data was collected between May 27th and June 7th, 2024. All participants explicitly gave their consent to participate in the study and to use their data for this study.

The eye-tracking data was analyzed using Tobii Lab Pro software, which provided detailed insights into participants' visual attention patterns. The questionnaires were systematically processed and analyzed using Microsoft Excel to extract and interpret the descriptive data. For the in-depth interviews, a thorough content analysis was conducted to identify and categorize recurring themes and insights. To ensure a comprehensive understanding, data triangulation was employed, integrating the results from the eye-tracking analysis, questionnaires, and in-depth interviews. This multi-method approach allowed for a deep and multifaceted interpretation of the findings.

## 4 Results

The CXT method facilitated the evaluation of both Instagram Reels. The results for the traditionally generated reel aligned with the pain points identified in the expert evaluation. Participants noted that the reel was too short, and the message was unclear. They struggled to understand the connection between the images and the product, leading to confusion about the product offering and reduced trust and interest in the brand. According to the questionnaire results, the reel did not spark interest in learning more about the brand. The results also indicated that a video without audio is less appealing.



Although participants did not immediately recognize the use of AI for text, image, and audio generation, their perceptions varied across different content types. Interestingly, participants found the AI-generated reel's message and overall presentation more attractive. The AI-generated text helped participants better understand the product offer and increased their interest in the brand and its offer. However, participants found the images less appealing; the images featured a couple created with character references to place them in various settings, but the couple appeared older than the requested mid-forties age. The participants perceived the AI-generated couple as “old” and “little attractive”. Regarding the audio, participants found the inclusion of a voice-over essential for following the information, initially not recognizing its artificial nature, though they later noted that the voice sounded somewhat unnatural. The inclusion of subtitles for the spoken part was perceived as positive and helpful.

Although having viewed the two different reels that included the brand's name, 71% of the participants did not recall the name. Participants stated that they remembered the second reel (the AI-generated reel) more, primarily due to auditory factors such as the voice-over and music.

Regarding whether participants noticed that one of the two videos was AI-generated, 13 out of 14 participants did not initially recognize the artificial nature of one of the videos. After asking them directly which one of the videos they thought was AI-generated, five out of 14 participants noticed that the second reel was AI-generated, one participant thought the first reel was AI-generated, and eight participants did not have a clear preference. After revealing which reel was AI-generated, 12 out of 14 participants indicated that the use of AI for content generation did not affect the brand's credibility. No one perceived the AI-generated content as a limitation to the company's credibility. Nearly half (6/14) would still buy a lottery ticket despite the AI-generated content. The participants who would not buy a lottery ticket expressed doubts about the brand's legitimacy, raising questions such as whether the money really goes to climate protection, if it is actually possible to win, and if the brand is legitimate.

Regarding improvements, participants stated that they would have liked to know more about the climate projects, have links to additional information, understand the percentage of money that goes to climate partner organizations, and receive information about the company itself. Furthermore, some participants mentioned the use of seals for trustworthiness.

## 5 Discussion and Conclusion

GenAI is a technological innovation transforming content marketing. However, despite its immense promise, there are still limitations that need to be acknowledged and addressed such as privacy concerns, ethical dimensions, and consumers' potentially negative attitudes towards GenAI in advertising. Overcoming these limitations is essential for marketers to effectively implement GenAI into their operations.

The objective of this research was to identify the impact of AI-generated text, images and audio on the customers' experience on Instagram. We tested ChatGPT, Midjourney and Elevenlabs for the created of an Instagram Reel and detected some limitations such as inaccurate text, image and speech output.

Although we support the idea of the disappearance of the quality-quantity trade-off in content marketing [3], [12], we acknowledge the gradual nature of this process. It still requires a

relatively high level of human verification, indicating that full automation of content creation is not yet recommendable. We especially highlighted the examination of different Midjourney outputs and their limitations for the use in advertising; this empirical use case helps to contribute to the call for research in content marketing, focusing especially on the use of GenAI in visual advertising [3].

Furthermore, we employed the CXT method to evaluate the differences between traditionally created and AI-generated Instagram Reels. The findings indicated that AI-generated content can capture higher levels of user attention, thereby enhancing brand interest. Particularly, the use of voice-overs was perceived positively, confirming the findings of [12] regarding the capability of GenAI-generated human-like speech to enhance the advertising experience. Negative consumer attitudes, such as mistrust or ethical concerns associated with the use of AI for generating Instagram Reels for a social lottery, were not substantiated in this study. With these insights, we shed light on the call for research to investigate consumer attitudes towards the use of GenAI in content creation [6].

The findings suggest five recommendation actions for integrating GenAI into social lottery communication.

- (1) Inspire Trust. Deliver relevant data about the core objectives of the social lottery. This can be supported by testimonials, numbers, data, and facts related to the usage of money and its impact on climate protection.
- (2) Include Seals of Trust. Integrate trust seals into communication content. It may be necessary to create a social lottery seal initiative, as such an instance does not currently exist.
- (3) Leverage Multimedia. Explain the social lottery using all commonly available media on Instagram, such as voice, music, videos, and text.
- (4) Align with Target Group Values. Identify the value attitudes of your target group and use appropriate signals for your new brand to achieve optimal value positioning for your social lottery offer.
- (5) Personalize Content with GenAI. Utilize GenAI based on user data files to personalize content communication. While this requires time and interactions, customers can be better attracted by delivering exactly the information they expect.

This study offers a preliminary assessment of the use and implementation of GenAI in social media marketing for social lotteries. Companies may find this information valuable as many are in the process of evaluating this technology. Consumer responses in a laboratory setting are an important consideration before deciding on implementation. Nonetheless, the research has its limitations. To achieve a more comprehensive understanding of the acceptance and reaction to AI-generated content, additional testing is required. We suggest a quantitative approach testing different concerns with respect to AI in content marketing and direct effects to the purchase intention or credibility of the brand. Furthermore, the number and variety of participants should be increased to cover different customer profiles of the brand.

## References

- [1] L. Tung, "ChatGPT can write code. Now researchers say it's good at fixing bugs, too | ZDNET." Accessed: Oct. 14, 2024. [Online]. Available:

- <https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too/>
- [2] J. Harjamäki *et al.*, “The Report of 85 AI Tools : GenAI Content Production: Enhancing Repeatability and Automation with ChatGPT,” May 2024, Accessed: Oct. 14, 2024. [Online]. Available: <https://trepo.tuni.fi/handle/10024/158346>
  - [3] M. Heitmann, “Generative AI for Marketing Content Creation: New Rules for an Old Game,” *NIM Marketing Intelligence Review*, vol. 16, no. 1, pp. 10–17, May 2024, doi: 10.2478/NIMMIR-2024-0002.
  - [4] H. Else, “Abstracts written by ChatGPT fool scientists,” *Nature*, vol. 613, no. 7944, p. 423, Jan. 2023, doi: 10.1038/D41586-023-00056-7.
  - [5] C. Campbell, K. Plangger, S. Sands, J. Kietzmann, and K. Bates, “How Deepfakes and Artificial Intelligence Could Reshape the Advertising Industry,” *J Advert Res*, vol. 62, no. 3, pp. 241–251, Sep. 2022, doi: 10.2501/JAR-2022-017.
  - [6] J. Paul, A. Ueno, and C. Dennis, “ChatGPT and consumers: Benefits, Pitfalls and Future Research Agenda,” *Int J Consum Stud*, vol. 47, no. 4, pp. 1213–1225, Jul. 2023, doi: 10.1111/IJCS.12928.
  - [7] K. B. Ooi *et al.*, “The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions,” *Journal of Computer Information Systems*, Oct. 2023, doi: 10.1080/08874417.2023.2261010.
  - [8] M. Jovanovic and M. Campbell, “Generative Artificial Intelligence: Trends and Prospects,” *Computer (Long Beach Calif)*, vol. 55, no. 10, pp. 107–112, Oct. 2022, doi: 10.1109/MC.2022.3192720.
  - [9] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 2023, Accessed: Oct. 14, 2024. [Online]. Available: <https://arxiv.org/abs/2303.08774v6>
  - [10] J. Jeon, S. Lee, and S. Choi, “A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models,” *Interactive Learning Environments*, Sep. 2024, doi: 10.1080/10494820.2023.2204343.
  - [11] B. Luo, R. Y. K. Lau, C. Li, and Y. W. Si, “A critical review of state-of-the-art chatbot designs and applications,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 12, no. 1, p. e1434, Jan. 2022, doi: 10.1002/WIDM.1434.
  - [12] B. Gao, Y. Wang, H. Xie, Y. Hu, and Y. Hu, “Artificial Intelligence in Advertising: Advancements, Challenges, and Ethical Considerations in Targeting, Personalization, Content Creation, and Ad Optimization,” *Sage Open*, vol. 13, no. 4, Oct. 2023, doi: 10.1177/21582440231210759/ASSET/IMAGES/LARGE/10.1177\_21582440231210759-FIG9.JPEG.
  - [13] Whitelist GGL, “Whitelist,” 2024.
  - [14] Jahresreport, “Jahresreports der Glücksspielaufsichtsbehörden der Länder.” Accessed: Oct. 14, 2024. [Online]. Available: <https://gluecksspiel-behoerde.de/de/gemeinsame-geschaeftsstelle/jahresreports>
  - [15] Aktion Mensch, “Die Aktion Mensch | Aktion Mensch.” Accessed: Oct. 14, 2024. [Online]. Available: <https://www.aktion-mensch.de/ueber-uns/die-aktion-mensch>
  - [16] Deutsche Postcode Lotterie, “Über uns | Deutsche Postcode Lotterie.” Accessed: Oct. 14, 2024. [Online]. Available: <https://www.postcode-lotterie.de/ueber-uns>
  - [17] Deutsche Fernsehlotterie, “Über uns - Deutsche Fernsehlotterie.” Accessed: Oct. 14, 2024. [Online]. Available: <https://www.fernsehlotterie.de/ueber-uns>
  - [18] ClimaClic, “Die erste Klimalotterie in Deutschland | ClimaClic.” Accessed: Oct. 15, 2024. [Online]. Available: <https://www.climaclic.de/die-klimalotterie>

[19] Midjourney, "Midjourney Overview." Accessed: Oct. 15, 2024. [Online]. Available: <https://www.midjourney.com/jobs/2de094d4-e38b-40c4-8bf4-1e588ad09709?index=1>

# Can Machine Learning and Explainable Artificial Intelligence Help to Improve an Expert Model for Predicting Thermomechanical Fatigue?

Stefan Glaser, Thomas Seifert, and Daniela Oelke

Offenburg University, Germany  
firstname.lastname@hs-offenburg.de

**Abstract.** Machine learning (ML) models are increasingly used for predictive tasks, yet traditional data-based models relying on expert knowledge remain prevalent. This paper examines the enhancement of an expert model for thermomechanical fatigue (TMF) life prediction of turbine components using ML. Using explainable artificial intelligence (XAI) methods such as Permutation Feature Importance (PFI) and SHAP values, we analyzed the patterns and relationships learned by the ML models. Our findings reveal that ML models can be trained on TMF data, but integrating domain knowledge remains crucial. The study concludes with a proposal to further refine the expert model using insights gained from ML models, aiming for a synergistic improvement.

**Keywords:** explainable artificial intelligence (XAI), TMF life prediction

## 1 Introduction

Predictive models using machine learning (ML) are increasingly applied across various fields. At the same time, data-based models themselves are not new in many of these areas. There are already models that have been manually created by experts using domain knowledge and, where necessary, adapted to the available data using statistical methods. Both approaches have their pros and cons. ML models have the advantage of uncovering unknown relationships, but they can also inherit data biases, especially with limited training data. Expert models, on the other hand, have the advantage that they can be based on known and proven correlations.

This paper explores improving a TMF expert model with an ML model by training it on data of the turbine-blade nickelbase superalloy MAR-M247. We evaluate the ML models using explainable artificial intelligence (XAI) and exploratory data analysis to understand the patterns and relationships it learns, aiming to improve the expert model effectively, thus combining the strengths and weaknesses of both approaches.

This work differs from related previous work in that the focus is not on training the best possible ML model (see e.g. [1,2,3]), but on gaining new insights by comparing the approaches of the expert model and the ML model.

## 2 Domain background

High-temperature turbine components are exposed to start-up and shut-down cycles in which the material experiences high thermal and mechanical loadings. These cyclic loadings result in a progressive damage of the material and after a certain number of

cycles to failure of the material. Different mechanisms are responsible for damage as fatigue crack growth (fatigue damage) that can be enhanced by creep of the material at higher temperatures (creep damage). For the assessment of the life of the turbine components, a model is required that allows engineers to predict the number of such start-up and shut-down cycles that the component can be used without failure. The number of cycles to failure  $N_f$  is the key result of models for TMF life assessment.

For the development of a TMF model and the fitting of the model to a certain material, different material tests are performed in which material specimens are cycled to failure using different load conditions, respectively. In isothermal low-cycle fatigue (LCF) tests, the specimens are exposed to constant amplitude cyclic mechanical loads at constant temperature, so that  $N_f$  is measured for different mechanical strain ranges and operation-relevant temperatures. In TMF tests, constant amplitude cyclic mechanical as well as cyclic thermal loads are applied to the specimen. Beside different mechanical strain and temperature ranges, different phase angle between the mechanical strain and the temperature are considered in the tests, since contributions of fatigue and creep damage and, thus,  $N_f$  strongly depends on the applied stress-temperature history the material is exposed to. In both tests, LCF and TMF, hold times at maximum load are often introduced as well. Since these material tests are time-consuming and expensive, usually the number of tests is rather small considering the number of factors influencing the TMF life.

Different features can be evaluated from the data of each test and used to generate a correlation to the number of cycles to failure  $N_f$ . We divide the available features into two categories: simple features and expert features (Table 1). Simple features are features that do not require any domain knowledge to create. They are either direct measurement results or simple combinations of features such as ratios or differences. The simple features include, for example, the maximum and minimum mechanical strain and the resulting mechanical strain range as well as maximum, minimum stress and the resulting stress range and stress ratio. When creating expert features, on the other hand, domain knowledge is required or additional information is included that is not part of the simple features.

In this work, a fracture-mechanics based model for TMF life prediction is considered that provides expert features. The first expert feature is the damage parameter, e.g. [4,5],

$$D_{\text{TMF}} = \left( 1.45 \frac{\Delta\sigma_{\text{eff}}^2}{\sigma_{\text{CY}} E} + 2.4 \frac{\Delta\sigma \Delta\varepsilon^{\text{P}}}{\sigma_{\text{CY}} \sqrt{1 + 3n'}} \right) F_{\text{creep}} \quad (1)$$

that includes besides the simple features also elastic, plastic and creep properties of the material (Table 1) and that allows the contributions of fatigue and creep damage to  $N_f$  to be evaluated. The effective stress range  $\Delta\sigma_{\text{eff}}$  is depending on the stress ratio  $R$  according to the empirical equation [6],

$$\Delta\sigma_{\text{eff}} = \left( 0.35 + (2.2 - R)^{-2} \right) \Delta\sigma, \quad (2)$$

and the function controlling creep damage  $F_{\text{creep}} \geq 1$  is a function of the stress-temperature history in the loading cycle. Using a fatigue crack growth law, the correlation

$$N_f = \frac{A}{D_{\text{TMF}}^B} \quad (3)$$

can be derived. The parameters  $A$  and  $B$  are fitted based on experimental data. This results in a linear relation in  $\log_{10} - \log_{10}$  scale:

$$\log_{10}(N_f) = \log_{10}(A) - B \log_{10}(D_{\text{TMF}}) \quad (4)$$

In the following, we refer to this calculation of  $N_f$  as the expert model or  $D_{\text{TMF}}$  regressor. Further expert features are the damage portions due to fatigue ( $0 \leq D_{\text{fat}} \leq 1$ ) and creep ( $0 \leq D_{\text{creep}} \leq 1$ ) adding up to a final damage of 1, i.e.  $D_{\text{fat}} + D_{\text{creep}} = 1$ . The latter is excluded because it is 100% correlated to  $D_{\text{fat}}$  in the data.

Feature	Type	Name	Further information
maximum mechanical strain $\varepsilon_{\text{max}}$	simple	<i>MeStrainMax</i>	
minimum mechanical strain $\varepsilon_{\text{min}}$	simple	<i>MeStrainMin</i>	
maximum stress $\sigma_{\text{max}}$	simple	<i>StressMax</i>	
minimum stress $\sigma_{\text{min}}$	simple	<i>StressMin</i>	
stress range $\Delta\sigma$	simple	<i>DeltaStress</i>	$\Delta\sigma = \sigma_{\text{max}} - \sigma_{\text{min}}$
stress ratio $R$	simple	<i>R</i>	$R = \sigma_{\text{min}}/\sigma_{\text{max}}$
plastic strain range $\Delta\varepsilon^{\text{p}}$	simple	<i>DeltaPlStrain</i>	$\Delta\varepsilon^{\text{p}} = \varepsilon_{\text{max}}^{\text{p}} - \varepsilon_{\text{min}}^{\text{p}}$ $\varepsilon^{\text{p}} = \varepsilon - \sigma/E$
maximum temperature $T_{\text{max}}$	simple	<i>TempMax</i>	
minimum temperature $T_{\text{min}}$	simple	<i>TempMin</i>	
hold time $t_{\text{hold}}$	simple	<i>HoldTime</i>	at maximum stress in LCF tests or temperature in TMF tests
heat time $t_{\text{heat}}$	simple	<i>HeatTime</i>	in TMF tests (0 s in LCF tests)
phase angle $\phi$	simple	<i>PhaseFactor</i>	phase shift of cyclic mechanical and thermal load in TMF tests
damage parameter $D_{\text{TMF}}$	expert	<i>DTMF</i>	includes simple features and material properties
damage portion due to fatigue $D_{\text{fat}}$	expert	<i>Dfat</i>	$D_{\text{fat}} = N_{\text{f,fat}}/N_f = 1/F_{\text{creep}}^B$
Young's modulus $E$	expert	<i>E</i>	elastic property
cyclic yield stress $\sigma_{\text{CY}}$	expert	<i>SigmaCY</i>	plastic property
cyclic hardening exponent $n'$	expert	<i>ROHardCy</i>	plastic property

Table 1: Used features;  $N_{\text{f,fat}}$  indicates the calculated fatigue life neglecting creep damage ( $F_{\text{creep}} = 1$ )

The  $D_{\text{TMF}}$  model contains only a low number of material properties to be determined on the basis of experimental data. To this end, fatigue tests at different strain amplitudes and different temperatures are required and additionally some fatigue tests including hold times are necessary to calibrate the creep-function  $F_{\text{creep}}$ . Basically, the properties for a certain material can be determined based on a set of approx. 25 fatigue tests. The material properties of the  $D_{\text{TMF}}$  model for the nickel-base superalloy MAR-M247, determined on the basis for 22 test, are available from [6]. For this paper, available data for MAR-M247 from 12 literature sources (research reports and scientific papers, see Acknowledgement) is used. This database comprises 185 isothermal LCF tests and 117 TMF tests, both in the temperature range between 20 and 1150 °C and for mechanical strain ranges from 0.0007 to 0.014. The number of cycles to failure are in the range from 17 to 1 820 000 cycles. The fatigue life data is complemented by stress-strain data which is obtained for all tests using a time- and temperature-dependent plasticity model calibrated to MAR-M247 [6].

### 3 Can a ML model learn the relationships from the data?

In order to answer the question of whether and how well ML is able to recognize the correlations in the data, various ML models were trained on the data. Among others,

ensemble methods such as Random Forest (RF), AdaBoost and XGBoost were tested, but also classic ML methods such as a Support Vector Regression (SVR), Decision Trees, Linear Regression or a K-Nearest-Neighbor (kNN) approach. Only a small multilayer perceptron (MLP) could be used as a neural network, as there was not enough training data available for a deep learning approach. All learning algorithms received scaled data. All prediction results (whether generated by the  $D_{\text{TMF}}$  regressor or an ML model) were post-processed so that values  $< 1$  are set to 1. This procedure takes into account the fact that there can be no negative lifetime, but also ensures that the logarithm in the error function can be calculated (see below).

The evaluation was carried out with a 5-fold cross-validation in which the logarithmic Root Mean Squared Error (RMSL10E) was used as the evaluation measure:

$$RMSL10E = \sqrt{\sum_i (\log_{10}(N_{fi}) - \log_{10}(\hat{N}_{fi}))^2} \quad (5)$$

where  $N_{fi}$  represents the number of cycles to failure of data instance  $i$  as measured in the experiment and  $\hat{N}_{fi}$  is its corresponding predicted value. The reason for choosing a logarithmic error measure is that, from a domain perspective, the error factor is more important than the absolute value of the error. This is also the reason why we train all ML models on predicting the logarithm of the number of cycles to failure rather than the number of cycles to failure directly.

The  $D_{\text{TMF}}$  regressor achieves an RMSL10E score of 0.496 when the material properties given in [6] are applied to the data of all literature sources. All, or when using simple features all but two ML-based approaches, were able to achieve better results than the  $D_{\text{TMF}}$  regressor. For all further analyses, we will work with the RF (expert features: 0.292 / simple features: 0.297) and the SVR (0.324 / 0.333) which were the best approaches for both the expert features and the simple features. From this initial experiment, it can be concluded that it is possible to train a machine learning model on the data of all literature sources even if expert knowledge has not already been incorporated into the ML model via the features provided.

## 4 Exploration of the ML models with methods of XAI

Our goal is to examine whether the existing expert model can be further improved by exploring the patterns and relationships found by the data-based models in the data. Such analysis could lead to a deeper understanding of the problem and thus also advance development and research in this area. In the following, two methods of Explainable Artificial Intelligence (XAI) are used to make the functionality of the ML model generated by the RF transparent. Where necessary, additional exploratory analyses are carried out to gain an even better understanding of the existing relationships or to verify or falsify assumptions made with the help of the XAI methods.

### 4.1 XAI method 1: Permutation Feature Importance

Permutation Feature Importance (PFI) is a model-agnostic, global XAI technique. It measures the contribution of a feature to the performance of a model. This value is determined as follows: For a single feature, the values observed in the dataset are randomly swapped. The model is then queried with the new instances created in this way. The



performance achieved is compared with the performance achieved on the unchanged instances. If the score hardly changes, this indicates that the feature was not important for the model’s decision. Conversely, a significant deterioration in the prediction means that the model relies on this feature [7]. Note that PFI is not a feature selection method. It does not judge the predictive power of a feature for the task, but rather measures how important the feature is for the particular model being analyzed. When interpreting the results, it must be considered that unrealistic data instances can be created in the permutation step if feature dependencies exist. Moreover, the presence of correlated features may result in the observed importance being distributed among multiple features, which can prove challenging to interpret [8].

Figure 1 shows the results of the PFI for the models that were calculated on simple features. For the RF model, the feature *DeltaPlStrain* is by far the most important. It is followed by *StressMin* and - again at a distance - *MeStrainMax* and *HeatTime*. In case of the SVR, *TempMax* is reported as most important feature followed by *StressMin*, *MeStrainMin* and *DeltaPlStrain*. The notable discrepancies in the relative significance of the features between the SVR and RF models can be attributed to the inherent differences in their learning algorithms. While the RF model employs an iterative, greedy approach, considering each feature independently, the SVR model adopts a more comprehensive, global perspective on the data.

From a domain perspective, the importance of features representing the plastic and mechanical strain range is understandable. In simple fatigue life prediction models, these features are applied as in the Coffin-Manson model that correlates the plastic strain range (*DeltaPlStrain*) with the number of cycles to failure or the Coffin-Manson-Basquin model that correlates the mechanical strain range (*DeltaMeStrain*) with the number of cycles to failure. While these simple models result in a good description of isothermal lower temperature data, additional information is required to account for temperature-dependency. With the further important features, namely the minimum stress (*StressMin*) and the maximum temperature (*TempMax*), the ML model seems to acquire information on temperature-dependency. The lower the absolute value of the minimum stress, the higher the temperature at minimum stress.

For the RF model based on expert features (see Figure 2),  $D_{\text{TMF}}$  is by far the most important feature. It is followed at a far distance by *Dfat*, *StressMin* and *R*. In case of the SVR model, *Dfat* is reported as most important feature, tightly followed by  $D_{\text{TMF}}$  and at some distance again *StressMin*.

From a domain perspective it is not surprising that the features related to the damage parameter and the damage mechanisms are considered important.  $D_{\text{TMF}}$  is seen as particularly important showing that expert knowledge has actually been used to create a feature that has a high information content with regard to the TMF life prediction. The damage portion due to fatigue (*Dfat*) is also important. For the RF, the PFI plot almost gives the impression that  $D_{\text{TMF}}$  alone is sufficient for the prediction. However, additional experiments showed that it is not possible to achieve good results with a RF model if only the  $D_{\text{TMF}}$  value is provided. Still it can be concluded that the RF considers it a very important feature.

Again, the minimum stress (*StressMin*) appears as an important feature. This is interesting from a domain perspective, as the stress information is already considered in the  $D_{\text{TMF}}$  feature (which is the basis for the expert model) via the stress range (*DeltaStress*)  $\Delta\sigma = \sigma_{\text{max}} - \sigma_{\text{min}}$  and the stress ratio (*R*)  $R = \sigma_{\text{min}}/\sigma_{\text{max}}$ .

None of the ML model considers the hold time (*HoldTime*) as important feature. Hence, effects of hold times, i.e. reduced life with longer hold times at higher temperatures

due to creep damage, are not described. Indeed, only around 5% of the tests in the experimental database contain hold times that show an effect in the fatigue live. While for the calibration of the expert model only a few (a minimum of two) tests with different hold times affecting the number of cycles to failure is required, modelling with ML may require more information for a reasonable description of the data. Hence, the training of the ML model requires a database with different tests than what is required for the calibration of the expert model. For the calibration of the ML model more tests with hold time are necessary such that the hold time becomes an important feature.

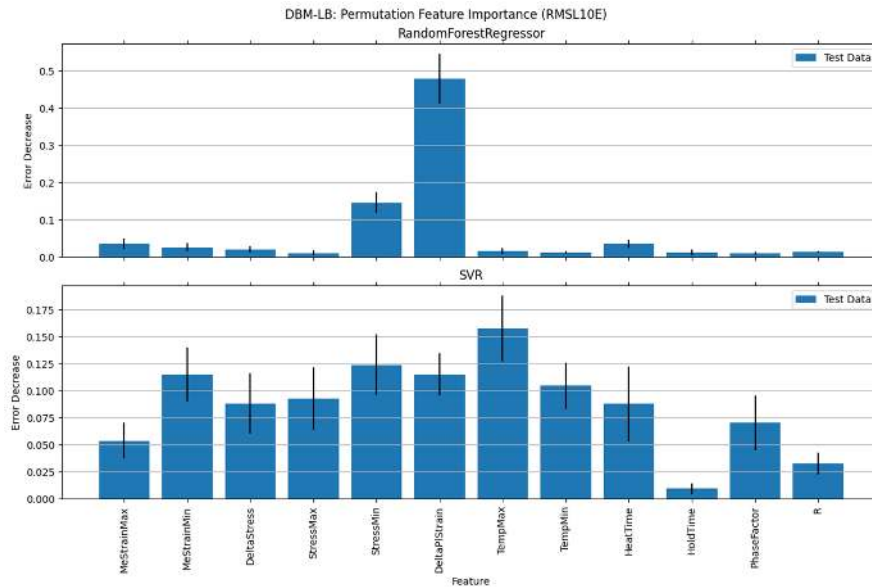


Fig. 1: Permutation Feature Importance (test data), Simple Features

## 4.2 XAI method 2: SHAP (SHapley Additive exPlanations) values

SHAP is an approximation method for calculating Shapely values. The methodology for calculating Shapely values has its origins in the field of game theory. The aim of the method is to determine the contribution of each player to the overall result. When applied to an ML model, the goal is to measure the contribution of each feature to the deviation of the predicted value from the expected value. In a regression task, the expected value is the average value of the target variable in the training data set [9].

To estimate the contribution of a feature to the overall result, a new model is trained for each possible subset of features. The deviation of the prediction quality with and without the feature gives an indication of its importance and, as a weighted sum of the marginal contributions, gives the contribution of the respective feature for the instance under consideration [9].

Since calculating Shapely values is a very computationally intensive process, several approximation methods have been developed to speed it up. We used KernelSHAP for

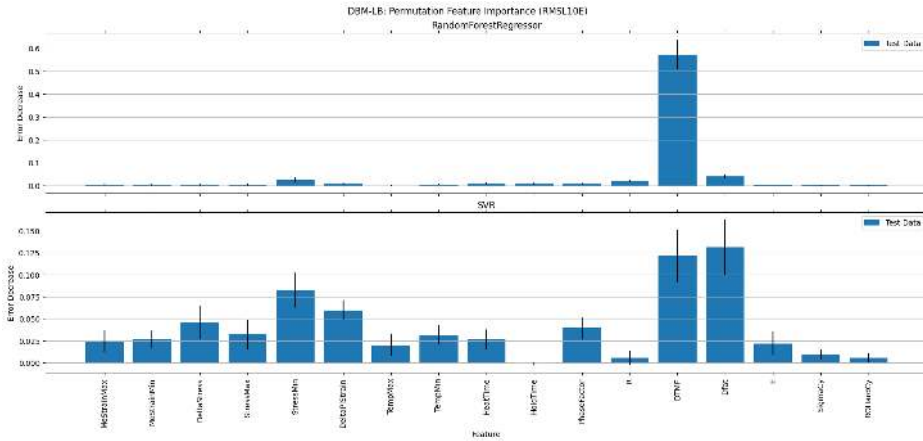


Fig. 2: Permutation Feature Importance (test data), Expert Features

the SVR models and TreeSHAP for the analysis of our RF models. TreeSHAP can only be applied to tree-based models and has the advantage of being significantly faster and handling feature dependencies well [8]. Nevertheless, in this paper we will only report the results for the SVR. The hierarchical structure of the tree-based models in the RF has the effect that some features are only applied to a small subset of the dataset. As a consequence, the results of the plots are difficult to interpret because it is unclear whether the described effect applies to all data or only to a small subset.

SHAP is a local method of XAI that makes a separate statement for each instance about how important each feature was to the decision made. In summary visualizations such as the Beeswarm plot (see Figure 3), each feature is represented as a line, with dots marking the SHAP value of a single data instance for that feature. Color indicates the value of the feature, making it possible to analyze the effect of the feature value on the target value. In this way, it is possible to get an impression of the global importance of the features - while retaining the advantage that the sometimes large differences for the individual instances are not lost by the formation of aggregation values.

According to the beeswarm plot for the SVR model based on simple features (Figure 3), a higher *StressMin* results in a higher life. This is especially true for tests with low plastic strain range and, thus, low  $D_{TMF}$  values, where only the stress range and the stress ratio are relevant. A higher minimum stress indicates a lower stress range resulting in a higher number of cycles to failure. However, higher *StressMin* can also reduce the life if the same stress range is applied and higher stress ratios occur. According to the SHAP analysis, lower values of the stress ratio  $R$  result in lower fatigue lives. From a domain perspective, the opposite should be true. Interestingly, this is only a global effect. The SHAP dependence plot (not shown here due to space limitations) for *StressMin* colored by  $R$  shows that higher  $R$  values tend to reduce fatigue life as expected, at least for higher *StressMin* values. However, the observed global effect again indicates that the database for the training of the ML model needs to be modified compared to the calibration of the expert model (i.e. more tests with different stress ratios). The XAI evaluation considers each feature separately. Changing *StressMin* (as well as the *StressMax*) should also change *DeltaStress* and stress ratio  $R$  that are not considered in the XAI evaluation. This could lead to uncertainties in the evaluation.

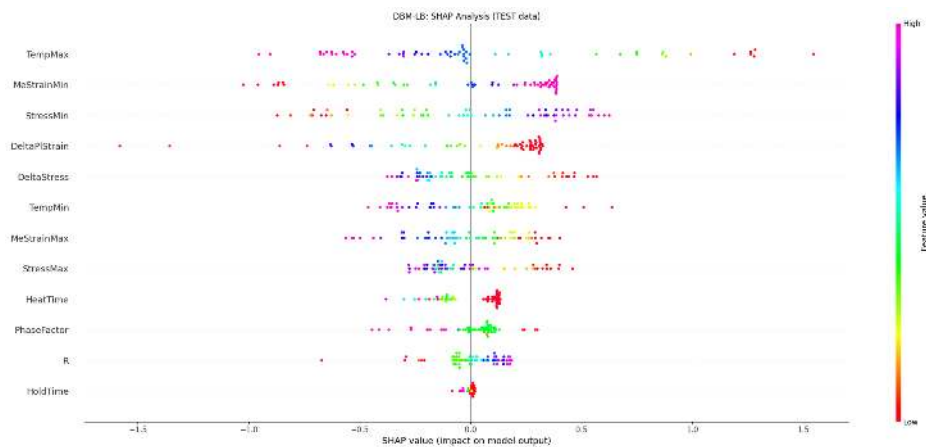


Fig. 3: SHAP beeswarm plot, SVR (test data), Simple Features

The SHAP plot for the expert features (not shown) did not provide additional insights and is therefore not shown here.

## 5 Conclusion

In summary, it can be said that the ML models were able to represent the data well. The analysis with the XAI methods showed that many of the relationships that the ML models detected are plausible from a domain perspective. Where this was not the case, this could often be explained by the fact that there was a bias in the data, as the experiments did not cover all possible cases or at least did not provide sufficient data for all cases. One lesson learned is thus that different data is required for training an ML-based model than is the case for a data-based expert model.

On the other hand, the analysis also revealed the limitations of the XAI methods. In particular, the fact that both methods assume that there are no dependencies between features limited the analysis, as there are many such dependencies in our dataset.

The principal trends identified by XAI are also predicted by the expert model in this way. This can also be seen as confirmation of the model derived by experts from the data. On the other hand, the additional use of further features by the ML model has also raised the question of whether this could be a key to further improvements of the expert model. It is possible that oversimplified assumptions are made when deriving the expert model, for example when calculating the effective stress range from the stress ratio or when assuming power-law hardening of the material via the hardening exponent  $n'$ . However, a direct "hint" as to how the expert model should be modified cannot be derived from the XAI results.

In this first analysis, we did not succeed in identifying the relationships which the ML models discovered in a way that they could be directly used to improve the expert model. Therefore, as a next step we intend to train a boosting model that directly builds on the  $D_{TMF}$  regressor and could give hints on what to add.

## Acknowledgement

The authors acknowledge Dr.-Ing. Christoph Schweizer of the Fraunhofer Institute for Mechanics of Materials IWM for providing digitized LCF and TMF material data from literature sources.

## References

1. Srinivasan, V., Valsan, M., Bhanu Sankara Rao, K., Mannan, S., Raj, B.: Low cycle fatigue and creep-fatigue interaction behavior of 316l(n) stainless steel and life prediction by artificial neural network approach. *International Journal of Fatigue* **25**(12) (2003) 1327–1338
2. Liu, Y., Wu, J., Wang, Z., Lu, X.G., Avdeev, M., Shi, S., Wang, C., Yu, T.: Predicting creep rupture life of ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Materialia* **195** (2020) 454–467
3. Zhang, X.C., Gong, J.G., Xuan, F.Z.: A deep learning based life prediction method for components under creep, fatigue and creep-fatigue conditions. *International Journal of Fatigue* **148** (2021) 106236
4. Seifert, T., Riedel, H.: Mechanism-based thermomechanical fatigue life prediction of cast iron. part i: Models. *International Journal of Fatigue* **32**(8) (2010) 1358–1367
5. Seifert, T., Schweizer, C., Schlesinger, M., Möser, M., Eibl, M.: Thermomechanical fatigue of 1.4849 cast steel – experiments and life prediction using a fracture mechanics approach. *International Journal of Materials Research* **101**(8) (2010) 942–950
6. Serrano, L., Scholz, A., Müller, F., Berger, C., Schweizer, C., Schlesinger, M., Eckmann, S.: Tmf crack behavior: Crack behavior under anisothermal loading conditions – calculation method for nickel-based alloys (in german). Technical report, Forschungsvereinigung Verbrennungskraftmaschinen e.V., project Nr. 959, issue 945 (2011)
7. scikit-learn User Guide: Permutation feature importance [https://scikit-learn.org/stable/modules/permutation\\_importance.html](https://scikit-learn.org/stable/modules/permutation_importance.html), accessed: 25.7.2024.
8. Molnar, C.: Interpretable machine learning - a guide for making black box models explainable (2024) <https://christophm.github.io/interpretable-ml-book/>, accessed: 18.9.2024.
9. Cooper, A.: Explaining machine learning models: A non-technical guide to interpreting shap analyses (2021) <https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/>, accessed: 25.7.2024.



# Semi-supervised mold differentiation using typical laboratory results as label data

Henrik Pichler<sup>1</sup>, Janis Keuper<sup>1</sup>, and Matthew Copping<sup>2</sup>

<sup>1</sup> Hochschule Offenburg  
hpichler@stud.hs-offenburg.de

<sup>2</sup> Biostates GmbH  
m.copping@biostates.de

**Abstract.** This study applies semi-supervised learning to automate the differentiation of mold colonies, thereby reducing the time and cost associated with air quality assessments. EfficientNet V2 and Normalization-Free Net (NfNet) were trained on a dataset of mold colony images, created in a semi-supervised way. NfNet demonstrated superior performance, particularly on non-padded images, with explainable AI techniques enhancing interpretability. The models exhibited generalization capabilities to environmental samples, indicating the potential for automating mold identification and streamlining air quality monitoring, thereby reducing manual effort and costs. Future work will focus on refining species handling and integrating the system into laboratory workflows.

**Keywords:** Artificial Intelligence; Biology; Semi-supervised learning; Object detection; Explainable artificial intelligence

## 1 Introduction

### 1.1 Problem Definition

Air quality in workplaces and production facilities affects employee well-being and operational efficiency. In accordance with VDI Guideline 6022, the quality of air entering premises via ventilation must be maintained [1]. Many factors affect indoor air quality, including VOCs, CO<sub>2</sub>, humidity, and mold spores [2]. Mold contamination is a key factor in maintaining a healthy indoor environment due to its ability to trigger allergic reactions and respiratory issues. This is crucial in vulnerable settings like hospitals [3]. To assess air quality, customers provide air samples using Petri dishes, which are incubated for 5-7 days to allow for mold growth. These colonies are then counted and differentiated to evaluate air quality. Especially the task of differentiating, often including the usage of a microscope, is time-consuming and costly.

### 1.2 Goal

To reduce the time and cost of evaluating the samples, a project was carried out with the specific goal of developing a more efficient method for differentiating macromorphologically distinguishable mold colonies. It was determined that colonies requiring microscopic analysis would be excluded from the project, as this would involve a more complex setup, that is beyond the scope of this task. A central objective of the project was to demonstrate the feasibility of training machine learning (ML) models on novel forms of data, even when such data is scarce and lacks comprehensive annotation. Furthermore, the

project aimed to demonstrate that existing ML models can be effectively extended and adapted to perform more complex tasks, rather than being entirely rebuilt from scratch. The models were trained and evaluated within this semi-supervised context to assess their effectiveness in achieving the goal of streamlining the differentiation process and enhancing overall efficiency.

### 1.3 Related Work

**Semi-supervised Learning** The subject of semi-supervised learning examines how to use partially labeled datasets for training machine learning models. [4] presents a method for classifying test data without prior exposure to some of the classes.

**Explainable AI** The field of explainable AI seeks to understand the behavior of complex machine learning and deep learning models. This is crucial, especially in biological and medical contexts, where such applications demand high levels of accuracy and trust. Consequently, numerous studies have been conducted with the aim of explaining a model’s behavior. [5] utilizes attribution maps to generate pseudo ground truths for semi-supervised semantic and instance segmentation.

**Transfer Learning** Transfer learning reduces training time and cost by using previously trained models. [6] reviews progress in transfer learning for classification, regression, and clustering.

**Machine learning models in a biological/medical context** There is limited research on the automatic differentiation of molds. However, numerous studies address biological and medical sample analysis, underscoring the significance of this topic. [7] discusses current approaches to AI in the field of nephrology. [8] utilizes a hyperspectral imaging device and a support vector machine (SVM) to identify the growth stage of molds on wheat grains.

## 2 Approach

### 2.1 Dataset Creation

To guarantee a minimal labeling effort and to quickly start a training on the novel data type, a dataset was constructed in a semi-supervised manner. The dataset comprises images of Petri dishes, captured from above with a camera with a resolution of 2,144x2,144 pixels and 17.9 pixels/mm. The Petri dishes were prepared using the three-point inoculation method, whereby individual known mold colonies were extracted from one dish and transferred onto another dish [9]. This enables the regulation of the quantity and variety of mold growth on a given sample. A total of 600 Petri dishes were prepared and incubated, with each dish containing three mold colonies, as per the method’s specifications. The samples comprised of five distinct mold species, with each distributed across 100 Petri dishes. The remaining 100 samples were colonized with 10 additional mold species, the actual species of which were inconsequential. This approach would enable the models to consider a more extensive set of mold morphologies. Consequently, mold colonies, for which an uncertainty would remain, could be classified as ”other” and examined for revision when required. Following the incubation and image capture, a YoloV7 model was applied to detect the mold colonies without any classification, as this was a problem on



which the model had previously been trained on [10]. This resulted in distinct bounding boxes for each mold colony. The predictions were then reviewed, to correct any errors that had occurred, which was a minor issue. As only one type of mold, which was also known, grew on each Petri dish, all annotations for one image could simply be assigned with the same class. The original dimensions of the mold colonies were maintained by utilizing this approach, with each image accurately representing the natural, unaltered dimensions of the colonies. This resulted in a diverse range of sizes for each mold image within the dataset. Henceforth, the dataset will be referred to as the "clean culture" dataset.

## 2.2 Classifier Training

Two classification models, EfficientNet V2 [11] and Normalization Free Net (NfNet) [12], were selected for training. Both models were initialized with pre-trained weights based on ImageNet [13]. The mold colonies were extracted from the clean culture dataset using the provided bounding boxes. Two training approaches were investigated. For the first approach, the mold colony crops were padded using zero padding or resized to a size of 1000x1000 pixels, depending on whether the image crop was larger or smaller than 1000x1000 pixels. The resulting images will be referred to as "padded" images. The second approach involved leaving the images unaltered, maintaining their original dimensions. However, they were made square. These images will be referred to as "non-padded" images. Following the padding/resizing, the images were fed through their respective models, resulting in a single classification per mold colony. In accordance with the concept of progressive learning as outlined in [3], adaptive image augmentation was employed to augment the training images and increase data variety. This meant that larger images, and consequently mold colonies, were augmented with a higher magnitude than smaller ones. Cross-entropy loss was utilized as the loss function throughout the training process. The models were subjected to continuous validation during training using a combination of the F1-Score and the loss on the validation data.

## 2.3 Feature Inspection

The field of explainable artificial intelligence (XAI) plays a crucial role in guaranteeing the transparency and reliability of model predictions, thus enabling the deployment of these models in practical applications [14]. In this work, Grad-CAM (Gradient-weighted Class Activation Mapping) [15] was used to provide visual explanations of the models' predictions. The method highlights the areas, that contributed most to the decision, enabling a visual inspection of the model through heatmaps.

## 2.4 Classifier Evaluation

The models were evaluated on both the clean culture Dataset and an "environmental" dataset, which represented real-world mold samples and included 640 additional annotated images with a total of 12,472 individual mold colonies. The environmental Dataset was employed exclusively for the purpose of evaluating the models' performance in more complex, real-world conditions. In a practical application, the dataset would not be built from the ground up, but rather developed over time to improve the model further. This served the purpose of a proof of concept, demonstrating the generalization ability of the trained models. The primary metrics, used to evaluate the classification performance on

the test datasets, were the accuracy, precision and recall. Particular interest was paid to the handling of unknown mold colonies and uncertain decisions respectively. These cases should be classified as "Other". The correct handling of such cases by a model would facilitate the refinement of predictions by a reviewer. The evaluation also considered the performance differences between models trained on padded and non-padded images, respectively. The evaluation of models trained on padded data was conducted exclusively using padded data, while the models trained on non-padded data were evaluated only with non-padded data.

### 3 Results

#### 3.1 Dataset Statistics

In total 600 images were taken for the clean culture dataset, which resulted in 2,116 single mold colonies from 589 Petri dishes.

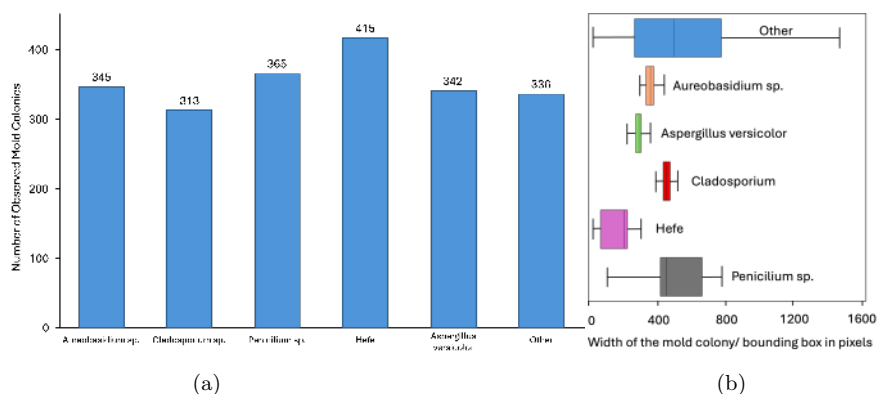


Fig. 1: Statistics for the clean culture dataset. (a) shows the distribution of classes for the dataset. The Acronym "sp." denotes "and related species". (b) shows the distribution of the different mold species based on their colony sizes, measured by the dimensions of the bounding boxes surrounding each colony.

After the creation and review of the clean culture dataset, the following distribution of mold species was found (Fig. 1a): 345 *Aureobasidium* sp., 313 *Cladosporium* sp., 365 *Penicillium* sp., 415 Hefe, 342 *Aspergillus versicolor*\* and 336 other mold colonies with 10 different species.

A size distribution of 1,500 pixels was found across the dataset (Fig. 1b). The largest observed mold colony was 1520 pixels wide ("Other"), the smallest 20 pixels ("Hefe"). The mean size of the colonies was 347 pixels. The size refers to the width of the bounding box around the mold colony.

A large size distribution was found across the dataset (Fig. 1b). This refers to the size of the bounding box around the colony. The largest observed mold colony was 1520 pixels large ("Other"), the smallest 20 pixels ("Hefe"). The mean size of the colonies was 347 pixels.

### 3.2 Training results

Both the EfficientNet V2 as well as the Normalization-Free Net (NfNet) were trained on the clean culture dataset (chapter 2.2).

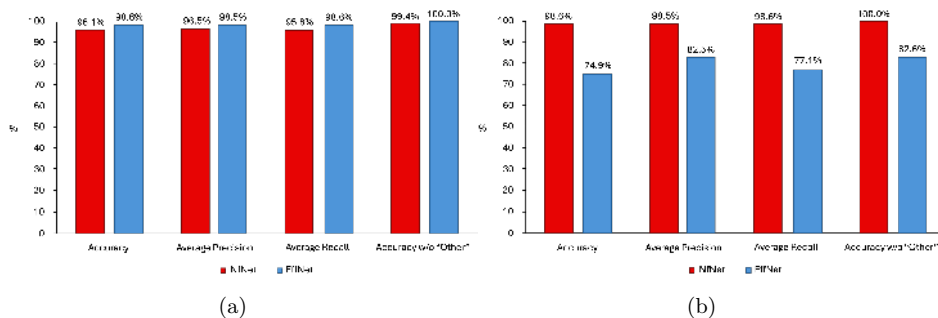


Fig. 2: Comparison of training both the EfficientNetV2 (EffNet) as well as the Normalization-Free Net (NfNet) on the clean culture dataset using both padded images (a), as well as non-padded images (b).

The training of the models on the clean culture dataset resulted in high accuracy across all models (Fig. 2). In the case of the padded images, the EfficientNet V2 demonstrated an accuracy of 98.6%, which represents a 2.5% improvement over the performance of the NfNet on the same dataset. In contrast, the NfNet demonstrated a higher accuracy of 98.6% for the non-padded images, representing a 23.7% improvement over the EfficientNet V2 on the same dataset. The NfNet demonstrated the same accuracy on the non-padded images, as the EfficientNet V2 on the padded images (98.6%).

### 3.3 Feature Inspection

A further inspection of the obtained results was performed, utilising Grad-CAM to identify the most important features for the models' predictions.

Fig. 3 shows the visualizations generated using Grad-CAM for both the EfficientNetV2 and NfNet models, with comparisons between predictions on a single mold colony sample from the class "Cladosporium sp.". The Grad-CAM visualizations illustrate that EfficientNet V2 produces coarse highlighted regions. In the padded input, it highlights the corners and edges of the image. The NfNet highlights specific regions of interest in both the non-padded and padded images, without highlighting the corners or edges of the padded input.

### 3.4 Evaluation on environmental Data

Fig. 4 shows a comparison between a sample from the clean culture dataset and a sample from the environmental dataset. It shows, that the mold colonies growing on the environmental sample are much smaller and occur in larger quantities.

Fig. 5 shows the evaluation of both models on the environmental dataset (chapter 2.4). EfficientNet V2 demonstrated superior performance on the padded images, achieving higher overall accuracy (43.3%) and accuracy without the class "Other" (60.0%) compared to NfNet (27.1% and 31.3%). In the case of the non-padded images, NfNet demonstrated

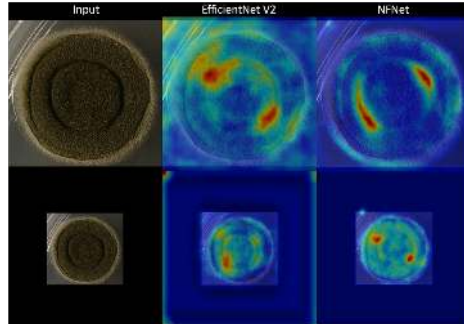


Fig. 3: Grad-CAM visualizations comparing the EfficientNet V2 and NfNet model predictions for a single mold colony image. The top row shows the results for the non-padded input, whereas the bottom row displays the results obtained with a padded input image. The heatmaps visualizations highlight the regions, on which the models focused. Dark red indicates a stronger focus of the model on that region, and conversely, dark blue indicates an uninteresting region. The depicted mold colony belongs to the class "Cladosporium sp."



Fig. 4: Comparison between a clean culture sample dish (a) with three colonies of *Penicillium* sp. and an environmental sample dish (b) with a combination of different mold species

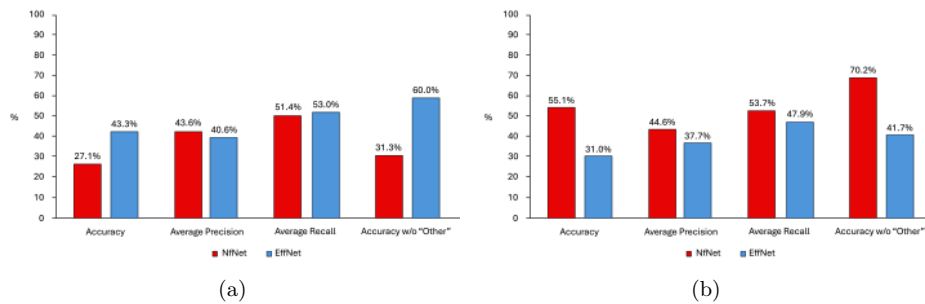


Fig. 5: Comparison of evaluating both the EfficientNetV2 (EffNet) as well as the Normalization-Free Net (NfNet) on the environmental dataset using both padded images (a), as well as non-padded images (b).

superior performance in comparison to EfficientNet V2. The former achieved a higher overall accuracy (55.1%) and accuracy without the class "Other" (70.2%) in comparison to the latter (31.0% and 41.7%). Overall, NfNet demonstrated an overall performance increase in comparison to EfficientNet V2 of 24.1% in accuracy, 6.9% in average precision and 5.8% in average recall.

## 4 Discussion

During the dataset creation (chapter 2.1) 11 samples had to be discarded due to contamination by additional mold growth. Furthermore, on several plates, the mold grew more extensively than anticipated, resulting in increased number of samples per dish. This resulted in a slight imbalance in the class distribution, which was effectively mitigated by applying class weights to the loss function during training.

The introduction of the "Other" class, with the creation of 100 additional samples, notably enhanced the diversity of the clean culture dataset, both in terms of mold size and morphology (Fig.1b). The introduction of a broader range of mold has led to an improvement in the classification task, as the models were able to gain a more complete understanding of the diversity of the dataset. Training on this expanded dataset yielded accurate and consistent results, particularly with NfNet, which demonstrated superior performance in handling data variability compared to EfficientNet V2. While EfficientNet V2 encountered challenges with non-padded images due to the size variance of the mold colonies, NfNet's resilience to such variability enabled it to outperform EfficientNet V2, particularly on the non-padded images.

Upon examination of the feature maps, it became evident, that there were notable discrepancies between the two models. EfficientNet V2, in particular, appeared to rely excessively on the padding of images, with a corresponding reduction in focus on the mold colonies themselves. This made it less suitable for the task at hand, as it did not consistently extract meaningful features from the mold colony images. In contrast, NfNet performed significantly better by concentrating on the actual features of the mold colonies, such as the characteristic rim of the colony, as seen in Fig. 3. This enabled NfNet to classify molds with greater reliability, especially on non-padded images.

To further evaluate the models' performances, especially in a real-world context, a supplementary annotation of 640 images was performed. These additional annotations were crucial for validating the models' real-world applicability. The results confirmed several key points. NfNet exhibited a clear preference for non-padded images (55.1% accuracy), whereas EfficientNet V2 demonstrated superior performance with padded images (43.3% accuracy). Nevertheless, Nfnet demonstrated superior overall performance, attaining higher accuracy on non-padded images than EfficientNet V2 on padded images. However, NfNet demonstrated difficulty in generalizing its knowledge to padded images, likely due to the introduction of artificial boundaries by padding, which interfered with its ability to recognize features across different scales. Conversely, EfficientNet V2 exhibited superior performance in handling padded images due to its reliance on padding, though this came at the cost of its feature extraction capabilities on non-padded images. A principal component of the evaluation process was the comparison of the accuracy of the models in question, with and without the class named "Other". This comparison proved highly beneficial in understanding the models' ability to handle cases of uncertainty or unfamiliar mold colony samples. Without the "Other" class, the models would have been forced to classify these ambiguous cases into one of the existing classes, which would have increased the likelihood of misclassification. The incorporation of the "Other"

class enabled the models to express uncertainty, thereby enhancing both generalization and precision. This proved to be a particularly beneficial approach in the environmental data, as it provided a safety net for uncertain cases and reduced the overall error rate. It is noteworthy that the highest accuracy achieved on the environmental data was 70% for the NfNet when the "Other" class was excluded. This illustrates that for predictions where the model is certain, the accuracy is already considerable, leaving only a few mold colonies that require manual review. This provides a robust basis for deploying the models on environmental data, specifically NfNet, as it suggests that the majority of classification work could be automated, with only a small portion requiring further inspection. It also offers a promising starting point for further improvements to the models, particularly in refining their handling of uncertain cases.

## 5 Future Work

One avenue for future research would be to only utilize the environmental data, which includes not only images and bounding boxes but also general class labels for mold colonies, to start training a model. It is important to note that the bounding boxes currently present in the environmental data set were manually annotated. In theory, the dataset would consist solely of the raw images and general class labels, without any bounding box information. The primary objective is to investigate the potential of an iterative approach to construct a comprehensive dataset from the available information, thereby eliminating the necessity for extensive manual annotation.

The classification model can be extended to more complex samples through training on single-class samples. In the initial stage, all colonies within a sample are identified and those belonging to known classes are excluded. If the remaining colonies belong to only a single unknown class, the sample can be utilized for further training. The model will classify the known colonies, leaving the unknown ones labeled as "other," which can then be assigned to the remaining class.

This method could result in a comprehensive dataset comprising all mold species, reducing the necessity for manual annotation and enabling the dataset to evolve. This approach requires a sufficient number of initial samples, particularly "clean culture" samples, that comprise only of mold colonies the same class, to learn effectively. Some human supervision is needed, but it offers a scalable solution that reduces manual effort.

## 6 Conclusion

This study shows that semi-supervised learning can be used to differentiate mold colonies. A clean culture dataset was developed and used to train classifiers that can identify mold species based on their macromorphological characteristics. The environmental dataset showed that the models can generalize to real-world conditions, although performance declined in more complex scenarios.

Automating this process could reduce costs. The ability to quickly identify mold colonies could cut air quality evaluation costs, making routine inspections more affordable and accessible. This could lead to more frequent assessments, contributing to cleaner air in workplaces and public buildings.

Semi-supervised learning can enhance the efficiency of the mold differentiation process, reducing time and cost. Future enhancements could prioritize the management of unknown species and integrate the system into standard lab workflows. This automation

could facilitate broader adoption of air quality monitoring, making cleaner air a more attainable goal.

## 7 Acknowledgements

Thanks to Melanie Priemer for creating the clean culture samples. Special thanks to Nadja-Stella Voce for the extensive annotation of the environmental data, which made the evaluation of the models possible.

## References

1. VDI: Vdi 6022 blatt 1 - raumlufttechnik, raumluftqualität - hygieneforderungen an raumlufttechnische anlagen und geräte (vdi-lüftungsregeln) (2018)
2. Agency, U.S.E.P.: Care for your air: A guide to indoor air quality. US EPA (aug 2024)
3. Held, J.: Measuring mold and the impact of mold on human health (August 2024)
4. Cao, K., Brbic, M., Leskovec, J.: Open-world semi-supervised learning
5. Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation (2021)
6. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**(10) (2010) 1345–1359
7. Niel, O., Bastard, P.: Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *American Journal of Kidney Diseases* **74**(6) (August 2019) 803–810
8. Sun, Y., Ye, Z., Zhong, M., Wei, K., Shen, F., Li, G., Yuan, J., Xing, C.: Rapid and nondestructive method for identification of molds growth time in wheat grains based on hyperspectral imaging technology and chemometrics. *Infrared Physics & Technology* **128** (December 2022) 104532
9. DTU Denmark: Team: Dtu-denmark/measurement - 2020 igem (2020)
10. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2023) 7464–7475
11. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. arXiv (April 2021)
12. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: Proceedings of the 38th International Conference on Machine Learning, PMLR (July 2021)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009)
14. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.: Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review* **55**(5) (2021) 3503–3568
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE (2017)





# Leveraging Generative AI for Enhancing Domain-Driven Software Design

Götz-Henrik Wiegand<sup>1</sup>, Filip Stepniak<sup>2</sup>, and Patrick Baier<sup>3</sup>

<sup>1</sup> Hochschule Karlsruhe University of Applied Sciences, Germany  
wigo1011@h-ka.de

<sup>2</sup> esentri AG, Ettlingen, Germany  
filip.stepniak@esentri.com

<sup>3</sup> Hochschule Karlsruhe University of Applied Sciences, Germany  
patrick.baier@h-ka.de

**Abstract.** Domain-Driven Design (DDD) is a key framework for developing customer-oriented software, focusing on the precise modeling of an application’s domain. Traditionally, metamodels that describe these domains are created manually by system designers, forming the basis for iterative software development. This paper explores the partial automation of metamodel generation using generative AI, particularly for producing domain-specific JSON objects. By training a model on real-world DDD project data, we demonstrate that generative AI can produce syntactically correct JSON objects based on simple prompts, offering significant potential for streamlining the design process. To address resource constraints, the AI model was fine-tuned on a consumer-grade GPU using a 4-bit quantized version of Code Llama and Low-Rank Adaptation (LoRA). Despite limited hardware, the model achieved high performance, generating accurate JSON objects with minimal post-processing. This research illustrates the viability of incorporating generative AI into the DDD process, improving efficiency and reducing resource requirements, while also laying the groundwork for further advancements in AI-driven software development.

**Keywords:** Generative AI, Domain-Driven Design, LoRA, QLoRA, Quantization, Consumer GPU, PEFT, Weighted Sum, Model Assessment

## 1 Introduction

Creating customer-oriented software demands efficient tools and methods. A promising approach is the Domain-Driven Design (DDD) pattern[1], a robust framework for software development emphasizing the understanding and modeling of the application’s domain. Initially, the software is described using Domain Specific Language (DSL) in JSON or UML, forming a Domain Model (DM) that underpins the iterative development process. From this DM, a code framework is derived, which is then endowed with logic to create a prototype. This prototype generates insights for refining the DM further.

The initial DM generation is typically a manual task performed by a system designer using a GUI tool. To enhance this process, we explore in this paper the possibility of partially automating it with the help of generative AI. We demonstrate how generative models can learn to create syntactically correct JSON objects for describing the DDD DM. Moreover, we show that being trained on real-world data from existing DDD projects, the AI model can automatically generate new parts of a DM through simple interactions with a system prompt. The AI model’s ability to produce syntactically correct JSON objects ensures machine readability, facilitating integration into existing DDD development tools.

Due to data confidentiality, the use of commercial Large Language Models (LLMs) are not an option, which led us to the constraint to develop the JSON code generator model on resource-restrictive hardware, specifically a single consumer-grade GPU.

The final results on the test dataset yielded impressively low loss on JSON generation and high BLEU[2] scores, underscoring the model’s proficiency. Most of the generated JSON objects exhibited syntactical correctness with minimal post-processing, and all JSON objects created from clear prompts were syntactically correct. The successful creation of a code generator for JSON objects in the DSL signifies a pivotal advancement towards incorporating generative AI into the DDD-based software development process, enhancing both efficiency and efficacy.

## 2 Related Work

The foundation for this work lies in the principles of DDD, as established by Eric Evans in his seminal works[1,3]. DDD provides a strategic approach to software development, emphasizing the modeling of complex systems based on their underlying business domains. The company internal framework used for this work builds upon these principles and knowledge regarding this were obtained from the internal documentation[4].

To address the challenges of efficient resource utilization in AI model training and deployment, techniques such as Parameter Efficient Fine-Tuning (PEFT) were employed, specifically the Low Rank Adaptation (LoRA) method introduced by Hu et al.[5]. Further refinements, including quantization methods like Quantized Low Rank Adaptation (QLoRA) by Dettmers et al.[6], played a crucial role in optimizing performance on resource-constrained hardware. The model in this work was quantized to 4-bit precision using the ‘BitsAndBytes’ library from Hugging Face[7]. This approach is supported by research on low-precision quantization, such as the work of Sun et al.[8] and Neshaei et al.[9].

The model used for the code generation component was *Code Llama* from *Meta*<sup>4</sup>, proposed by Rozière et al.[10]. Other models relevant to this field include StarCoder[11] and CodeT5[12]. Additionally, commercial AI code generation tools such as *GitHub Copilot*[13] and *Amazon CodeWhisperer*[14] provide further context and reference in evaluating the landscape of AI-assisted software development.

For the evaluation of the importance of the hyperparameters after hyperparameter Tuning we used a permutation importance analysis referencing the *Random Forest Regressor* from Louppe [15].

For performance evaluation, two key metrics, *BLEU*[16] and Loss, were used to assess the quality of the fine-tuned model. These metrics have been adopted in the evaluation of code generation models, as discussed in the works of Chen et al.[17] and Yetiştirten et al.[18], providing a foundation for assessing syntactic and semantic alignment in generated outputs.

## 3 Methods

In this section, a comprehensive outline of the methodological approach is presented, detailing the processes and techniques used for data handling, model development, and evaluation.

---

<sup>4</sup> <https://www.llama.com/code-llama/>

**Goals and Constraints:** This work explores integrating Generative AI into the software development process within a DDD framework, focusing on automating early-stage development by generating DMs from business requirements. The prototype uses causal language modeling to produce Unified Modeling Language (UML) representations in JSON format, aligning with the iterative nature of DDD.

Technical constraints include the use of open-weight models due to data privacy regulations, prohibiting commercial AI models and requiring local hosting and fine-tuning. Additionally, the project operates under a €1000 budget for external computational resources, demanding resource-efficient model selection and training. The limited dataset further challenges the generation of unbiased, generalizable results, requiring mitigation of data-induced biases. The research aims to evaluate the feasibility of AI-driven code generation within these constraints, emphasizing model performance, resource management, and compliance.

**Data Basis:** The dataset utilized for this study comprises 1,022 files, each containing a single JSON object. Of these, 821 files—accounting for 80% of the dataset—originate from a customer project, while the remaining 20% are derived from a test project. The data represent hierarchically structured DDD logic, encoded in JSON format.

Each JSON object consists of specific key-value pairs, which are defined within a specialized framework. These key-value structures are inherited from a metamodel, which serves as the basis for the framework’s logic. However, the metamodel itself is not included within the dataset, limiting direct access to the underlying inheritance structure.

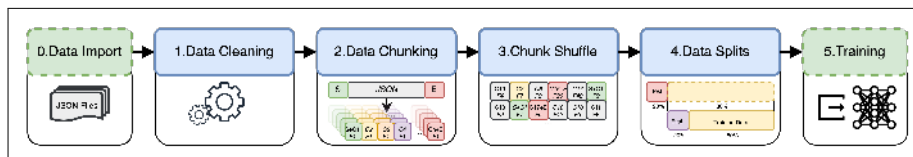
**Data Pre-Processing:** The data pre-processing step is essential for developing a robust code generator, particularly given the dataset’s significant bias, with approximately 80% of the dataset sourced from a customer project and 20% from a test project. This dataset, comprising 1,022 files containing completed JSON[19] objects, necessitates careful handling to ensure effective model training.

The pre-processing process began with data import, followed by cleaning and abstraction, where high-variability keys were replaced with placeholder values to anonymize customer-specific information. This step not only protects sensitive data but also simplifies the dataset’s complexity, allowing for a clearer focus on the JSON structure. Subsequently, the data was chunked into non-overlapping segments of 2,048 tokens, which were shuffled to enhance randomness. The final step involved a double 80:20 split[20] of the data into training, evaluation, and test sets, resulting in 64% for training, 16% for evaluation, and 20% for testing. This structured approach to data pre-processing ensures that the dataset is well-prepared for effective model training while maintaining compliance with data privacy standards. After exportation, the datasets were versioned for future use, solidifying the pre-processing phase as a foundational element in the overall development process. A full process flow of data pre-processing is displayed in Figure 1.

**Training and Setup:** The foundation of the code generator utilizes the Code Llama 7B model, released by *Meta*<sup>5</sup> proposed by Rozière et al.[10]. With a VRAM size of approximately 25 GB, it necessitated adaptations for the limited hardware available, including a local PC with an RTX 2080 GPU (11 GB VRAM) and a *Lambda Cloud*<sup>6</sup> instance with an RTX A6000 GPU (48 GB VRAM). Due to financial constraints, the

<sup>5</sup> <https://www.llama.com/code-llama/>

<sup>6</sup> <https://lambdalabs.com/service/gpu-cloud>



**Fig. 1.** Abstracted visualization of the various steps of data pre-processing with data cleaning, chunking and splitting to the various data sets for training.

cloud instance was primarily used for hyperparameter tuning. To facilitate fine-tuning on the local setup, a 4-bit quantization was applied, reducing the model’s size to around 4 GB VRAM. The LoRA method, part of PEFT, was selected due to its proven effectiveness when combined with model quantization. The *Hugging Face Transformers*<sup>7</sup> framework’s Trainer[21] was employed alongside a LoRA adapter to optimize training on constrained hardware. This was used to fine-tune *Code Llama 7B* to generate JSON with the help of Next Token Prediction from the dataset. Key training arguments were established to manage resource use, including batch size, gradient accumulation steps, and mixed-precision training. Evaluation metrics were critical for assessing model performance; while the built-in loss function was utilized, metrics such as *BLEU*[22] and *ROUGE-L-F1*[23] were employed to guide the training process. Memory overflow issues were addressed by implementing a custom function for pre-processing logits, ensuring efficient metric evaluation. Overall, the training utilized both local and cloud resources.

**Hyperparameter Tuning:** Hyperparameter tuning is crucial for optimizing model performance by selecting the most effective values for key hyperparameters. In this process, two categories of hyperparameters were identified for tuning: basic training parameters—learning rate, number of training epochs, and warm-up steps—and adapter-specific parameters such as the rank (R-value) and alpha value of the LoRA adapter[21,24,25]. These parameters were chosen due to their significant impact on the model’s performance, particularly in hardware-constrained environments.

To guide the tuning process, initial ranges were defined: a learning rate of  $1e-5$  to  $5e-5$ , 1 to 5 training epochs, 200 to 1200 warm-up steps, an R-value of 4 to 16, and an alpha value of 4 to 16[24,25]. These ranges were informed by engineering practices and recommendations from existing documentation. After 100 trials, adjustments were made to refine the search, particularly for the number of training epochs and the LoRA rank. In the second phase of tuning, the R-value range was expanded to 4 to 32, and the number of training epochs increased to 5 to 12, allowing further exploration of these critical parameters while leaving the other ranges unchanged. This iterative approach helped maximize model efficiency under the available hardware constraints.

To determine the optimal hyperparameters for final training, a multi-objective weighted sum approach was used, following Bazgan et al.[26]. The weighted sum function  $f(x)$  was initially defined for three evaluation metrics (Equation (1)).

However, hyperparameter tuning results showed that the *ROUGE-L-F1* metric was outside the expected range. As a result, the weight for *ROUGE-L-F1* was set to zero in Equation (2), removing its influence on  $f(x)$ .

Finally, introducing the Inverse Loss  $\tilde{L}(x)$  as  $1 - L(x)$ , the weighted sum was simplified to  $f(x)$  in Equation (3).

<sup>7</sup> <https://huggingface.co/docs/transformers/index>

$$f(x) = w_{\text{Loss}} \cdot (1 - L(x)) + w_{\text{BLEU}} \cdot B(x) + w_{\text{ROUGE-LF1}} \cdot R(x) \quad (1)$$

$$f(x) = w_{\text{Loss}} \cdot (1 - L(x)) + w_{\text{BLEU}} \cdot B(x) + 0 \cdot R(x) \quad (2)$$

$$f(x) = w_{\text{Loss}} \cdot \tilde{L}(x) + w_{\text{BLEU}} \cdot B(x) \quad (3)$$

**Model Assessment** To comprehensively evaluate the model’s performance for a generative DDD system, a three-phase assessment approach was used, as traditional metrics alone offer limited insight. In the first phase, the evaluation metrics Loss and *BLEU*[16] from both the training and test datasets were reviewed. The second phase assessed the syntactic correctness and machine-readability of the generated JSON objects. Here, 100 JSON samples were generated from 10 *clear* and 10 *experimental* prompts. *Clear* prompts specify a distinct DDD class object, guiding the model to create a corresponding JSON object, while *experimental* prompts progressively reduce detail, giving the model more room for errors and issues. If any of the generated samples exceeded the token length limit of 4,000, post-processing was applied to ensure completeness, followed by verification through a JSON parser. The final phase involved a qualitative review of the generated JSON objects to identify potential errors and issues. This multi-step evaluation offers a more detailed understanding of the model’s quality and its suitability for real-world applications.

## 4 Results and Discussion

In this section, the results of the Hyperparameter Tuning, Final Model Training, and Model Assessment are summarized and discussed.

**Hyperparameter Tuning:** Table 1 displays the results of the top three values for the different evaluation metrics (objectives) from the hyperparameter tuning process.

The results from the hyperparameter tuning also allow for the derivation of the importance of individual hyperparameters. A Permutation Importance analysis was conducted using a *Random Forest Regressor*[15] to assess the influence of each parameter on the evaluation metrics (objectives) shown in Figure 2. This method helps to quantify how changes in specific hyperparameters affect the model’s performance.

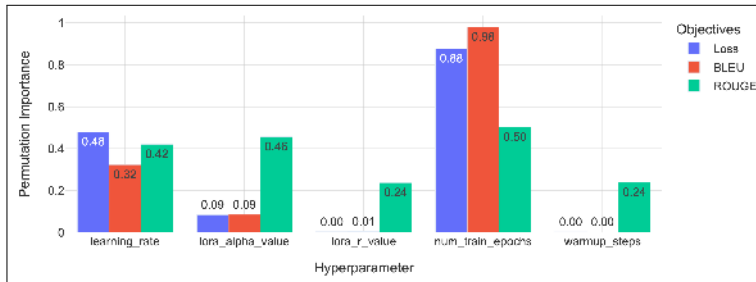
In the analysis of the results from hyperparameter tuning, a noticeable discrepancy was observed between the expected and actual values of the ROUGE-*L-F1* score. The ROUGE-*L-F1* score was anticipated to approach 1. During hyperparameter tuning, it reached a maximum of only approximately 0.062 in the second trial (see Table 1). Due to this significant deviation, the ROUGE-*L-F1* metric was excluded from the determination of the optimal hyperparameters.

Subsequently, the weighted sum method described in Section 3 was applied and calculated for each trial. Assuming that all evaluation metrics converge towards 1 (using the inverse loss as  $1 - \text{Loss}$ ), it can be inferred that the trial with the maximum weighted sum defines the optimal hyperparameters, denoted as  $\theta^*$ . Table 2 presents the top five trials, ranked by their weighted sum along with their respective objectives. Figure 3 illustrates the convergence of the weighted sum towards 1, in relation to BLEU and inverse loss, providing a visual representation of this progression.

The maximum of the weighted sum  $f(x)$  with  $\theta^* = \max(f(x)) \mid x \in \text{Trials}$  is reached at trial 116. Concluding to  $\theta^* = \theta_{f(\text{Trial}_{116})}$ . Therefore, trial 116 defines the optimal

Trial Number	Ranking	Loss↓	BLEU↑	ROUGE-L-F1↑	Learning Rate	LoRA Alpha	LoRA R	Train Epochs	Warmup Steps
116	1. Loss	<b>0.031224</b>	0.991329	0.046125	3.4e-05	30	10	6	448
128	2. Loss	0.03168	0.990963	0.04753	3.5e-05	29	13	6	419
127	3. Loss	0.031686	0.991554	0.046529	4.2e-05	22	5	6	1044
110	1. BLEU	0.034611	<b>0.991905</b>	0.047163	3.8e-05	17	10	12	1194
124	2. BLEU	0.032367	0.991696	0.047422	3.4e-05	24	10	9	974
125	3. BLEU	0.033012	0.991665	0.04615	3.3e-05	27	10	9	968
2	1. ROUGE	0.054714	0.987621	<b>0.062322</b>	1.3e-05	16	4	2	1033
130	2. ROUGE	0.038864	0.99047	0.061189	3e-05	11	9	6	675
42	3. ROUGE	0.04477	0.989132	0.05983	3.1e-05	12	11	2	852

**Table 1.** List of the top three Trials of hyperparameter tuning for each evaluation metric (objective) along with the marking of the best values for each objective.



**Fig. 2.** Parameter importance for multiple evaluation metrics (*Objectives*) with importance calculated using permutation importance with *Random Forest Regressor*[15].

hyperparameters  $\theta^*$  for the final training as follows:  $\theta_{\text{learning\_rate}}^* = 3.4e - 5 = 0.000035$ ,  $\theta_{\text{num.train.epochs}}^* = 6$ ,  $\theta_{\text{warmup.steps}}^* = 448$ ,  $\theta_{\text{lora.r.value}}^* = 10$  and  $\theta_{\text{lora.alpha.value}}^* = 30$ .

**Final Model Training:** The final model training was conducted using the optimal hyperparameters, denoted as  $\theta^*$ , which were determined during the hyperparameter optimization phase. This training was performed on an NVIDIA RTX 2080 GPU with 11 GB of VRAM. Key statistics related to training times and memory usage are summarized in Table 3.

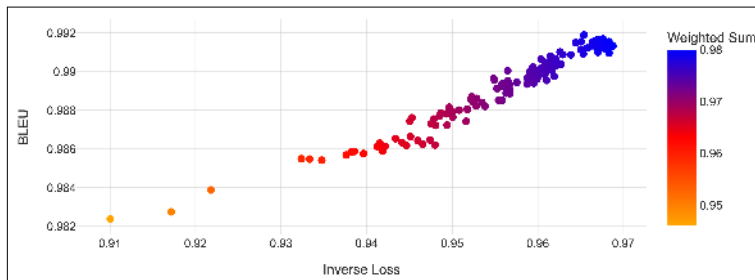
Figure 4 illustrates the progression of the loss function for both the evaluation and test datasets. The training loss shows a high degree of fluctuation, while the evaluation loss remains stable throughout the process. Given that the evaluation loss follows a similar trend to the training loss without significant deviation, it can be inferred that overfitting did not occur during the training.

**Model Assessment** In addition to the final training evaluation metrics (Table 3), the results on the test dataset are shown in Table 4. A slight improvement in loss by 0.0028 and a minor decrease in BLEU by 0.0006 were observed, both considered negligible, indicating stable model performance.

During model assessment, hardware limitations restricted generation to 4,000 tokens per sample. Of 100 JSON samples from 20 prompts, only one terminated correctly within this limit. After post-processing to remove incomplete key-value pairs and close JSON objects, 81 out of 100 samples were successfully parsed.

Trial Number	Loss↓	Inverse Loss↑	BLEU↑	ROUGE-L-F1 ↑	Weighted Sum $f(x)$ ↑
116	0.0312	0.9688	0.9913	0.0461	<b>0.9801</b>
127	0.0317	0.9683	0.9916	0.0465	0.9799
124	0.0324	0.9676	0.9917	0.0474	0.9797
108	0.0319	0.9681	0.9912	0.0488	0.9797
102	0.032	0.968	0.9913	0.0482	0.9797

**Table 2.** List of the top five trials with the highest results for the weighted sum  $f(x)$  sorted in descending order.



**Fig. 3.** Weighted sum in reference to inverse Loss  $\tilde{L}(x)$  and BLEU  $B(x)$ .

Two prompt types were used: *experimental* and *clear* (see Section 3). All 50 JSON samples from the *clear* prompts were parsed without errors after post-processing. However, 19 parsing errors occurred in samples from the *experimental* prompts.

The third phase involved qualitative analysis. Most samples were content-wise comparable to the original dataset, but limitations emerged. In some cases, the model repeated certain sections (e.g., Field Model) until reaching the 4k token limit. This repetitive behavior is permitted in JSON structure but unrealistic for real-world applications, requiring further investigation with more computational resources.

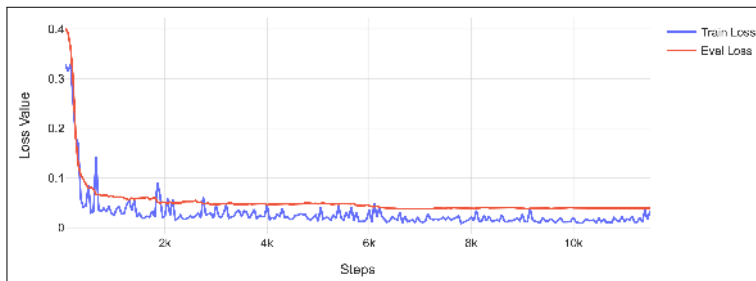
The analysis of the erroneously generated samples revealed two main types of issues:

1. Some generated JSON objects began within another JSON object, such as within a key-value pair. This led to parsing problems due to violations of JSON syntax and structure. This error is likely caused by data chunking during preprocessing.
2. The generation of unwanted characters, such as the zero-width space Unicode symbol (Unicode U+200B [27]), was observed. Since these characters were not present in the training data, it is assumed that they are artifacts originating from the Code Llama 7B model.

**Results Conclusion:** We can conclude that the model is capable of generating machine-readable JSON objects when given appropriate prompts. Fine-tuning on a GPU with sufficient memory produced strong results, and the abstraction of datasets reduced complexity, allowing for the use of customer data. This first version is already capable of being integrated into possible applications with respect to the described bias.

Training Duration	Training Steps	Loss↓ (Train Data)	Loss↓ (Eval Data)	BLEU↑ (Eval Data)
36.43hr	11.54k	0.0337	0.0393	0.9924
Model Size (Quantized)	LoRA File Size	Train Dataset Size	Eval Dataset Size	Available VRAM
4.0046GB	21MB	30.3MB	7.4MB	11GB

**Table 3.** Summary of the results of the final training.



**Fig. 4.** The development of the Training Loss (blue) and Evaluation Loss (red) are plotted over the training steps of the final training, with an update every 50 steps.

Loss ↓	BLEU↑	ROUGE-1-F1 ↑	ROUGE-2-F1 ↑	ROUGE-L-F1 ↑	ROUGE-L-F1 ↑
0.0309	0.9918	0.0565	0	0.0565	0.0565

**Table 4.** Results of evaluation of the model from final training using the test dataset. In this case, the ROUGE values were not within the expected range and were included only for completeness.

## 5 Conclusions

The rapid development of AI technology and the increasing prevalence of LLMs have created opportunities for new applications and tools. With continuous improvements in the efficiency of LLMs, they are increasingly being used in the consumer and Small and Medium-Sized Enterprises (SME) sectors. This work demonstrates the potential and capabilities of open LLMs, the extent to which the development of efficient model trainings with approaches such as LoRA and quantization has already progressed and how these can be combined under restrictive resources.

The work of this paper mark an initial step towards a generative AI assistance system showed the potential for further development into an assistant system for DDD software development framework. In particular, the ability to generate machine-readable JSON objects enabled the use of the final model of this paper in potential tool chains and systems. The results and findings as well as the limitations and challenges, form a broad basis for further development.

Thus, the work of this thesis is a functional code generation model prototype that offers further possibilities and learning, paving the way for the development of an “*Artificial Intelligent*” assistant system that meets the requirements and needs of DDD software development.

The code we used to train and evaluate our models is available at <https://github.com/Tr33Bug/DomainlifecyclesCodeGenerator>.



## References

1. Evans, E.: Domain-Driven Design: Tackling Complexity in the Heart of Software. Addison-Wesley (2004)
2. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., Lin, D., eds.: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics (July 2002)
3. Evans, E.: Domain-Driven Design Reference: Definitions and Pattern Summaries. Dog Ear Publishing (2014)
4. esentri, c.: Internal Company Documentation (2024)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
6. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms (2023)
7. Hugging Face: Bits and bytes documentation v0.44.1. <https://huggingface.co/docs/bitsandbytes/main/en/index> (2024)
8. Sun, X., Wang, N., Chen, C.y., Ni, J.m., Agrawal, A., Cui, X., Venkataramani, S., El Maghraoui, K., Srinivasan, V., Gopalakrishnan, K.: Ultra-low precision 4-bit training of deep neural networks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Red Hook, NY, USA, Curran Associates Inc. (2020)
9. Neshaei, S.P., Boreshban, Y., Ghassem-Sani, G., Mirroshandel, S.A.: The impact of quantization on the robustness of transformer-based text classifiers (2024)
10. Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C.C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., Synnaeve, G.: Code llama: Open foundation models for code (2024)
11. Li, R., Allal, L.B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T.Y., Wang, T., Dehaene, O., Davaadorj, M., Lamy-Poirier, J., Monteiro, J., Shliazhko, O., Gontier, N., Meade, N., Zebaze, A., Yee, M.H., Umapathi, L.K., Zhu, J., Lipkin, B., Oblokulov, M., Wang, Z., Murthy, R., Stillerman, J., Patel, S.S., Abulkhanov, D., Zocca, M., Dey, M., Zhang, Z., Fahmy, N., Bhattacharyya, U., Yu, W., Singh, S., Luccioni, S., Villegas, P., Kunakov, M., Zhdanov, F., Romero, M., Lee, T., Timor, N., Ding, J., Schlesinger, C., Schoelkopf, H., Ebert, J., Dao, T., Mishra, M., Gu, A., Robinson, J., Anderson, C.J., Dolan-Gavitt, B., Contractor, D., Reddy, S., Fried, D., Bahdanau, D., Jernite, Y., Ferrandis, C.M., Hughes, S., Wolf, T., Guha, A., von Werra, L., de Vries, H.: Starcoder: may the source be with you! (2023)
12. Wang, Y., Wang, W., Joty, S., Hoi, S.C.H.: Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation (2021)
13. GitHub Inc.: Github copilot. <https://github.com/features/copilot> (2024)
14. Amazon Web Services (AWS): AWS codewhisperer. <https://aws.amazon.com/de/codewhisperer/> (2024)
15. Louppe, G.: Understanding random forests: From theory to practice (2015)
16. Tran, N., Tran, H., Nguyen, S., Nguyen, H., Nguyen, T.: Does bleu score work for code migration? In: 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC). Volume 10., IEEE (May 2019) 165–176
17. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam,

- J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating large language models trained on code (2021)
18. Yetiştirilen, B., Özsoy, I., Ayerdem, M., Tüzün, E.: Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt (2023)
  19. Bourhis, P., Reutter, J.L., Suárez, F., Vrgoč, D.: Json: data model, query languages and schema specification (2017)
  20. Joseph, V.R.: Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **15**(4) (April 2022) 531–538
  21. Hugging Face: Trainer class for fine-tuning models (2023)
  22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., Lin, D., eds.: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics (July 2002) 311–318
  23. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*, Barcelona, Spain, Association for Computational Linguistics (July 2004) 74–81
  24. Hugging Face: Performance considerations for training on gpus. [https://huggingface.co/docs/transformers/main/en/perf\\_train\\_gpu\\_one](https://huggingface.co/docs/transformers/main/en/perf_train_gpu_one) (2024)
  25. Hugging Face: Lora (low-rank adaptation) for training diffusion models. <https://huggingface.co/docs/diffusers/training/lora> (n.d.)
  26. Bazgan, C., Ruzika, S., Thielen, C., Vanderpooten, D.: The power of the weighted sum scalarization for approximating multiobjective optimization problems. *Theory of Computing Systems* **66**(1) (November 2021) 395–415
  27. The Unicode Consortium: The unicode standard, version 15.0. <https://www.unicode.org/versions/Unicode15.0.0/UnicodeStandard-15.0.pdf> (2022)

# Prediction of PV Power Production with Neural ODEs on the base of Weather data

Lukas Schwab<sup>1</sup>, Louis Emier<sup>1</sup>, Paul Machauer<sup>1</sup>, Michael Quarti<sup>2</sup>, and Rainer Gasper<sup>2</sup>

<sup>1</sup> University of Applied Science Offenburg  
{lschwab,lemier,pmachau}@stud.hs-offenburg.de  
<sup>2</sup> INES – University of Applied Science Offenburg  
{michael.quarti,rainer.gasper}@hs-offenburg.de

**Abstract.** Predicting energy production from photovoltaics (PV) is crucial for efficient energy management. In order to apply different operating strategies, it is necessary to predict the expected amounts of PV energy. The operating strategies are typically optimized with regard to economic or technical goals or a combination of both. Within this work, we show a possibility to predict PV power production using local weather data and Neural Ordinary Differential Equations (NODE). Based on the measured values from the PV system and an associated weather station, the NODE is trained and validated with regard to PV production. The measurement data are collected from the PV system of the former Campus North of Offenburg University of Applied Sciences.

**Keywords:** PV Power Production, time series modeling, NODE

## 1 Introduction

Predicting energy production from photovoltaics (PV) is crucial for efficient energy management. In order to be able to apply different operating strategies, it is necessary to forecast the expected generation of PV power. Models for forecasting PV generation can be categorized by the forecast horizon like (1) short-term forecasting for minutes or an hour [1, 2], (2) middle-term forecasting for about a day [3, 4] and (3) long-term forecasting for a couple of days up to months [5, 6]. It is common to use neuronal networks for all of the forecasting horizon. Within this work, we show a possibility to predict PV power production using local weather data using Neural Ordinary Differential Equations (NODE) [7]. The authors of Ref. [8] modelled a lithium-ion battery to predict the behavior of a lithium-ion battery using NODE and grey-box modelling. Based on the measured values from the PV system and an associated weather station, our model is trained and validated with regard to PV production. The measured data is collected from the PV system of the former Campus North of Offenburg University of Applied Sciences from the years 2017-2021. The PV system consisted of 3 strings with a nominal output of 2.16 kWp each. The strings have a collector inclination of 30° and 40° and an orientation of 18° south.

The article will be organized as follows. In the Introduction we will give an overview of the state of the art for PV power prediction and the use of Neural ODEs for modeling of physical systems. In the second chapter we give a short introduction to Neural ODEs for the prediction of time series. After that, we present the data and will describe the data preparation steps. The data preparation includes the resampling to a half hourly base. We also applied Principal Component Analysis and Fourier Transform to the data. After the preparation of the data we will discuss the hyper parameter tuning for the

NODE. In the next step we present the training and validation results of the NODE for the PV power prediction. We will conclude the article with a conclusion and discussion.

## 2 Overview Neural Ordinary Differential Equations

We will use Neural Ordinary Differential Equations (NODEs) [7] for the prediction of the PV power production. NODEs can be understood as a Residual Neural Networks, given by

$$x(t_{k+1}) = hf(x(t_k), t, \Theta) + x(t_k) \text{ where } h = \Delta t, \quad (1)$$

where the discretization  $h = \Delta t$  goes to zero and the function changes from discrete to continuous

$$\frac{dx}{dt} = f(x(t), t, \Theta) \quad (2)$$

where the Neural Network  $f(x(t), t, \Theta)$  with the parameters  $\Theta$  represents the continuous change of the hidden state  $x(t)$ . Or in other words Residual Neural Networks are the Euler discretization of NODEs [9–11]. In contrast to traditional neural network architectures with discrete layers, NODEs model the continuous evolution of hidden states with ordinary differential equations. More concrete, the input features don't pass through a fixed number of layers, the hidden states evolve over a time interval by using a solver. The differential equations themselves are represented by Neural Networks and describe the rate of change of the hidden states over time. The Neural Network is then integrated with a solver to obtain the trajectory of the hidden states over the time. In the context of solving ordinary differential equations, the input features can be interpreted as initial values for the hidden states. For the training of the NODEs we are using the so-called discretize-then-optimize approach where we backpropagate through the internal operation of the ode solver to obtain the gradients [12]. Because of the mentioned properties, NODEs are very well suited for the modeling of time series. For the same reasons NODEs are also well suited for the modeling of physical systems that can be expressed as differential equations. The NODEs defined above solve an initial value problem and the number of inputs/features has to be the same as the number of outputs. In the prediction of the PV power, the system depends not only on the initial values but also on additional inputs, most of all the weather. Therefore, the number of inputs is greater than the number of outputs. We consider a so called forced nonlinear system [13]

$$\frac{dx}{dt} = f(x(t), u(t), t, \Theta) \quad (3)$$

where additional information about the weather, e.g., is cumulated in the input vector  $u(t)$ . Note that the solution of the nonlinear system does not depend only on the initial condition of the states  $x(0)$  but also on the input vector  $u(t)$ . It is important to note that the forced nonlinear system with the additional input vector  $u(t)$  can be solved and trained with the same methods as the initial value problem. It has to be taken care that the input vector  $u(t)$  is an external function of time and is not computed by the ode solver (as the state vector  $x$ ) and if the inputs  $u(t)$  are given as data points at discrete time steps  $t_k$  instead as an explicit function of time, an appropriate interpolation method has to be chosen.

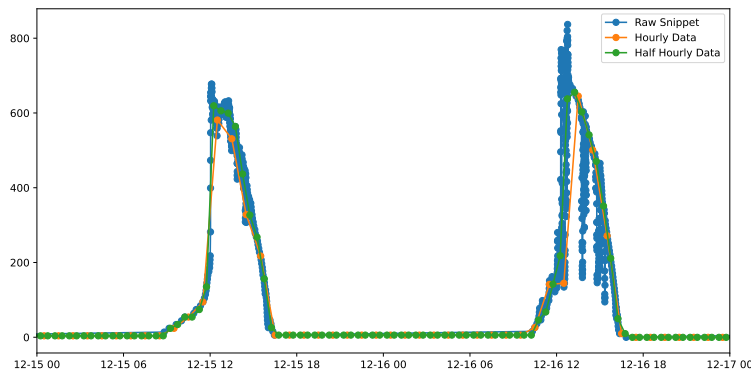
### 3 Data Preparation and Analysis

Before we can train and test the NODE, we have to do the post processing of the data. That includes the resampling to a half hourly and hourly base. We also applied Principal Component Analysis and Fourier Transform to the data.

#### 3.1 Resampling

To process the data, it must first be converted to the equidistance format. To this end, a resampling procedure was devised for the purpose of incorporating hourly and half-hourly values.

In the case of hourly sampled data at 11 a.m., the sampled values are calculated by aggregating all values between 10:30 a.m. and 11:30 a.m., thereby avoiding any consideration of future values. In order to guarantee the inclusion of data within the resampled data set at all times, any absent values are incorporated through the utilisation of a forward fill procedure. This is only feasible due to the fact that the recorded values of the sensors are only stored when a change is detected; otherwise, there are no values at specific points in time. It is important that the forward fill does not use the final value



**Fig. 1.** Comparison between raw data, half-hourly re-sampled data and hourly re-sampled data for 48 hours, so that you can recognize the noon high twice.

from the resampling, but rather incorporates the last known value of the raw data into the missing values. This guarantees the precision of the resampled value over the course of the night. The resampled value is the average of the last hour with sunshine. As the measured values remain constant during the night, the period without sunshine is not considered. By using the last raw data value, the values over the night are accurate. Figure 1 shows an example of the resampled data for 32 hours.

#### 3.2 Principal Component Analysis (PCA)

For further insight on the relevance of the features and comparison of the resampling rates (3.1) principal component analysis (PCA) is used. It is for data analysis purposes

only and is not used to transform the inputs of the NODE. PCA is suitable because the maximum variance of the data is retained in the principal components. The highest eigenvalues in table 1 thus show the relevance of the individual features in the data set from the point of view of variance.  $S$  represents the covariance matrix and  $u_1$  the highest eigenvector [14].

$$u_1^T S u_1 = \lambda_1 \quad (4)$$

where  $\lambda_1$  is the first eigenvalue. In the hourly data set, the three largest principal components explain 92.11 % of the variance and in the half-hourly 93.21 %. Since the three principal components are able to explain the data set with little loss in variance, the mean of  $u_1 - u_3$  is used for each feature. Considering the variance to be obtained, the relevance and order of the features of the hourly and half hourly dataset is similar. So the hourly data set contains hardly any less information.

**Table 1.** Each column shows the values for the separate features. The values are calculated from the three largest eigenvectors like described in the section above. The table also serves the purpose of comparing the hourly and half hourly dataset.

Parameter	Hourly sampled data	Half hourly sampled data
Wind direction	0.508	0.515
Moist	0.493	0.486
Power	0.322	0.324
String 40° irradiation	0.247	0.270
Global solar irradiation	0.224	0.230
Pressure	0.198	0.105
Temperatur	0.139	0.075
Wind speed	0.093	0.033

Due to the similarity of the two datasets it is more reasonable to only focus on one dataset.

When observing the results of table 1 it is noticeable that wind direction, moist and power are considered more important than irradiation or temperature. This does not meet the intuitive expectations. One has to keep in mind, that the PCA components explain the variance in the data, and not a causal effect or dependency.

As soon as there are two similar features, usually one feature is left out. This was not done in the case of String 40° irradiation and global solar irradiation. Both features describe something similar.

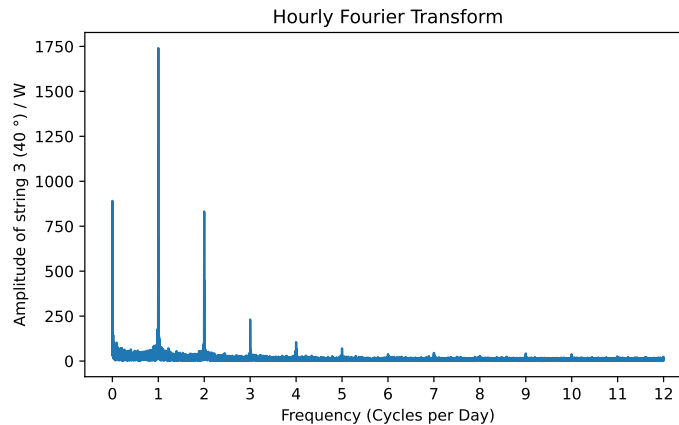
In conclusion, PCA is used to recognise which features may be most important in order to test which data is necessary to collect in the future. However, reduced data sets would first have to be tested and compared with the actual application on the NODE. Additionally, it is beneficial to note that both resampling rates are good to use.

### 3.3 Fourier Transform

To further validate that the derived equidistant timeseries accurately represent the non-equidistant raw data, we generate spectrograms using Fourier transformations. The spectrograms are then utilized to gain insights into the prevalent frequencies within the

equidistant timeseries. We anticipate observing the daily solar cycle reflected in our derived timeseries of power production.

Different sampling frequencies are used for hourly and half-hourly timeseries to ensure good interpretability of frequencies across all spectrograms. For the hourly timeseries we use a sampling frequency of  $1/3600$  Hz which equals 1 h and for half-hour timeseries we use  $1/1800$  Hz which equals half a hour.



**Fig. 2.** Spectrogram of the fourier transformation using the hourly sampled data and a sampling frequency of  $1/3600$  Hz.

Figure 2 shows the resulting spectrogram for the hourly timeseries. Because of our previously chosen sampling frequency we are able to interpret the frequency-axis in cycles per day. The highest amplitude appears at a frequency of 1, a strong indication for the presence of daily sun cycle within our dataset. Similar results are obtained when using the half-hourly timeseries.

### 3.4 Hyperparameter tuning

The time-consuming step of finding hyperparameters is automated with Optuna’s Bayesian optimization [15] in combination with MLflow [16] for tracking. This allows for the collection of hyperparameters that result in a low loss within a reasonable time frame.

The optimization takes place in a separate validation loop. An average loss for a whole epoch is calculated from all weekly accumulated losses. Furthermore, the search space is defined by the following parameters: *tanh* and Rectified Linear Unit (ReLU) as activation functions for the neural network. ADAM, RMSprop and SGD for gradient descent. The number of neurons from 30 to 300. Learning rate in the of magnitude from  $10^{-5}$  to  $10^{-3}$ .

We used the optimization algorithms from PyTorch’s `torch.optim` package [17].

In the majority of runs, Bayesian optimisation utilizes the rectified linear unit (ReLU) activation function and the adaptive moment estimation (ADAM) optimizer. For this reason NODE seems to work best with ReLU and ADAM in the current use case. The number of neurons and the learning rate don’t differ significantly for the best runs.

**Table 2.** The columns 1 to 3 show the three best runs of the hyperparameter tuning under consideration of the weekly accumulated loss.

	1.	2.	3.
Activation function	ReLU	ReLU	ReLU
Optimizer	ADAM	ADAM	ADAM
Neurons	171	192	140
Learning rate	$8.77 \cdot 10^{-4}$	$6.68 \cdot 10^{-4}$	$5.84 \cdot 10^{-4}$
Weekly accumulated loss	94803	95228	95334

Notably, the row learning rate could be replaced by a clever choice of learning rate scheduler.

## 4 Results for the prediction of PV power production

The following results were achieved using our best hyperparameter configuration (see table 2). For our experiments we split both, the hourly and half-hourly, dataset into training (70 %) and validation (30 %) sets. A single training example contains one week of data. The neural network receives 6 weather features (wind direction, wind speed, temperature, global irradiation, moisture & pressure) as well as two positional encodings (time of day & current month) and the last preceding prediction. We observe promising convergence results with just a few weeks of training data, reaching diminishing convergence results within the first epoch of training (167 given example weeks).

Training is stopped after the first epoch to prevent overfitting on the dataset. The trained model achieves a mean error of 265.40 W and a median error of 30.24 W. These errors amount to 12.29 % and 1.40 % respectively when measured in relation to the 2.16 kWp of installed PV capacity over the validation set of 41 weeks. The large differences between these metrics can be explained by outliers which are caused by sudden cloud shading. It’s important to mention that the focus of the project is not to predict the impact of individual clouds on the power production but rather larger timespans (e. g. > 3 h).

The results for predicting one week of PV power are shown in figure 3. As mentioned above, the model is not able to predict the peaks due to cloud shading. However, the overall prediction is in good agreement with the measurement.

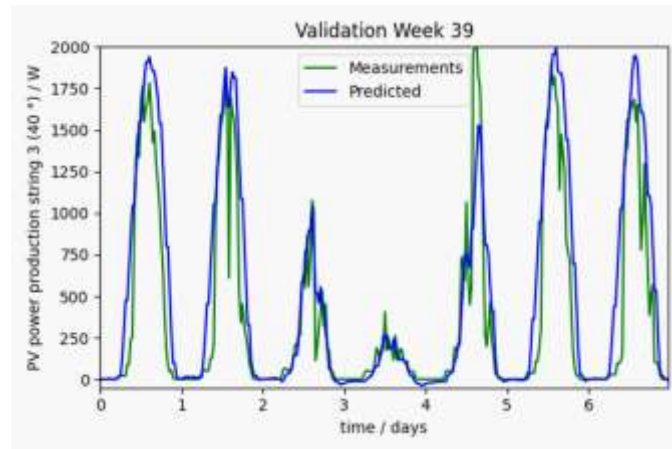
A more informative (conclusive) way to measure the error might be to exclusively evaluate the day time phases where the model is not able to receive a “free lunch” by predicting a power output of 0 W. In this case the mean error is 401.37 W (18.58 %) and the median error 214.12 W (9.91 %).

### 4.1 Conclusion

The study has demonstrated that it is feasible to forecast PV power data using a neural ordinary differential equation (ODE). The model is capable of accurately forecasting the photovoltaic (PV) output of individual days.

There are minor discrepancies between the original and predicted values, particularly at the peak output observed at midday. Furthermore, the model does not anticipate abrupt declines in output, which may be attributed to cloud cover obstructing the photovoltaic panels. Furthermore, in exceptional circumstances, the model may occasionally





**Fig. 3.** Comparison of predicted and measured PV power production of string 3 with a collector inclination of  $40^\circ$

predict negative PV power values for the night. It is not possible for negative power values to exist in the real world. Despite the aforementioned limitations, a generalised prediction of PV output for the day can be made.

A comparison of the prediction results with those of other methods and models was not conducted. This represents a potential avenue for further investigation. At present, the model is configured to receive only the most recent power value in order to facilitate the next prediction. One potential avenue for further investigation would be to modify the model in a way that incorporates a greater quantity of historical data into the next prediction step. The current model is only capable of making predictions one step into the future at a time. The incorporation of a multi-step output could facilitate the prediction of PV output at an extended temporal horizon.

#### 4.2 Acknowledgement

We gratefully acknowledge financial support by the Carl-Zeiss Foundation through foundation professorship Mechatronic Systems Engineering.

#### References

1. Wolff, B., Kühnert, J., Lorenz, E., Kramer, O., Heinemann, D.: Comparing support vector regression for pv power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. *Solar Energy* **135** (October 2016) 197–208
2. Mellit, A., Pavan, A.M., Benghanem, M.: Least squares support vector machine for short-term prediction of meteorological time series. *Theoretical and Applied Climatology* **111**(1–2) (May 2012) 297–307
3. Ekici, B.B.: A least squares support vector machine model for prediction of the next day solar insolation for effective use of pv systems. *Measurement* **50** (April 2014) 255–262
4. Lima, F.J., Martins, F.R., Pereira, E.B., Lorenz, E., Heinemann, D.: Forecast for surface solar irradiance at the brazilian northeastern region using nwp model and artificial neural networks. *Renewable Energy* **87** (March 2016) 807–818

5. Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petković, D., Sudheer, C.: A support vector machine–firefly algorithm-based model for global solar radiation prediction. *Solar Energy* **115** (May 2015) 632–644
6. Sobri, S., Koochi-Kamali, S., Rahim, N.A.: Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management* **156** (January 2018) 459–497
7. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations (2019)
8. Brucker, J., Behmann, R., Bessler, W.G., Gasper, R.: Neural ordinary differential equations for grey-box modelling of lithium-ion batteries on the basis of an equivalent circuit model. *Energies* **15**(7) (April 2022) 2661
9. Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations (2020)
10. Haber, E., Ruthotto, L.: Stable architectures for deep neural networks. *Inverse Problems* **34**(1) (December 2017) 014004
11. Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations (2018)
12. Kidger, P.: On neural differential equations (2022)
13. Vidyasagar, M.: *Nonlinear Systems Analysis*. Society for Industrial and Applied Mathematics (January 2002)
14. M., C.: *Pattern Recognition and Machine Learning*. 1 edn. Information Science and Statistics. Springer, New York, NY (August 2006)
15. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework (2019)
16. Zaharia, M.A., Chen, A., Davidson, A., Ghodsi, A., Hong, S.A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Xie, F., Zumar, C.: Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.* **41** (2018) 39–45
17. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhersch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S.: PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24), ACM (April 2024)

# A Commissioning-Oriented Fault Detection Framework for Building Heating Systems Using SARIMAX Models

Parantapa Sawant<sup>1</sup>, Ralph Eismann<sup>1</sup>

<sup>1</sup>Institute for Sustainability in Energy and Construction (INEB),  
University of Applied Sciences Northwest Switzerland (FHNW)  
`parantapa.sawant@fhnw.ch`

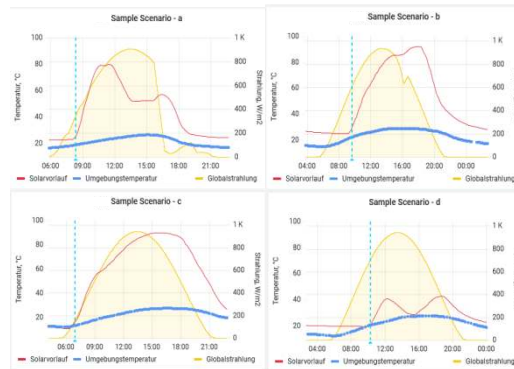
## Abstract

A scalable and rapidly deployable fault detection framework for building heating systems is presented. Unlike existing data-intensive machine learning approaches, a SARIMAX-based concept was implemented to address challenges with limited data availability after commissioning of the plant. The effectiveness of this framework is demonstrated on real-world data from multiple solar thermal systems, indicating potential for extensive field tests and applications for broader systems, including heat pumps and district heating.

**Keywords:** Building Technologies; Data-Driven Fault Detection; SARIMAX.

## 1 Introduction

From 2019 to 2023, LoRaWAN-based temperature sensors were installed to monitor the outlet of approximately 450 building solar thermal systems [1]. We developed a rule-based algorithm (RBA) for fault detection and diagnosis (FDD) by leveraging extensive operational expertise and achieved around 95% accuracy. However, the RBA faced scalability issues due to the variations in plant characteristics, dependent on type of installation or control strategies amongst other factors, that cannot be captured with just one sensor per plant. While it meets practical requirements, our current goal is to develop a complementary data-driven algorithm capable of rapid fault detection (FD) across various installations, and potentially extendable to heat pumps and district heating systems. This work presents the results of the first proof-of-concept. The ability of machine learning (ML) algorithms for FDD in building HVAC systems especially solar thermal systems is well documented [2–9]. These studies employed extensive process-history data from multiple sensors, simulation data, and used complex models such as random-forest-regression to detect numerous faults and in certain cases, diagnose them. We also attempted to train ML models using results of the RBA but were hindered by a lack of labeled data representing ground truth and the intrinsic limitations of training ML models to replicate the RBA [10]. Instead, we propose a novel approach utilizing a time series forecasting model that strikes a balance between scalability and minimal data available post-commissioning, facilitating preliminary FD using only the single data-point per installation.



**Fig. 1.** Examples for different scenarios that may occur in a solar thermal plant.<sup>1</sup>

**Fig. 1** shows the solar output temperature for four different installations under similar environmental conditions. For example, scenario a and scenario b: A monotonic, concave temperature increase over a period until a rapid temperature change is observed due to cloud cover or peak load, etc. Scenario c: A completely monotonic, concave temperature curve until sunset or lower solar radiation. Scenario d: A special case where none of the above occurs, although there is an initial temperature rise and a fault in the system could be the cause. These are just a few examples among many others, including night cooling or stagnation in solar thermal systems, for which the reader is referred to previous work by the authors [1]. Different types of systems will behave differently due to their design and control measures for similar ambient temperatures and global irradiance. A single model trained on all data will be limited in predicting failures for all plants. We want to explore an approach that uses minimal and readily available historical data from a plant to be implemented for that plant and easily integrated into a pipeline for widespread implementation. The concept of implementing this pipeline is also discussed in this paper.

## 2 Methodology

The data was collected from solar thermal systems at 10-minute intervals, and for this project we have one to two years of data per system. However, since future installations may not have extensive historical data, the focus is on developing a model that requires minimal data for fitting. So, for the exploratory analysis and visual analysis of the time series we used the entire data set of a plant and for the model training and fitting phase we assume 3 past days are enough for predicting the target day. This assumption was empirically evaluated by comparing training models using past 1 and 2 days and the results are discussed later in this paper. The data handling tasks were performed using *Python-Pandas v2.1.1*, the model fitting and diagnosis was performed using the *statsmodels v0.14.0* [11] and the *pmdarima v2.0.4* [12] package on a laptop equipped with an 11th Gen Intel(R) Core(TM) i9-11950H @ 2.60GHz and 32GB Ram. Duplicate timestamps were identified and removed, with only the first occurrence retained to maintain temporal accuracy. Missing values were addressed using linear interpolation for gaps up to four hours, reducing the percentage of missing data from 0.30% to 0.13%, thus preparing the dataset for subsequent analysis.

<sup>1</sup> Solarvorlauf/Collectorvorlauf = Collector outlet temperature, Umgebungstemperatur = Ambient temperature, Globalstrahlung = Global irradiation

### 2.1 Exploratory Data Analysis (EDA)

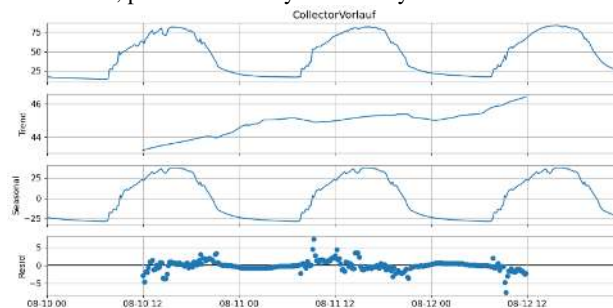
Each dataset comprises four key variables: collector outlet temperature (in Celsius), representing the temperature of the fluid exiting the solar thermal collectors; ambient temperature (in Celsius), which measures the surrounding environmental temperature; cloud cover (in percentage), indicating the percentage of cloud coverage, with 100% representing fully cloudy conditions; and global irradiation (in Watts per square meter), measuring the total solar radiation received per square meter. A statistical analysis was conducted on these variables, calculating their mean, minimum, and maximum values, and producing box plot visualizations for the purpose of assessing data quality, identifying outliers, and evaluating the spread of values. It was found that all sensor readings fell within the expected operational ranges for Switzerland, thereby confirming the validity of the dataset for further analysis. The correlation between variables was assessed using *Spearman's rank correlation coefficient*, as shown in **Table 1**. A strong positive linear correlation was detected between collector outlet temperature and both ambient temperature and global irradiation. The effect of cloud cover was deemed superfluous in the context of fitting, given that its influence is presumed to be encapsulated in the global irradiation effect. In consideration of the considerable discrepancies in the magnitude of these variables, a scaling process was undertaken to ensure a mean of 0 and a standard deviation of 1.

**Table 1.** Correlation between collector output and other variables

	Dependant ↓ / Independent →	Cloud Cover	Ambient Temp.	Global Irradiation
Entire data set	Collector outlet temp.	-0.21	0.78	0.82
3 days subset	Collector outlet temp.	-0.21	0.92	0.87

### 2.2 Model Training

Prior to establishing the production pipeline, the *Box-Jenkins methodology* was utilized to ascertain, quantify and diagnose the model. The identification of the three-day time series was conducted through the application of additive decomposition, with the resulting data for a single set illustrated in the **Fig. 2** below. Despite the absence of an overall trend, which was anticipated for solar thermal collectors, pronounced daily seasonality was discernible.



**Fig. 2.** Additive decomposition of the 3-day training data set

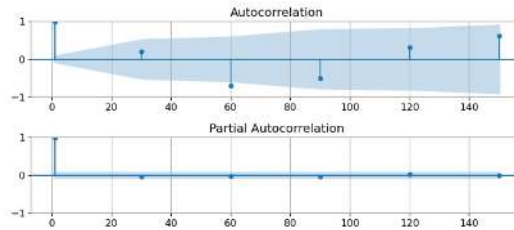
It was determined that a *Seasonal Autoregressive Integrated Moving Average with Exogenous Variables* (SARIMAX) model would be the most appropriate means of capturing the identified seasonal patterns and relationships between variables. An *Augmented Dickey-Fuller* (ADF) test was performed to check for stationarity. The results are shown in **Table 2**, with a single seasonal

differencing applied. Significant reduction in the ADF-statistic with the seasonal differencing informed the decision to use single differencing for the seasonality.

**Table 2.** Results of an Augmented Dickey-Fuller test for stationarity

Series	ADF-statistic	p-value
Original	-3.8	0.002
First Difference	-2.9	0.04
Seasonal Difference	-5.5	1.6e-06

Model parameters were further refined using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, which informed the selection of a first order autoregressive (AR) model. The ACF and PACF plots for the training data, plotted for 5-hour lags (equivalent to 30 lags for 10 minute resolution data), show a clear seasonal pattern. This seasonality is evident from the cyclical nature of the ACF plot, while the PACF plot shows a significant cut-off after the first lag.



**Fig. 3.** ACF and PACF plots for training data set

These initial estimates were used as a basis for parameter tuning by grid search, using the *pmdarima* package to automate the selection of an optimal SARIMAX model based on minimizing the *Akaike Information Criterion* (AIC). The refined SARIMAX model was then applied to predict the outlet temperature for a given target day, using solar radiation and ambient temperature forecasts as exogenous regressors. The model's prediction was evaluated against actual measurements using the *Root Mean Squared Error* (RMSE) criterion, and the target day was classified as either a 'fault day' (F-Day) or a 'no fault day' (NF-Day) based on predefined accuracy thresholds (10 K in our case). This process was integrated into a machine learning (ML) pipeline designed to run once a day for each plant. For each iteration, the last three NF-days were used as the training data set for the SARIMAX model, ensuring continuous adaptation to changing conditions.

The methodology described above is illustrated in the framework shown in FIGURE. In this framework, the user first defines a training dataset consisting of as few as three consecutive NF-days. This data set is used to fit the SARIMAX model, which generates predictions for the following day. If the predicted and observed values for the target day are sufficiently close, the day is labelled an NF-day, otherwise it is labelled an F-day. The training dataset is then updated with the latest NF-days to ensure that the SARIMAX model is always trained on the latest operational data. In practical applications, this loop is run once a day for each sensor, and the initial training data only needs to be defined once for each sensor.

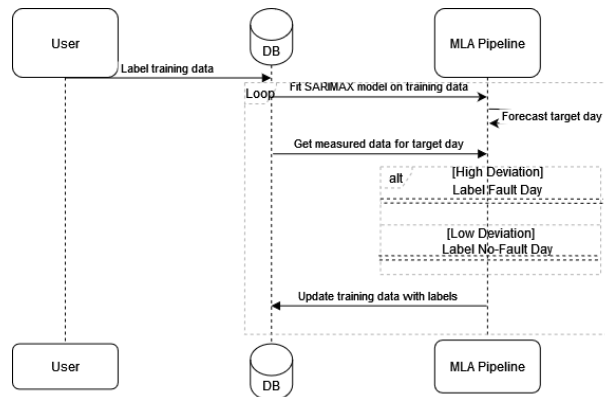


Fig. 4. Exemplary framework for implementing the MLA pipeline in the production system of industry partners

Results of this approach are shown in the next section.

### 3 Results

To facilitate concise presentation, the results presented here are from a single plant that exhibited the most stable performance, characterized by minimal NF-days. This allows effective model validation. A manual review of all historical data to identify NF-days would be impractical, and the limitations of the ground truth are discussed in the Introduction section. The results of the various tests, in which the dataset lengths were either 2 or 3 days, demonstrate the performance across different types of days selected to represent variations in global irradiation and ambient temperature (see Fig. 5). The analysis demonstrates that the automated fitting process identifies distinct optimal parameters for the SARIMAX model contingent on the length of the dataset. It is notable that while the fitting time of the grid search is shorter with a 2-day dataset, this is accompanied by a reduction in RMSE accuracy. In contrast, the 3-day tests, while taking longer to fit, tend to yield more reliable predictions, indicating a trade-off between model complexity, accuracy, and computational efficiency. These findings highlight the necessity of meticulously determining the optimal dataset length to achieve a balance between the necessity for a robust model performance and the practical considerations of fitting time and computational resources.

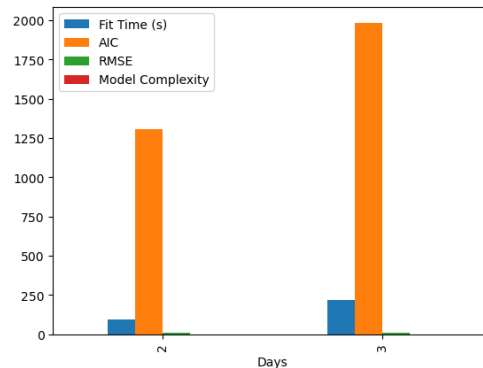
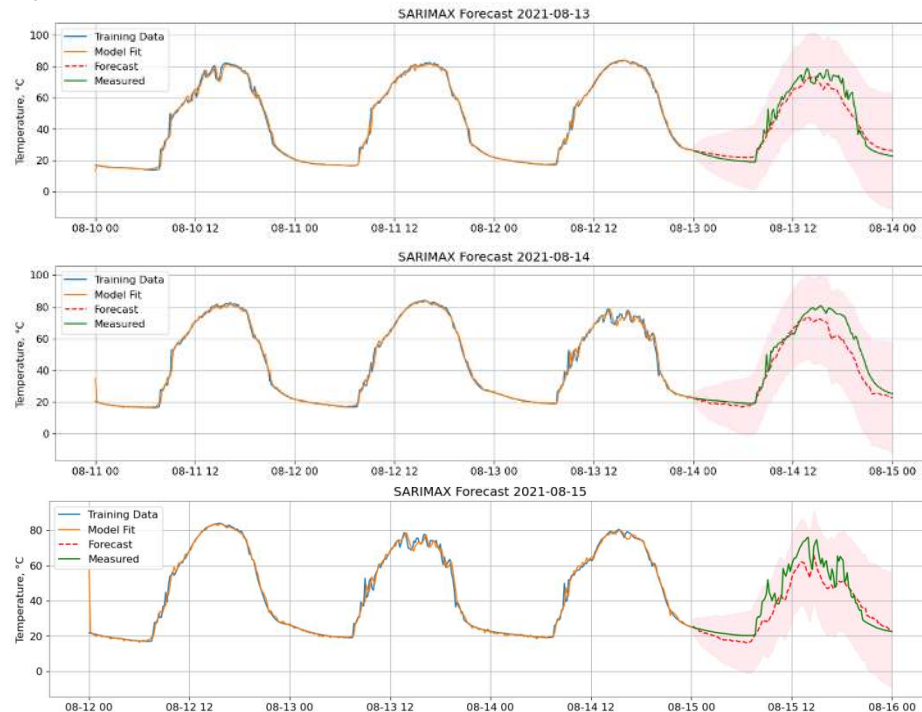


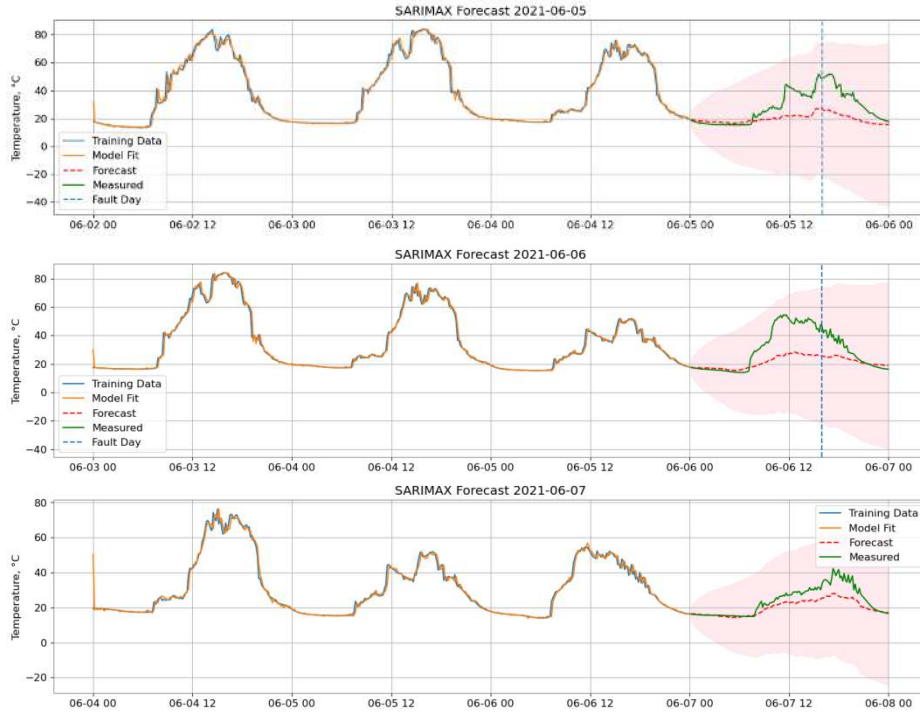
Fig. 5 Summary of hyperparameter search and results for various test data sets

**Fig. 6** and **Fig. 7** illustrate the iterative deployment of the ML pipeline designed to analyze three target days, utilizing temperature data from their preceding three NF-Days for training. The autofitting of the SARIMAX model is employed in this context. In **Fig. 6**, all three forecasted days are classified as NF-Days and subsequently appended to the NF-Day list, which is then utilized for further model fitting and refinement. Conversely, in **Fig. 7**, the algorithm effectively identifies F-Days, which are characterized by significant deviations from the expected behavioral patterns of the system. Such deviations may arise due to various factors, including faulty hydraulic configurations, suboptimal controller settings, inefficient pump operations, or adverse environmental conditions. While diagnosing these anomalies was beyond the scope of the current algorithm, it is critical to acknowledge that some deviations could be attributed to rapidly fluctuating global irradiation, often influenced by cloud cover dynamics. The complexity of SARIMAX model fitting underscores the necessity for rigorous evaluation and validation of the model's performance. Selecting appropriate training days is a critical step that requires expert judgment to ensure the representativeness and relevance of the data utilized. The expert selection process helps mitigate potential biases and enhances the model's capacity to generalize across different operational conditions. Future enhancements could involve implementing the algorithm primarily on days characterized by minimal cloud cover to reduce variability in the data. Additionally, adjusting the weights of exogenous variables within the SARIMAX model fitting could further improve the model's responsiveness to external influences. Furthermore, exploring alternative algorithms that better capture the effects of rapidly changing global irradiation remains a valuable avenue for future research, although it lies outside the current study's objectives.



**Fig. 6.** Results for three iterations of the ML-Pipeline on non-cloudy days in August for plant ST4048

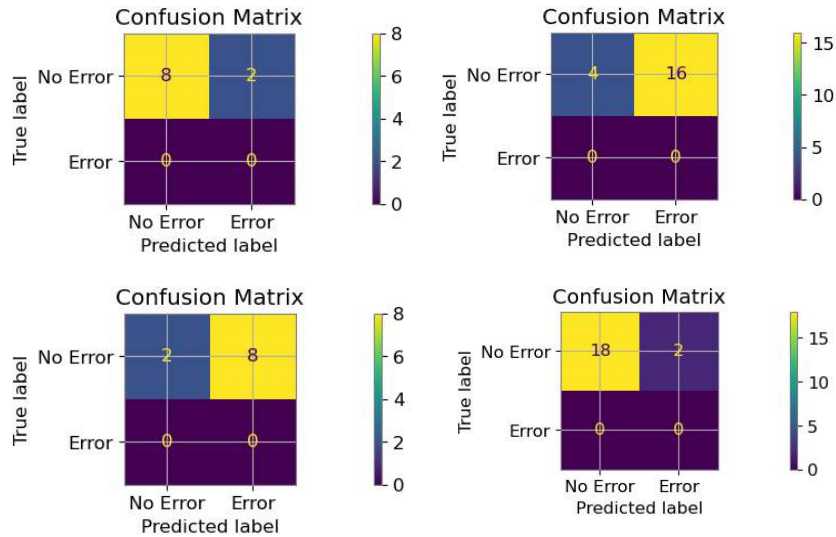




**Fig. 7.** Results for three iterations of the ML-Pipeline on cloudy days in June for plant ST4048

For the quantification of the results, a confusion matrix was constructed for various tests, utilizing subsets of the entire data that predominantly contained NF-days, as identified by field experts. This approach was necessary since ground truth verification could not be conducted otherwise, and the datasets lacked a specific variable to decisively classify a day as either F-day or NF-day.

The results, summarized below, reveal that the autotuned SARIMAX model underperformed when compared to the manually tuned model, which used fixed parameters across all iterations. The autotuned model tended to falsely identify more F-days, indicating higher rates of false positives, while the manually tuned model with fixed order exhibited better consistency and fewer misclassifications. This highlights the challenges in automating model tuning and the importance of expert input, particularly when ground truth data is not readily available for validation.



**Fig. 8.** Confusion matrix results for various tests. Clockwise: (a) SARIMAX(1,1,1)(1,1,0)[144] for 10 cloudy days, (b) Autotuned for 20 mixed days, (c) SARIMAX(1,1,1)(1,1,0)[144] for 20 cloudy days, (d) Autotuning for 10 cloudy days

#### 4 Conclusion

This study has demonstrated the effectiveness of a SARIMAX-based framework for fault detection in solar thermal systems that uses minimal historical data while still providing reliable results. By incorporating solar irradiance and ambient temperature as exogenous variables, the proposed model successfully adapts to dynamic operating conditions with only three fault-free days required for training. This makes the solution scalable and applicable to installations where limited data is available after commissioning.

Automated model updating through a machine learning pipeline further enhances the system's ability to adapt to changing conditions with minimal recalibration. This lays the groundwork for practical applications not only in solar thermal systems, but also in broader systems such as heat pumps and district heating, which can benefit from this streamlined approach.

Future work should focus on extending the framework to other heating systems, including more diverse test environments, refining the accuracy of error detection, and incorporating additional exogenous factors such as collector orientation and refined weather data filters. The results of this study highlight the potential for a scalable, low-data solution to monitor and optimize building heating systems, contributing to their overall efficiency and reliability.

#### 5 Acknowledgements

We would like to express our gratitude to *Verein Energie Zukunft Schweiz* for providing the data used in this study, which was collected as part of the LoCoSol+ project.

## References

- [1] P. Sawant, B. Sintzel, R. Eismann, J.W. Hofmann, B. Sitzmann, Low Cost Monitoring thermischer Solaranlagen mit IoT-Sensor und maschinellem Lernen, (n.d.). <https://www.aramis.admin.ch/Texte/?ProjectID=49446> (accessed July 11, 2024).
- [2] G. Faure, M. Vallée, C. Paulus, T.Q. Tran, Fault detection and diagnosis for large solar thermal systems: A review of fault types and applicable methods, *Solar Energy* 197 (2020) 472–484. <https://doi.org/10.1016/j.solener.2020.01.027>.
- [3] L. Feierl, V. Unterberger, C. Rossi, B. Gerardts, M. Gaetani, Fault detective: Automatic fault-detection for solar thermal systems based on artificial intelligence, *Solar Energy Advances* 3 (2023) 100033. <https://doi.org/10.1016/j.seja.2023.100033>.
- [4] Z. Liu, Y. Liu, D. Zhang, B. Cai, C. Zheng, Fault diagnosis for a solar assisted heat pump system under incomplete data and expert knowledge, *Energy* 87 (2015) 41–48. <https://doi.org/10.1016/j.energy.2015.04.090>.
- [5] M.S. Mirnaghi, F. Haghghat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review, *Energy and Buildings* 229 (2020) 110492. <https://doi.org/10.1016/j.enbuild.2020.110492>.
- [6] H. He, T.P. Caudell, D.F. Menicucci, A.A. Mammoli, Application of Adaptive Resonance Theory neural networks to monitor solar hot water systems and detect existing or developing faults, *Solar Energy* 86 (2012) 2318–2333. <https://doi.org/10.1016/j.solener.2012.05.015>.
- [7] R. Kicsiny, Z. Varga, Real-time nonlinear global state observer design for solar heating systems, *Nonlinear Analysis: Real World Applications* 14 (2013) 1247–1264. <https://doi.org/10.1016/j.nonrwa.2012.09.017>.
- [8] C. De Keizer, S. Kuethe, U. Jordan, K. Vajen, Simulation-based long-term fault detection for solar thermal systems, *Solar Energy* 93 (2013) 109–120. <https://doi.org/10.1016/j.solener.2013.03.023>.
- [9] S. Jiang, M. Lian, C. Lu, S. Ruan, Z. Wang, B. Chen, SVM-DS fusion based soft fault detection and diagnosis in solar water heaters, *Energy Exploration & Exploitation* 37 (2019) 1125–1146. <https://doi.org/10.1177/0144598718816604>.
- [10] J.W. Hofmann, B. Sitzmann, J. Dickinson, D. Kunz, R. Eismann, Use of machine-learning for monitoring solar thermal plants, *Journal of Physics: Conference Series* 2042 (2021) 012007. <https://doi.org/10.1088/1742-6596/2042/1/012007>.
- [11] Josef Perktold, others, statsmodels/statsmodels: Release 0.14.1, (2023). <https://doi.org/10.5281/ZENODO.593847>.
- [12] T.G. Smith, others, pmdarima: ARIMA estimators for Python, (2017). <http://www.alkaline-ml.com/pmdarima>.



# Artificial Intelligence for Quality Assurance and Troubleshooting in Industry

Rudolf Hoffmann<sup>1</sup>, Slimane Arbaoui<sup>2</sup>, Léa Charbonnier<sup>3</sup>, Amel Hidouri<sup>2</sup>, Ali Ayadi<sup>2</sup>, Franco Giustozzi<sup>2</sup>, Thomas Heitz<sup>4</sup>, Julien Saunier<sup>3</sup>, Frédéric Pelascini<sup>4</sup>, Christoph Reich<sup>1</sup>, Ahmed Samet<sup>2</sup>, Cecilia Zanni-Merk<sup>3</sup>

<sup>1</sup> Institute for Data Science, Cloud Computing and IT Security; Furtwangen University; 78120 Furtwangen, Germany  
contact: {rudolf.hoffmann or christoph.reich}@hs-furtwangen.de

<sup>2</sup> ICube, CNRS (UMR 7357) INSA Strasbourg, University of Strasbourg, 67000 Strasbourg, France

<sup>3</sup> INSA Rouen Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France

<sup>4</sup> CETIM – Centre Technique des Industries Mécaniques, 67402, Illkirch-Graffenstaden

## Abstract.

This paper presents the new X-Quality conceptual framework, that applies Artificial Intelligence (AI) to contribute to the improvement of quality assurance and troubleshooting in manufacturing. The goal is to identify and resolve quality issues effectively using AI techniques, applying Explainable AI (XAI) and stream reasoning to ensure transparency and comprehensibility to find causes for predicted quality defects. There are mainly three approaches of the framework described, that are tackling typical industry challenges. The first approach combines Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for time series quality prediction with SHAPley Additive exPlanations (SHAP) to explain the LSTM-CNN. The second method combines Machine Learning (ML) and Fault Tree Analysis (FTA) methods for comprehensive fault detection and analysis. The third technique applies semantic reasoning for real-time contextualization and root cause identification.

**Keywords:** AI; Machine Learning; XAI; Time Series; Root Cause Analysis; Fault Tree Analysis; Stream Reasoning; Ontology; Quality Assurance; Manufacturing

## 1 Introduction

In the present globalized economic era, industry competition demands continuous quality and reliability. Monitoring manufacturing processes is essential to prevent failures and maintain product quality. Artificial Intelligence (AI) enhances quality assurance by automating tasks traditionally handled by humans, using Machine Learning (ML) and Deep Learning (DL) to improve accuracy and consistency in defect detection [1]. Traditional troubleshooting methods like Root Cause Analysis (RCA) are extended by AI techniques to analyze vast amounts of data from multiple sources, improving defect detection and prediction [2]. For the stable deployment of AI-based systems and their acceptance by experts and regulators, it is crucial that the decisions and results produced by these systems are comprehensible, interpretable, and transparent, in other words, “Explainable” [3].

Our work aims to leverage AI techniques to enhance quality assurance and troubleshooting processes in various industries by developing methods for precise defect detection, predictive maintenance, and effective RCA.

The rest of this paper is organized as follows. Section 2 presents the related work, discussing previous research and approaches relevant to our study. Section 3 introduces the proposed X-Quality conceptual framework, and section 4 details the approaches that are integrated within this framework. Section 5 provides a critical discussion, evaluating the strengths and challenges of this framework and provides an outlook for future work. Finally, Section 6 concludes this paper.

## 2 Related Work

Recent research in quality assurance and troubleshooting in manufacturing has increasingly turned to AI and data-driven approaches. This section reviews related works that explore these approaches. While many works focus on automation, predictive capabilities, and explanations, our X-Quality conceptual framework focuses on a holistic view of the production line for a more comprehensive explainability and traceability of quality issues to determine the causes of the occurrences.

Several studies have explored AI-driven approaches for quality assurance in manufacturing, addressing challenges like data complexity, lack of transparency, and adaptability. The work [4] reviews these challenges and proposes a functional software architecture using Automated ML (AutoML) for automated model training and advanced data preparation to handle diverse data sources. Similarly, the study [5] focuses on the potential of ML and DL techniques for predictive quality, clustering existing methods by manufacturing processes, data sources, and ML methods. [6] proposes the Hybrid Digitization Approach to Process Improvement (HyDAPI) methodology that utilizes key elements of the Six Sigma Define-Measure-Analyse-Improve-Control (DMAIC) and the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodologies to enhance decision-making and operational efficiency. [7] and [8] address the need for transparency through explainable AI (XAI), using various XAI techniques to enhance transparency of ML models and improve preventive maintenance.

While these studies contribute valuable insights into AI applications for quality assurance in manufacturing, our X-Quality conceptual framework differentiates by integrating inductive and deductive AI, XAI, and expert knowledge across multiple production stages. Unlike [4], who emphasizes automation, or [7], who focus on improving transparency of ML models, our approach provides a more holistic perspective. By using ontology-based analysis, it enables more effective RCA and proactive actions, in order to prevent quality issues throughout the entire manufacturing workflow.

## 3 X-Quality Conceptual Framework

In multi-stage manufacturing processes, such as milling, grinding, and assembly, various operational parameters, such as cutting conditions, tool wear, and surface quality, must be monitored. Failures at one process can propagate, affecting subsequent processes and the final product. The X-Quality conceptual framework provides a more comprehensive view of the manufacturing workflow by monitoring data across multiple processes rather than focusing on a single manufacturing

process, enabling the system to identify final product issues and trace them back to the process where the deviation occurred.

In traditional manufacturing production lines, each operator does manual inspection at their respective machine (milling, grinding, assembly) and a quality manager supervises the overall production process for total quality control (Figure 1a). In the X-Quality conceptual framework (Figure 1b) from each machine, data is collected and different AI/XAI methods are applied to the collected data, in order to predict the quality after the manufacturing process and additionally provide the explanation for the prediction for the corresponding operator. Data streams, predictions and explanations are used to enrich an ontology. When a quality issue is predicted, the ontology is used to trace the root cause by linking machine failures to the quality issue. The quality manager supervises the entire manufacturing process using this ontology. This allows the quality manager to take proactive steps to maintain the overall quality of the production line, ensuring more effective quality assurance and troubleshooting in manufacturing. For example, if a defect, like a misaligned component, is detected in the assembly stage, the system can trace this issue back through the earlier processes, identifying that the problem arises from surface roughness during milling due to excessive tool wear. By taking proactive actions, such as replacing worn tools, similar defects can be prevented from occurring in future production cycles.

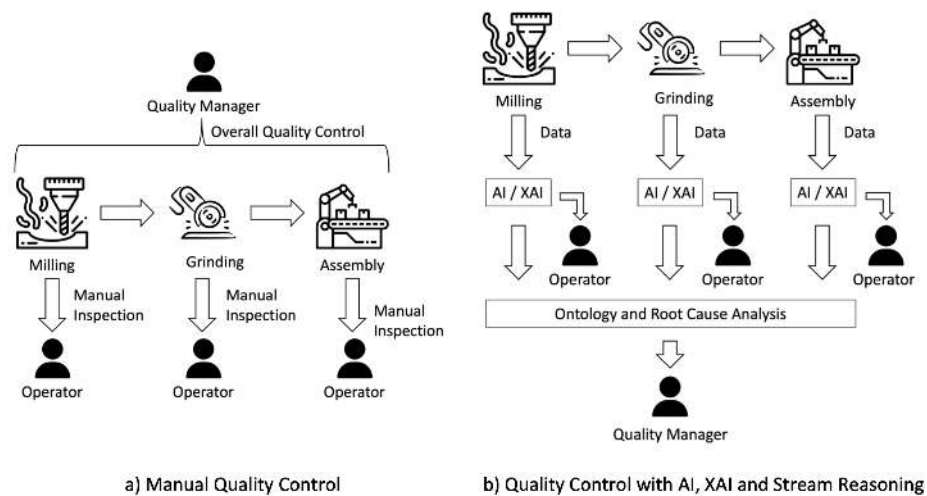


Figure 1: X-Quality Conceptual Framework

#### 4 Three Approaches for X-Quality Conceptual Framework

The following sections introduce the approaches that are integrated in the X-Quality conceptual framework.

#### 4.1 Time Series Data used for Quality Prediction

To predict product quality using time series sensor data, DL models such as Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) are commonly employed due to their ability to handle this type of data effectively. Combining these models can leverage their complementary strengths. CNNs reduce the dimensionality of the input data while capturing spatial features, whereas LSTMs excel at capturing temporal dependencies, leading to more accurate predictions of product quality.

The proposed model is expected to yield strong results in estimating product quality. However, like most DL models, it suffers from the "black box" nature, making its decision-making process difficult to interpret. To address the lack of transparency, a well-known post-hoc explanation model, named SHapley Additive exPlanations (SHAP) [9] is incorporated, which provides explanations in the form of relative importance values, being commonly referred to as SHAP scores. These scores highlight the features that most influence the model's predictions and reveal how each feature contributes to the final output. This enhances the interpretability of the model while maintaining its predictive power.

The important scores provided by the SHAP library will be used to select the features that most affect the prediction of our model. These features associated with their SHAP values that provide a link between the input and output, will be used as input to a Multidimensional Matrix Profile (MMP) [10] that allows to identify numerous structural elements within time series data, such as repeated behaviors, known as motifs, as well as anomalies, referred to as discords.

Our focus is on identifying discords which must appear simultaneously in both time series, named as natural anomalies, and represent data points that are most different among all the time series. Notably, the model is able to capture and explain these irregularities that correspond to product quality loss. The architecture of the proposed method is illustrated in Figure 2.

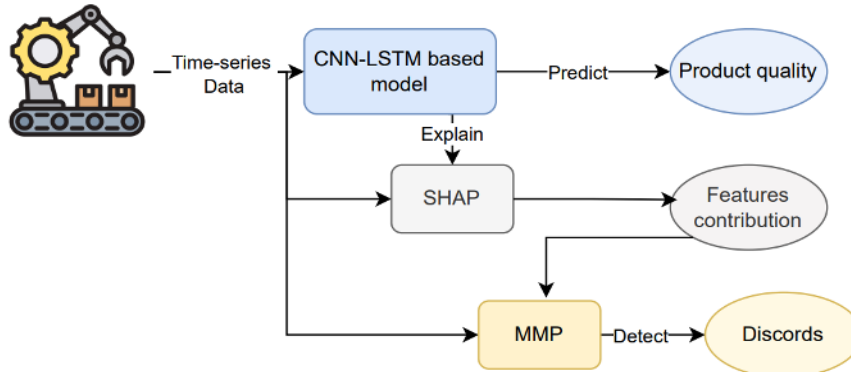


Figure 2: Quality prediction model with explainability for time series data



#### 4.2 Combination of Artificial Intelligence and Fault Tree Analysis

The approach combines AI, specifically ML models, with Fault Tree Analysis (FTA) to enhance the prediction and understanding of system failures. ML models predict the probabilities of Basic Events (BEs) in a fault tree, which are then used to calculate the likelihood of the Top Event (TE). This enables the identification of underlying failure causes. The deductive structure of fault trees helps to determine the TE, thereby enhancing the explainability of the TE predictions.

Figure 3 illustrates the implementation of this approach. To implement this approach, an expert first constructs a fault tree based on their domain knowledge. ML models are then trained to predict the probabilities of the BEs within this fault tree. Once these probabilities are obtained, FTA is used to determine whether the TE will occur. If the TE is predicted to occur, the system can analyze the fault tree to explain which BEs or combinations of BEs are responsible for triggering the TE. Furthermore, it can provide recommendations on which BEs should be mitigated or eliminated to prevent the occurrence of the TE, offering actionable insights to improve system reliability.

The proposed approach offers several potentials, particularly in the context of explainability and understandability in the manufacturing domain. One strength is the ability of this approach to provide an explainable occurrence of the TE. Since fault trees are a well-established method in the industry for representing the logical relationships between different system states, stakeholders in manufacturing are more familiar with this explanation. This familiarity enhances the understanding in the prediction and makes the explanation more intuitive for stakeholders in the manufacturing. This understanding provides a better decision-making and intervention in manufacturing processes. However, there are also limitations that has to be considered. While the occurrence of TEs can be explained, the occurrence of BEs remains opaque due to the "black box" nature of the DL models that are used to predict these events. This lack of transparency in DL models is a significant challenge, because it restricts our ability to fully understand and interpret the underlying mechanisms that lead to BE occurrences. Additionally, the construction of fault trees requires expert knowledge and is highly time-consuming. This dependency on domain experts not only limits the scalability of the approach but also increases the required resource. The need for expert involvement in developing, refining, and validating fault trees can create bottlenecks, particularly in dynamic or rapidly evolving manufacturing environments where quick adaptation is crucial.

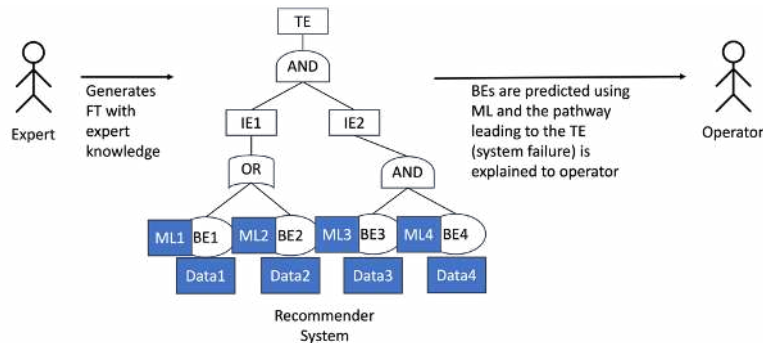


Figure 3: Combining ML with FTA for explainable TE occurrence

### 4.3 Stream Reasoning

Stream Reasoning is used to continuously query heterogeneous data streams from multiple sources in real time and apply logical reasoning to the data. The aim is to detect situations associated to quality problems in finished products and to understand the causes that lead to them. During the process, an ontology is populated to contain all the collected information to reason about the detected situations and causes.

To apply this approach in manufacturing, it is necessary to define the constraints that apply to the machines and products in question. This will enable the identification and categorization of abnormal situations. A constraint represents a rule on a property of a machine or a product and a situation is composed of sets of constraints. Defined constraints and situations are added into the ontology to support querying and detection. The ontology contains information on machines and sensors used in the production line. Indeed, as each query is composed of constraints representing one situation, numerous references are made to the ontology to obtain links between machines, sensors and values. Once an abnormal situation is detected, the ontology is updated with the related constraint values, sensor and machine or product. As quality issues are often from machine-related causes, identifying the origin of the problem is essential. Reasoning over the ontology helps trace quality issues back to the initial machine failures, providing insights into the root cause of the problem.

Since the input data is not formatted for use in an ontology, it is transformed into W3C standards such as Resource Description Framework (RDF). These statements exchange data on the web as triples: subject-predicate-object [11]. RDF statements can be combined into a dataset which can be queried using a query language such as SPARQL [12]. A SPARQL query is a tuple composed of a SPARQL algebra expression, an RDF dataset and a query form. As the data must be continuously treated in real time, it is therefore processed as streaming data. RDF streams, which are unbounded sequences of timestamped RDF statements, are used for this purpose.

The goal is to contextualize data streams composed of raw data and prediction results and explanations from predictive models (see Figure 4). Data processed by these models is collected and used in a Stream Generator to create RDF streams, which are continuously queried with a Stream Reasoner for pertinent information. An ontology containing expert knowledge is used to contextualize the streams. Since streaming data cannot be queried directly, the streams are parsed into finite pieces using time windows. A time window is defined by two-time stamps, such that any event within that interval is included in it [13]. To select a stream piece, the time window uses the timestamp of the events. Once parsed, streams can be queried like static data.

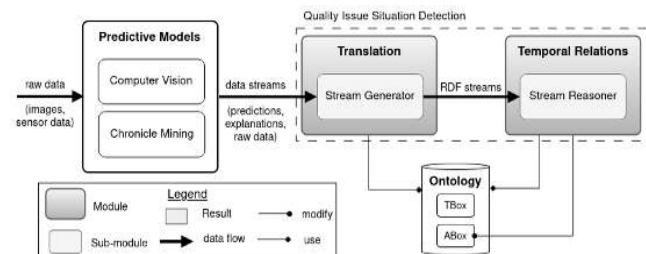


Figure 4: Quality issue detection with stream reasoning

## 5 Critical Discussion

The X-Quality conceptual framework presents a holistic approach to integrate inductive and deductive AI, XAI, and expert knowledge to improve quality prediction and RCA in manufacturing. However, several considerations arise when evaluating the framework, particularly in terms of practical deployment, scalability, and maintainability in real-world environments.

One of the key strengths of the X-Quality conceptual framework is its comprehensive integration of various methods. By using different AI/XAI methods for time series data, FTA, and stream reasoning, the framework is able to effectively address predictive maintenance, defect detection, and RCA. This approach ensures that the framework not only forecasts or detects potential issues but also provides actionable explanations and contextual insights, allowing operators and the quality manager to make informed decisions. This enhances both the product quality and the process reliability. Another strength of this framework is its adaptive capability, which is enabled by the use of stream reasoning. This enables the framework to continuously update the system based on data streams and to respond to changes in machine conditions or product quality. A further strength is the ontology that capitalizes expert knowledge to provide a structured formal model of the manufacturing environment. The ontology establishes a meaningful relationship between different machines or products, sensors, and related constraint values, enabling contextualized analysis and RCA.

Despite its strengths, the framework faces several challenges, particularly in terms of scalability and maintainability. The integration of this framework to larger manufacturing plants that provide high-frequency data streams, requires high computational resources. Moreover, the ontology requires continuous updates to remain relevant as machine configurations, sensor types, and production lines evolve. This maintenance and updating of the ontology present another significant challenge. In dynamic manufacturing environments the need for regular updates could become a bottleneck, since it requires expert dependence. Another challenge is to provide heterogeneous data streams, because they are coming from various sources and could also be perturbed by noise. A further potential challenge is the interpretability of the explanations provided by the framework, because the explanations may still be complex or difficult for operators or the quality manager to interpret.

Future work should focus on improving scalability by automating ontology updates to reduce the expert dependence, and improving the interpretability and comprehensibility of the explanations provided by the framework. Developing efficient methods for processing high-frequency data and simplifying complex outputs will be essential for real-world deployment in dynamic manufacturing environments.

## 6 Conclusion

In conclusion, the X-Quality conceptual framework combines machine data with AI and XAI methods to predict future quality issues and trace potential failures back to their root causes. By offering transparent explanations for these predictions, the system enables operators and the quality manager to understand and address the root causes of defects and thus provide more effective quality assurance and troubleshooting in manufacturing. This data-driven approach reduces downtime,

improves operational efficiency, and contribute to cost reduction, leading to better product quality and more reliable production processes.

## References

- [1] H. a. K. Y. Yang, "Design and implementation of machine learning-based fault prediction system in cloud infrastructure.," *Electronics*, vol. 11, no. 22, 2022.
- [2] I. Ahmed, G. Jeon and F. Piccialli, "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where.," *IEEE Transactions on Industrial Informatics*, vol. 18, p. 5031–5042, 2022.
- [3] R. Hoffmann and C. Reich, "A Systematic Literature Review on Artificial Intelligence and Explainable Artificial Intelligence for Visual Quality Assurance in Manufacturing.," *Electronics*, vol. 12, p. 4572, 2023.
- [4] Y. Wilhelm, U. Schreier, P. Reimann, B. Mitschang and H. Ziekow, "Data Science Approaches to Quality Control in Manufacturing: A Review of Problems, Challenges and Architecture," in *Service-Oriented Computing*, Cham, 2020.
- [5] H. Tercan and T. Meisen, "Machine learning and deep learning based predictive quality in manufacturing: a systematic review," *Journal of Intelligent Manufacturing*, vol. 33, no. 7, p. 1879–1905, 2022.
- [6] R. Clancy, D. O'Sullivan and K. Bruton, "Data-driven quality improvement approach to reducing waste in manufacturing," *The TQM Journal*, vol. 35, no. 1, pp. 51-72, 2023.
- [7] G. Sofianidis, J. M. Rožanec, D. Mladenčić and D. Kyriazis, "A Review of Explainable Artificial Intelligence in Manufacturing," *ArXiv*, vol. abs/2107.02295, 2021.
- [8] B. Hrnjica and S. Softic, "Explainable AI in Manufacturing: A Predictive Maintenance Case Study," in *IFIP International Conference on Advances in Production Management Systems (APMS)*, Novi Sad, Serbia, 2020.
- [9] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions.," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [10] C. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. Dau, D. Silva, A. Mueen and E. Keogh, "Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords, and shapelets.," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016.
- [11] R. W. Group, "Resource Description Framework (RDF)," W3C Recommendation, 2014. [Online]. Available: <https://www.w3.org/2001/sw/wiki/RDF>.
- [12] A. Seaborne and S. Harris, "SPARQL 1.1 Query Language.," W3C Recommendation, 2013. [Online]. Available: <https://www.w3.org/TR/2013/REC-sparql11-query-20130321/>.
- [13] D. Dell'Aglio, E. Della Valle, J.-P. Calbimonte and O. Corcho, "RSP-QL Semantics: a Unifying Query Model to Explain Heterogeneity of RDF Stream Processing Systems.," *International Journal on Semantic Web and Information Systems*, vol. 10, pp. 17-44, 2014.

# Exploration of Neural Network Architectures for Inertia Parameter Identification of a Robotic Arm

Thomas Granser, Maximilian Giessler, Stefan Glaser, Bernd Waltersberger and Stefan Hensel

Offenburg University of Applied Sciences

{thomas.granser, maximilian.giessler, stefan.glaser, bernd.waltersberger, stefan.hensel}@hs-offenburg.de

**Abstract.** We propose a machine learning based approach for identifying inertia parameters of robotic systems. We evaluate the method in simulation and compare it against classical methods. Specifically, we implement parameter identification based on numerical optimization and test it using ground truth data. For a case study, we set up a physical simulation of a four-degree-of-freedom robot arm, formulating the problem with Newton-Euler equations as opposed to the conventional Lagrangian formulation at the joint level. Additionally, we derive a test methodology for assessing various Artificial Neural Network architectures.

**Keywords:** Inertia parameters identification, robotics, numerical optimization, Newton-Euler, Neural Networks

## 1 Introduction

Inertia parameter identification is essential in robotics for precise motion planning and control [1]. The actual inertia parameters of robots often deviate from those predicted by CAD models due to unmodeled components, production tolerances, or modifications introduced during manufacturing [2]. Several methods have been proposed for inertia parameter identification, as reviewed by Leboutet et al. [3]. The most widely used approach involves modeling the system with an inverse dynamic model, measuring joint torques, and using motor encoder signals to obtain the system’s kinematic parameters at specific time points while the robot follows an excitation trajectory.

Traditionally, the equations of motion (EOM) are derived from the Lagrangian formulation, yields linear equations in terms of kinematic parameters (angular position, velocity, and acceleration) at the joint level. However, this method depends heavily on motor encoder signals to estimate dynamic parameters, making it susceptible to noise and errors from numerical differentiation when calculating joint velocities and accelerations. Additionally, the conventional approach relies on torque values from the joint motors, which are often indirectly measured via electrical current and voltage. Since direct torque measurements are often unavailable, accounting for nonlinear friction, thermal losses, and electromagnetic effects becomes challenging, limiting the practical application of these measurements in real-world scenarios [1].

## 2 Background

In the context of *inertia parameter identification* for a multi body-system of  $n$  rigid links connected by actuated rotational joints, the classical approach utilizes the Lagrangian

formalism to construct an inverse dynamic model. This model is represented by a set of  $n$  equations for the generalized forces — specifically the joint torques ( $\tau_i, i = 0..n$ ) of the joints connecting the links:

$$\boldsymbol{\tau}(t) = \mathbf{f}_i(q_i(t), \dot{q}_i(t), \ddot{q}_i(t)), \quad (1)$$

This equation can be expressed as a direct relationship between the system’s kinematic parameters  $\mathbf{Y}(q_i(t), \dot{q}_i(t), \ddot{q}_i(t))$  and the inertia parameters of the system  $\boldsymbol{\theta}$ :

$$\boldsymbol{\tau}(t) = \mathbf{Y}(q_i(t), \dot{q}_i(t), \ddot{q}_i(t))\boldsymbol{\theta}, \quad (2)$$

with  $i=0..n$ . Here  $\boldsymbol{\theta}$  represents the ten unknown inertia parameters for each link. These parameters are comprised by the mass  $m_i$ , the three first moments (mass times the center of mass vector)  $m \cdot \mathbf{r}_{\text{com}[x|y|z],i}$  and the six components of the inertia tensor for each of the  $n$ -links. The first moments thereby represent a nonlinear relation between the masses and center of mass vectors of the links. The first moments introduce a nonlinear relationship between the mass and the center of mass vectors of the links, making inertia parameter optimization an inherently nonlinear problem. Additionally, selecting appropriate excitation trajectories poses another challenge, as it presents an optimization problem in itself, as discussed by Lee et al. [4]. For serial robotic systems, such as robotic arms, issues related to non-, low-, or linked-observability of certain inertia parameters can arise, depending on the system’s kinematics and excitation. This may result from inadequate excitation or from the nonlinear relationship between the first moments and the weak influence of the off-diagonal elements of the inertia tensor, depending on the robot’s configuration [5]. Furthermore, the single-axis actuation of the initial links reduces the system’s ability to observe dynamic effects such as Coriolis and centrifugal forces, which are critical for estimating the complete set of inertia parameters. As a result, parameters like the off-diagonal inertia tensor components  $I_{xy}, I_{xz}, I_{yz}$  and certain components of the center of mass vectors  $\mathbf{r}_{\text{com}}$  often remain unidentifiable [5].

### 3 Methodology

In contrast to the traditional approach, we propose using the Newton-Euler equations in an inertial frame, leveraging direct measurements of angular velocities and accelerations with a newly developed sensor concept at Offenburg University [6, 7]. Additionally, we employ an external force-torque measurement unit mounted at the robot’s base to measure the total forces and moments resulting from the system’s motion on the robot’s fixture. By directly measuring both the kinematic parameters and the reaction forces and moments at the robot’s base, we eliminate the need to model complex joint-motor interactions, such as nonlinear friction. This approach also avoids the requirement for numerical differentiation or integration of sensor data. The Newton-Euler equations can be derived using either the recursive Newton-Euler Algorithm or the general formulation in a non-inertial frame for a multi-body system, as described by Giessler et al. [8, 7]. The resulting equations describe the relationship between the motion of the entire system and the resulting forces and moments at the robot’s base, rather than the traditional equations of motion (EOM) on joint level.

This approach yields a set of six equations for any given time, compared to the  $n$  equations produced by the EOM at the joint level, where  $n$  equals the number of joints in the robot. Similar to the Lagrangian formulation, these equations place the forces and

moments on the left-hand side, while the right-hand side contains a function representing the interaction between the system’s kinematics and inertia parameters [9].

$$\mathbf{F}_{\text{ext}} = \xi(q_i(t), \dot{q}_i(t), \ddot{q}_i(t))\boldsymbol{\theta}, \quad (3)$$

$$\mathbf{M}_{\text{ext}} = \phi(q_i(t), \dot{q}_i(t), \ddot{q}_i(t))\boldsymbol{\theta}, \quad (4)$$

where  $\mathbf{F}_{\text{ext}}$  represents the three external forces and  $\mathbf{M}_{\text{ext}}$  represents the three moments measured by the force-torque sensor.

Our contribution consists of two distinct solutions: first, classical numerical optimization; and second, the exploration and development of a machine learning approach employing neural networks. Both methods are evaluated on a synthetic ground-truth dataset, which includes inertia parameters and dynamic states (excitation frames) for various robotic configurations. Using the Newton-Euler equations, we compute the reaction forces and moments corresponding to the movements of a robot configuration.

We perform numerical optimization using solvers for both nonlinear and linearized problems, with the optimization results serving as benchmarks for our artificial intelligence (AI) methods. The AI approach is based on various fully connected feed-forward neural network architectures, which are systematically tested in different configurations. Additionally, we designed a more sophisticated AI approach using Siamese network architectures. Finally, we tested a custom loss function that incorporates physical constraints by embedding the analytical equations of the system into the loss function.

### 3.1 Data Generation

Identifying the inertia parameters of the initial links – such as the base and shoulder – in robotic arms, particularly serial manipulators, presents a specific challenge. These links are typically actuated along only a single axis, resulting in limited excitation of the system’s dynamics. This limitation makes it difficult to fully excite the degrees of freedom necessary for identifying all inertia parameters, especially the off-diagonal terms of the inertia tensor and the components of the center of mass vector.

In this work, we use the four link, four degrees-of-freedom robotic arm of the humanoid robot *Sweaty* from Offenburg University as a case study to evaluate the presented approaches. For this robotic arm, we take the inertia parameters of the first two links from the CAD drawings as given base values, which we do not attempt to estimate since the first two segments are not fully actuated. We focus on the remaining two links, after the third joint, to test the outlined strategies. Consequently, there are twenty unknown inertia parameters in the system that we aim to estimate.

As a first step, we formulate the Newton-Euler equations for this system. For the proposed method, it is not necessary to generate physically sensible excitation trajectories since the approach relies on directly measuring all system-relevant information at a given point in time – without the need for numerical differentiation and is currently confined to simulation. While this may lead to suboptimal excitation, it significantly simplifies the data generation process.

The generated data include measurements of angular positions, velocities, and accelerations. Additionally, the generated inertia parameters are checked against boundary constraints to ensure their validity in comparison to real life robots. In particular, we verify that the inertia tensor is positive definite, which is a crucial prerequisite for any real physical inertia tensor.

To generate the data, we first prepared kinematic parameter sets representing physically sensible motions, as well as arbitrary inertia parameter sets for different robotic arm configurations. Our aim was to train the model on varying robot arm configurations to achieve abstraction capabilities and to have multiple datasets for testing the numerical optimization. To this end, we identified boundaries for the kinematic parameters to define the range of values they can take during normal operations of the robotic arm. Furthermore, we used the inertia parameters of the last two links of the arm, collected from the CAD drawings, as base values to create randomly sampled robot configurations.

We decided to generate the data separately and not as part of the training process of the networks – firstly, to save time during training, and secondly, to be able to use the same dataset for the numerical optimization approach. After generating one hundred different inertia parameter sets for each of the two links in question, we applied a full factorial combination of these sets to generate a total of ten thousand unique robot configurations. Each configuration was paired with one hundred randomly generated kinematic parameter sets to create one hundred excitation trajectory frames for each robot configuration. Using the previously derived Newton-Euler equations, we then calculated the instantaneous reaction forces and moments for each of the excitation frames.

### 3.2 Numerical Optimization

For each frame, there are twenty unknown inertia parameters to be estimated but only six equations. To obtain a reasonable estimate for the parameters, we need at least four such observations to have a total of twenty-four equations. For the experiments, we used MATLAB’s `lsqnonlin` (least squares nonlinear) solver function to estimate the parameters, treating the first moments as separate constituting parameters – this represents the nonlinear problem. Afterwards, we combined the first moments (the linearized problem) and performed the estimation using the `lsqlin` (least squares linear) solver function – treating the first moments as individual atomic parameters.

### 3.3 Machine learning setup

For the machine learning setup, we used the generated ground-truth data to train different model architectures for this regression problem.

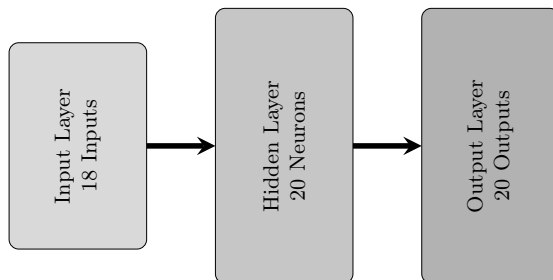
Our work can be divided into multiple iterative steps. Initially, we designed the simplest feed-forward neural network possible for this data. We used only one frame at a time, resulting in twelve measured kinematic parameter values – the angular positions, velocities, and accelerations of all four joints –and six parameter values for the reaction forces and moments, totaling eighteen input parameters  $\mathbf{X}$ .

Since we aim to estimate the inertia parameters of the last two links of the sequential robotic arm, we try to estimate a combined twenty output values  $\mathbf{y}$ . In this first step, we aimed to test whether the overall setup is functional and capable of learning the underlying data distribution. To evaluate this, we selected only ten of the generated robot configurations to limit the training time. As shown in Figure 1, the simple neural network structure has a single hidden layer.

After testing this network, we expanded the network architecture by adding additional layers and neurons per layer, as well as exploring different overall shapes of the studied networks. In addition, we tested different activation functions.

To explore the impact of network depth and width on the performance, we evaluated several architectures. We expanded our set of neural networks to include the following configurations:





**Fig. 1.** A simple neural network structure with one hidden layer. The diagram shows an input layer with 18 inputs, a hidden layer with 20 neurons, and an output layer with 20 output values.

- **Deeper Networks:** Networks with increased numbers of hidden layers ( $L$ ):
  - $L = 8$
  - $L = 12$
  - $L = 16$
- **Wider Networks:** Networks where each hidden layer has a width ( $W$ ) that is a multiple of the number of desired outputs ( $N_{\text{out}}$ ):
  - $W = k \times N_{\text{out}}$ , where  $k \in \{1, 2, 4, 8, 16\}$
- **Funnel-Shaped Network:** A network with 6 hidden layers where the number of neurons decreases in each layer:
  - Layer widths:  $[64, 32, 16, 8, 4, 2]$

This is summarized in Table 1.

**Table 1.** Extended Neural Network Architectures with Various Depths and Widths

Network	Description	Hidden Layers ( $L$ )	Width Factor ( $k$ )	Layer Width ( $W$ )
Network 1	Baseline width	8	1	$W = N_{\text{out}}$
Network 2	Double width	8	2	$W = 2 \times N_{\text{out}}$
Network 3	Increased depth	12	1	$W = N_{\text{out}}$
Network 4	Further increased depth	16	1	$W = N_{\text{out}}$
Network 5	Quadruple width	8	4	$W = 4 \times N_{\text{out}}$
Network 6	Octuple width	8	8	$W = 8 \times N_{\text{out}}$
Network 7	Sixteen times width	8	16	$W = 16 \times N_{\text{out}}$
Funnel Network	Decreasing layer sizes	6	Variable	$W = [64, 32, 16, 8, 4, 2]$

Building upon those experiments, we expanded the network structure to simultaneously receive multiple excitation frames of the same robot configuration as inputs – an approach inspired by the numerical optimization method. As a variation of this technique, we also implemented Siamese Networks. The key characteristic of Siamese Neural Networks is the use of two identical neural networks that share the same architecture and weights [10]. These subnetworks process different inputs in parallel, but because they

share the same weights, the overall network is more directly inclined to find outputs that satisfy all given input combinations simultaneously, which is particularly beneficial in our case where the expected output for one robot configuration is an identical set of inertia parameters. Lastly, we attempted to integrate the symbolically formulated Newton-Euler equations into the loss function of the networks, in an effort to create a physics-inspired neural network.

**Custom Loss Function for Physics-Inspired Neural Networks.** To integrate physical consistency into the neural network training, we designed a custom loss function that combines the prediction error of the inertia parameters with the discrepancy in reaction forces and moments computed via the Newton-Euler equations. The loss function is defined as:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \alpha \cdot \frac{1}{n} \sum_{i=1}^n \left( \hat{\theta}_i - \theta_i^{\text{true}} \right)^2 + \beta \cdot \frac{1}{N} \sum_{j=1}^N \left\| \mathbf{R}_{\text{meas}}^{(j)} - \boldsymbol{\Psi}^{(j)} \hat{\boldsymbol{\theta}} \right\|^2, \quad (5)$$

where:

- $\hat{\boldsymbol{\theta}}$  is the vector of predicted inertia parameters (outputs of the neural network).
- $\theta_i^{\text{true}}$  is the  $i$ -th ground truth inertia parameter.
- $n$  is the total number of inertia parameters.
- $N$  is the number of excitation frames.
- $\mathbf{R}_{\text{meas}}^{(j)} = \begin{bmatrix} \mathbf{F}_{\text{meas}}^{(j)} \\ \mathbf{M}_{\text{meas}}^{(j)} \end{bmatrix}$  is the vector of measured reaction forces and moments for the  $j$ -th excitation frame.
- $\boldsymbol{\Psi}^{(j)} = \begin{bmatrix} \boldsymbol{\xi}^{(j)} \\ \boldsymbol{\phi}^{(j)} \end{bmatrix}$  is the regressor matrix derived from the Newton-Euler equations for the  $j$ -th excitation frame.
- $\alpha$  and  $\beta$  are weighting coefficients that balance the importance of the two terms.

In this formulation the first term represents the Mean Squared Error (MSE) between the predicted inertia parameters and the ground truth parameters. The second term, penalizes the discrepancy between the predicted and measured reaction forces and moments, ensuring that the predicted inertia parameters lead to physically consistent behavior according to the Newton-Euler equations.

By minimizing this loss function, the neural network adjusts  $\hat{\boldsymbol{\theta}}$  to not only match the ground truth inertia parameters but also to ensure that these parameters result in reaction forces and moments that align with the physical measurements.

The functions  $\boldsymbol{\xi}^{(j)}$  and  $\boldsymbol{\phi}^{(j)}$  represent the contributions of the kinematic states to the reaction forces and moments, respectively, and are derived from the Newton-Euler equations for the  $j$ -th excitation frame. This form of the loss function emphasizes the importance of the predicted inertia parameters in reproducing the measured reaction forces and moments according to the Newton-Euler equations, thus ensuring physical plausibility.

For the exploration of the networks we constricted the training on a selection of three-thousand of the ten-thousand generated robot configurations, and trained the models for at least one-hundred epochs each, to limit training time.

## 4 Results

The results of the classical optimization demonstrated that the masses and the center of mass vector values could be correctly identified by this approach for both the linearized and nonlinear problem formulations. The linearized problem could also identify all other parameters correctly, whereas the nonlinear solver produced a wide range of accuracies – between approximately 50% and 95% – for the remaining parameters (inertia entries), depending on the selected robot configuration and especially the selected excitation frames. The iterative exploration of neural network architectures revealed varying degrees of effectiveness among the examined models. A variety of different network architectures were trained using various scaling methods and activation functions, with performance assessed through training loss and prediction accuracy. Under otherwise identical constraints, the *tanh* activation function, the robust scaler, and the SGD optimizer produced the most promising results for the given architectures. Despite extensive testing, the AI-based methods struggled to achieve the precision of classical numerical optimization, given the limited training time and data used to explore the presented network architectures. Although the approach of a physics-inspired neural network worked in principle, it led to unmanageable and impractical training times. This was because TensorFlow could not sufficiently optimize the backpropagation step for the custom loss function, which prevented us from performing this attempt with the limited processing power available at the time. In this preliminary exploration, the most promising result was produced by the Siamese network with eight hidden layers and a width factor of eight. However, it still could not approach the performance of the numerical optimization approach.

## 5 Conclusions

Our findings underscore the potential of combining direct dynamic measurements with numerical optimization and machine learning methods for inertia parameter identification. The performance of the studied neural networks, compared to numerical solvers, reveals both their limitations and their potential when training time is extended and the dataset is expanded. Since the influence of noisy sensor data was not studied in this work, this remains an important area for future research. Incorporating noisy data could provide a more relevant and realistic use case, where a well-designed model can demonstrate its generalization and robustness capabilities.

## References

1. Hollerbach, J., Khalil, W., Gautier, M.: Model Identification. Springer Handbooks. Springer International Publishing (2016)
2. Lages, W.F.: Parametric identification of the dynamics of mobile robots and its application to the tuning of controllers in ros. In Koubaa, A., ed.: Robot Operating System (ROS). Volume 707 of Studies in Computational Intelligence. Springer International Publishing (2017) 191–229
3. Leboutet, Q., Roux, J., Janot, A., Guadarrama-Olvera, J.R., Cheng, G.: Inertial parameter identification in robotics: A survey. Appl. Sci. **11**(9) (May 2021) 4303
4. Lee, T., Lee, B.D., Park, F.C.: Optimal excitation trajectories for mechanical systems identification. Automatica **131** (2021) 109773
5. Gayral, T., Daney, D.: A sufficient condition for parameter identifiability in robotic calibration. In Thomas, F., Perez Gracia, A., eds.: Computational Kinematics, Dordrecht, Springer Netherlands (2014) 131–138

6. Giefler, M., Werth, J., Waltersberger, B., Karamanidis, K.: A wearable sensor and framework for accurate remote monitoring of human motion. *Commun. Eng.* **3**(1) (Jan. 2024) 20
7. Giessler, M., Waltersberger, B.: Computational advantages in robotics by evaluating newton-euler equations with respect to a moving reference point in a non-inertial frame. In Review (Jun 2022) preprint.
8. Featherstone, R. In: *The Recursive Newton-Euler Algorithm*. Springer Berlin Heidelberg, Berlin, Heidelberg (2020) 1–5
9. Atkeson, C.G., An, C.H., Hollerbach, J.M.: Estimation of inertial parameters of manipulator loads and links. *The International Journal of Robotics Research* **5**(3) (1986) 101–119
10. de Rosa, G.H., Papa, J.P.: Chapter 7 - learning to weight similarity measures with siamese networks: a case study on optimum-path forest. In Falcao, A.X., Papa, J.P., eds.: *Optimum-Path Forest*. Academic Press (2022) 155–173

# AI-Guided Noise Reduction for Urban Geothermal Drilling

Daniel Ladwig<sup>1</sup>, Martin Spitznagel<sup>1</sup>, Jan Vaillant<sup>1,2</sup>, Klaus Dorer<sup>1</sup>, and Janis Keuper<sup>1</sup>

<sup>1</sup> Institute for Machine Learning and Analytics (IMLA)  
Offenburg University, Germany  
`firstname.lastname@hs-offenburg.de`  
<sup>2</sup> Herrenknecht AG

**Abstract.** Urban geothermal energy production plays a critical role in achieving global climate objectives. However, drilling operations in densely populated areas generate significant noise pollution, posing challenges to community acceptance and regulatory compliance. This research presents an artificial intelligence-driven approach to dynamically reduce noise emissions during geothermal drilling. We integrate Deep Reinforcement Learning (DRL) with generative neural network models to provide real-time recommendations for optimal drilling parameters. Specifically, the Drill-LSTM model forecasts future machine states, while the Sound-GAN framework predicts sound propagation based on varying operational conditions. These models feed into a DRL-Agent that learns to balance drilling efficiency with noise minimization. Additionally, an interactive assistance system GUI presents predictions, forecasts, and recommendations to human operators, facilitating informed decision-making. Our system demonstrates significant potential in reducing noise levels, enhancing operational efficiency, and fostering greater acceptance of urban geothermal projects. Future work will focus on refining the models and validating the system in real-world drilling scenarios.

**Keywords:** Geothermal Drilling, Noise Reduction, Deep Reinforcement Learning, Generative Models, AI-Assisted Control

## 1 Introduction

Urban geothermal energy production is increasingly recognized as a pivotal component in achieving global climate goals. However, the deployment of geothermal technology in urban areas is not without challenges, particularly the significant noise pollution generated during drilling processes. Current methods for mitigating this issue are largely manual and often inadequate for maintaining noise levels within legal urban limits. In densely populated areas, continuous deep drilling operations required for geothermal energy can severely disrupt local communities. Legal requirements often cap noise levels at 35dB during nighttime, posing a substantial challenge given the 24/7 operational needs of these projects. Current solutions, including temporal shifting of operations and physical barriers, provide limited relief.

This research introduces an application of artificial intelligence to overcome the constraints of traditional noise reduction techniques in geothermal drilling. By integrating Deep Reinforcement Learning (DRL) with generative neural network models, we dynamically suggest drilling parameters based on continuous feedback. Our system utilizes two models trained on real-world data: one forecasting noise outcomes and the other predicting drilling scenarios. A DRL model uses these simulations to learn optimal drilling

strategies that minimize noise while maintaining drilling efficiency. The system’s performance is planned to be further refined through real-world application, aiming to ensure its effectiveness across various urban geothermal sites.

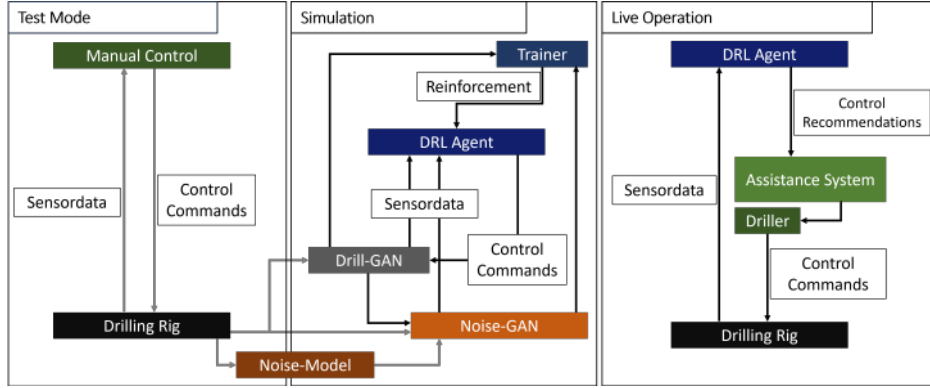


Fig. 1: Architecture of the AI models during the development stages Test Mode, Simulation and Live Operation.

**Implementation Prototypes:** The project is structured into three distinct prototypes, each illustrated in Figure 1. In *Test Mode*, generative models simulate sound propagation, drilling rig behaviour and predict drilling progress from specific states and commands. In the *Simulation* phase, a DRL model uses input from generative models to learn the control of the simulated drilling rig through reinforcements to provide operation recommendations, which then are validated against real-world scenarios. Finally, in *Live Operation*, the DRL system operates as an assistant that recommends strategic control changes during the drilling process. The human operators stay in control, providing oversight and evaluating AI-recommended commands. The system will be further improved by continuously learning from actual data and human feedback.

## 2 Related Work

**Generative models for complex physical problems.** Integrating physical principles into generative models is a rapidly advancing field. Models such as PUGAN [1] and FEM-GAN [2] have successfully merged GANs with physical modeling, enhancing performance in environments governed by complex physical laws. Physics-guided GANs have notably improved efficiency and precision in areas like fluid dynamics and structural system identification by incorporating physics-based loss functions and simulations [3, 4]. Additionally, machine learning has made significant strides in understanding and predicting physical interactions, as demonstrated by models that grasp the dynamics of block towers beyond simple memorization [5], applications such as fall detection through body part tracking [6], and the generation of physically plausible human animations [7]. Furthermore, physics-guided AI approaches, including grey-box models, have effectively incorporated physical laws into model training, enhancing performance and reliability [8, 9].

**Generative models and deep reinforcement learning.** Integrating generative models with deep reinforcement learning was first introduced by [10] with a case study in

the domain of near-field/far-field communication. This work mentions how generative AI could improve deep reinforcement learning considering data and policy. Simulating DRL environments of real-world scenarios with generative models could open up more applications for DRL.

### 3 Assistance System

The predictions, forecasts and recommendations of the developed models are presented to the driller in a continuously updated assistance system GUI. Figure 2 shows a Demo of the application appearance. The interface is separated into three areas, the status area (top), the analysis area (middle) and the recommendation area (bottom). The status area displays the two most important indicators: the rate of penetration and the sound level. This allows a quick overview on the current state. The analysis area shows the historical, the current and forecast trends of the monitoring variables. These dataplots help to explain a given recommendation to adjust the current machine control parameters and assists the driller in the decision to apply the recommendation or to decline. In particular, it shows the forecast of the machine behaviour with and without accepting the recommendation. The recommendation area displays new control recommendations given by the deep reinforcement learning agent and also allows to give feedback on their quality. A central middleware retrieves the machine data, preprocesses them for the various models, awaits results and synchronously sends updates to the UI.

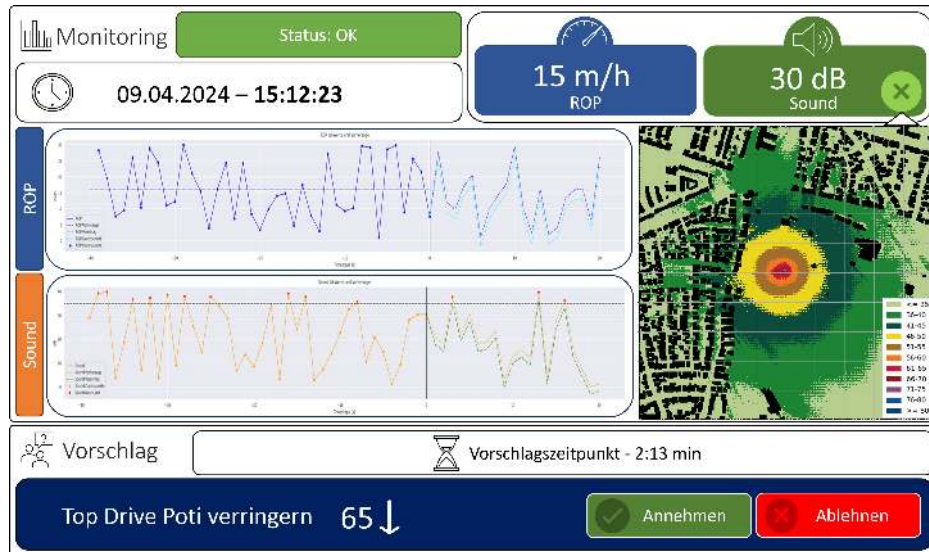


Fig. 2: Assistance system GUI Mockup

### 4 Drill-LSTM

Accurate prediction of machine states is vital for informed decision-making in drilling operations. The Drill-LSTM model serves as the foundational component for simulating

machine behaviour, enabling the DRL agent to understand and anticipate the outcomes of various control actions. In this study, we used a sequence-to-sequence encoder-decoder Long Short-Term Memory (LSTM) network, to predict future machine states based on defined action features. Our dataset comprises about 700 features collected throughout the entire drilling procedure, with one feature intentionally shifted to represent human actions that control machine behaviour. The training dataset encompasses one month of drilling operations, while the test dataset covers one week.

Data preprocessing involved standardizing the logging interval to a fixed 60-second interval and normalizing features. For the forecasting task, the model was trained to predict the next time step based on a defined history window, with current experiments focusing on 1-step forecasting. The model was trained for 50 epochs using the Mean Squared Error (MSE) loss.

Figure 3 visualizes the forecasting performance of the Drill-LSTM model on four anonymised machine features. The model successfully captures general trends. These results establish a foundation for future multi-step forecasting, which will require architectural enhancements to maintain prediction accuracy over longer horizons. The main purpose of the forecasting model is to serve as a machine model for the assistance system, enabling the prediction of future machine states based on potential user actions.

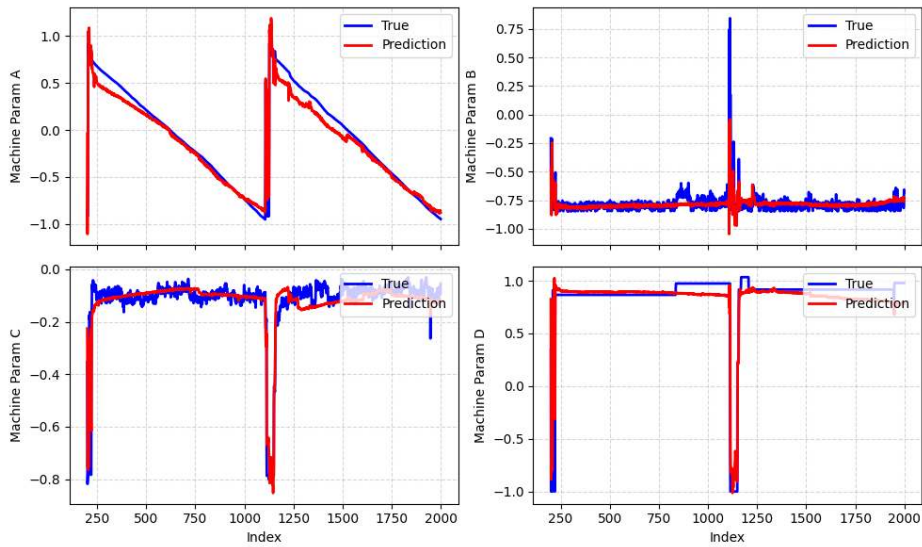


Fig.3: Forecasting performance of the sequence-to-sequence LSTM model on four anonymised machine features. The true machine parameter values (blue lines) are compared with the predicted values (red lines) over the test dataset.

## 5 Sound-GAN

Understanding and predicting sound propagation is essential for mitigating noise pollution during drilling operations. The Sound-GAN framework leverages generative models to efficiently simulate sound distribution, providing real-time feedback that the DRL agent can use to recommend noise-reducing actions to human operators.



Table 1: Model vs. Simulation Performance Comparison for Single Sample Processing. The complex source is a single test sample for a more complex source with 28 descriptive sound signal sources for the simulation, illustrating how the processing time increases significantly with more complex signal sources. It is important to note that this analysis may not provide a completely fair assessment from a theoretical perspective, as no efforts were made to optimize the simulation codes for GPU execution.

Model - Condition	Mean Runtime (s)	Std. Dev. (s)
UNet	0.0126	0.0012
GAN	0.0095	0.0012
Diffusion	4.1560	0.0061
Simulation Single Source	186.2295	16.8491
- 3rd Order Reflections		
Simulation Machine Setup	540.0000	-
- Complex Source		

**Simulation vs. Generative AI:** Building on our previous work, where we evaluated the effectiveness of generative models for predicting sound propagation [11], we generated over 15,000 data samples using the *NoiseModelling v4* framework [12], compliant with *CNOSSOS-EU* standards [13]. Each sample was defined by unique drilling parameters. This simulation data serves as the foundation for three different generative image-to-image models: **GANs** based on [14], **UNets** [15] and **DDPM diffusion** models [16]. Generative models significantly outperform traditional sound propagation simulations in processing speed, achieving up to a  $50k$  factor improvement in runtime, with a mean absolute error of 0.55 dB in their predictions.

**Simulation Setup:** The dataset for this study was generated using the Noise Modelling Framework v4 [12], We systematically altered five key machine parameters, each representing distinct components that influence the noise distribution around a stationary drilling machine. The parameters were anonymised, with load values ranging from 1.0, representing the maximum dB output of a component, to 0.5, indicating a -20 dB linear reduction. Each simulation was based on the drilling machine’s fixed location and initial noise measurements.

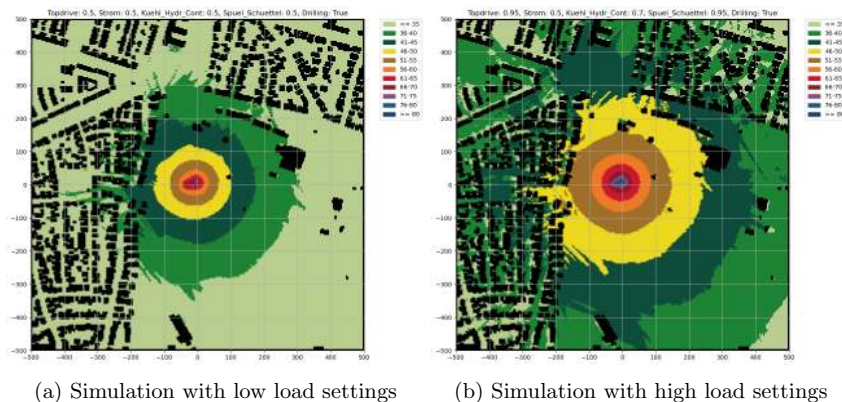


Fig. 4: Comparison of simulated sound propagation maps under different machine settings: (a) low load and (b) high load operational states.

The final dataset comprises over 15,000 data points, each depicting a unique combination of machine settings while keeping the noise source stationary across all simulations. Figure 4 provides a comparison of sound propagation maps under different operational states, with (a) illustrating low load settings and (b) depicting high load settings, both color-coded according to dB levels. This simulation setup allows the model to capture how varying machine loads affect the spatial distribution of noise.

**Sound Prediction Results:** Building on this simulation setup, we evaluated the performance of three models — UNet, GAN, and DDPM — in predicting sound propagation from a fixed grayscale OpenStreetMap (OSM) input, where buildings are represented by black pixels and open spaces by white pixels. The task was framed as a conditional image-to-image translation, where each model received the same OSM image as input, and the objective was to predict the interpolated sound distribution map based on the machine parameter settings.

Each model architecture was trained for 50 epochs, with 20% of the dataset held out for testing. The UNet and Diffusion models were trained using Mean Squared Error (MSE) loss, while the GAN employed a combination of Binary Cross Entropy (BCE) loss and L1 loss. The evaluation of the models was based on two key metrics: Mean Absolute Error (MAE), which quantifies the average magnitude of prediction errors, and Weighted Mean Absolute Percentage Error (wMAPE), which penalizes incorrect predictions behind and inside of buildings.

Table 2: Evaluation of all architectures on the LoS and NLoS tasks using MAE and wMAPE metrics.

Model	LoS		NLoS		LoS		NLoS	
	MAE	wMAPE	MAE	MAE	wMAPE	wMAPE	wMAPE	wMAPE
UNet	0.70	12.78	0.58	0.85	5.32		21.87	
GAN	<b>0.48</b>	<b>3.42</b>	<b>0.32</b>	<b>0.68</b>	<b>1.93</b>		<b>5.24</b>	
DDPM	1.19	24.94	1.07	1.35	14.23		37.98	

**Results:** The results of the evaluation are summarized in Table 2, which shows the performance of each model for both Line of Sight (LoS) and Non-Line of Sight (NLoS) regions. The GAN model consistently outperformed the other architectures, achieving the lowest MAE and wMAPE scores in both LoS and NLoS conditions. To further analyze the performance of the GAN model, we generated a heatmap of the wMAPE across the test dataset, as shown in Figure 5. This heatmap provides a pixel-wise visualization of the prediction error. The highest errors occur in NLoS areas.

In addition to the quantitative analysis, Figure 6 provides a visual comparison of the predicted sound distributions from all three models for a single datapoint. The GAN model shows the most accurate spatial distribution of sound. By contrast, the Diffusion model exhibits larger deviations with visible over-prediction, especially around building edges and occluded spaces.



Fig. 5: Heatmap of the Mean Absolute Percentage Error (MAPE) over the entire dataset for the UNet prediction.

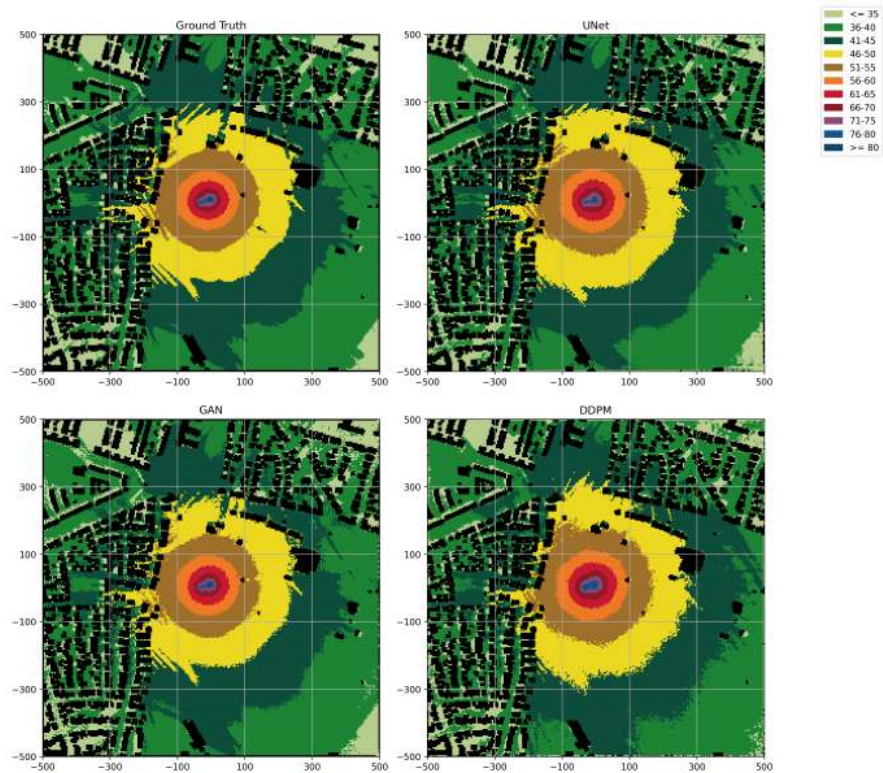


Fig. 6: Comparison of predicted sound distributions for a single datapoint across three models: UNet, GAN, and DDPM. The ground truth simulation is shown in the top left.

## 6 DRL-Agent

Recommending beneficial control changes requires an effective control strategy. Deep reinforcement learning will be utilized to train an agent to fulfill this task. Since training on the real drilling machine is no reasonable option, a simulation of the machine (DRILL-LSTM) is being developed to emulate the necessary machine behavior to train on. To be able to monitor and react to the resulting noise of a machine state after applying control changes on the simulated machine, a model named NOISE-GAN is being developed. These two models build the prerequisite for developing a DRL-Agent, which is a part of future work. The combination of generative models with deep reinforcement learning is promising especially for domains where training an agent on the real machine is not feasible.

The second crucial issue to approach is how to recommend control changes in a timely manner that are processable by a human operator. A deep reinforcement agent usually operates with a high frequency and has direct access to the environment to quickly react to changes when necessary. Since the control of the drilling machine still depends on a human operator to make the final decision, the drilling recommendations have to fulfill special requirements. The recommendations have to be as sparse as possible to not annoy or distract the driller but with the most impact within a reasonable timeframe. Current recommendations have to be monitored to be dismissed if the elapsed time since occurring or machine state changes turn them ineffective. The upcoming challenge is to produce recommendations for humans based on a high frequency observing DRL-Agent.

## 7 Conclusions

In this work we have shown that the prediction of sound propagation can be sped up by a factor of 50.000 when using generative AI models to do the prediction. This can be used to effectively train a DRL model to efficiently - in the sense of speed and noise - drill in urban areas. The so reduced sound emissions from geothermal drilling operations will enhance community tolerance and broader acceptance of urban geothermal energy projects. This contributes to the economic viability of such projects and their potential impact on achieving climate targets. Integrating AI to control noise in geothermal drilling presents a transformative solution to one of the most pressing challenges facing urban renewable energy projects, promising not only to reduce noise pollution effectively but also to streamline drilling operations. The advancements in generative AI could open ways to solve optimization problems with DRL where training on the real environment is not an option. In future work the practicality of the presented integration of DRL and the simulated environment with generative AI will be tested on real world data. Furthermore the efficiency of DRL as an assistance system in human processable speed is yet to prove in a real world scenario.

## Acknowledgements

The authors acknowledge the financial support by the German Federal Ministry of Education and Research (BMBF) in the program “Forschung an Fachhochschulen in Kooperation mit Unternehmen (FH-Kooperativ)” within the joint project ”KI-Bohrer” under grant 13FH525KX1.

## References

1. Cong, R., Yang, W., Zhang, Weiand Li, C., Guo, C., Huang, Qingmingand Kwong, S.: Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators. *IEEE Transactions on Image Processing* (2023)
2. Argilaga, A.: Fem-gan: A physics-supervised deep learning generative model for elastic porous materials. *Materials* (2023)
3. Kim, B., Azevedo, V.C., Thürey, N., Kim, T., Gross, M., Solenthaler, B.: Deep fluids: A generative network for parameterized fluid simulations. *Computer Graphics Forum* **38** (2018)
4. Yu, Y., Liu, Y.: Physics-guided generative adversarial network for probabilistic structural system identification. *Expert Systems with Applications* **239** (2024) 122339
5. Lerer, A., Gross, S., Fergus, R.: Learning physical intuition of block towers by example (2016)
6. Li, W., Azimi, S., Leonardis, A., Fritz, M.: To fall or not to fall: A visual approach to physical stability prediction (2016)
7. Zhang, Y., Li, W., Gao, Y., Li, S.: Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2106.03704* (2021)
8. Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V.: Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:2212.02500* (2022)
9. Chen, Y., Wu, Q., Sui, Y., Wang, Y., Zhu, S.C.: Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561* (2023)
10. Sun, G., Xie, W., Niyato, D., Mei, F., Kang, J., Du, H., Mao, S.: Generative ai for deep reinforcement learning: Framework, analysis, and use cases (2024)
11. Spitznagel, M., Keuper, J.: Urban sound propagation: a benchmark for 1-step generative modeling of complex physical systems (2024)
12. Bocher, E., Guillaume, G., Picaut, J., Petit, G., Fortin, N.: Noisemodelling: An open source gis based tool to produce environmental noise maps. *ISPRS International Journal of Geo-Information* **8**(3) (2019) 130
13. Kefhalopoulos, S., Paviotti, M., Anfosso-Lédée, F.: Common noise assessment methods in europe (cnossos-eu). Technical Report EUR 25379 EN, Publications Office of the European Union, Luxembourg (Luxembourg) (2012) JRC72550.
14. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *CoRR* **abs/1611.07004** (2016)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. *CoRR* **abs/1505.04597** (2015)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *CoRR* **abs/2006.11239** (2020)



# Supervised Transfer Learning Framework for Fault Diagnosis in Wind Turbines

Kenan Weber and Christine Preisach

Hochschule Karlsruhe - University of Applied Sciences

kenan.weber@h-ka.de

christine.preisach@h-ka.de

**Abstract.** Common challenges in fault diagnosis include the lack of labeled data and the need to build models for each domain, resulting in many models that require supervision. Transfer learning can help tackle these challenges by learning cross-domain knowledge. Many approaches still require at least some labeled data in the target domain, and often provide unexplainable results. To this end, we propose a supervised transfer learning framework for fault diagnosis in wind turbines that operates in an *Anomaly-Space*. This space was created using SCADA data and vibration data and was built and provided to us by our research partner. Data within the Anomaly-Space can be interpreted as anomaly scores for each component in the wind turbine, making each value intuitive to understand. We conducted cross-domain evaluation on the train set using popular supervised classifiers like Random Forest, Light-Gradient-Boosting-Machines and Multilayer Perceptron as metamodels for the diagnosis of bearing and sensor faults. The Multilayer Perceptron achieved the highest classification performance. This model was then used for a final evaluation in our test set. The results show, that the proposed framework is able to detect cross-domain faults in the test set with a high degree of accuracy by using one single classifier, which is a significant asset to the diagnostic team.

**Keywords:** Condition Monitoring, Wind Turbines, Anomaly Detection, Fault Detection, Fault Diagnosis, Transfer Learning

## 1 Introduction

In Germany, electricity generated from wind turbines (WTs) makes up a large portion of the total generated energy from renewable energy sources [1]. In order to increase the total energy yield, it is very important to reduce total downtimes by monitoring critical aspects of WTs. With condition monitoring, faults can be detected early and maintenance times and measures can be planned accordingly. This further reduces the risk of total failure due to the propagation of the faults to other areas of the machine. The total amount of WTs is increasing, with each of them equipped with more and more sensors over the years. Hence, this raises the amount of components, that can be measured which leads to higher costs and more signals, that need to be monitored. Furthermore, when a fault has been detected, we still need to infer the type of fault and localize it, e.g. which component(s) is/are affected. With an increasing amount of signals, there is need for more and more highly specialized personnel in order to monitor these machines manually. Therefore, we propose a solution that automates this whole process from end-to-end.

Supervised learning is a method for performing intelligent fault diagnosis on WTs. In literature, numerous solutions based on supervised learning are available [2, 3, 4]. However, these solutions typically focus solely on diagnosing faults in one particular machine, resulting in the development of separate models for each machine. Transfer learning is a promising approach to use knowledge extracted from a subset of WTs to multiple other ones. This can decrease the total model count needed for reliable fault diagnosis. Additionally, transfer learning based fault diagnosis is not limited to detect only those fault types, that have been occurred in the past on that particular WT, since fault data from other WTs can also be used for the diagnosis. Many transfer learning solutions for WTs exist. Zhang et al. [5] built a fully connected neural network which is able to detect, whether ice are on the WT blades by only using SCADA data. A small data set from another WT was used to fine-tune the model. Yang et al. [6] are able to detect blade defects by segmenting blade images with the otsu threshold segmentation algorithm and then using a pre-trained Alexnet classifier for the feature extraction. A Random Forest was used for the fault diagnosis in the last step. Li et al. [7] pre-trained a convolutional autoencoder on SCADA data from 14 WTs and fine-tuned the model on data from the 15th WT in order to detect fault types like high temperature in gearbox or generator or low pressure of the hydraulic system. A stacked autoencoder was employed by Deng et al. [8] by pretraining the model on source data and then utilizing and fine-tuning a fully connected layer for the diagnosis in the target domain.

The aforementioned solutions have the limitation, that some labeled data in the source domain have to be available. Furthermore, features extracted with the help of neural networks are mostly abstract and not interpretable for diagnosticians. If the fault diagnosis system indicates that a fault is present, then it has to be clear to the diagnostician, *how* the decision was made. To this end, we propose a fault diagnosis framework, which operates in an *Anomaly-Space*. This new feature space provides several normalized anomaly scores for each WT component, for every available WT. Values above 1.0 are considered anomalous. It was built and provided to us by our research partner EnBW Energie Baden-Württemberg AG. Both SCADA data and vibration data were used for the creation of the Anomaly-Space. A supervised classifier takes data from the Anomaly-Space as input and provides fault diagnosis results. These results can easily be interpreted by diagnosticians, since features in the Anomaly-Space represent deviations from the normal behavior of the WT. This can be seen as a feature-based transfer learning approach, where the Anomaly-Space represents the domain-shared feature space.

In summary, the contributions of this paper are the following:

1. Fault diagnosis based on derived signals from SCADA data and vibration data, that are easily interpretable, in contrast to many other transfer learning approaches.
2. Extensive model training and evaluation with stratified cross-validation from real data across 5 WTs from 4 wind parks and comparing classification performance of popular supervised classifiers, such as Random Forest (RF), Light-Gradient-Boosting-Machines (LightGBM) and Multilayer Perceptron (MLP).
3. Showing transfer learning capabilities by applying the best performing classifier from the aforementioned analysis on a new test set, which consists of 2 WTs from 2 wind parks, one of which is a completely different wind park compared to the train set.

This paper is organized as follows: In section 2, we give an overview about the dataset. This includes a brief description and explanation of the Anomaly-Space. In section 3,



some background information about transfer learning in fault diagnosis is given. Our supervised fault diagnosis framework is introduced in section 4. Results are shown in section 5 and conclusions are made in section 6.

## 2 Dataset Description

The dataset we used contains fault-types and anomaly scores, which are deduced from SCADA data and vibration data. Two common faults can be found within the data: bearing fault and sensor fault. Bearing faults are common and severe faults in WTs. Ignoring these can result in a total failure of the machine and leads to substantial downtimes and repair costs. Therefore, there is a huge interest in detection of bearing faults as early as possible.

Faulty sensors are also very common in WTs, but are very cheap to replace in terms of raw material cost and are not damaging the WT itself. At a first glance, this might not be as important, but these type of faults can lead to several problems. Unnecessary maintenance works could be performed if sensor faults are confused with a more serious disturbance. Consequently, diagnosticians and technicians might increasingly distrust the fault diagnosis application and/or mistake a serious fault with a sensor fault.

SCADA data are typically used for condition monitoring in WTs. In general, the term SCADA stands for "Supervisory Control and Data Acquisition" and refers to the monitoring and control of technical processes using data that originates from sensors, actuators and other field devices and is sent to a control system. Among other things, process variables such as temperature, pressure and similar values are recorded. Each recorded 10-minute window is aggregated into four scalar values: minimum, maximum, standard deviation and average. Vibration sensors are able to capture data in a much higher sample rate than SCADA and are commonly used to identify early signs of wear, imbalance, or misalignment in rotating machinery.

The data originates from a total of 7 WTs, across 5 different wind parks. All fault cases with further information are listed in Table 1. Our train set (case 1 to 6) consists of data from 5 WTs, which are from 4 different wind parks. The test set (case 7 and 8) has data from 2 WTs, one of which is from a completely different wind park.

**Table 1.** More information about the data. The train data is considered the source data, validation and test data are considered target data. P = park, U = unit, (N)DE = (Non-)Drive End.

Park/Unit	Fault-Type	Fault-Location	Dataset	Case-No.
P1/U1	sensor fault	temperature generator phase 3	train/validation	1
P2/U1	sensor fault	temperature transformer phase 3	train/validation	2
P2/U1	sensor fault	temperature generator phase 1	train/validation	3
P3/U1	bearing fault	Fast Shaft Bearing DE	train/validation	4
P3/U2	bearing fault	Fast Shaft Bearing NDE	train/validation	5
P4/U1	bearing fault	Fast Shaft Bearing DE	train/validation	6
P1/U2	sensor fault	temperature generator phase 2	test	7
P5/U1	bearing fault	Fast Shaft Bearing DE	test	8

## 2.1 Anomaly-Space

The Anomaly-Space refers to a feature space, that has been constructed by using multiple proprietary algorithms, denoted as detectors. The input of these detectors are SCADA data and vibration data.

Detectors are monitoring critical components of the WT and provide anomaly scores, which represent a deviation from the normal behavior. These values are normalized such that measurements with values above 1.0 are considered anomalous.

Broad-Band-Characteristic-Value (bbcv) is one of the detectors. This detector first captures multiple windows of vibration data when pre-defined conditions have been met (e.g. approximately constant wind speeds). Several features are then extracted from the raw vibration data and from the frequency domain, after applying the Fast-Fourier-Transformation (FFT), such as skewness, kurtosis and average values. In the last step, the trendiness of the aforementioned features with the help of hypothesis testing is being calculated. Bearing faults will usually result in an increased trendiness in multiple features.

Another detector is the tuple detector. This detector is designed to detect SCADA data deviations from groups of semantically similar components, for example the generator temperature of all three voltage supply phases. This is being achieved by monitoring the variance of the measurements of these components. A sensor fault in any of these components significantly increases the variance. A statistical test quantifies the difference to the expected null hypothesis, null hypothesis being a variance value of 0.

The procedure of both detectors are depicted in Figure 1 and Figure 2.

The Anomaly-Space is provided to us by our research partner EnBW Energie Baden-Württemberg AG and is used as input data to our preprocessing steps and metamodels.

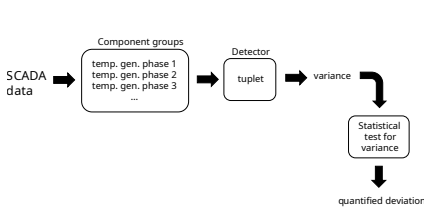


Fig. 1. tuple detector procedure

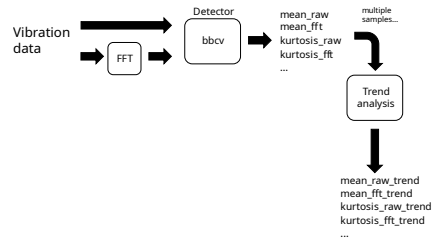


Fig. 2. bbcv detector procedure

## 3 Transfer Learning in Fault Diagnosis

Fault diagnosis can be decomposed into fault detection and fault classification. Fault detection (e.g. using anomaly/outlier detection) is the detection of deviations from normal behavior in the data. These deviations come from either a fault within the monitored system, or from a faulty measuring device. The transformation of the input data can help in detecting faults, that are otherwise not visible.

To be able to diagnose the type of fault, the transformed data can then be classified into pre-defined fault-classes (labels) using supervised classification algorithms.

One of the challenges of fault diagnosis is the lack of labels, since faults occur rarely. In the context of WTs, there might be many fault-labels available across all available WTs. However, there are many reasons why it is still not eligible to train a ML model on one WT and apply it on another without some modifications. For example, WTs may come from different manufacturers and therefore have different components, which results in non-similar signal patterns. Consequently, at least one model for each WT needs to be created.

Transfer learning aims to capture domain-invariant knowledge from just few available domains and apply this knowledge onto other domains. In the context of WTs, for instance, the objective is to leverage insights gained from a few WTs and extend them to others.

There are 4 types of transfer learning methods according to Lei et al. [9]: **feature-based** approaches, **GAN-based** approaches, **instance-based** approaches and **parameter-based** approaches. Feature-based approaches map cross-domain data into a common feature space and decrease the distribution discrepancy before applying a classifier. GAN-based approaches utilize the GAN framework in order to learn the distribution of the target data and generate new ones to improve the classifier. Instance-based approaches reweight misclassified instances from source and target domain, increasing/decreasing the influence of those instances on the fault diagnosis classifier. Parameter-based approaches train models (e.g. neural networks) on the source data and fine-tune the learned model parameters on the target data. Instance- and parameter-based approaches assume, that few labeled samples of the target domain are available. Our solution can be regarded as a feature-based TL approach without the step of further decreasing the distribution discrepancy. This space is designed to represent deviations from the individual normal behavior, making each value intuitive to understand.

## 4 Supervised Fault Diagnosis Framework

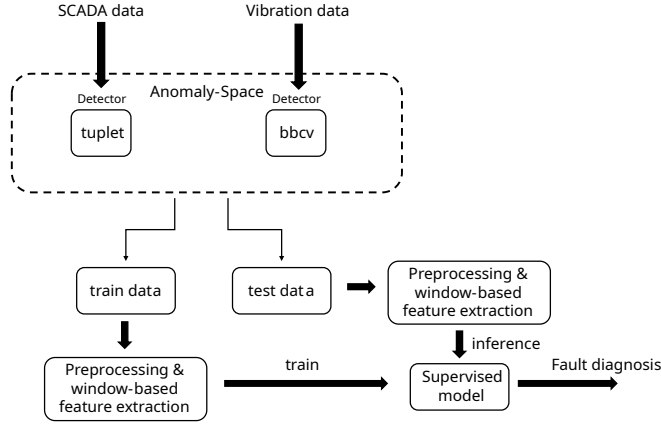
A general overview of the proposed fault diagnosis framework is depicted in Figure 3.

The dataset labels are generated using fault time frames (appearance dates and repair dates) provided by diagnosticians. Data from within the fault time frames are labeled with the corresponding fault type. Data outside these time frames are labeled as "Normal". Data that are not within the normal operating-mode of the WT (e.g. stillstand or wind speeds below a specified threshold) have been omitted. A time-based forward fill was applied, filling missing values for up to 3 hours after the first occurrence. Remaining data gaps are filled with the value 0.0.

The resulting data frame has data from each available WT component as observations (rows) with each detector output being a feature (column).

The amount of features provided by the bbcv detector is reduced to one single feature, by only keeping the feature with the largest variance. Consequently, both detectors provide only a single feature each, resulting so far in 2 total features.

Sliding-window based feature extraction has been employed in order to capture the relationship between neighbored data samples, with a window-size of 144 and a stride of 1. The extracted features are trend-certainty (tc) and variance (var). The Mann-Kendall



**Fig. 3.** Proposed fault diagnosis framework

test is a statistical test for determining trends in data. We used Equation 1 to determine the value of our trend-certainty feature, with  $p_{mk}$  being the p-value for having a positive trend within a window. The values of both features have been manually set to 0.0, if the window contains only values below 1.0. The final feature count has been increased to 6 after extracting those features.

For MLP, the base features, e.g. the outputs of both detectors, have been normalized with a min-max scaler.

$$tc = \begin{cases} 1 & \text{if } p_{mk} < 0.001, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

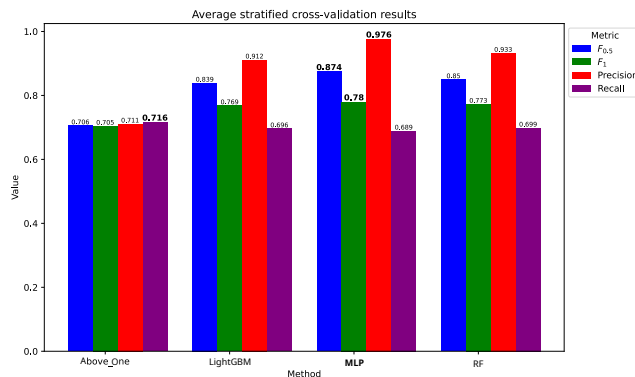
## 5 Results & Discussion

Stratified 3-fold cross-validation was used for evaluation on the train data. We chose the  $f_\beta$ -score with  $\beta = 0.5$  as our evaluation metric. This way, precision has a larger impact than recall. This choice was made because trusting the fault diagnosis is crucial, which can be achieved by minimizing false positive predictions; therefore, precision should be weighted more heavily. Additionally, measurements from faulty sensors could occasionally resemble healthy ones, if the root-cause is loose contact. There could be multiple days of data within the fault time frame, where the sensor delivers non-faulty measurements. Weighting precision and recall equally (e.g. with  $\beta = 1$ ) would provide overly pessimistic evaluation values in these cases.

### 5.1 Comparison between different classifiers on train data

Several popular classifiers were used for our evaluation on the train data: RF, LightGBM and MLP. We created a baseline model which classifies each instance with bbcv values above 1.0 as bearing fault and tuple values above 1.0 as sensor fault, which we simply termed *Above-One*. The results are depicted in Figure 4. The best performing model

is MLP, achieving a  $F_{0.5}$  score of **0.874** with the following hyperparameters: ReLU activation function, Adam optimizer, learning rate of 0.001 and 1 hidden layer with 5 neurons.



**Fig. 4.** Average stratified cross-validation results

## 5.2 Evaluating the best classifier on test data

The MLP was trained on the whole train data with the best parameters and then applied on the test data. Table 1 shows the results. The model achieved a  $F_{0.5}$  score of **0.937**.

**Table 2.** Evaluation results on test data.

Method	$F_{0.5}$ score	$F_1$ score	Precision	Recall
mlp	<b>0.937</b>	0.871	0.992	0.789

## 5.3 Discussion

It can be seen from Figure 4 and Table 2, that higher evaluation scores have been achieved in the test data (0.874 vs. 0.937). This is due to data quality issues, that are present in the train data. More precisely, there is a loose contact sensor fault case, in which multiple signal segments appear normal, within the fault time frame. These segments can be present multiple days. This complicates the fault diagnosis evaluation, since our window size for our feature extraction methods is approximately one day. Possible solutions would be to increase the window size or to split the fault time frame into multiple smaller ones to increase the coverage of the visible fault pattern.

## 6 Conclusion & Future Work

In this paper, a fault diagnosis framework based on an Anomaly-Space is proposed. The Anomaly-Space is a feature space, in which deviations from normal behavior (anomaly

scores) for each WT component are encoded. Window-based features are then extracted from the Anomaly-Space, such as trend values extracted with the Mann-Kendall test. This new feature space provides intuitive values which can help explain fault diagnosis results, since these values represent deviations from the normal behavior in contrast to many other approaches. This framework can be regarded as a feature-based transfer learning method without further decreasing the distribution discrepancy. Supervised classifiers such as Random Forest, Light-Gradient-Boosting-Machines and Multilayer Perceptron are compared on the train data with stratified cross-validation. The Multilayer Perceptron achieved the highest classification performance in diagnosing bearing and sensor faults and was tested on 2 new WTs, one of which stems from a different wind park, compared to the train data. This final evaluation also showed good results, making this a promising fault diagnosis approach for cross-domain fault diagnosis. Future work could include Out-Of-Distribution (OOD) detection to the framework, in order to detect previously unseen fault types.

## Acknowledgement

This work was conducted as part of the research project *AutoDiagCM - Automatisierte Diagnose von Schäden an Windenergieanlagen* (grantnumber 03EE2046B) funded by the German Federal Ministry of Economic Affairs and Climate Action and in cooperation with our research partner EnBW Energie Baden-Württemberg AG, who kindly provided us with data from the Anomaly-Space.

## References

- [1] Geschäftsstelle der Arbeitsgruppe Erneuerbare Energien-Statistik (AGEE-Stat) am Umweltbundesamt. *Erneuerbare Energien in Deutschland Daten zur Entwicklung im Jahr 2023*. Accessed: 7-19-2024. URL: <https://www.umweltbundesamt.de/publikationen/erneuerbare-energien-in-deutschland-2023>.
- [2] Guoqian Jiang et al. “Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox”. In: *IEEE Transactions on Industrial Electronics* 66.4 (2019), pp. 3196–3207. DOI: 10.1109/TIE.2018.2844805.
- [3] Yanting Li, Shujun Liu, and Lianjie Shu. “Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data”. In: *Renewable Energy* 134 (2019), pp. 357–366. ISSN: 0960-1481. DOI: 10.1016/j.renene.2018.10.088.
- [4] Zhi-Xin Yang, Xian-Bo Wang, and Jian-Hua Zhong. “Representational Learning for Fault Diagnosis of Wind Turbine Equipment: A Multi-Layered Extreme Learning Machines Approach”. In: *Energies* 9.6 (2016). ISSN: 1996-1073. DOI: 10.3390/en9060379.
- [5] Chengkai Zhang, Junchi Bin, and Zheng Liu. “Wind turbine ice assessment through inductive transfer learning”. In: *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018, pp. 1–6. DOI: 10.1109/I2MTC.2018.8409794.
- [6] Xiyun Yang et al. “Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier”. In: *Renewable Energy* 163 (2021), pp. 386–397. DOI: 10.1016/j.renene.2020.08.125.
- [7] Yanting Li et al. “Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data”. In: *Renewable Energy* 171 (2021), pp. 103–115. DOI: 10.1016/j.renene.2021.01.143.

- [8] Ziwei Deng et al. “A deep transfer learning method based on stacked autoencoder for cross-domain fault diagnosis”. In: *Applied Mathematics and Computation* 408 (2021), p. 126318. DOI: 10.1016/j.amc.2021.126318.
- [9] Yaguo Lei et al. “Applications of machine learning to machine fault diagnosis: A review and roadmap”. In: *Mechanical systems and signal processing* 138 (2020), p. 106587. DOI: 10.1016/j.ymssp.2019.106587.





# The Potential of Generative AI for Systematic Engineering Innovation

Pavel Livotov, Mas'udah

Offenburg University of Applied Sciences, Badstr. 24, 77652 Offenburg, Germany  
[pavel.livotov@hs-offenburg.de](mailto:pavel.livotov@hs-offenburg.de)  
[masudah@hs-offenburg.de](mailto:masudah@hs-offenburg.de)

**Abstract.** Generative AI offers a new path for engineering innovation by automating idea generation and evaluation. This study explores its effectiveness in addressing complex and inventive engineering challenges. Using automated multi-directional and systematic prompt generation, the paper investigates the ability of AI chatbots to autonomously generate and evaluate innovative solution ideas and concepts. Experiments with various LLMs revealed their potential to accelerate the innovation process but also highlighted limitations in generating feasible, ready-to-use solution concepts. To address these challenges, the paper proposes mixed AI innovation teams, where different generative chatbots can complement and monitor each other. This collaborative approach can improve the quality and feasibility of AI-generated solutions. Case studies demonstrate the practical application of these findings and strategies for effective human-AI collaboration in the innovation process. While generative AI holds significant promise, future research should focus on refining AI models and developing frameworks for effective human-AI interaction to ensure the practical feasibility of AI-generated engineering design solutions for inventive problems.

**Keywords:** Generative AI; Problem-Solving; Engineering Innovation, Inventive Design.

## 1 Introduction – Background and Related Work

In recent years, generative artificial intelligence (AI) has gained significant attention in engineering due to its ability to autonomously generate content, solve complex problems, and assist in decision-making. In engineering design, it shows potential for enabling rapid prototyping, optimizing designs, and streamlining iterative processes. However, fully realizing its potential requires exploring new methods for creatively and autonomously solving engineering problems beyond traditional paradigms.

The impact of generative AI on engineering innovation has been extensively documented, with tools like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Large Language Models (LLMs) demonstrating their efficiency in generating innovative solutions based on training data. Brad [1] explores how inventive principles can enhance activation functions in AI models, increasing their creative capacity. Similarly, Ayaou and Cavallucci [2] propose a framework integrating AI with TRIZ principles to formalize knowledge and link disparate sources for innovative problem-solving. Human-AI collaboration remains crucial in addressing complex engineering challenges. Memmert and Bittner [3] highlight the opportunities of hybrid teams, while Qiu and Jin [4] emphasize the integration of AI with human expertise in enhancing design support systems. Müller, Roth, and Kreimeyer [5] outline barriers to AI-product development integration, such as the lack of standardized processes and documented best practices. Zhu et al. [6] show the success of GPT models in early-stage design concept generation, while Gomez et al. [7] and Ege et al. [8] examine the benefits and limitations

of LLMs in complex system design and ideation. Generative AI tools, such as Open AI ChatGPT, Google Gemini, Anthropic Claude or others, can generate ideas and provide guidance, but human intervention is often required for practical implementation. Excessive or insufficient human involvement poses risks of bias or unfeasible designs. The need for behavioral science integration into AI systems is stressed by Van Rooy and Vaes [9], while Boussioux et al. [10] explore scalable human-AI collaboration for sustainable business innovation.

Xu et al. [11] compare ChatGPT's performance with human evaluators in engineering design tasks, highlighting the need for alignment in judgment confidence to improve decision-making. Chiarello et al. [12] discuss the theoretical and practical benefits of LLMs in automating design tasks, increasing efficiency, and balancing computational and human-centric design. Ranscombe et al. [13] evaluate image generative AI for design inspiration boards, noting differences in quantity, variety, and accuracy compared to traditional methods. Studies [14] and [15] emphasize AI's success in generating a variety of ideas during brainstorming and solving technical problems in process engineering.

However, significant challenges remain in creating detailed, practical solution concepts, especially in fields like mechanical engineering. AI-generated designs often lack the technical precision necessary for implementation, requiring clear instructions and technical drawings. Current text-to-image tools typically produce unsatisfactory results for engineering purposes, underscoring the need for more advanced AI capabilities to bridge the gap between concept generation and practical design implementation.

This paper advocates an integrative approach to automated multidirectional prompt generation, drawing from methodologies such as design theory [16], theory of inventive problem solving TRIZ [17], biomimetics, process intensification [18], and other approaches to systematic innovation. The goal is to improve generative AI chatbots' effectiveness in structured collaborative ideation and problem-solving for engineering design. It also explores AI strategies for developing comprehensive solutions by integrating multiple ideas and evaluating their practical applicability. Through controlled experiments, this study identifies patterns in prompting strategies that enhance the creative potential of AI chatbots, whether operating autonomously or in collaborative groups. However, current research limitations may affect the generalizability of the findings, as challenges remain in objectively assessing engineering creativity.

## 2 Methodology

### 2.1 Multidirectional Prompting

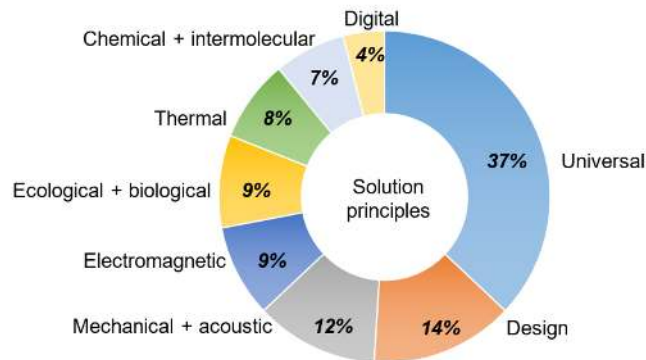
There are various approaches to formulating prompts for generative AI chatbots. This paper introduces Multidirectional Prompting (MDP), which applies elementary solution principles to generate innovative solutions. These solutions are defined by the novel, practical, and feasible combination of one or more ideas, specifically tailored to the problem or objectives. MDP explores multiple directions by addressing sub-problems and applying inventive stimuli, allowing AI to generate holistic solutions. This approach enhances the AI's generative capacity by combining solution ideas suited to the specific problem. In MDP, the multiple directions, sub-problems, and inventive stimuli can be selected either by the user or autonomously by the AI chatbots. Typical MDP techniques, including random, systematic, collaborative, and multi-problem prompting, are presented in Table 1.

**Table 1.** Techniques of Multidirectional Prompting (MDP)

MDP technique	Brief description
1. Random	The AI chatbot simultaneously applies multiple solution principles (SPs) and/or predefined engineering domains without a specific order, generating solution ideas and concepts in a single step.
2. Systematic	The AI chatbot applies a set number of solution principles (SPs) sequentially, generating ideas for each SP. It then combines complementary ideas to create comprehensive solution concepts.
3. Collaborative	Several AI chatbots (e.g., ChatGPT, Google Gemini) independently generate ideas - either randomly or systematically - and then exchange them to develop combined solution concepts.
4. Multiproblem	For complex problems, AI chatbots address prioritized sub-problems either one by one or simultaneously, generating ideas for each sub-problem and combining them into comprehensive solution concepts.

## 2.2 Automated Formulation of Elementary Creative Stimuli

This paper employs an automated method for generating creative stimuli for product and process design across various engineering domains [15]. Validated in both industrial and educational settings, the method has proven its efficacy in generating innovative solutions and improving the design process. The knowledge base is built on 160 elementary inventive principles [18], enhanced by selected TRIZ tools, including the 40 inventive principles, trends of technical evolution, and standard solutions, along with methodologies like biomimetics, process intensification, and others. Automated idea generation operates at multiple levels, such as: a) improving or transforming system components, b) enhancing useful actions, c) eliminating harmful effects, and d) resolving engineering contradictions. A proposed application for automated prompt generation uses 200 predefined inventive principles and allows user customization for solution search across engineering domains, as illustrated in Figure 1. The prompt composition process follows four key steps, as detailed in Table 2.



**Fig. 1.** Variety of the elementary solution principles for multidirectional prompting [15]

**Table 2.** Main steps of automated prompt generation

Prompting step	Brief description
1. Interactive problem definition	AI chatbot and users collaboratively define the problem, ensuring all relevant information and constraints are considered while avoiding biases. The chatbot refines the problem statement, highlighting key data like the ideal outcome, system components, and undesirable effects. Analyses, such as root cause identification, can be conducted as well.
2. Pre-selection of inventive stimuli and engineering domains	Users / AI chatbots select elementary inventive principles and engineering domains based on the initial problem. There's no limit to the number of principles, and new ones can be added at any stage. Users can also customize their choice based on their expertise or preferences.
3. Step-by-step idea generation	After the problem definition is confirmed, the chatbot generates ideas for each selected solution principle. Users can request more ideas or pause the process to move on to concept creation, ensuring the ideation remains flexible and aligned with user preferences.
4. Solution concept creation	The chatbot combines generated ideas into innovative solution concepts (e.g., five or more), leveraging its broad knowledge base. Users can guide the AI with specific strategies, such as focusing on feasibility, novelty, or a particular technology or core ideas, to ensure the solutions meet the problem's inventive objectives and specific requirements.

### 3.3 Experimental Approach

This section outlines two series of experiments investigating AI-based automated ideation and solution concept generation in engineering design. The experiments involved student projects from a course on AI-aided inventive design at the Offenburg University of Applied Sciences, Germany. Five groups of graduate students and one group of undergraduate students participated, along with a control group of the study's authors. The undergraduates were in their 5th or 6th semester, while the graduate students were pursuing a Master's in Mechanical Engineering and Robotics. All students received training in systematic new product development and TRIZ methodology. Working in groups of 2-3, they used generative AI chatbots with automated multidirectional prompt generation to tackle design challenges.

The experiments were conducted in two series. In the first educational series, all groups were assigned the same problem, the "Twist-off Screw Cap" and same initial prompt composition to use with the AI chatbots. This problem concerns jars or bottles with twist-off caps, which are difficult to open due to the high torque in the thread and the vacuum inside. The goal is to find solutions that make opening easier without additional tools or causing issues for manufacturers or consumers.

In the second series of experiments, six student groups already experienced in generative AI applied their skills to a design problem of their choice. They defined the problem, selected appropriate solution principles for multidirectional prompting, and used AI chatbots of their choice. Participants were free to choose their concept generation strategies, including selecting promising combinations of principles, identifying strong ideas, and setting evaluation criteria. For example, focusing on developing concepts around a core idea appropriate for targeted

improvements. Table 3 presents the details of these experiments, including the number of selected solution principles, generated ideas, evaluated concepts, and applied AI chatbots.

**Table 3.** Experimental scope in the second series of experiments using different AI chatbots.

Gr.	Design problem name	Number of			AI tools applied	Concept creation strategy
		solution principles	generated ideas	solution concepts		
1	Smoke detector	10	50	10	ChatGPT Gemini	a) autonomous proposal by AI b) based on 10 strongest AI ideas selected by AI c) based on strongest AI ideas selected by the users
2	Hot drink cup	10	45	12		
3	Quick release	14	50	30	ChatGPT Gemini Claude	
4	Shape adaptive gripper	10	196	18		
5.	Barbeque grill	22	150	33		
6.	Cable winder	10	150	30		

### 3 Results and Discussion

#### 3.1 AI-driven Ideation and Solution Concept Creation

Generative AI, utilizing multi-directional prompting with elementary solution principles, is capable of formulating up to 100–200 distinct ideas, substantially enhancing both productivity and the diversity of generated ideas, surpassing traditional approaches. By systematically exploring a wide range of potential solutions, this technique significantly strengthens the innovation process, fostering the generation of more varied and inventive concepts. In experiments ChatGPT and Claude performed best with multi-directional prompting, while Gemini, though requiring more interaction, tends to provide more objective evaluations.

In the phase of concept creation different generative AI tools perform variably across tasks, presenting an opportunity for mixed AI teams. Tools like ChatGPT, Google Gemini, and Anthropic Claude complement each other, with each offering unique insights, even though the solutions are often similar. However, AI often introduces hidden biases, so monitoring and adjustments are needed for workable solutions. Human judgment remains crucial, especially in addressing subtle aspects of innovation.

Preliminary results indicate that the most robust AI-driven strategy for solution concept creation involves identifying the most promising core ideas and developing multiple concepts based on them. This approach prioritizes inventive goals, such as usefulness or value (as key metrics for goal achievement), over novelty and feasibility during the initial stages of concept development.

#### 3.2 Limitations in AI Evaluation of Ideas and Concepts

During the evaluation phase, the chatbots autonomously assessed their ideas and solution concepts using the following criteria: Feasibility (0 = unviable, 1 = feasible with effort, 2 = easily implementable), Novelty (0 = common, 1 = moderately novel, 2 = highly original), and Usefulness (0 = irrelevant, 1 = moderately useful, 2 = highly useful). In both experimental series,

the chatbots consistently overvalued their concepts compared to participant ratings, particularly in usefulness and feasibility. The use of finer rating scales (e.g., 5- or 10-point scales) in the second series did not significantly reduce overestimation with ChatGPT, whereas Claude demonstrated better accuracy. It's also noted that AI evaluations can vary slightly, typically by  $\pm 1$  point, between repeated evaluations in the same or separate chat sessions. Additionally, individual concept ratings may differ from aggregate ratings of multiple concepts. Finer scales, such as 10-point ratings, provide more nuanced and consistent assessments, helping to mitigate this variability.

Moreover, generative AI models appear to exhibit a moderate "Not Invented Here" effect when evaluating solutions proposed by other AI chatbots or engineers in concept evaluation across all experiments. For example, in a pairwise comparison, both ChatGPT4.0 and Gemini rate the usefulness of their own concepts higher. The authors consider this phenomenon useful, as it promotes a more balanced assessment of ideas and concepts when different AI chatbots operate as virtual teams of specialists, either autonomously or in collaboration with engineers.

### 3.3 Feedback and Observations from Experiments

Different generative AI tools perform variably across tasks, presenting an opportunity for mixed AI teams. Tools like Open AI ChatGPT, Google Gemini, and Anthropic Claude complement each other, with each offering unique insights despite similar solutions. ChatGPT and Claude perform best with multi-directional prompting, while Gemini, though requiring more interaction, tends to provide more objective evaluations. However, AI often introduces hidden biases, so monitoring and adjustments are needed for workable solutions.

The results of an anonymous survey conducted among the 17 participants at the end of the second series of experiments are particularly interesting. The participants rated their responses using a 10-point scale: 1-2 (very low), 3-4 (low), 5-6 (medium), 7-8 (high), and 9-10 (very high), with the mean values presented in Table 4.

**Table 4.** Results of an anonymous survey on the performance of generative AI

No.	Survey question: How do you rate the following aspects in application of generative AI ...	Mean values (17 participants)
1	contribution of AI to increasing your personal inventive CREATIVITY?	7.1 SD=1.6
2	performance of AI in terms of the ideas USEFULNESS?	6.1 SD=2.1
3	performance of AI in terms of the ideas NOVELTY?	6.7 SD=1.9
4	performance of AI in terms of the ideas FEASIBILITY?	4.7 SD=1.9
5	overall performance of AI in the solution concept development phase?	6.1 SD=2.0
6	level of detail of the solution concepts proposed by the AI, so that designers can quickly implement a solution concept?	4.7 SD=2.0
7	accuracy of the evaluation of solution concepts by AI?	4.2 SD=1.7

AI chatbots received the highest rankings for enhancing participants' personal inventive creativity, but the lowest for evaluation accuracy and the level of detail in solution concepts needed for quick implementation.

Interestingly, participants also reported difficulties in personally evaluating the large number of ideas generated by AI. Subsequent analysis of the students' protocols by supervisors revealed that many novel and useful ideas were not recognized as promising and were excluded from concept creation. This highlights a key challenge in applying AI to the innovation process: engineers or students often expect ready-to-use solutions and struggle to thoroughly process numerous ideas generated by AI. This challenge highlights the need for a systematic exploration of collaborative frameworks and models for AI and human interaction in the inventive design process. Future research should prioritize developing and refining these collaboration models to optimize the integration of AI technologies in human-centered innovation, while also fostering the acceptance of design concepts created autonomously by AI.

#### 4 Concluding Remarks and Outlook

The results of this study reveal key insights for applying generative AI in inventive engineering design. First, multi-directional prompting with elementary solution principles greatly boosts productivity and variety in idea generation, surpassing traditional methods like brainstorming or classical TRIZ. The challenge now shifts to selecting strong ideas and developing effective solution concepts. A key question is finding the optimal balance of human involvement in AI-assisted problem-solving. Second, AI chatbots tend to overestimate the feasibility of their concepts, highlighting the need for better self-evaluation algorithms. Bridging the gap between AI and human evaluation is crucial for real-world application. Third, varying degrees of overestimation between AI models (e.g., ChatGPT versus Gemini) show that model architecture impacts assessment accuracy. Future research should focus on minimizing these biases. Finally, the gap between AI-generated ideas and practical implementation remains a challenge. Advances in AI's ability to produce technically feasible solutions, including text-to-CAD tools will be essential for improving AI's role in engineering design and inventive problem solving.

#### References

1. Brad, S.: Enhancing Creativity in Deep Learning Models with SAVE-Inspired Activation Functions. In: Cavallucci, D., Livotov, P., Brad, S. (eds) Towards AI-Aided Invention and Innovation. TFC 2023. IFIP Advances in Information and Communication Technology, vol 682. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-42532-5\\_12](https://doi.org/10.1007/978-3-031-42532-5_12)
2. Ayaou, I., Cavallucci, D.: Multi-domain and Heterogeneous Data Driven Innovative Problem Solving: Towards a Unified Representation Framework. In: Cavallucci, D., Livotov, P., Brad, S. (eds) Towards AI-Aided Invention and Innovation. TFC 2023. IFIP Advances in Information and Communication Technology, vol 682. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-42532-5\\_10](https://doi.org/10.1007/978-3-031-42532-5_10)
3. Memmert, L., Bittner, E.: Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. Proceedings of the 55th Hawaii International Conference on System Sciences, pp 378-387 (2022). <http://hdl.handle.net/10125/79376>
4. Qiu, Y., Jin, Y.: ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. Intelligent Systems with Applications, 21, 200308 (2024). <https://doi.org/10.1016/j.iswa.2023.200308>

5. Müller, B., Roth, D., Kreimeyer, M.: Barriers to the Use of Artificial Intelligence in the Product Development – A Survey of Dimensions Involved. *Proceedings of the Design Society*, 3, 757-766 (2023). doi:10.1017/pds.2023.76
6. Zhu, Q., Luo, J.: Generative transformers for design concept generation. *Journal of Computing and Information Science in Engineering*, 23(4), Article 041003 (2023). DOI: 10.1115/1.4056220
7. Gomez, A.P., Krus, P., Panarotto, M., Isaksson, O.: Large language models in complex system design. *Proceedings of the Design Society*, 4, 2197-2206 (2024). doi:10.1017/pds.2024.222
8. Ege, D.N., Øvrebø, H.H., Stubberud, V., Berg, M.F., Steinert, M., Vestad, H.: Benchmarking AI design skills: insights from ChatGPT's participation in a prototyping hackathon. *Proceedings of the Design Society*, 4, 1999-2008 (2024). doi:10.1017/pds.2024.202
9. Van Rooy, D., Vaes, K.: Harmonizing human-AI synergy: behavioral science in AI-integrated design. *Proceedings of the Design Society*, 4, 2287-2296 (2024). doi:10.1017/pds.2024.231
10. Boussioux, L., Lane, J.N., Zhang, M., Jacimovic, V. and Lakhani, K.R.: The Crowdless Future? Generative AI and Creative Problem Solving. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-005, July 01, 2024). <http://dx.doi.org/10.2139/ssrn.4533642>
11. Xu, W., Kotecha, M.C., McAdams, D.A.: How good is ChatGPT? An exploratory study on ChatGPT's performance in engineering design tasks and subjective decision-making. *Proceedings of the Design Society*, 4, 2307-2316 (2024). doi:10.1017/pds.2024.233
12. Chiarello, F., Barandoni, S., Majda Škec, M., Fantoni, G.: Generative large language models in engineering design: opportunities and challenges. *Proceedings of the Design Society*, 4, 1959-1968 (2024). doi:10.1017/pds.2024.198
13. Ranscombe, C., Tan, L., Goudswaard, M., Snider, C.: Inspiration or indication? Evaluating the qualities of design inspiration boards created using text to image generative AI. *Proceedings of the Design Society*, 4, 2207-2216 (2024). doi:10.1017/pds.2024.223
14. Mas'udah, Livotov, P.: Nature's lessons, AI's power: sustainable process design with generative AI. *Proceedings of the Design Society*, vol. 4, pp. 2129–2138, Cambridge University Press (2024). doi:10.1017/pds.2024.215
15. Livotov, P.: Enhancing Engineering Creativity with Automated Formulation of Elementary Solution Principles. *Proceedings of the Design Society*, vol. 3, 1645-1654 (2023). doi:10.1017/pds.2023.165
16. Kannengiesser, U., Gero, J.S.: Can Pahl and Beitz' systematic approach be a predictive model of designing? *Design Science*, vol. 3, e24, Cambridge University Press (2017). doi:10.1017/dsj.2017.24
17. Altshuller, G.S.: *Creativity as an exact science: the theory of the solution of inventive problems*, Gordon and Breach Science Publishers, New York, ISSN 0275-5807 (1984)
18. Livotov, P., Chandra, S.A.P., Mas'udah, Law, R., Reay, D., et al.: Eco-innovation in process engineering: Contradictions, inventive principles and methods, *Thermal Science and Engineering Progress*, Volume 9, Pages 52-65 (2019). <https://doi.org/10.1016/j.tsep.2018.10.012>



# Evaluating AI-Generated Solution Ideas: A Comparative Study of AI and Human Assessments for Sustainable Process Design

Mas'udah<sup>1</sup>, Pavel Livotov<sup>1</sup>, Saptadi Nugroho<sup>1,2</sup>

<sup>1</sup>Offenburg University of Applied Sciences, Badstr. 24, 77652 Offenburg, Germany

<sup>2</sup>Albert Ludwig University of Freiburg, Georges-Köhler-Allee 101, 79110 Freiburg im Breisgau, Germany

`masudah@hs-offenburg.de`

`pavel.livotov@hs-offenburg.de`

`saptadi.nugroho@hs-offenburg.de`

**Abstract.** The integration of generative artificial intelligence (AI) in sustainable process design has gained substantial traction, with AI increasingly employed to generate innovative solutions. However, the efficacy of these AI-generated ideas requires rigorous evaluation to ensure their quality. This study examines the dual role of GPT-4o in generating and evaluating solution ideas for sustainable process design at the concept development stage. Focusing on process engineering, the research applies these methods to a case study involving froth flotation for nickel recovery. By comparing AI-driven assessments with those from human experts, the research aims to determine the alignment between AI and human evaluations across key criteria: novelty, feasibility, usefulness, and sustainability. The results reveal strong alignment in most areas, though notable discrepancies in novelty suggest that human expertise remains essential for nuanced judgments on uniqueness. These findings highlight GPT-4o's potential as a preliminary evaluation tool, while also underscoring the need for a hybrid approach that combines AI insights with human expertise.

**Keywords:** Generative AI; Sustainability; Process Design; AI-Human Evaluation.

## 1 Introduction

Recent advancements in artificial intelligence (AI) have had a transformative impact across various sectors, enabling rapid innovation and the generation of solutions for complex challenges. Among these developments, Generative Pre-trained Transformer (GPT) models developed by OpenAI [1], have demonstrated significant potential in automating ideation processes, which is especially valuable in the context of sustainable process design. The versatility of these models lies in their ability to process extensive datasets and generate contextually relevant responses, making them well-suited for early-stage concept development aimed at addressing environmental and technical issues [2, 3].

Despite the promise of generative AI, challenges remain in ensuring the quality and applicability of AI-generated solutions. Current approaches often rely heavily on human evaluation to assess key criteria, such as novelty, feasibility, usefulness, and sustainability. These human-driven evaluations, while effective, are time-consuming and often subjective, which highlights the need for AI systems capable of autonomously assessing the quality of their outputs.

Prior studies on generative AI have explored its application in eco-innovation, but limitations persist regarding the AI's ability to evaluate its outputs. This research addresses this gap by investigating GPT-4o's capacity not only to generate but also to evaluate solution ideas. The

research aimed to explore how effectively GPT-4o could autonomously assess its outputs and compare these AI-driven evaluations with those conducted by human experts. By concentrating on established evaluation criteria such as novelty, feasibility, usefulness, and sustainability, the study sought to determine the extent to which GPT-4o's assessments aligned with expert evaluations. The research seeks to address the following questions:

- 1) Can GPT-4o autonomously evaluate its generated ideas?
- 2) How do AI evaluations align with human expert assessments across key metrics in sustainable process design?

## 2 Background and Related Work

### 2.1 Generative AI in Sustainable Innovation

Generative AI, such as GPT models, has emerged as a powerful tool in the drive toward sustainable innovation by enabling the rapid generation of ideas that address pressing environmental and technical challenges [4]. Since the introduction of GPT-2 in 2019, these models have evolved significantly, with GPT-3 and subsequent versions, including GPT-4o, demonstrating advanced capabilities for generating complex and contextually relevant ideas from extensive datasets [1, 5, 6]. This versatility makes GPT models particularly valuable in fields where rapid innovation and sustainability are essential, such as process engineering, where early-stage concept development plays a pivotal role in tackling environmental issues and technical obstacles.

Despite the promise of generative AI, significant challenges remain in ensuring the quality and applicability of AI-generated solutions. Assessing these qualities has largely depended on human evaluations, which present scalability issues and introduce subjectivity into the assessment process. For instance, Zu et al. [2, 3] utilised GPT-3 to generate biologically inspired design concepts but relied solely on a design team to evaluate feasibility and novelty. While human expertise is invaluable, it often requires substantial time and resources, thus limiting the potential for large-scale, efficient evaluations. Similarly, Li et al. [7] explored multiple generative AI tools, including ChatGPT (GPT-3.5) [1], Midjourney (via Discord) [8], and Stable Diffusion (via WebUI) [9], to facilitate sustainable design projects. Although the study incorporated topics aligned with Sustainable Development Goals (SDGs), it focused exclusively on usability and depended merely on human assessment. While these previous studies successfully demonstrate that generative AI can support sustainable innovation, they fall short in exploring AI-driven evaluation mechanisms, leaving a significant gap in objectively assessing the quality of AI-generated ideas.

Building on this foundation, the previous research by authors [4], integrated nature-inspired principles into a GPT-3.5 model to generate eco-innovative solutions for challenges in process design. This study not only used GPT-3.5 to generate solution ideas but also to evaluate them through self-assessment, with comparisons drawn against human evaluations. However, the study was limited by a lack of in-depth sustainability assessments. The evaluation metrics were relatively general and did not delve into the specifics, leaving a gap in comprehensively assessing the AI-generated solution ideas. This gap underscores the need for more robust evaluation frameworks that incorporate detailed sustainability metrics, enabling AI to perform more rigorous self-assessments in alignment with sustainable development goals.

## 2.2 Idea Evaluation

Evaluating ideas is an essential component of the innovation process, as it helps in selecting concepts that are most likely to succeed and contribute meaningfully to their intended goals. Common criteria for idea evaluation typically include novelty, feasibility, usefulness, and sustainability [10–12]. These criteria provide a structured basis for assessing the quality of ideas and determining their potential impact.

Various methods have been developed to evaluate these criteria, ranging from expert assessments to crowdsourcing and AI-assisted frameworks. Dean et al. [11] focused on measuring ideas through constructs such as novelty, workability, relevance, and specificity using multi-dimensional scales rated by human experts. The study noted that human raters often interpret these criteria differently, leading to variability in results. This variability underscores a broader challenge in maintaining consistency across evaluations, especially when relying on subjective human assessments.

In another study, Baffo et al. [12] employed utility value analysis to score ideas based on environmental, economic, and social sustainability parameters, allowing for a comprehensive sustainability assessment. This method provides a structured approach to sustainability but tends to be labour-intensive and primarily suited for detailed project-level evaluations, making it less practical for rapid, large-scale screening in early-stage concept development.

Furthermore, a recent study by Mesbah et al. [13] integrated crowdsourced evaluations with machine learning models to improve scalability and consistency. In this framework, crowd raters are guided by well-defined scales to measure criteria such as novelty and usefulness. The machine learning component adjusts for variability in crowd ratings, providing a more reliable aggregate score. However, while the study enhances scalability, it remains limited by the variability and potential biases introduced by crowd raters.

## 3 Methodology

### 3.1 Research Design

This study employed a dual approach in which GPT-4o was used to both generate and evaluate solution ideas for sustainable process design, focusing on a specific case study outlined in Table 1. GPT-4o was selected for its enhanced contextual understanding and advanced generative capabilities [14]. Compared to previous versions, GPT-4o has demonstrated improved performance in generating contextually relevant responses to complex prompts, making it particularly suitable for generating innovative solutions [1, 14].

**Table 1.** Case study: problem and ideal final result

<b>Case study</b>	<b>Problem description</b>	<b>Ideal final result</b>
Froth Flotation for Nickel Recovery	The utilisation of chemicals in the process results in water pollution and the generation of solid waste. Additionally, the low efficiency of the process contributes to elevated production expenses.	Develop a sustainable and effective procedure that reduces the utilisation of chemicals and waste, thereby decreasing production expenses.

### 3.2 Solution Idea Generation

In alignment with the methodology outlined in the authors' prior research [4], this study employed a solution-driven approach (SDA) in biomimetics, combined with AI-automated prompting to generate solution ideas. The SDA framework leverages principles derived from natural ecosystems exposed to environmental stress, such as mangroves, rainforests, caves, etc. as a basis for developing innovative solutions. By analysing the adaptation strategies within these ecosystems, the approach identifies inventive principles that can be applied to engineering challenges. AI-automated prompting complemented the basic prompts, guiding GPT-4o to achieve more refined results by structuring the input queries and enhancing the output iteratively. The idea generation process followed four key phases, as outlined in Table 2.

**Table 2.** Idea generation using SDA with generative AI [4]

SDA Phase	Basic prompting	AI-automated prompting
1. Identification and analysis of the natural solution in ecosystem exposed to environmental stress (e.g. mangroves, rainforest, etc.)	Identify the key components and adaptation strategies that help the following ecosystems survive in hostile environments, and explain how they aid survival under environmental stress.  <i>[list all-natural ecosystems that are intended to be identified and analysed]</i>	Basic prompting for phase 1 +  Follow the instructions below: 1) Revise the prompt to be clear, concise and easily understood by you. Ask any relevant questions needed to improve the prompt.
2. Extraction of natural solution principles	Based on the adaptation strategies identified, distil abstract, non-biomimetic principles that capture the essence of how these ecosystems survive in hostile environments. Name these principles succinctly, ensuring they are generalized enough to apply across different engineering challenges.	2) Execute the revised prompt upon approval, and provide iterative feedback to refine the information generated. Ask: 'Do you agree with the revised prompt (please type NO or YES)?' If 'NO', ask for more detail to refine the prompt. If 'YES', proceed to provide the answer, and then ask: 'Do you want more [e.g. information, ideas, examples, etc] (please type MORE) or go to the next step (type NEXT)?' If 'MORE', suggest 5 more [e.g. information, ideas, examples, etc] and repeat the feedback question. If 'NEXT', ask:
3. Ideas generation with nature-inspired principles to solve the problem	Using the extracted inventive principles, generate 5 ideas that address the sustainable process design challenge to the following problem. Ensure that the ideas are distinct and not direct copies of natural processes.  <i>[Describe the problem and the desired result]</i>	'What else can I assist you?' Wait for the feedback. If I give you an inquiry, you repeat instructions 1 to 2 above until the job is done.
4. Concepts generation	Create 5 inventive solution concepts by combining different complementary ideas to address the problem. These solution concepts should offer comprehensive solution approaches that incorporate various elements to enhance useful action and mitigate harmful effects.	

In Phase 1, GPT-4o was prompted to identify and analyse adaptation strategies within selected ecosystems. The initial basic prompts were designed to elicit information about the ecosystem components and their survival mechanisms. The AI-automated prompting then provided additional structured steps, enabling GPT-4o to iteratively refine the prompts based on feedback, ensuring a deeper and more accurate understanding of each ecosystem’s adaptation strategies.

In Phase 2, the focus shifted to extracting natural solution principles from these adaptation strategies. GPT-4o was directed to distil these principles into abstract, non-biomimetic forms that could be applied across various engineering contexts. The AI was prompted to avoid direct copying of natural components and instead translate the underlying inventive principles into a generalised format that could inspire innovative solutions.

Phase 3 involved using these nature-inspired principles to generate specific solution ideas addressing the sustainable process design challenge outlined in the case study. GPT-4o was directed to use these principles to produce a diverse range of ideas, generating up to 50 ideas to address the problem. These ideas were then evaluated using the criteria outlined in Section 3.3.

In the final phase, GPT-4o was tasked with creating inventive solution concepts by combining complementary top-scoring ideas from phase 3. These concepts incorporated various elements to enhance beneficial actions and mitigate harmful effects, aligning closely with sustainability goals. In this study, GPT-4o was asked to create up to five solution concepts to address the problem. To ensure consistency and minimise variations in AI responses, both generation and evaluation of the idea were conducted on the same day and at the same time.

### 3.3 AI-generated Solution Idea Evaluation

The AI-generated solution ideas were evaluated by both AI and human experts using metrics as shown in Table 3. The definitions and rationale for these criteria were derived from the literature [3, 10, 11]. A total of 50 solution ideas for the case study were assessed using these metrics on a scale of 0 to 2.

- *AI self-evaluation* – ChatGPT was asked to evaluate its solution ideas.
- *Human-assisted evaluation* – Two experts, specialising in process engineering and database system engineering, were trained to perform a manual evaluation of the solution ideas independently.

**Table 3.** Assessment criteria for generated solution idea

Parameter	Description	Rating scale
<b>Novelty</b> - assesses the uniqueness of the solution idea from existing solutions		
• Originality	The solution introduces a unique approach that is not found in existing solutions, such as those found in patent databases, published articles, and other sources	0 - Not novel (common or existing solution) 1 - Moderately novel (introduces some new aspects) 2 - Highly novel (completely new or unique)
• Inventiveness	The solution combines existing concepts in new ways	
• Paradigm shift	The solution transforms existing norms, conventions, or models	

<b>Feasibility</b> - evaluates the viability of implementing the solution idea		
• Technical viability	The necessary technologies and resources are available for implementation	0 - Unviable (highly impractical) 1 - Moderately feasible (possible but requires effort) 2 - Highly feasible (easily implementable)
• Financial viability	The solution has reasonable cost implications compared to the expected benefits	
• Scalability	The solution can be effectively adapted for large-scale or varying needs	
<b>Usefulness</b> - measures the solution idea's potential to address the identified problem effectively		
• Effectiveness	The idea directly addresses the core issues	0 - Useless (does not address the problem) 1 - Moderately useful (resolves a few issues) 2 - Highly useful (completely addresses the problem)
• Practicality	The solution is applicable in real-world settings or targeted process	
• Relevance	The solution meets the expectations or needs of users	
<b>Sustainability</b> - estimates the environmental, social and economic impact of the solution idea		
• Environmental impact	The solution minimises harm to the environment	0 - Unsustainable (significant negative impacts) 1 - Moderately sustainable (minor negative impacts) 2 - Highly sustainable (major positive impacts)
• Social impact	The solution contributes positively to social well-being	
• Economic impact	The solution provides economic benefits and promotes financial stability	

To assess the level of agreement between AI and human evaluators, Cohen’s Kappa [15] was used to measure inter-rater agreement between individual human evaluators and GPT-4o. This statistic provides insight into how consistently the AI ratings align with each human rater independently. Additionally, Fleiss’ Kappa [16] was applied to evaluate overall agreement across all evaluators, including GPT-4o and the two human experts, thereby offering a comprehensive measure of consensus. These analyses help quantify the reliability of the AI’s evaluations in comparison to human judgments across the assessment categories.

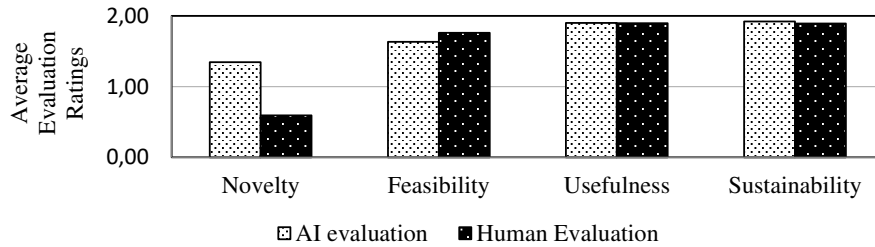
**Table 4.** Interpretation of kappa values [17]

<b>Kappa value</b>	<b>Agreement reliability</b>
≤0	No agreement
0.1 – 0.20	Slight Agreement
0.21 – 0.40	Fair Agreement
0.41 – 0.60	Moderate Agreement
0.61 – 0.80	Substantial Agreement
0.81 – 1.00	Almost Perfect Agreement

## 4 Results and Discussion

### 4.1 Idea evaluation

Figure 1 compares AI-driven and human expert evaluations across four criteria: novelty, feasibility, usefulness, and sustainability. Overall, the results show a strong alignment between AI and human assessments, particularly in feasibility, usefulness, and sustainability, supporting GPT-4o as a preliminary evaluation tool where human judgment is resource-intensive. A notable difference appears in novelty, where AI-rated ideas are higher than human evaluators. This suggests that GPT-4o perceives greater originality, likely due to its ability to draw upon a vast dataset and combine existing concepts in new ways. In contrast, human experts may apply stricter criteria, expecting more distinct deviations from conventional solutions. This divergence underscores a need for refinement to bring AI's novelty assessments closer to human standards. These results imply that, while AI can be a valuable asset for initial assessments, human input remains essential for a more nuanced understanding, particularly in the evaluation of novelty.



**Fig. 1.** The average evaluation rating of generated ideas: AI vs human

To further illustrate these findings, Table 5 presents a sample of ideas generated by GPT-4o to address the challenges in the case study, along with their practical implications. The table compares AI and human expert evaluations across key assessment categories. Overall, the results reveal a high level of alignment between AI and human evaluations, with closely matching ratings in most areas, except for novelty, where some discrepancies were observed. This outcome underscores GPT-4o's potential as an effective preliminary evaluation tool. While AI closely mirrors human assessments in feasibility, usefulness, and sustainability, the differences in novelty indicate that human expertise remains essential for nuanced judgment. These findings highlight the value of a hybrid approach that combines AI-driven assessments with human insights, particularly for evaluating aspects such as originality, inventiveness, and paradigm shift.

Table 6 presents the inter-rater agreement across key assessment categories using Cohen's Kappa and Fleiss' Kappa. Overall, the strongest alignment is observed in sustainability, particularly in the environmental and social aspects, where Cohen's Kappa values show higher consistency. In contrast, lower agreement scores are found in novelty, especially for originality and inventiveness, indicating differences in how AI and humans assess uniqueness. Fleiss' Kappa values confirm this trend, showing that sustainability has the highest overall agreement, while novelty displays more variability. These results suggest that GPT-4o aligns well with human assessments in endurance-related aspects but may require refinement in evaluating novelty.

**Table 5.** Generated idea examples and practical implications for case study according to AI and Human ratings on novelty (N), feasibility (F), usefulness (U), and sustainability (S) (in fragment)

ID	Idea description	Practical implication	Rated by	N				F				U				S				Total
				Originality	Inventiveness	Paradigm shift	Average ( $\bar{N}$ )	Technical	Financial	Scalability	Average ( $\bar{F}$ )	Effectiveness	Practicality	Relevance	Average ( $\bar{U}$ )	Environmental	Social	Economic	Average ( $\bar{S}$ )	
1	Use biodegradable frothing agents that break down naturally after use	Reduces chemical pollution in discharge water	AI	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	8
			Human rater 1	1	1	1	1	2	1	1	1.33	1	2	2	1.67	2	2	2	2	6
			Human rater 2	0	0	0	0	1	1	1	1	1	2	2	1.67	2	2	2	2	4.67
...	...	...	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
50	Implement a multi-stage flotation process	Reduce chemical dosages and enhance ore recovery rates	AI	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	7
			Human rater 1	0	0	0	0	2	2	2	2	2	2	2	2	2	2	2	2	6
			Human rater 2	0	0	0	0	2	1	2	1.67	2	2	2	2	2	2	2	2	5.67

**Table 6.** Inter-rater agreement among raters on novelty (N), feasibility (F), usefulness (U), and sustainability (S) using Cohen's kappa and Fleiss' kappa

Ratings	Cohen's kappa value											
	N			F			U			S		
	Originality	Inventiveness	Paradigm shift	Technical	Financial	Scalability	Effectiveness	Practicality	Relevance	Environmental	Social	Economic
AI – Human rater 1	0.160	0.022	0.239	0.393	0.517	0.506	0.407	0.638	-0.056	0.396	0.558	0.694
AI – Human rater 2	0.093	-0.057	0.153	0.132	0.322	0.330	0.225	0.260	-0.056	0.396	0.457	0.390
Human rater 1- Human rater 2	0.680	0.701	0.696	0.651	0.786	0.737	0.675	0.485	0.728	0.811	0.779	0.336
<b>Fleiss' kappa value (Overall)</b>	<b>0.191</b>	<b>0.164</b>	<b>0.212</b>	<b>0.362</b>	<b>0.541</b>	<b>0.471</b>	<b>0.442</b>	<b>0.464</b>	<b>0.250</b>	<b>0.556</b>	<b>0.606</b>	<b>0.481</b>



## 4.2 Limitation and Future Work

This study demonstrates GPT-4o's effectiveness as a preliminary evaluation tool, though some limitations are evident. The AI showed discrepancies in assessing novelty, possibly due to its limited ability to fully interpret human standards for uniqueness. Additionally, the reliance on a single AI model and a specific case study restricts the generalisability of the findings, as results may vary with different models or contexts. The limited sample of two human experts for comparison also suggests that a larger, more diverse panel would provide a more comprehensive benchmark for AI assessments. Furthermore, the solution-driven approach (SDA) used in this study employed AI-automated prompting, where ChatGPT revised each prompt based on user instructions. While effective, this iterative process requires manual confirmation at each step. In the future, developing a system like a large language model (LLM) prompter could automate this refinement, reducing manual intervention and improving efficiency.

Future research could explore other AI models and apply the evaluation framework to multiple case studies to broaden the scope of findings. Integrating AI tools trained on creative or sustainability-specific datasets might improve AI's accuracy in evaluating subjective criteria like originality. Furthermore, a hybrid evaluation model that combines AI with a larger, diverse panel of human experts could enhance understanding of AI-human alignment, particularly in complex areas such as originality and paradigm shifts. Expanding stakeholder involvement to include sustainability practitioners and domain experts could also enhance the applicability of AI-based evaluations in sustainable process design.

## 5 Conclusion

This study finds that GPT-4o can autonomously evaluate its generated ideas, showing strong alignment with human assessments in feasibility, usefulness, and sustainability. However, discrepancies in novelty assessment indicate that human expertise is essential for nuanced judgments on originality, inventiveness, and paradigm shift. These results demonstrate that GPT-4o can function as a preliminary evaluation tool, but combining AI-driven assessments with human insights provides a more robust framework. This study suggests that while GPT-4o aligns well with human evaluations in most criteria, further refinement is needed for novelty assessments. Future research should explore broader applications of AI in sustainable innovation and focus on refining AI models to improve alignment in subjective evaluations like originality.

## References

1. ChatGPT Homepage, <https://chat.openai.com/>, last accessed 2024/09/24.
2. Zhu, Q., Zhang, X., Luo, J.: Generative Pre-Trained Transformers for Biologically Inspired Design. In: 34th International Conference on Design Theory and Methodology (DTM), American Society of Mechanical Engineers (2022). <https://doi.org/10.1115/DETC2022-90366>. 3.
3. Zhu, Q., Zhang, X., Luo, J.: Biologically inspired design concept generation using generative pre-trained transformers. *Journal of Mechanical Design* 145(4), art. 041409 (2023).
4. Mas'udah, Livotov, P.: Nature's lessons, AI's power: sustainable process design with generative AI. In: *Proceedings of the Design Society*, vol. 4, pp. 2129–2138, Cambridge University Press (2024). DOI: <https://doi.org/10.1017/pds.2024.215>

5. Radford, A., et al.: Language Models are Unsupervised Multitask Learners. Technical report, OpenAi (2019).
6. Brown, T. et al.: Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, 33, pp. 1877-1901 (2020).
7. Li, M. et al.: Generative AI for Sustainable Design: A Case Study in Design Education Practices. In: Kurosu, M., Hashizume, A. (eds) *Human-Computer Interaction. HCII 2024. Lecture Notes in Computer Science*, vol 14687. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-60441-6\\_5](https://doi.org/10.1007/978-3-031-60441-6_5)
8. Midjourney Homepage: <https://discord.com/>, last accessed 2024/09/24.
9. Stable Diffusion Homepage: <https://stablediffusionweb.com/WebUI>, last accessed 2024/09/24.
10. Ma, K., Grandi, D., McComb, C., Goucher-Lambert, K.: Conceptual Design Generation Using Large Language Models. In: *Proceedings of the ASME Design Engineering Technical Conference*, vol 6, Boston, Massachusetts, USA (2023). <https://doi.org/10.1115/detc2023-116838>.
11. Dean, D.L., Hender, J.M., Rodgers, T.L., Santanen, E.L.: Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information System*, vol. 7, Iss.10 (2006). <https://doi.org/10.17705/1jais.00106>.
12. Baffo, I., Leonardi, M., Bossone, B., Camarda, M.E., D'Alberti, V., Travaglioni, M.: A decision support system for measuring and evaluating solutions for sustainable development. *Sustainable Futures*, vol. 5, 100109 (2023). <https://doi.org/10.1016/j.sfr.2023.100109>.
13. Mesbah, S., Arous, I., Yang, J., Bozzon, A., Bozzon, A.: HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale. In: *Proceedings of the ACM Web Conference 2023*. Association for Computing Machinery, pp. 3837–3848, New York, NY, United States (2023). <https://doi.org/10.1145/3543507.3583496>
14. Shahriar, S. et al.: Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. (2024). <https://doi.org/10.48550/arXiv.2407.09519>
15. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37–46. <https://doi.org/10.1177/001316446002000104>
16. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382 (1971).
17. Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174 (1977). <http://dx.doi.org/10.2307/2529310>.

# Specification and Identification of Relationships between Products in the Food Segment

Sian Brumm, Rolf Krieger, and Christoph Brosch

Trier University of Applied Sciences, Umwelt-Campus Birkenfeld

s.brumm@umwelt-campus.de  
r.krieger@umwelt-campus.de  
c.brosch@umwelt-campus.de

**Abstract.** There are numerous relationships between products in retail. Understanding these relationships is particularly valuable for optimizing e-commerce, master data management, and marketing, among other applications. However, these relationships are often not explicitly known. In this work, we address the problem of automatically detecting a wide range of product relationships within the food segment. To achieve this, we provide a detailed specification of relationships specific to food products. Additionally, we describe multi-class classification models for automatically determining these relationships.

**Keywords:** Product Relations, Product Matching, Machine Learning

## 1 Introduction

Retail companies often manage data for hundreds of thousands, and sometimes millions, of products in their information systems. Understanding the relationships between different products is crucial for maintaining and utilizing this product data across various functional areas of the company. There are many types of relationships. For example, two products may be considered related if they share the same brand and content but differ in package size, or if they share the same brand, content type, and quantity but come in different flavors. Many other types of relationships are also of interest.

Knowledge about product relationships can be leveraged in various ways to support retail business processes. For example, certain relationships may be useful in sales to recommend complementary products, such as accessories or those that present an up-sell opportunity. Product recommendations are a key tool in online stores for boosting sales. Competitive pricing analysis, on the other hand, requires a detailed comparison of product prices between competitors. To perform such an analysis, one must first identify the products to compare, which involves determining the relationships between products in the assortments of both competitors. In master data management, these relationships can simplify or even partially automate the maintenance of master data. For instance, the process of creating new product records can be facilitated by suggesting attribute values from similar products, and significant discrepancies in these attributes might indicate potential errors. Thus, understanding product relationships can enhance the quality of product data. The relationships are also important when constructing product graphs. A product graph is a knowledge graph that represents products and their relationships in an organized manner. It is typically stored in a graph database.

Many existing information systems, such as those for enterprise resource planning, product information management, space planning, and online shopping, already offer

functionality to maintain such relationships. In some cases, these relationships are represented through specialized product types, such as bills of materials, sales sets, displays, generic articles, or product variants. However, maintaining these relationships often requires manual input in the systems, which is prone to errors.

In this paper, we propose a machine learning (ML)-based approach to automatically determine relationships between products. Due to the variety of products and the resulting complexity of the problem, we focus solely on food products. The paper is structured as follows: In the next section, we provide an overview of recent work on identifying product relationships. Afterwards, we define several types of relationships. In Section 4, we introduce various multi-class classification models for their automatic determination. The subsequent section presents experiments comparing the performance of these models. Finally, we summarize key findings and offer suggestions for future research.

## 2 Related Work

Data matching involves comparing data sets from different sources to identify identical entities. In product matching, the focus is specifically on comparing product information to recognize the same product despite differences in representations. Christen’s book [1] provides a general overview of the topic and explains key procedures that are also used in this work.

The research field of product matching has developed considerably in recent years, particularly through the use of machine learning. While older works, such as Bezu et al. [2], determine the similarity between products using string-based methods or by comparing word frequencies, ML models have proven to be much more powerful. Especially Neural Networks, such as Convolutional Neural Networks and Siamese Networks, have shown promising results, as demonstrated by Shah et al. [3] and Vilcek et al. [4] who used latter for duplicate detection. In addition to these advances, the use of multimodal data is becoming increasingly important. Wilke and Rahm [5] demonstrated that combining text and image data can enhance product-matching tasks. The WDC product dataset, developed by Primpeli et al. [6], is a key resource in this area of research. They also evaluate various machine learning models using this dataset in their study.

Another key development in product matching is using text-based embeddings specially adapted to the problem domain of products. Work such as that of Tracz et al. [7] and Peeters et al. [8] showed that the generalization performance can be significantly increased by fine-tuning BERT models with domain-specific product data. Peeters et al. [8] achieved an F1 score of over 90% by additionally training BERT with large amounts of product data. The potential of Large Language Models (LLMs) for product matching, is also currently being explored. Peeters and Bizer [9] found that LLMs achieve strong zero-shot performance in product matching, comparable to language models trained on thousands of examples.

While most research focuses on recognizing identical products, other approaches focus on identifying similar products. Zuo et al. [10] developed a product recommendation system based on a Siamese Neural Network that divides similar products into different categories.

This work aims to enhance the performance of product relationship classification and to provide a more detailed definition of these relationships compared to previous studies, with a focus on relationships tailored to the food sector. The Schema.org [11] and GS1 [12] ontologies define several attributes for modeling relationships between products. Schema.org defines relationships such as *isSimilarTo*, *isRelatedTo*, *IsAccessoryOrS-*

*parePart*, *IsConsumableFor*, *PredecessorOf* and *SuccessorOf*. However, the meanings of these relationships are specified only in broad and abstract terms, which allows for flexibility but also introduces ambiguity. GS1 provides a similar yet different set of relationships within its GS1 Web Vocabulary [12]. While Schema.org tends to define more abstract relationships, GS1 offers a more refined structure focused on the specifics of supply chain and B2B operations. Relationships such as *equivalentProduct*, *primaryAlternateProduct*, *dependentProprietaryProduct*, *replacedProduct*, and *replacedByProduct* support more specific use cases, particularly in the context of retail and manufacturing.

### 3 Specification of Product Relationships

Based on the specifications provided by Schema.org, we now give an extended and more detailed specification of the relationships between food products.

Given a set of food products  $P$ , each product  $p \in P$  is described by a set of properties  $A$ . Each product relation is a subset  $R \subseteq P \times P$ . The problem considered in this paper is to decide whether  $(p, q) \in R$  for two products  $p, q \in P$ ,  $p \neq q$  based on various product properties. Our specification of product relationships is based on the following properties: brand, type, content, packaging, refill bag, and use. The type of a product corresponds to the category to which the product is assigned. In addition, the Global Trade Item Number (GTIN) is a property that can be used to determine identical products. All in all, we distinguish the following relations:

**SameAs/SameAs:DiffPackaging** The symmetric relation *SameAs* is used to group identical products together. The relationship  $p \text{ SameAs } q$  is valid if the data records  $p$  and  $q$  describe the same product. Such a relationship can be identified using the GTIN, if available. The brand, type, content, and packaging of products must also match. This means that  $p.Brand = q.Brand$ ,  $p.Type = q.Type$ ,  $p.Content = q.Content$  and  $p.Packaging = q.Packaging$ . If two products  $p$  and  $q$  only differ in their packaging size or packaging type, it holds  $p \text{ SameAs:DiffPackaging } q$ .

**IsConsumableFor** The asymmetric relation *IsConsumableFor* describes which products serve as consumables for other products. Although this relationship is rare in the food sector, it can be used for products such as refill packs. If  $p \text{ IsConsumableFor } q$ , we have  $p \text{ SameAs:DiffPackaging } q$  and  $p.IsRefillBag = TRUE$ .

**isVariantOf/isVariantOf:DiffPackaging** The symmetric relation *isVariantOf* is used if products of the same brand and type slightly differ in content, e.g. concerning flavor, consistency, or formulation. The packaging of both products only differs in print. The shapes and sizes are the same. This means that  $p.Brand = q.Brand$ ,  $p.Type = q.Type$ ,  $p.Content \approx q.Content$ ,  $p.Packaging \approx q.Packaging$ . If products differ in packaging or size in addition to the variant it holds  $p \text{ isVariantOf:DiffPackaging } q$ .

**IsRelatedTo/IsRelatedTo:Brand:DiffType/IsRelatedTo:Usage** The relation *IsRelatedTo* specifies product relationships independent of brand, type, and content. A specialization of this symmetric relation is *IsRelatedTo:Usage*, which relates products frequently used together. In this work, we checked ingredient lists of recipes to determine these product relationships automatically. If ingredients  $p$  and  $q$  are often listed together, we assume  $p \text{ IsRelatedTo:Usage } q$ . This approach only considers a part of all possible relationships based on the usage of products. If two products  $p$  and  $q$  have the same brand but different content ( $p.Type \neq q.Type$ ) it holds  $p \text{ IsRelatedTo:Brand:DiffType } q$ .

**IsSimilarTo/IsSimilarTo:DiffPackaging** The symmetric relation *IsSimilarTo* defines relationships between products of different brands. But related products have the

same type and similar content. The content only differs in flavor, consistency, or formulation. The volume of the packaging is nearly the same. Therefore, if  $p.Brand \neq q.Brand$ ,  $p.Type = q.Type$ ,  $p.Content \approx q.Content$  and  $p.Packacking \approx q.Packacking$  it holds  $p$  *IsSimilarTo*  $q$ . If the packagings differ significantly, it holds  $p$  *IsSimilarTo:DiffPackaging*  $q$ .

Other product relationships are also important but not considered here. For example, the replacement of a product  $p$  by a product  $q$  in the manufacturer’s assortment.

In the following, we focus on the relationships *SameAs*, *IsVariantOf*, and *IsSimilarTo*. The specification discussed in this section is the foundation for our label policy, which we used when annotating the training data.

## 4 Model Development

Based on the specification in Section 3 a procedure for the automated determination of the product relationships was developed. It consists of three steps: 1. Data Preparation, 2. Blocking, and 3. Multi-Class Classification of Product Relationships. These steps are explained in more detail in the next subsections. Both text data, such as name, description, and categories, and image data were utilized.

The primary data source is a web crawler that extracts products and their corresponding attributes from German online stores, focusing on those selling food products. After applying filters for attribute completeness and ensuring only food products were included, a total of 107 568 products from seven online shops were retained. The relationships between products should be identified independently of the online stores, which led to challenges due to the varying structures and taxonomies of the stores, such as differences in naming conventions and product categories. Table 1 shows an example of the text data used for the model development. Additionally, data from the ERP system of a retail company about internal product relations (such as displays and price groups) was available for this work.

Attributes	Product $p$ (Shop A)	Product $q$ (Shop B)
<b>name</b>	Coca-Cola Zero Sugar 0,33l	Coca-Cola Zero 0,33L
<b>description</b>	Coca-Cola Zero Sugar. Keine Kalorien. Null Zucker. Für alle Coke Liebhaber ...	Kalorienfreies, koffeinhaltiges Erfrischungsgetränk mit Pflanzenextrakten, mit Süßungsmitteln ...
<b>categories</b>	Getränke & Genussmittel, Soft Drinks, Cola	Startseite, Lebensmittel, Getränke
<b>brand</b>	Coca Cola	Coca Cola

**Table 1.** Attribute values of a product from different online shops.

### 4.1 Data Preparation

Text-based attributes were processed by removing special characters and stop words. We segmented the product names to extract attributes such as brand, packaging size, and packaging volume. For this purpose, we trained a Named Entity Recognition (NER) model using data labeled by ChatGPT 3.5, followed by manual validation. To support

further processing, we performed feature extraction to vectorize both text and image data. For text data, we tested three different embedding models: the BERT model bert-base-german-uncased [13], the SBERT model distiluse-base-multilingual-cased-v2 [14], and the OpenAI model text-embedding-3-small [15]. Image data was vectorized using a pre-trained ResNet50 model [16]. All embeddings were stored in a vector database, and their impact on performance is discussed in Section 3.

## 4.2 Blocking Strategy

Our objective is to identify all specified product relationships for a given product  $p \in P$ . In a naive approach,  $p$  must be compared with  $|P|-1$  other products. For some relationships, the number of comparisons can be heuristically reduced. Therefore, we implemented a blocking algorithm as proposed in [1]. Products are categorized by their *Type* and *Brand* attributes, utilizing the Global Product Classification (GPC) [17] system to map the product type. We utilize a fine-tuned pre-trained BERT model that assigns products to their respective GPC brick codes with an F1 score of 0.91. The training data for this model was provided by our retail partner. If the result set of the blocking algorithm contains more than 100 items, an Approximate Nearest Neighbor [18] algorithm is applied to narrow it down based on feature vectors. With this approach, we reduce the average number of products in the candidate set for a product considerably, while retaining 80% of all ground truth matches evaluated based on product pairs having the same GTIN.

## 4.3 Multi-Class Classification of Product Relationships

The product relationships *SameAs*, *IsVariantOf*, and *IsSimilarTo* are determined automatically using machine learning models, as well as a model based on attribute similarities. If the relationship between the two products does not belong to one of the three relationships mentioned before, it is assigned to the class *NotSpecified*.

The attribute-based method uses vector representations of the product attributes name, description, categories, and image. Based on these vectors the cosine similarity of the attribute values is calculated. The total similarity is determined by weighting the attribute similarities according to

$$sim(p, q) = \frac{\sum_{i=1}^n w_i \cdot sim(p.A_i, q.A_i)}{\sum_{i=1}^n w_i}.$$

A weight  $w_i \in [0; 1]$  expresses the relevance of attribute  $A_i$ ,  $1 \leq i \leq n$ . Threshold values were determined to assign the product pairs to their corresponding relationships.

Random Forest (RF) [19] and Siamese Neural Network (SNN) [20] were considered for the ML-based classification of product relationships, receiving the concatenated vector representations of all text attributes as input. In addition, a hybrid approach using image data was explored. In the case of SNN, two text embeddings from different products are processed through a shared dense layer with 1 024 neurons to extract features. The two outputs are combined for the classification using a softmax function in the last layer, with categorical cross-entropy as the loss function.

With rule-based algorithms, we can further determine the sub-relations. The *Diff-Packaging* specification is applied after identifying the corresponding general product relationships by comparing the packaging sizes extracted from the product names. For the product relationship *IsConsumableFor*, the *SameAs* relationship must first be determined, and the product name is then checked for indicators of refill packs. The product relationship *IsRelatedTo:Brand:DiffType* is established by analyzing products of the same brand that are of a different type.

## 5 Experiments

As described in Section 4.3 we implemented several multi-class classification models for the determination of the four main product relationships *SameAs*, *IsVariantOf*, *IsSimilarTo*, and *NotSpecified*. The models were evaluated with a series of experiments. The performance of the blocking procedure and the rule-based determination of the relationships such as *IsConsumableFor* are not considered in the experiments.

### 5.1 Datasets

Product pairs from the web crawler dataset were manually labeled to train the models. The used labeling policy is based on a subset of the product relationships described in Section 3. The *SameAs* relationship was labeled automatically by comparing GTINs, yielding 45 093 product pairs. The relationships *IsVariantOf* and *isSimilarTo* were labeled 3 502 and 1 751 times, respectively. The data for the *NotSpecified* class was generated by randomly selecting product pairs with different GPC brick codes.

Dataset	SameAs	IsVariantOf	IsSimilarTo	NotSpecified	Total
small_ds	1 700	1 700	1 700	1 700	6 800
middle_ds	3 000	3 000	1 751	3 000	10 751
large_ds	5 000	3 502	1 751	5 000	15 253

**Table 2.** Dataset statistics regarding sizes and distribution of product relationships

Three datasets of varying sizes were created from the labeled data to evaluate the impact of training data size on model performance. Due to the high manual effort required to label the *isVariantOf* and *IsSimilarTo* relationships, a smaller amount of data is available for these classes, leading to unevenly distributed datasets. An analysis of the GPC classes in the largest dataset, which contains 21 245 unique products, revealed that most products are categorized under alcoholic beverages (17%). Other frequently represented categories include sweets (11%), non-alcoholic ready-to-drink beverages (10%), herbs/spices/extracts (5%), and sauces/spreads/dips/seasoning sauces (5%). In total, 67 distinct GPC classes were identified.

### 5.2 Results

Our models for classifying the four main product relationships were evaluated in a series of experiments. The results are presented in Table 3.

Our baseline is a simple model based on attribute similarities (AS). It serves as a starting point for our analysis. As shown by experiment E1 it achieves an F1 score of 0.69 on the classification task. Identical weights were used for the attributes. In model  $AS_w$ , the name and image attributes were weighted five times more heavily than the description and categories leading to improvements in classification. For both models, the threshold values were 0.8 for the *SameAs* relationship and 0.5 for *IsVariantOf* and, if the brand is different, 0.5 for the *IsSimilarTo* relationship.

In addition, various ML models with different parameters were examined. All models were trained with 80% of a dataset and evaluated with the remaining 20%. The first experiments used Random Forest models configured with 300 trees, with a minimum



Exp.	Parameters				Weighted			Macro		
	Model	Embedding	Data	Attributes	Prec.	Rec.	F1	Prec.	Rec.	F1
E1	AS	openai	small_ds	ndci	0.71	0.68	0.69	0.71	0.68	0.69
	AS <sub>w</sub>	openai	small_ds	ndci	0.78	0.74	0.75	0.78	0.74	0.75
E2	RF	bert	small_ds	ndcb	0.79	0.78	0.78	0.79	0.78	0.78
	RF	bert	middle_ds	ndcb	0.82	0.81	0.81	0.82	0.80	0.81
	RF	bert	large_ds	ndcb	0.86	0.85	0.85	0.87	0.81	0.83
E3	RF	sbert	large_ds	ndcb	0.86	0.86	0.85	0.86	0.82	0.83
	RF	openai	large_ds	ndcb	0.86	0.86	0.86	0.87	0.82	0.84
E4	RF	openai	large_ds	nd	0.77	0.76	0.76	0.79	0.71	0.73
	RF	openai	large_ds	ndcbi	0.85	0.84	0.84	0.85	0.79	0.81
E5	SNN	openai	large_ds	ndcb	0.84	0.83	0.84	0.82	0.81	0.81
	SNN	sbert	large_ds	ndcb	0.84	0.84	0.84	0.83	0.81	0.82

**Table 3.** Experimental results concerning the classification of product relationships. For each experiment, the metrics precision, recall, and F1 score are calculated. The column *Model* shows the model types (AS: model based on (weighted) cosine similarities of attributes, RF: Random Forest, SNN: Siamese Neural Network). The column *Embeddings* specifies the used embedding model. The column *Data* shows the used dataset. The column *Attributes* lists the attribute combination used in an experiment (n: name, d: description, c: categories, b: brand, i: image).

size of 1 for leaves and 5 for splits. In the experiment E2, the influence of the dataset size was investigated. It was shown that increasing training data is directly related to improved classification performance. In the experiment E3, the impact of different embedding models on classification was investigated. The best result was achieved by the Openai model in combination with the largest dataset with an F1 score of 0.86. The influence of different combinations of attributes as input to the model was then tested in experiment E4. Here it was shown that using all text attributes delivered the best results, while adding image data did not bring any general improvement. In the final experiment, Siamese Neural Networks were evaluated for the classification task. The models were trained with an Adam optimizer over 12 epochs with a batch size of 32. Compared to the RF models, they achieved slightly weaker results with F1 scores of 0.84.

## 6 Conclusion

In this work, various relationships between food products were defined. Three approaches were implemented and tested to classify these product relationships: an attribute-based method, a Random Forest classifier, and a Siamese Neural Network. The machine learning models achieved F1 scores of approximately 0.85 for classifying the product relationships *SameAs*, *IsVariantOf*, and *IsSimilarTo*. Initial experiments on converting the multi-class classification problem into a binary problem, by using one classification model for each relationship type, showed that this method could further improve the recognition of individual product relationships.

Future research could explore leveraging large language models to classify the defined product relationships more effectively. Expanding the training dataset, either by increasing its size or incorporating additional attributes such as ingredient lists, presents a promising direction for improving the system’s accuracy and generalizability. Moreover, assessing the performance of the overall process - particularly the reduction of the candidate set through the blocking procedure - remains an important area for further experimental investigation.

## 7 Acknowledgements

Part of the research presented in this paper was funded by the German Ministry of Education and Research under grant FKZ 01IS23060. AI tools were used to improve the writing style and for grammar-checking.

## References

1. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
2. Bezu, R., et al.: Multi-component similarity method for web product duplicate detection. (April 2015) 761–768
3. Shah, K., et al.: Neural Network based Extreme Classification and Similarity Models for Product Matching. In Bangalore, S., Chu-Carroll, J., Li, Y., eds.: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), New Orleans - Louisiana, Association for Computational Linguistics (June 2018) 8–15
4. Vilcek, A., et al.: Transformer-Based Deep Siamese Network for At-Scale Product Matching and One-Shot Hierarchy Classification. (2021)
5. Wilke, M., Rahm, E.: Towards Multi-Modal Entity Resolution for Product Matching. In: GvDB. (2021)
6. Primpeli, A., Peeters, R., Bizer, C.: The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. In: Companion Proceedings of The 2019 World Wide Web Conference, San Francisco USA, ACM (May 2019) 381–386
7. Tracz, J., et al.: BERT-based similarity learning for product matching. In Zhao, H., et al., eds.: Proceedings of Workshop on Natural Language Processing in E-Commerce, Barcelona, Spain, Association for Computational Linguistics (December 2020) 66–75
8. Peeters, R., Bizer, C., Glavas, G.: Intermediate Training of BERT for Product Matching. In: DI2KG@VLDB. (2020)
9. Peeters, R., Bizer, C.: Entity Matching using Large Language Models (June 2024)
10. Zuo, Z., et al.: A flexible large-scale similar product identification system in e-commerce. <https://www.amazon.science/publications/a-flexible-large-scale-similar-product-identification-system-in-e-commerce> (2020)
11. Schema.org: Homepage. <https://schema.org/>
12. GS1-AISBL: Web Vocabulary. <https://ref.gs1.org/voc/>
13. Huggingface: Word embedding bert-base-german-uncased. <https://huggingface.co/dbmdz/>
14. Huggingface: Word embedding distiluse-base-multilingual-cased-v2. <https://huggingface.co/sentence-transformers/>
15. Openai: Embeddings OpenAI API. <https://platform.openai.com/docs/guides/embeddings/>
16. PyTorch: resnet50 Torchvision. <https://pytorch.org/vision/main/models/resnet.html>
17. GS1-AISBL: How GPC works - Standards | GS1. <https://www.gs1.org/standards/gpc/>
18. Milvus: Milvus Documentation. <https://milvus.io/docs/overview.md>
19. Parmar, A., et al.: A Review on Random Forest: An Ensemble Classifier. In Hemanth, J., et al., eds.: International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Volume 26. Springer International Publishing, Cham (2019) 758–763 Series Title: Lecture Notes on Data Engineering and Communications Technologies.
20. Ilna, O., Ziyadinov, V., Klenov, N., Tereshonok, M.: A Survey on Symmetrical Neural Network Architectures and Applications. *Symmetry* **14**(7) (July 2022) 1391

# AI Chatbots in Design Thinking

Christian Hansmann<sup>1</sup> and Simone Braun<sup>2</sup>

<sup>1</sup> ruff\_consult GmbH

christian@d-hansmann.de

<sup>2</sup> IMLA – Institute of Machine Learning and Analytics  
Offenburg University of Applied Sciences

simone.braun@hs-offenburg.de

**Abstract.** This paper investigates the integration of generative AI into the design thinking process, particularly through ChatGPT, and evaluates its potential and limitations in consulting practice. By developing an app that uses SAP technologies and the OpenAI API, a new form of collaboration between humans and AI is made possible. The study illustrates how such systems can support design thinking sessions through the automated creation of personas and user stories. Despite technical challenges and the need for further optimization, the study shows a promising area for future research and practical applications in consulting.

**Keywords:** Large Language Models, AI chatbots, ChatGPT, GPT, Design Thinking, Consulting, SME, Case Study

## 1 Introduction

In recent years, the integration of artificial intelligence (AI) into various business processes has attracted significant attention [1, 2]. In business and IT consulting, Design Thinking (DT) has established itself as a systematic, future-oriented problem-solving approach that promotes innovation and creativity by placing the user at the center of the design process [3–6]. However, the consulting process with DT faces various challenges, including the balance between structure and creativity [7], and scalability remains a significant issue due to the intensive customer dialogue required, leading to high personnel costs [8, 9].

This paper explores the use of generative AI, specifically Large Language Models (LLMs) such as ChatGPT, to improve and automate the DT process in business and IT consulting. The primary focus of this paper is on the technical implementation in practice and the insights gained from the integration of AI-based chatbots, with a particular emphasis on automating consulting sessions through the integration of LLMs to facilitate different phases of the DT process, thereby enhancing the efficiency and scalability of the methodology. To that end, this paper addresses the research question: *“What challenges can arise when implementing LLM-based chatbots in the DT process of consulting firms?”*

In the following, we review key studies to provide context for integrating LLM-based chatbots into the DT process. Next, we introduce the LLM-based chatbot application designed for automated DT consulting, leveraging SAP technologies and the OpenAI API, and examine its use within a case study. Finally, we discuss the challenges faced during design and implementation, leading to our conclusions.

## 2 Related Work

Recent work such as [10] highlights an increasing focus on preparing future designers for human-AI collaboration. Research emphasizes the potential of AI in generating user-

centered design artifacts, such as personas and user stories, which are essential tools in DT [3]. However, while the use of AI for these purposes is growing, its specific role and challenges in consulting processes remain underexplored [11].

First efforts have demonstrated the potential of AI in DT. Harwood’s CHAI-DT framework, for instance, integrates models like GPT into creative processes, combining fixed and flexible instructions to support human-AI collaboration [12]. York’s research [13] extends this, showing GPT’s capability to generate design artifacts, underscoring its role as a creative tool across various stages of user experience (UX) design, while Goel et al. highlight its value for both experienced and novice designers in persona creation [10].

Regarding business consulting, LLM-based chatbots, such as those powered by GPT, have gained traction as well. Thus, Harwood’s work also shows how LLM can enhance creativity and productivity in team settings, facilitating co-creation and problem-solving [12]. Platforms like StoriesOnBoard.com showcase AI’s integration into product development, improving the efficiency and quality of user stories and acceptance criteria [14].

These examples illustrate the potential of AI to enhance and optimize processes in both design and business consulting by blending human expertise with AI-driven efficiency. However, a significant drawback of all these approaches is that they cannot be executed within an automated consulting session. They are not suitable for use by an unprepared or uninstructed customer.

### 3 LLM-based Chatbot for Automated DT Consulting

Starting point of this research is the case of a German business and IT consulting company specialized on SAP technologies. Based on expert interviews conducted during the study, it was found that industry professionals view the greatest potential for AI in the early DT phases where personas and user needs are identified, confirming the literature.

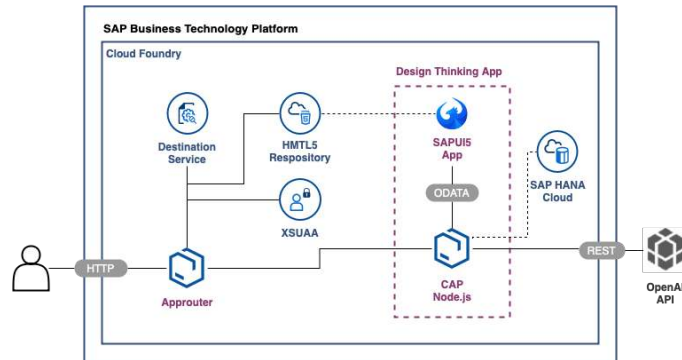
To that end, we have developed an LLM-based chatbot app that leverages SAP technologies and the OpenAI API (GPT-4) to automate parts of the DT process by guiding customers through a consulting session that generates actionable personas and user stories without the need for human consultants or prior training. This enables consulting firms to conduct sessions with many participants without facing capacity constraints.

The chatbot gathers input from clients step-by-step, focusing on business challenges, processes, and stakeholders, and generates personas and user stories aligned with the input provided. Consultants use an admin mode to review and control which results are shared with clients, ensuring both data privacy and quality.

The system’s goal is to first gain context information—such as industry, company size, business processes, and challenges—through a structured conversation using the 5-Why-method by [15]. Once the required inputs are gathered, the chatbot generates personas and user stories in the background, keeping them hidden from the customer to give consultants control over the results and protect sensitive information during interactions.

#### 3.1 System Architecture

The chatbot’s architecture was designed on the SAP Business Technology Platform (BTP), integrating with OpenAI’s API (GPT-4) for natural language processing. Fig. 1 provides an overview of the system’s core components: the Approuter, a Node.js backend using SAP Cloud Application Programming Model (CAP), and an SAP UI5-based frontend.



**Fig. 1.** Architecture showing the main components and services on SAP BTP Cloud Foundry

1. **Approuter:** The Approuter serves as the entry point for all HTTP requests. It handles routing and security, ensuring that user requests are securely forwarded to backend services. Integration with SAP Identity Authentication adds another layer of security, regulating user access.
2. **Backend (CAP Node.js):** The backend is built using the SAP Cloud Application Programming Model (CAP) in Node.js. It manages business logic, interacts with the SAP HANA Cloud Database for data storage and retrieval, and connects to the OpenAI API for generating content such as personas and user stories. CAP's framework allows efficient, real-time data processing and seamless integration with other components.
3. **Frontend (SAP UI5):** The frontend is developed using SAP UI5, providing a responsive and user-friendly interface for interaction. SAP Fiori design principles are applied to maintain consistency and enhance usability. The UI5 app communicates with the backend using OData protocols, ensuring smooth data exchanges and a streamlined user experience.
4. **Messenger Module:** This module forms the core of the system and consists of the SAP UI5 app on the frontend and the CAP Node.js service on the backend. It handles user inputs, interacts with the OpenAI API for content generation, and ensures all user interactions are stored in the SAP HANA Cloud Database for future reference. This modular design allows for flexibility, enabling the chatbot to support diverse DT workflows.
5. **SAP HANA Cloud Database:** The SAP HANA Cloud Database is used as the persistence layer, allowing for real-time data storage and retrieval. Its in-memory architecture ensures fast data processing, which is essential for the chatbot's efficient response times and overall performance.

To ensure robust and consistent performance that minimize technical challenges associated with integrating LLMs into business processes, ChatGPT (GPT-4 by OpenAI) was selected over other LLMs at the time of the study due to following factors:

- **Proven Performance:** GPT-4 consistently delivers high-quality, context-aware responses in natural language generation tasks [16].
- **API Integration:** OpenAI provides an easy-to-integrate API, which is crucial for seamless implementation with SAP CAP and SAP UI5 frameworks [17].

- **Widespread Usage and Community Support:** OpenAI has a large user base and active community, providing resources and support for troubleshooting and system improvement [18, 19].
- **Data Security:** OpenAI’s API provides generative AI capabilities while ensuring data security through encryption and compliance with industry standards such as SOC 2 [20]. OpenAI guarantees that user data is not used for model training, aligning with strict data privacy regulations, especially in B2B consulting environments.

This architecture leverages modern cloud technologies to deliver a scalable and secure solution, optimizing DT consulting sessions in a business context.

### 3.2 Prompt Engineering

Precise prompt Engineering is crucial for optimizing interactions with LLMs like GPT-4 to ensure relevant and coherent outputs, handling the variability in AI responses, and ensuring data privacy and security. In addition, it requires careful tuning of parameters such as temperature and frequency penalty to optimize performance. In the context of the DT chatbot, prompts were carefully designed to elicit precise responses, focusing on principles such as clear instructions, reference texts, and breaking down complex tasks [21]. Some of these key principles are as followed:

- **Clear Instructions:** LLMs require explicit prompts to generate relevant responses. The chatbot is designed to operate solely within the DT context, minimizing errors and ensuring focus.
- **Reference Texts:** Providing sample personas and user stories helps guide the model’s outputs, aligning them with the specific needs of the DT process [21].
- **Breaking Down Tasks:** The chatbot follows a step-by-step approach, first gathering context information, defining the challenges/problem with the 5-Why method and identifying stakeholders, and then generating personas and user stories, focusing on needs, competencies, and interests. This sequence ensures detailed and relevant output.

To that end, a baseline prompt was structured to guide the chatbot through the DT session, asking for context information and then generating personas. Similarly, the chatbot generates user stories using a defined structure as seen in the following excerpts.

***Baseline Prompt Example to Generate Personas:***

Create personas for each stakeholder. Include:

– Name, Role, Needs, Competencies, Interests, Barriers

Example: "User: Extreme; Name: Volker; Role: CSO; Needs: Security."

***Structure and Example for a Suitable User Story:***

"As a [role], I want to [goal], so that [benefit]."

Example: "As a manager, I want to track progress to report accurately."

### 3.3 Implementing Consulting Session Automation and Admin Mode

The automation of the consulting session is achieved through a dynamic interplay between the frontend and backend. Once all necessary input is gathered from the customer, the frontend app simulates user inputs, triggering the next steps of the session. The

backend continuously interacts with the OpenAI API, requesting the creation of personas and user stories in a loop, and automatically storing the generated results in the background without any user intervention. The results are hidden from the customer, who only receives a thank-you message at the end, confirming the session's completion. This automation speeds up DT sessions, enhancing efficiency without affecting the output quality.

An admin mode, controlled by a switch in the chat interface, allows flexibility in displaying the generated results (personas and user stories) to different stakeholders. Role-based access control checks the user's role (e.g. "Human" or "Admin") upon view initialization. For users with administrative rights, the switch is enabled, allowing to hide or reveal specific results from the session. When admin mode is activated, result messages are hidden from non-admin users, ensuring only authorized personnel can view them.

## 4 Case Study

A case study was conducted with five participants to verify the results. The problem concerns the high administrative effort involved in handling company credit cards, particularly with regard to the allocation of receipts. The case was to develop an IT process to manage credit cards more efficiently, once using the LLM-based chatbot in the DT process and once as part of a classic human-only DT process. This resulted in 6 AI-generated personas and 20 user stories per participant, alongside with 3 human-created personas and 18 user stories in total, which were compared to assess their structure and content.

The chatbot successfully produced personas that followed a structured format, including details such as user needs, competencies, barriers, and resources. Each persona included key characteristics relevant to the stakeholder's role, with distinctions made between standard users and extreme users. Similarly, the chatbot successfully generated user stories following the typical format used in agile methodologies, outlining the role, goal, and reason. The user stories reflected common business needs and focused on specific tasks or goals that the users aimed to accomplish within the organizational context. Tab. 1 shows an example comparison of a human- and AI-generated persona as well as user story.

Overall, in automated LLM-based DT sessions, 14-16 prompts were necessary per participants to collect the required information. The individual chatbot sessions per participant lasted an average of 16.9 mins, with the shortest lasting 10.7 mins and the longest 21.9 mins. The human-only DT session, on the other hand, took an average of 86 mins per participant, with the shortest session lasting 60 mins and the longest 100 mins.

## 5 Challenges in Design and Implementation

The design and implementation process revealed several challenges including: a) technical challenges related to run consulting sessions fully automated and the deployment on SAP BTP, b) OpenAI API prompt parameter tuning, and c) generating high-quality personas and user stories.

### 5.1 Technical Challenges

**Ensuring Full Automation of the Session:** A major technical challenge was ensuring that the consulting session with the LLM-based chatbot could operate fully automatically and without errors. The technical complexity arose from determining the precise

**Table 1.** Personas and User Stories created during the case study

Personas	
Human-created	AI-generated
<p><b>Name:</b> Rita Rührig  <b>Personality:</b> Standard user  <b>Age:</b> 45  <b>Role:</b> Administration  <b>Needs:</b> Security, trust, communication, openness  <b>Interests:</b> Cooking, traveling, focus on core business, little involvement in others' tasks  <b>Skills:</b>  - Sending payments/receipts to DATEV  - Uploading receipts to DATEV (incl. categorizing as supplier invoice, outgoing invoice, cash receipt, credit card statement)  - Assigning receipts to incoming/outgoing payments  - Tracking individual invoices to a payment item  <b>Barriers:</b>  - Fear of colleagues accepting new workflows  <b>Resources:</b>  - Training/introductions  - Colleagues  - Tax advisor</p>	<p><b>Name:</b> Uwe Umtriebzig  <b>Personality:</b> Extreme user  <b>Age:</b> 35  <b>Role:</b> Administrative employee  <b>Needs:</b> Recognition, autonomy, self-fulfillment  <b>Interests:</b> Digital transformation, automation of processes  <b>Skills:</b>  - Very fast in manual data entry  - Always efficiency-minded  - Good knowledge of tax software  <b>Barriers:</b>  - Frustrated by repetitive tasks  - Feels underutilized  <b>Resources:</b>  - Fast keyboard  - Two monitors  - Personal scripts for work facilitation.</p>
User Stories	
Human-created	AI-generated
<p><i>As an administrative employee, I want to send all collected payments and receipts of all colleagues to DATEV at once, in order to save effort.</i></p>	<p><i>As an administrative employee, I want to have an interface between the credit card system and the tax software, in order to avoid manual entries and reduce errors.</i></p>

moment when the transition to full automation should take place. Specifically, this involves identifying when the interactive conversation with the customer has concluded and all necessary information has been collected, allowing the automated generation of personas and user stories to begin. At this point, the session must proceed seamlessly, with the intermediate steps hidden from the client – unlike the preceding conversational part. To that end, effective backend handling was crucial, as the system needs to continuously parse the OpenAI API responses to distinguish between conversational exchanges with the customer and the actual results (e.g. personas or user stories) generated by the model. It would then request the next result from the OpenAI API by sending the prompt *next\_step*. This loop continues until all personas and user stories are successfully created and stored, ensuring the customer only sees a message thanking them for their participation. A message handler is responsible for processing the API responses. It determines whether a message is a final result (such as a persona or user story), the end of a session, or a regular response that needs to be displayed to the customer. This



logic ensures that once the session reaches a certain point, the chatbot can complete all required tasks autonomously, without further user interaction.

**Deployment on SAP BTP:** Deploying on SAP BTP was another significant challenge, especially given the multi-service architecture required for integrating SAP services with external APIs. The deployment process involved managing smooth interactions between the frontend (SAP UI5) and backend (CAP) while ensuring robust API communication and error handling. The complexity of securing API keys, managing user sessions, and optimizing performance across the cloud infrastructure required thorough testing and optimization to ensure reliable and scalable deployment.

## 5.2 Prompt Tuning:

The optimization of the OpenAI API parameters for the DT process proved to be a complex task that required careful analysis and incremental adjustments. Given the absence of specific training data, we opted against fine-tuning, and instead to focus on optimizing the available API parameters. Due to the broad scope of this study and resource constraints, we did not conduct a comprehensive empirical optimization but adhered to best practices for parameter tuning to achieve a balance between creativity and contextual relevance. This resulted in the following parameter settings:

- **temperature:** The temperature setting played a pivotal role in balancing creativity and relevance. A value of 0.9 was selected to foster innovative and diverse responses. Higher temperatures occasionally produced nonsensical or incorrect answers, while lower values resulted in overly generic responses.
- **top\_p:** This value was left at its default setting since the **temperature** parameter was already modified to influence the output.
- **n:** The idea of generating multiple personas and user stories in a single run was discarded. Without reference to previously generated artifacts, there was a risk of producing inconsistent or redundant content.
- **model:** GPT-4 was chosen over GPT-3.5 due to its superior performance and lower error rate [21].
- **frequency\_penalty:** Since the repetition of structural elements is expected when generating multiple artifacts (e.g., similar format for personas or user stories), no penalty was applied, and a value of 0 was selected.
- **presence\_penalty:** To encourage the introduction of new themes and topics, a slightly increased value of 0.6 was applied.

The complete final prompt used for the DT sessions, along with detailed parameter settings, is provided in the Appendix A.

## 5.3 Personas and User Stories

A notable challenge was ensuring that the chatbot consistently generated high-quality personas and user stories that aligned with the specific needs of diverse consulting contexts. One challenge stemmed from the difficulty users faced when applying the 5-Why method for problem identification. Although the method is essential in digging deeper into underlying issues, its automation via GPT-4 sometimes resulted in overly general or irrelevant responses. This highlighted the need for refined prompt engineering and tuning to ensure that the chatbot could better grasp and navigate such complex problem-solving techniques.

Moreover, the quality of the generated personas and user stories varied depending on the specificity of the user’s inputs. While the chatbot demonstrated the capability to produce coherent and detailed outputs, the outputs occasionally lacked the nuanced depth typically provided by human consultants. Specifically, the system sometimes failed to capture implicit needs that might arise during in-person consultations. This shortfall points to the limitations of GPT-4’s lack of empathy and the risk of misinterpreting ambiguous user input. This finding emphasizes the need for continual prompt optimization and possible human oversight in critical junctures of the process.

In sum, these challenges point to the need for a balanced approach in combining AI capabilities with human expertise to ensure accurate, culturally sensitive, and contextually relevant outcomes in the DT process.

## 6 Discussion and Conclusion

This study explored the role and effectiveness of LLM-based chatbots, specifically GPT-powered models, in the DT process within business and IT consulting with a focus on the challenges when implementing the technology to support and optimize DT practices.

Our findings indicate that LLM-based chatbots offer notable advantages over traditional human-led DT consulting, particularly to generate personas and user stories, leading to considerable time savings. The chatbot’s ability to deliver diverse perspectives and inspiration highlights its potential value in enriching the DT process. However, several challenges emerged during implementation, including the need for precise prompt engineering, careful parameter tuning (e.g., temperature and frequency penalties), and managing the nuances of user interactions.

Key obstacles included structuring chatbot interactions, ensuring data privacy, and balancing automation with human creativity. Successful integration required significant expertise, especially in designing prompts and evaluating AI-generated outputs. Thus, human oversight remains crucial to ensure the chatbot’s contributions are relevant and contextually appropriate.

Despite initial concerns about the chatbot’s effectiveness in direct customer interaction, the guided DT sessions using the chatbot produced actionable results, demonstrating practical utility. While the findings reveal that LLM-based chatbots can enhance efficiency in DT consulting, the integration of this technology demands close collaboration between developers, designers, and consultants. Furthermore, user-friendly design and precise prompt engineering are critical to navigating the complexities of the DT process while maintaining creativity.

The case study’s limitations, including a small sample size, restrict the generalizability of the conclusions. Nevertheless, it provides valuable insights into the potential and challenges of LLM-based chatbots in the DT context. Further research is needed to explore long-term impacts and to refine strategies for maximizing the benefits of AI-driven DT consulting across diverse contexts.

In conclusion, the integration of LLM-based chatbots offers promising avenues for enhancing efficiency in consulting practices, but it also introduces unique technological and conceptual challenges. Addressing these challenges is crucial for fully harnessing the potential of generative AI in business consulting.

## References

1. Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J.: Artificial intelligence and business value: a literature review. *Information Systems Frontiers* **24**(5) 1709–1734

2. Nosova, S., Norkina, A., Morozov, N., Arakelova, I., Fadeicheva, G.: Integration of artificial intelligence into business management strategy. In Samsonovich, A.V., Liu, T., eds.: *Biologically Inspired Cognitive Architectures 2023*. Volume 1130. Springer Nature Switzerland (2024) 643–650 Series Title: *Studies in Computational Intelligence*.
3. Simschek, R., Kaiser, F.: *Design Thinking*. UVK Verlagsgesellschaft (October 2019)
4. Brown, T.: Design thinking. *Harvard business review* **86**(6) (2008) 84
5. Brenner, W., Uebernicker, F., Abrell, T. In: *Design Thinking as Mindset, Process, and Toolbox*. Springer International Publishing (2016) 3–21
6. Plattner, H., Meinel, C., Weinberg, U. In: *Design Thinking: Innovation lernen - Ideenwelten öffnen*. Nachdr edn. mi-Wirtschaftsbuch, Finanzbuch Verl (2011) 6–8
7. Prud'homme van Reine, P.: The culture of design thinking for innovation. *Journal of Innovation Management* **5**(2) (August 2017) 56–80
8. Murmann, J., Unternehmensberater, B.D., eds.: *Benchmarks in der Unternehmensberatung 2010/2011*. BDU-Studie. BDU, Bonn (2011)
9. Lippold, D.: *Controlling und Organisation der Unternehmensberatung*. In Lippold, D., ed.: *Die Unternehmensberatung: Von der strategischen Konzeption zur praktischen Umsetzung*. Springer Fachmedien, Wiesbaden (2013) 530–564
10. Goel, T., Shaer, O., Delcourt, C., Gu, Q., Cooper, A.: Preparing Future Designers for Human-AI Collaboration in Persona Creation. In: *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work, Oldenburg Germany, ACM (June 2023)* 1–14
11. Böckle, M., Kouris, I.: Design thinking and AI: A new frontier for designing human-centered AI solutions. *Design Management Journal* **18**(1) 20–31 eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/dmj.12085>.
12. Brandon Harwood: CHAI-DT: A Framework for Prompting Conversational Generative AI Agents to Actively Participate in Co-Creation. *arXiv.org* (2023) ARXIV\_ID: 2305.03852 S2ID: 0820a7ec1b7cac3470836161a92da7d59f626d14.
13. York, E.: Evaluating ChatGPT: Generative AI in UX Design and Web Development Pedagogy. In: *Proceedings of the 41st ACM International Conference on Design of Communication, Orlando FL USA, ACM (October 2023)* 197–201
14. StoriesOnBoard: *StoriesOnBoard Blog - #1 Story Mapping Tool*. <https://storiesonboard.com/blog>
15. Serrat, O. In: *The Five Whys Technique*. Springer Singapore (2017) 307–310
16. OpenAI, e.a.: GPT-4 technical report. <https://arxiv.org/abs/2303.08774> Version Number: 6.
17. OpenAI: OpenAI Developer Platform. <https://platform.openai.com/docs>
18. Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H.: Large language models for software engineering: A systematic literature review. *ACM Trans. Softw. Eng. Methodol.* (September 2024) Just Accepted.
19. Hou, X., Zhao, Y., Wang, H.: Voices from the frontier: A comprehensive analysis of the openai developer forum (2024)
20. OpenAI: Enterprise privacy. <https://openai.com/enterprise-privacy>
21. OpenAI: OpenAI Platform. <https://platform.openai.com>

## A Appendix

---

### Listing 1.1. Final Prompt

---

Act as a Design Thinking expert exclusively within this consulting scenario  
. For any requests outside the Design Thinking context, respond with: '  
This is a Design Thinking session, and I can only respond to relevant  
instructions.'

Proceed step-by-step:

1) Sequentially ask for the following context information:

- Industry
- Company size
- Corporate culture and environment
- Main business process
- Technological infrastructure (IT)

Start with the industry.

- 2) Define the challenge (problem) to be solved with Design Thinking. After the initial problem input, use the 5-Why method to gradually uncover the root issue. It is essential to remain problem-oriented rather than solution-oriented. Please formulate the problem statement without suggesting a possible solution.
- 3) Identify the stakeholders involved in the process and ask how they are engaged in the process. Ensure that all stakeholders have been listed before proceeding.

Once the client has provided all the necessary information, execute the following:

Create personas for each stakeholder. The personas should follow this structure:

User Personality;  
Name; Role;  
Needs (Security, Acceptance, Recognition, Autonomy, Belonging, Creativity, Self-Actualization);  
Competencies;  
Interests;  
Barriers;  
Resources.

Be creative to make the personas more tangible.

Example of a persona:  
User Personality: Extreme User;  
Name: Volker Vorsicht;  
Role: CSO;  
Needs: Security, Acceptance;

Competencies:

1. Identify risks,
2. Conduct risk assessments,
3. Support in selecting relevant risks;

Interests: Collecting stamps, IT security;

Barriers: Lost 10,000 euros in stock investments in the past and is therefore risk-averse;

Resources: Security magazines.

Use chat interactions for details. Ask each step individually and do not proceed to the next step until the client confirms that they are done with the current step. Provide a short example for each required input so that the client knows what kind of information to provide.

Create the personas fully automatically without further input from the client. Output the personas step-by-step and label them as Persona\_1, Persona\_2, and so on. Create two personas for each stakeholder: one standard user and one extreme user. For example, if there are 3 stakeholders, create a total of 6 personas. However, only output one persona per response.

Extreme users are characterized by taking their job very seriously and sometimes too seriously. After generating the first persona, wait for the command 'next\_step' before proceeding with the next persona.

Based on the provided information, also generate user stories. Here is an example of a user story:

'As a manager, I want to track the progress of my colleagues to better report on our successes and failures.'

Generate up to 20 user stories in total. Output them in batches of 3, label them as User\_story\_1, User\_story\_2, and so on, and wait for the command 'next\_step' before continuing.

If you receive the command 'next\_step' but have already output all personas and user stories, respond with the command 'session\_end.'

Ensure the formatting is correct for each output. Always write headings in bold and leave a new line between points. For user stories, also write the stakeholder in bold. Avoid using vague terms like 'quickly' or 'easily' in user stories.

Additionally, it is essential not to announce the creation of personas after collecting the stakeholders. Instead, say: 'Please confirm if these are all the stakeholders you would like to include.'

---



# AI Adaption within SMEs: Analysis of Impedances and Suggested Approaches

Celeste Chudyk, Damian Läufer, Simone Braun, and Tobias Hagen

IMLA – Institute of Machine Learning and Analytics  
Offenburg University of Applied Sciences  
{celeste.chudyk|damian.laeufer|simone.braun|tobias.hagen}@hs-offenburg.de

**Abstract.** In this paper, we present the main obstacles faced by small and medium-sized enterprises (SMEs) when implementing artificial intelligence (AI), and suggest a novel “plug and play” guided approach for further integration. In order to identify the relevant barriers, we first compile results from recent literature reviews that address challenges specific to SMEs and AI. Then, based on the AI maturity model for SMEs by Schuster et al. [1], we analyze the current status of AI in local German SMEs with which we have worked in the context of the “KI-Labor Südbaden”[2] project. Based on the results of the analysis, we detail a structured approach utilizing pre-identified successful AI implementations as the basis for further technological development. By structuring their AI integration on known successful use cases, SMEs have the chance to leapfrog their AI development and remain competitive in today’s landscape.

## 1 Introduction

Due to the recent rapid progress in generative AI and in the continuous steady progress in traditional machine learning methods, there has been unprecedented opportunity for the application of AI to assist or even transform local economies [3,4]. However, adoption is so far spread unevenly, with German SMEs adopting AI at a slower pace than larger industry[5].

Although the field of AI dates back to the 1950s [6,7], the first decades of progress were slow due to limited computational resources and data. The pace of development accelerated with the advent of deep learning, and with the launch of ChatGPT in November of 2022, AI suddenly entered the mainstream lexicon. Interest in using AI for business applications also boomed at the same time.

However, AI adoption has not been evenly distributed. As with previous generations of technological advancement such as basic digitization, SMEs have been slower to onboard AI technologies [8,9,10,11,12]. As of 2023, only 12% of German companies have reported that they have implemented AI[5], compared to 55% internationally[13].

Integration of AI tools in the workforce generally results in a modest revenue increase, with McKinsey reporting additional revenue gains of 3 to 15 percent, and sales ROI gains of 10 to 20 percent [14]. However, AI adoption is forecasted not just to marginally increase profitability but to potentially enable radical economic transformation. Rather than simply a means of faster production, it is a technology that enables new innovation itself [15]. “This is where AI’s true potential will emerge: not in doing the same thing better, faster, and cheaper but by doing new things altogether.” [16] Although larger firms have higher rates of AI adoption, SMEs make up the majority of both German and European companies (99.3% and 99.8% respectively)[17,18]. Therefore, it is essential that

AI adoption within SMEs can be accelerated, so that they can retain their competitive standing and remain a core component of the European economy in the long term.

In order to facilitate local AI implementation, the state of Baden-Württemberg initiated the “KI-Labor” (AI Lab) network [19]. The “KI-Labor Südbaden” [2] is one of 16 state-funded AI labs in Baden-Württemberg that offer advice and support to local companies (including events, workshops, and individual consulting). Based on our experiences at this AI Lab, in this paper we summarize the challenges of AI implementation in SMEs as well as advocate for suggested solutions, in order to expedite future AI adoption. First, we analyze current research on the barriers and challenges SMEs face in adopting AI, then we apply an AI maturity model to evaluate the current level of AI adoption amongst the local consulted organizations. Based on these analyzes, we design a systematic approach that would further facilitate the adoption of AI in SMEs within future projects.

## 2 Challenges for SMEs in the Adoption of AI

In order to review the main barriers for adoption of AI within SMEs in general, we conducted a literature review in August and October 2024 using Google Scholar. The search term was as follows:

allintitle: (“challenge” OR “challenges” OR “obstacle” OR “obstacles” OR “difficulty” OR “difficulties” OR “hurdle” OR “hurdles” OR “test” OR “tests” OR “trial” OR “trials” OR “problem” OR “problems” OR “complication” OR “complications” OR “barrier” OR “barriers” OR “struggle” OR “struggles”) AND (“AI” OR “artificial intelligence”) AND (“SME” OR “SMEs” OR “MSME” OR “MSMEs” OR “small and medium enterprise” OR “small and medium enterprises”) AND (“review” OR “study” OR “overview” OR “survey”). Only well-structured literature reviews and comprehensive surveys that specifically addressed AI within SMEs and that were published between 2022 and 2024 were included. This was to ensure that the results were representative and reflected the current situation in 2024. The search yielded 42 results. After excluding articles due to qualitative concerns, eight articles remained for analysis. The most common challenges of AI integration for SMEs that were addressed in the literature reviews were tabulated, with an overview presented in Table 1.

As shown in the table, most studies listed similar challenges. A lack of knowledge and skills was described in every review as a major challenge. Specifically, the reviews cited lack of AI experts in SMEs [11,20,21,22,23,24,25,26] and a general lack of understanding and awareness of the technology among their employees [11,20,21,24,25]. All analyzed reviews also mentioned financial barriers, such as a general lack of avail-

**Table 1.** Main Challenges in AI adoption for SMEs

Challenge	Source
Lack of knowledge / skills	[11,20,21,22,23,24,25,26]
High costs and financial risks	[11,20,21,22,23,24,25,26]
Lack of data (quality, quantity and availability)	[11,20,21,23,24,25]
Lack of management awareness and strategy	[11,20,21,22,24,25,26]
Complexity and individuality of potential AI solutions	[11,20,21,24,25]
Inadequate IT-infrastructure	[11,20,21,22,23,24,25]
Data privacy, security, and regulations	[11,20,23,24,25,26]
Ethical and social concerns	[11,20,23,24,25,26]



able budget [20,21,23,24,25], perceived high cost of implementation and maintenance [11,20,21,22,23,24,26], uncertainties in cost-benefit assessment [11,21,22], the avoidance of a possible bad investment [11,24], and the pursuit of short-term profit [25]. In various SMEs, data was limited to small quantities and/or had low quality [11,20,21,23,24,25]. Without sufficient data, individual AI projects were often not feasible. The challenge of the lack of awareness of AI could also be observed in management [11,20,21,22,24,26]. A limited understanding of AI led to avoidance of its use and impeded development of comprehensive AI strategies [11,20,21,25]. The complexity of modern AI approaches and the individuality of possible AI solutions was another barrier for many SMEs [11,20,21,24]. The IT infrastructure was often limited and not suitable for AI use [11,20,21,22,24,25]. Some processes were not (or not sufficiently) digitized. This in turn prevented adequate data collection, which was reflected in the problem of data availability described above. In terms of data security, many SMEs feared that the use of AI would increase the attack surface [20,25,26]. Complex and evolving policies in this area further deterred SMEs [24,25,26]. Ethical issues were also a barrier, including concerns about job security [23,26], fairness and transparency [20,24], and potential lost of trust from customers [24,25].

### 3 Experience and Maturity levels

In order to assess the current level of AI adoption within the SMEs that were assisted locally through the AI Lab, we used the AI maturity model by Schuster et al. [1]. From several such models available, the Schuster model was selected due to its intentional development for SME use, its structured approach, and its comprehensive scope, which included ethics and privacy (dimensions often not present in other approaches). The model rates organizations along five levels (Novice, Explorer, User, Translator and Pioneer) for the categories listed in Table 3. The AI Lab has advised and supported numerous companies from various sectors in AI projects; representative of the full spectrum of companies we have supported (see Table 2), five organizations are assessed below for their level of AI maturity, with results shown in Table 3.<sup>1</sup> All companies were graded as falling within in the lower two maturity levels in all dimensions.

In October 2024, we also asked nine SMEs from the southern Black Forest region to self-assess their AI maturity. An overview of the self-assessments can be found in Figure 1. The self-assessments confirmed our experience, that local SMEs remained mostly at the “Novice” or “Explorer” level. All but two companies gave a self-evaluated rating below the “User” level in all categories. Below, we summarize the state of AI adoption for each dimension in the consulted organizations.

**Table 2.** Considered organizations and related AI projects

Organization	Sector	AI Project
Org 1	Service	Chatbot
Org 2	Engineering	Predictive Maintenance
Org 3	Engineering	Predictive Maintenance
Org 4	Sales	Churn Prediction
Org 5	Consulting and Assessment	AI workshops

<sup>1</sup>Not all organizations analyzed here meet the full EU definition of an SME. Nevertheless, we assume that the results are transferable to local SMEs due to the sectors and business models of the organizations.

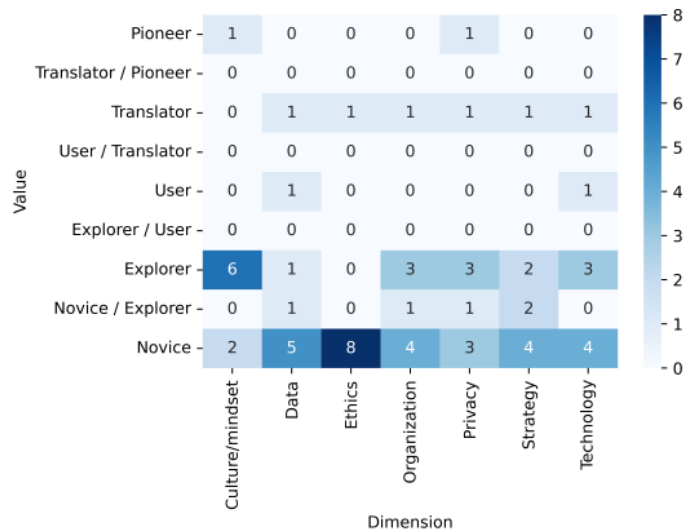
**Table 3.** Maturity ratings of the five considered companies

Category	Org 1	Org 2	Org 3	Org 4	Org 5
<b>Culture/mindset</b>	Explorer	Explorer	Explorer	Novice / Explorer	Explorer / User
<b>Data</b>	unknown	Novice	Novice	Explorer	Novice
<b>Ethics</b>	Novice	Novice	Novice	Novice	Novice / Explorer
<b>Organization</b>	Explorer	Novice	Novice	Novice	Novice
<b>Privacy</b>	Novice	Novice	Novice	Explorer	Explorer
<b>Strategy</b>	Novice/Explorer	Novice	Novice	Novice	Novice / Explorer
<b>Technology</b>	Novice	Novice	Novice	Novice / Explorer	Novice / Explorer

**Culture and Mindset:** The majority of managers in all companies surveyed show a strong interest in AI, as evidenced by their cooperation with the AI Lab. Companies saw potential in the use of AI, and in some cases considered it a necessity to remain competitive. However, the organizations lacked any noticeable AI culture that promoted innovation.

**Data:** The data provided by companies often had qualitative and quantitative shortcomings. The data were never collected explicitly for AI usage, resulting in low temporal resolution and imprecise output values. Additionally, there were often factual contradictions between related data sets and other inconsistencies. In other cases, high-quality data were only available for a limited time period or for a small sample size.

**Ethics:** Most organizations had no AI-specific ethics policy. However, in multiple cases, an active knowledge of consumer privacy laws (regarding General Data Protection Regulation – GDPR [27]) was demonstrated.



**Fig. 1.** Overview of the self-assessed AI maturity levels of nine SMEs from the southern Black Forest region.

**Organization:** In terms of organizational structures, companies were not ready for AI integration. The available resources (in particular: data, computing power, and personnel) were not yet adequate.

**Privacy:** The topic of data privacy was not initially prioritized by most of the companies surveyed, however they did show awareness of basic concepts, having worked with GDPR regulations previously.

**Strategy:** Although many managers were interested in incorporating AI technologies into a future strategy, and also had ideas for their first AI projects, these ideas were not yet part of official company policy.

**Technology:** Despite interest, AI applications were still rarely used in production. There was a lack of understanding about the technology along with its possibilities and limitations. In some cases, there were unrealistic ideas of AI, and uncertainty of when applying AI would be appropriate.

**Overall AI Maturity of SMEs:** The companies surveyed were still in the early stages of AI adoption. Despite a strong interest and a generally positive attitude towards AI, there was a lack of practical education concerning the technology along with a lack of necessary data. Working on building an understanding of the basics would be essential to progressing further on the maturity model.

## 4 Structured Approach for the Introduction of AI projects

In order to overcome the barriers to AI adoption listed in Table 1, and in consideration of feedback from managers of local SMEs, we have developed a suggested “plug and play” approach for quickly prototyping and integrating AI solutions for SMEs new to AI. Where continuing education resources already exist, they have been well received [28]. To expand on the concept of educational AI workshops, we propose a hands-on approach consisting of not only explanations of fundamental AI concepts, but also of guided implementation of a predetermined sample AI project. Industry leader Andrew Ng advises companies “to work on a *concrete idea*, meaning a specific product envisioned in enough detail” rather than following the typical design-thinking approach of brainstorming and developing strategy first [29]. Developing “plug-and-play” solutions that could be easily applicable to most companies would ensure faster implementation as well as a faster learning process to identify the company’s challenges regarding the onboarding of AI technologies. These quick implementations could then be used as building blocks on which to foster wider AI adoption within the whole business. This would allow for an experiential education process that addresses all eight key challenge areas during project implementation. This approach would include:

- structured workshops / introduction to AI
- a list of specific predefined use cases that organizations can adapt as part of the project
- accompanying prerequisites and suggestions for when each use case is appropriate
- lists of suggested software frameworks to use for each case and instructions on how to use them (with more than one option for each functionality so that there is room for individualized choices based on requirements)
- guidance through the first implementation steps of the selected use case, with details of these steps shown in Figure 2.

This approach would not only serve as a foundation for the company’s understanding of AI technology, but also act as the first stepping stone for developing the capacity for the

## PLUG AND PLAY IMPLEMENTATION STEPS

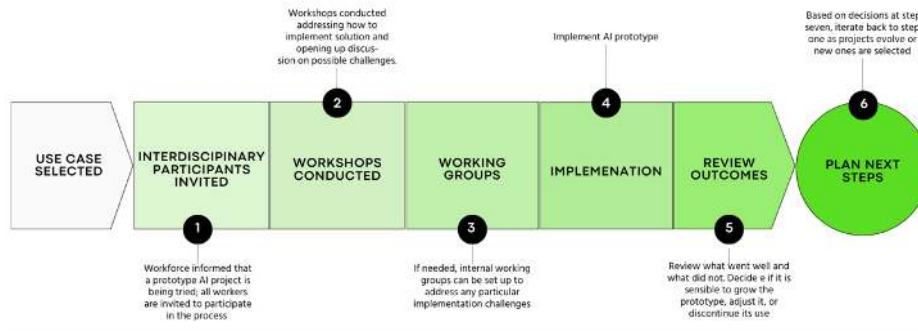


Fig. 2. Suggested Implementation Steps

organization to identify, select, and run AI projects more independently in the future, without (or with less) reliance on third parties.

A guided implementation of an already mostly-finished solution would address the eight challenges in Table 1 in the following way:

- **Lack of knowledge / skills:** Workshops included along with the solutions would both introduce basic AI concepts as well as link the theoretical concepts to their real-life counterparts.
- **High costs and financial risks:** Because the solution would be either free or have minimal costs (depending on the exact implementation details the company chooses i.e., are they combining the project with a cloud service), it would mitigate financial risk.
- **Lack of data:** Through either the implementation of using pretrained models as a solution, or using a limited machine learning solution, what data is needed would be limited, using only a predefined set of data (based on what has been found to be the minimum data necessary to solve the problem).
- **Lack of management awareness and strategy:** Similar to the first issue, the accompanying workshops will address management awareness. The development of a comprehensive AI strategy for the company can be delayed until after the plug and play prototype launch, at which point management will be better prepared through the first-hand experience to development an appropriate strategy.
- **Complexity and individuality of potential AI solutions:** By using a predefined prototype, the complexity is greatly reduced. Although the underlying technology remains complex, the implementation steps are clear and the uncertainty of which AI project to start with is removed.
- **Inadequate IT-infrastructure:** By starting with a relatively simple AI example, IT-infrastructure problems are minimized. By working from a predetermined list of possible solutions, which would include the use of software as a service solutions that are chosen for user-friendliness, there are no complicated IT decisions to be made upfront.
- **Data privacy, security, and regulations:** Data privacy, security, and regulation concerns are preemptively considered within the possible use cases. How these issues are addressed within the prototype can act as a useful reference for future projects.

- **Ethical and social concerns:** Ethical and social concerns are addressed throughout the project by maintaining transparency within the organization. Moreover, the process acts as a continuing education opportunity.

This structured approach could significantly mitigate overall risk of the implementation failing. Although larger industry players have demonstrated faster adoption rates of newer technology over time, studies estimate over an 80% failure rate of data science and AI projects [30,31,32]. The rates of success for purely generative AI applications are not yet clear; Gartner estimates a 30% failure rate by 2026 [33]. Therefore, we can look to research to find examples that have already demonstrated a successful track record, and select those examples as first implementations. Based on a McKinsey study, the top three use cases with the greatest financial success are: 1. Sales, 2. Software engineering, and 3. Marketing [4]. Considering their demonstrated success, plug and play solutions based in these areas can be used in SMEs as their first AI prototypes, with use cases as follow:

- **Sales:** Automated churn prediction, customer lead prioritization, chatbot integration for product recommendation
- **Software engineering:** Augmented software development through the use of LLMs or products such as GitHub Copilot.<sup>2</sup>
- **Marketing:** Generative AI can be used for the expedited creation of marketing materials such as landing pages and social media posts.

It is possible that later AI adoption could even be advantageous, in that initial R&D investments and trial and error of testing various methodologies can be entirely skipped over, following the economic development concept of leapfrogging. New adopters also have the advantage of avoiding the technical debt of older implemented systems. In the most ideal situations, these leapfroggers “not only jump over some phases of technological development but also create new paths to follow.” [34]

## 5 Conclusion

Recent rapid AI technological progress has great economic potential, with a bare minimum effect of productivity gains in the near-term, along with the more long-term potential of fundamental industry transformation. However, deployment of AI solutions has been uneven according to business resources. In order for SMEs not to fall behind, we advocate for an assisted approach, providing guidance through the implementation of a first AI prototype. By implementing manageable, already tested solutions, SMEs could leapfrog past the need for local R&D, and instead implement current state-of-the-art solutions, meanwhile gaining the knowledge and experience necessary for further independent AI integration and innovation.

## References

1. Schuster, T., Waidelich, L., Volz, R.: Maturity models for the assessment of artificial intelligence in small and medium-sized enterprises. In: Digital Transformation, Cham, Springer International Publishing (2021) 22–36

---

<sup>2</sup>AI augmented code generation is relevant for more than firms selling software solutions. Examples include AI-assisted G-Code generation for CNC machines in the manufacturing sector, and AI-assisted HTML editing for homepages.

2. KI-Labor Südbaden: Hochschule Offenburg KI-Labor Südbaden. <https://ki-suedbaden.de/> (2024) [Accessed 10-09-2024].
3. Acemoglu, D.: The simple macroeconomics of ai. Technical report, National Bureau of Economic Research (2024)
4. Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., Zimmel, R.: The economic potential of generative ai: The next productivity frontier. Technical report, McKinsey (2023)
5. Rammer, C., Doherr, T., Kinne, J., Lenz, D.: Ki-einsatz in unternehmen in deutschland. Technical report, Bundesministerium für Wirtschaft und Klimaschutz (2024)
6. Turing, A.M.: Computing machinery and intelligence. *Mind* **59**(236) (1950) 433–460
7. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine* **27**(4) (2006)
8. Barton, M., Budjac, R., Tanuska, P., Gaspar, G., Schreiber, P.: Identification overview of industry 4.0 essential attributes and resource-limited embedded artificial-intelligence-of-things devices for small and medium-sized enterprises. *Applied Sciences* **12**(11) (2022)
9. Cotrino, A., Sebastián, M.A., González-Gaya, C.: Industry 4.0 roadmap: Implementation for small and medium-sized enterprises. *Applied Sciences* **10**(23) (2020)
10. Bauer, M., van Dinther, C., Kiefer, D.W.: Machine learning in sme: An empirical study on enablers and success factors. In: *Americas Conference on Information Systems*. (2020)
11. Oldemeyer, L., Jede, A., Teuteberg, F.: Investigation of artificial intelligence in smes: a systematic review of the state of the art and the main implementation challenges. *Management Review Quarterly* (2024)
12. Tominc, P., Oreški, D., Čančer, V., Rožman, M.: Statistically significant differences in ai support levels for project management between smes and large enterprises. *AI* **5**(1) (2024) 136–157
13. Chui, M., Hall, B., Singla, A., Sukharevsky, A., Yee, L.: The state of ai in 2023: Generative ai’s breakout year. Technical report, McKinsey (2023)
14. Deveau, R., Griffin, S.J., Reis, S.: Ai-powered marketing and sales reach new heights with generative ai. Technical report, McKinsey (2023)
15. Cockburn, I.M., Henderson, R., Stern, S.: The impact of artificial intelligence on innovation. Working Paper 24449, National Bureau of Economic Research (2018)
16. Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B., Kiron, D.: Winning with ai. *MIT Sloan Management Review* (2019)
17. Statistisches Bundesamt: 55% in kleinen und mittleren unternehmen tätig. <https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Kleine-Unternehmen-Mittlere-Unternehmen/aktuell-beschaefigte.html> (2024) [Accessed 14-10-2024].
18. GROW.A.2: 2024 sme fact sheet. Technical report, European Commission (2024)
19. Ministerium für Wirtschaft, Arbeit und Tourismus: Regionale Labore für Künstliche Intelligenz in Baden-Württemberg. <https://www.wirtschaft-digital-bw.de/ki-made-in-bw/regionale-ki-labs/standorte-der-regionalen-ki-labs> (2024) [Accessed 10-09-2024].
20. Lu, X., Wijayarathna, K., Huang, Y., Qiu, A.: Ai-enabled opportunities and transformation challenges for smes in the post-pandemic era: a review and research agenda. *Frontiers in Public Health* **10** (2022)
21. De Simone, V., Pasquale, V.D., Miranda, S.: An overview on the use of ai/ml in manufacturing msms: solved issues, limits, and challenges. *Procedia Computer Science* **217** (2023) 1820–1829
22. Govori, A., Sejdija, Q.: Future prospects and challenges of integrating artificial intelligence within the business practices of small and medium enterprises. *Journal of Governance and Regulation* **12**(2) (2023) 176 – 183
23. Schönberger, M.: Artificial intelligence for small and medium-sized enterprises: Identifying key applications and challenges. *Journal of Business Management* **21** (2023) 89–112

24. Yusuf, S.O., Durodola, R.L., Ocran, G., Abubakar, J.E., Echere, A.Z., Paul-Adeleye, A.H.: Challenges and opportunities in ai and digital transformation for smes: A cross-continental perspective. *World Journal of Advanced Research and Reviews* **23**(3) (2024) 668–678
25. Grünbichler, R., Salimbeni, S.: Artificial intelligence in small and medium-sized enterprises: Requirements and barriers. In: *Latest Advancements in Mechanical Engineering*, Springer Nature Switzerland (2024) 173–184
26. Iyelolu, T.V., Agu, E.E., Idemudia, C., Ijomah, T.I.: Driving sme innovation with ai solutions: overcoming adoption barriers and future growth opportunities. *International Journal of Science and Technology Research Archive* **7**(1) (2024) 036–054
27. Council of European Union: Regulation (eu) 2016/679 of the european parliament and of the council. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L:2016:119:FULL> (2016)
28. Reyes, C.C., Eisele, O., Lennings, F., Sandrock, S., Schüth, N.J., Terstegen, S.: Künstliche intelligenz erkennen, bewerten und einführen. Technical report, ifaa (2023)
29. Ng, A.: OpenAI Shrinks GPT-4o, Meta Withholds Models From Europe, and more — deeplearning.ai. <https://www.deeplearning.ai/the-batch/issue-259/> (2024) [Accessed 29-07-2024].
30. VentureBeat: Why do 87% of data science projects never make it into production? <https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/> (2019) [Accessed 01-10-2024].
31. Ryseff, J., Bruhl, B.F.D., Newberry, S.J.: The root causes of failure for artificial intelligence projects and how they can succeed: Avoiding the anti-patterns of ai. Technical report, RAND Corporation (2024)
32. Bojinov, I.: Keep your ai projects on track. *Harvard Business Review* **11** (2023)
33. Gartner Inc.: Gartner predicts 30 <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025> (2024) [Accessed 08-10-2024].
34. Kopka, A., Fornahl, D.: Artificial intelligence and firm growth — catch-up processes of smes through integrating ai into their knowledge bases. *Small Business Economics* **62**(1) (2024) 63 – 85





# Data augmentation for pathogen segmentation in vinewood fluorescence microscopy images

Julie Munsch<sup>1,2</sup>, Sonia Ouali<sup>1</sup>, Jean-Baptiste Courbot<sup>1</sup>, Romain Pierron<sup>3</sup>  
and Olivier Haerberlé<sup>1</sup>

<sup>1</sup>IRIMAS, UR 7499, Université de Haute-Alsace, Mulhouse, France

<sup>2</sup>Eiffage Energie Systèmes - Centre Expertise IA, Mulhouse, France

<sup>3</sup>LVBE, UR 3991, Université de Haute-Alsace, Colmar, France

**Abstract.** In this paper, we address the problem of segmentation of vine pathogens within fluorescence microscopy images. To our knowledge, the quantification from such images is an original problem. As a consequence, there is no available database to rely upon in order to use supervised machine learning techniques. We provide a workaround by creating realistic images containing the desired filamentary pattern and variable blur effect. Numerical results show the interest of this data augmentation technique, especially on images corresponding to difficult segmentation.

**Keywords:** data augmentation, image segmentation, fluorescence microscopy deep learning, machine learning

## 1 Introduction

### 1.1 Context

Grapevine trunk diseases have become a serious problem for vinegrowers around the world. In France, approximately 13% of the vineyard is unproductive each year, leading to losses of 1 billion euros [1]. Esca is one of these diseases and the oldest described. It affects 4.4% of vines causing deterioration or even complete dieback of the wood. The behavior of pathogens believed to be responsible for this disease is until today poorly known, and no treatment exists once the wood is contaminated. Understanding the colonization process is key to the

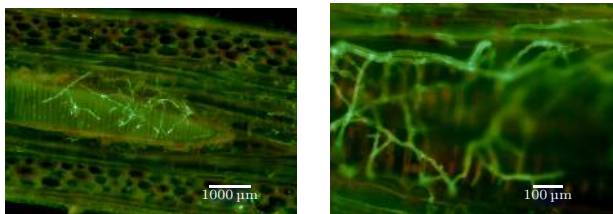


Fig. 1: Examples of pathogen in vinewood images observed with a fluorescence microscope, at x10 magnification on the left and x40 on the right. The pathogen fluorescence appears as green-yellow filaments, while wood auto-fluorescence appears as yellow-brown.

development of sustainable cures. To gain a better understanding of the pathogens' behavior, inoculation experiments are performed and the resulting plant is observed in fluorescence microscopy (see Fig. 1). Then, image segmentation techniques can help quantify the pathogen's presence in vinewood.

Obtaining real microscopic grapevine images, together with expert-segmented images (here by Dr. R. Pierron), is however a time-consuming process. The resulting images form a relatively limited database that might be too small to use robust supervised learning techniques. Thus, data augmentation is a key step to fully make use of these methods.

### 1.2 Related works

Data augmentation techniques are, in general, a helpful tool to enlarge databases. This is in particular the case for medical images, for which the patient set is often limited [2].

When handling images, basic augmentation relies on some simple transforms, such as rotation, flip, cropping, and dilation. These techniques might be improved using advanced deep learning approaches, such as adversarial training or neural style transfer (see [3] for a review). Those are mostly based on image textures, and do not incorporate a model for the image formation process. We can make a similar observation regarding segmentation in microscopy images, as highlighted in [4].

Hence, there is a lack for a data augmentation technique that specifically accounts for the image formation model. This is particularly striking when handling the case of fungi in vinewood images, which can be affected by a varying blur depending on the region of the image.

In this article, we propose to generate synthetic images that mimic real fluorescent microscopy grapevine images in order to train supervised algorithms. This article is organized as follows. First, we describe the image formation process of both the real fluorescence microscopy images and the synthetic ones (Section 2). Then, in Section 3 we present the segmentation methods and the results obtained on the synthetic and real images.

## 2 Image formation

### 2.1 Real images

Colored grapevine cuttings are observed using a wide-field fluorescent microscope. The resulting images are characterized by the filamentary light green fluorescent pathogen (see Fig. 1). However, the wood also exhibits auto-fluorescence, which makes it challenging to distinguish between the pathogen and the wood. The latter is a complex plant tissue, which results in heavily textured background images.

Furthermore, the images acquired using a fluorescence microscope suffer from a blurring effect. The blur is due to the convolution of the sample with the system’s Point Spread Function (PSF). The latter is known at an instrumental level, so the blur is partially known. The remaining unknown is the sample’s depth, which influence on the width of the PSF to take into account.

Given an un-blurred image  $\mathbf{I}$  containing  $S$  pixels, corresponding to an ideal sample, we formalize this model as

$$(\mathbf{H}(\mathbf{v})\mathbf{I})_s = \langle \mathbf{h}_{s,[v_s]}^{3D}, \mathbf{I} \rangle, \quad (1)$$

where  $\mathbf{h}^{3D} \in \mathbb{R}^{S \times S \times P}$  is known only on a 3D grid, see Fig. 2.  $P$  represents the depth of the PSF,  $\mathbf{h}_{s,[v_s]}^{3D} \in \mathbb{R}^S$  represents the PSF of the system at depth  $v_s$  in the pixel  $s$ , and  $\mathbf{v}$  is a depth map associated to the image. In other words,  $\mathbf{H}$  forms a depth-selection operator to model the blur. Thus, we consider that an image  $\mathbf{y} \in \mathbb{R}^S$  is formed as:

$$\mathbf{y} = \mathbf{H}(\mathbf{v})\mathbf{I} + \mathbf{b}, \quad (2)$$

with  $\mathbf{b}$  an additional background term. Note that we consider  $\mathbf{I}$  the ideal image containing only the pathogen, including the colors. Later on, we note its binary counterpart  $\mathbf{x}$  to model the segmentation mask.

### 2.2 Synthetic images

In this section, we present how we mimic the real images to augment the available database. As part of the grapevine inoculation process, some plants were part of a control group, *i.e.* there were no inoculated fungi. These plants were also observed with fluorescence microscopy. The resulting images capture well what is considered as a background in our segmentation framework ( $\mathbf{b}$  in Eq. (2)).

Besides, there is a need to generate synthetic pathogens in the image, that have a filamentary texture and are also corrupted by a varying blur. Filamentary texture is well represented in retina images, which have been investigated for many years (see [6] for an overview). Several databases are publicly available, and we selected the DRIVE (Digital Retinal Images for Vessel Extraction) [7] database as filament models in our approach. We divided the masks of the segmented retina image into two sub-parts, first vertically and then horizontally, resulting in two sub-parts measuring 128x256 and 256x128, respectively. We reconstruct an image of size 256x256 by filling the missing spaces with a black (empty) rectangle of dimensions 128x256 or 256x128. Additionally, for data augmentation, we perform for each picture rotations of 90°, 180° and 270°. So we obtain images containing pieces of segmented retina. The synthetic images  $\tilde{\mathbf{y}}$  are designed such that:

$$\tilde{\mathbf{y}} = \mathbf{H}(\mathbf{v})(\mathbf{r} + \sigma\mathbf{w}), \quad (3)$$

where  $\mathbf{w}$  represents images of healthy grapevine wood,  $\mathbf{r}$  are images containing pieces of segmented retina.  $\sigma$  allows to define the SNR, according to :

$$\text{SNR} = 10 \log_{10} \left( \frac{\|\mathbf{H}\mathbf{r}\|^2}{\sigma S} \right). \quad (4)$$

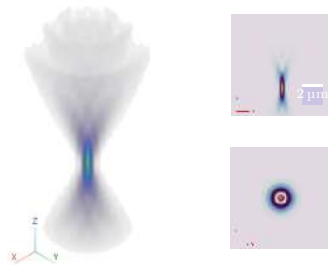


Fig. 2: Typical PSF  $\mathbf{h}^{3D}$  in fluorescence microscopy generated via PSF generator [5], with axial and longitudinal slices of the central plans.

$\mathbf{H}(\mathbf{v})$  is the 3D depth-selection operator for the 3D PSF, as described in (1). It relies on a continuous field  $\mathbf{v} \in \mathbb{R}^S$ , for which we make a Gaussian Markov Random Field assumption:

$$p(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\mathbf{v}^T \boldsymbol{\Sigma}^{-1} \mathbf{v}\right) \quad (5)$$

We assume that  $\boldsymbol{\Sigma} \in \mathbb{R}^{S \times S}$  is a circulant covariance matrix, whose basis is parameterized by a Gaussian correlation function.

Summarizing, the image generation is as follows:

- select a snippet  $\mathbf{w}$  from a healthy wood image.
- select a same-sized snippet  $\mathbf{r}$  from the DRIVE database, color and crop it.
- sample a GMRF realization  $\mathbf{v}$  (5).
- blend the images according to (3) and a given target SNR (4).

Fig. 3 illustrates the process.

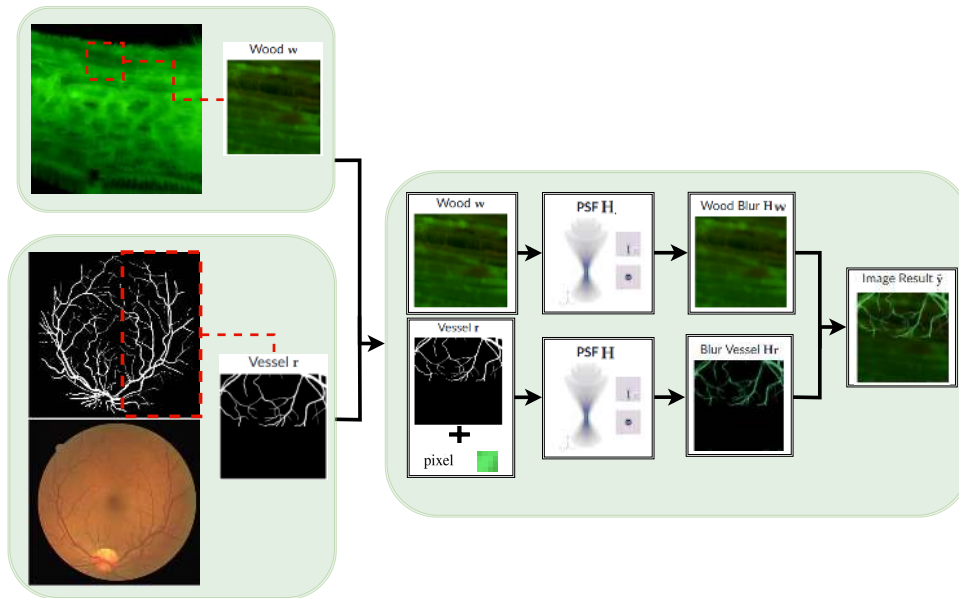


Fig. 3: Illustration of the synthetic image formation process.  $\mathbf{y}$  correspond to a SNR = -9dB.

### 3 Numerical results

To evaluate our approach, we test how its use impacts the segmentation of images, depending on the choice of the training and testing databases. To measure the quality of the segmentation, we will use the Accuracy metric, which allows to measure the total percentage of correctly predicted pixels. To show performance, we only present the accuracy score to keep the reading straightforward.

#### 3.1 Segmentation methods

- *Pre-processing.* Prior to segmentation, we extract relevant features as a pre-processing. The latter were identified in a preliminary random forest-based study using the Ilastik software [8]. The features were sorted along their Gini importance [9], and the 14 first features were retained (see Table 1). In addition, preliminary experiments have shown that including classical image filters (namely Gabor, Sobel, Roberts, Scharr, and Prewitt filters) also improves segmentation. Then, for any 3-channel RGB input image, the pre-processed version contains 57 channels.
- *Random Forests.* The random forest algorithm [10] requires several parameters to be selected. The number of decision trees was selected from grid search and we set 80 trees in this model. The other RF parameters are set to have fully-grown and unpruned trees.

Identity	Hessian	Gaussian $\sigma = \{3; 5; 7\}$	Gradient Magnitude $\sigma = 2$
Difference of Gaussian $(\sigma_i, \sigma_j) =$ $\{(1, 3.5); (1, 12); (1, 30)\}$	Laplacian of Gaussian $\sigma =$ $\{0.5; 1.6; 3\}$	Eigenvalues of Tensor Structure $\sigma = 0.7$	Hessian of Gaussian Eigenvalues $\sigma = 3.5$

Table 1: 14 features retained based on their Gini importance.

- *U-Net*. The U-Net model is widely recognized for its effectiveness in image segmentation and is commonly used for this task [11]. To implement a U-Net model, we selected from preliminary experiment 4 depth levels, from  $256 \times 256$  input images to  $16 \times 16 \times 256$  bottleneck, and a block convolution parameter  $\alpha$  of 0.2, 0.2, 0.1 and 0.1 for each level. The architecture weights were estimated with the Adam optimizer, using the binary cross entropy criterion. The optimizer runs for at most 1000 epochs, stops when accuracy decreases, and handles each image one by one (the batch are one-sized). We also set the input layer depth as 16, *i.e.* 16 features are first generated. The former takes RGB images as input, while the latter takes the pre-processed images as input.

### 3.2 Synthetic and real image databases

We design and make use of several databases (Fig. 4) in order to validate the proposed approach:

- The first dataset contains only synthetic images, produced along the procedure detailed in Subsection 2.2. We name it *dataset A* in the following, and it contains 427 images.
- The second dataset exclusively contains real images together with expert-labeled ground truths. We label it as *dataset B*. We split this dataset depending on the content of the image:
  - with fungi in lower quality image, presenting notably a higher blur: *dataset B1*, 247 images.
  - without fungi: *dataset B2*, 312 images.
  - with fungi : *dataset B3*, 569 images
- *Dataset C* contains images from both datasets A, B1 and B3, with 128 images.

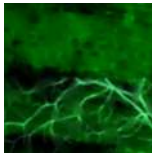
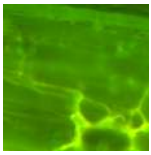
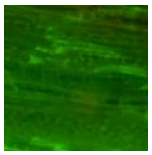
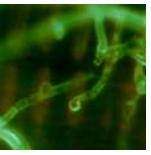
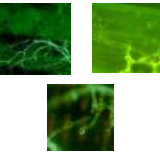
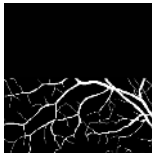



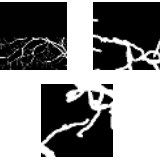
		B : real images				
dataset name	A	B1	B2	B3	C	
description	synthetic images	with fungi in lower quality image	without fungi	with fungi	mix A, B1 and B3	
number of images	427	247	312	569	128	
images						
segmentation						

Fig. 4: Overview of our 5 datasets.

The database resulting from datasets A and B will be made available online.

Note that we made the choice of relatively small images ( $256^2$  pixels) to ensure diversity in learning, yielding a relatively high number of thumbnails. This is to be contrasted with the low number of acquisitions (15 images) used to form thumbnails in dataset B.

Then, we use datasets A, B1, and C for training. For testing, we target at first dataset B1 to cover difficult cases specifically, then dataset B to verify the assessment of true negatives, *i.e.* in the absence of fungi, and then dataset B3 to assess the presence of fungi. The dataset overlaps and use is depicted in Fig. 5.

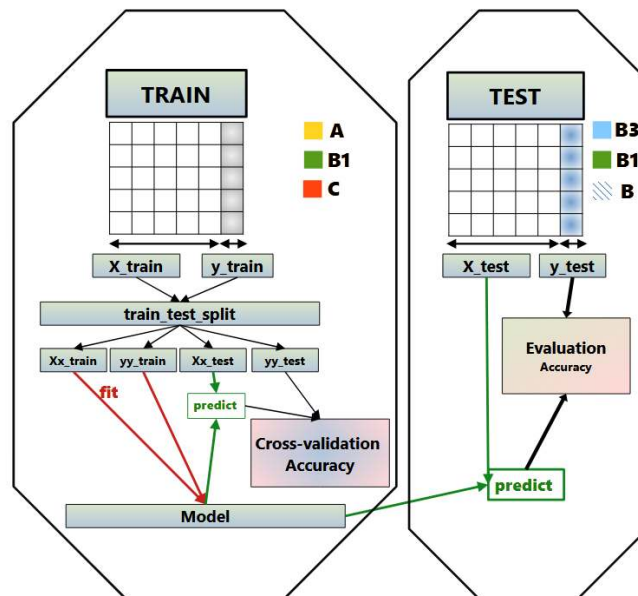


Fig. 5: Datasets split for the different evaluation processes run in this paper.

### 3.3 Results

Table 2 depicts the results (accuracy) we obtained for the three models of each learning method, using cross validation with 5 folds. For each dataset, the best method result is highlighted in bold font. Fig. 6 also depicts some results. The main outcomes can be summarized as follows:

- U-Net 57D yields better results than U-Net 3D for the segmentation of dataset B1 (93.5% instead of 90.1%). This shows that the pre-processing is helpful in the process, despite the change of dimensions.
- On B1 again, we observe an improvement of the results when incorporating synthetic images: the latter do allow an average 1.57% improvement of accuracy. An example of improvement can be seen in the first column of Fig. 6.
- Noteworthy, training on lower-quality images of B1 yields for the best segmentation results of the higher-quality datasets B and B3.
- Generally, the random forest remains a very powerful learning method, even when only training on synthetic image, as shown in columns 2–4 in Fig 6.
- Note: On retinal segmentation, it is common for the distribution of pixels between the two binary classes (e.g., "background" and "structures of interest," such as blood vessels) to be unbalanced. In other words, the majority of the pixels often belong to the background of the image, while a much smaller number corresponds to the specific structures we aim to segment.

		Random forest			U-Net 3D		
Training database:		Dataset A	Dataset B1	Dataset C	Dataset A	Dataset B1	Dataset C
Test	CV accuracy	92.96%	96.16%	97.06%	89.68%	93.84%	90.96%
	Dataset B1	2.69%	94.13%	<b>95.70%</b>	18.98%	90.10%	83.25%
	Dataset B	41.00%	<b>95.70%</b>	92.45%	10.47%	94.17%	85.73%
	Dataset B3	6.58%	92.14%	88.08%	10.83%	<b>93.79%</b>	89.00%

Table 2: Segmentation accuracies, as a function of the segmentation method, the training database, and the testing database.

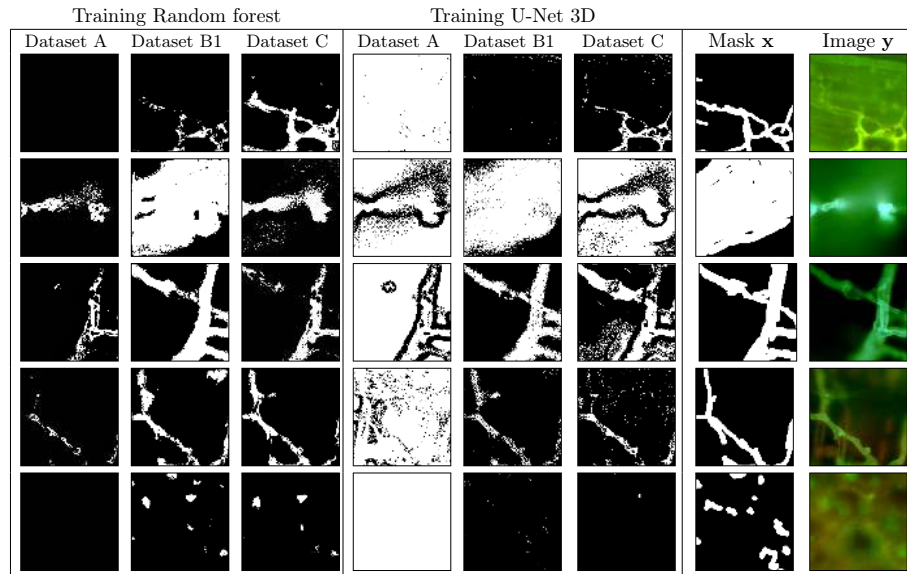


Fig. 6: Example of results on real vine fungi images.

#### 4 Conclusion

In this work, we proposed a data augmentation technique dedicated to fluorescence microscopy images. We have shown that for the cleaner images, supervised segmentation is feasible even on a small database, *i.e.* without augmentation. However, we also showed that for lower-quality images, the addition of the synthetic images was indeed helpful, leading to noticeable accuracy improvements.

This work stems a few perspectives on the topic, as it could be generalized to other segmentation problems in fluorescence microscopy (such as tubulin networks in cells), as well as other imaging techniques (such as retina imaging).

#### References

1. Del Frari, G., Gobbi, A., Aggerbeck, M.R., Oliveira, H., Hansen, L.H., Ferreira, R.B.: Characterization of the wood mycobiome of vitis vinifera in a vineyard affected by esca. spatial distribution of fungal communities and their putative relation with leaf symptoms. *Frontiers in Plant Science* **10** (2019) 910
2. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology* **65**(5) (2021) 545–563
3. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of big data* **6**(1) (2019) 1–48
4. Ma, J., Hu, C., Zhou, P., Jin, F., Wang, X., Huang, H.: Review of image augmentation used in deep learning-based material microscopic image segmentation. *Applied Sciences* **13**(11) (2023) 6478
5. Kirshner, H., Aguet, F., Sage, D., Unser, M.: 3-D PSF fitting for fluorescence microscopy: implementation and localization application. *Journal of Microscopy* **249**(1) (2013) 13–25
6. Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunović, H.: Artificial intelligence in retina. *Progress in retinal and eye research* **67** (2018) 1–29
7. Challenge, G.: Drive: Digital retinal images for vessel extraction (Accessed on January 21<sup>st</sup>, 2024)
8. Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., et al.: Ilastik: interactive machine learning for (bio) image analysis. *Nature methods* **16**(12) (2019) 1226–1232
9. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA (2001)
10. Breiman, L.: Random forests. *Machine learning* **45** (2001) 5–32
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* **18**, Springer (2015) 234–241

# Enhancing Phishing Email Detection with Context-Augmented Open Large Language Models

Fabian Nicklas<sup>1</sup>, Nicolas Ventulett<sup>2</sup>, and Prof. Dr.-Ing. Jan Conrad<sup>3</sup>

<sup>1</sup> University of Applied Sciences Kaiserslautern  
fani1001@stud.hs-kl.de

<sup>2</sup> University of Applied Sciences Kaiserslautern  
nive1002@stud.hs-kl.de

<sup>3</sup> University of Applied Sciences Kaiserslautern  
jan.conrad@hs-kl.de

**Abstract.** Large Language Models offer a promising approach for improving phishing detection through advanced natural language processing. This paper evaluates the effectiveness of context-augmented open LLMs in identifying phishing emails. An approach was developed that combines the methods of Few-Shot Learning and Retrieval-Augmented Generation (RAG) to remarkably improve the performance of LLMs in this area. On this basis, it has been shown that the presented approach can significantly improve the recognition rate even for smaller models.

**Keywords:** Artificial Intelligence, AI, Cybersecurity, Large Language Models

## 1 Introduction

Phishing is a significant and increasing threat to cybersecurity. Attacks using constantly evolving techniques aim to tempt people into revealing sensitive personal information. It is estimated that 90 percent of all successful cyberattacks have phishing as an initial vector of attack [1]. The rise of Large Language Models (LLM) has revolutionized the field of Natural Language Processing (NLP). First popular representatives as the model GPT (Generative Pretrained Transformer) by OpenAI [2] have showcased the power of Large Language Models for language generation and understanding. They are trained across diverse datasets of large text corpora and their application beyond the original task of text generation for machine learning problems is an increasingly addressed research question [3]. LLMs with their deep understanding of natural language are a promising starting point for the detection of phishing emails. This paper presents an approach of combining the in-context learning and augmentation methods Few-Shot Learning and Retrieval Augmented Generation (RAG) for phishing email classification. It dynamically augments the context of LLMs in a problem-specific way at the time of inference without the need for intensive, task-specific training or the use of a dedicated model. The approach is evaluated in experiments across different open models and compared to more common state-of-the-art prompting techniques.

## 2 Related Work

Phishing email detection has been an area of active research for decades, evolving from the application of rule-based systems, lexical analysis and machine learning algorithms,

including SVM and tree-based classifiers, up to leveraging deep learning methods like recurrent and convolutional neural networks as well as transformers [4][5]. The use of Large Language Models for identifying email phishing is still an emerging field with a sparse number of research publications.

A majority of recent studies based their work on the GPT models of OpenAI [6][7][8]. The model family achieved a high level of popularity with the release of its derivative ChatGPT. Rosa et al. [7] achieved an overall accuracy of 75.75 % for binary phishing email classification by feeding emails to GPT 3.5. With their high number of active parameters the GPT models proved a strong performance across many application areas, however, the GPT models are proprietary and closed-source [9]. This paper focuses on the use of open models, that are free to use and are meeting higher demands regarding data privacy. While some studies on phishing detection use open LLMs solely as upstream feature extractors for other machine learning methods [3], Koide et al. [10] employs the model Llama 2 to classify emails and achieves an overall accuracy of 88.61 % through prompt engineering. Their study contrast this with the use of the much-larger GPT-4 model showing 99.70% accuracy.

Baumann et al. [11] proposes a combination of RAG and FSL to generate models for domain-specific languages (DSLs) finding application in the field of software engineering. Their approach uses RAG to retrieve relevant examples from a knowledge base, enabling FSL to generate synthetic models for underrepresented DSLs lacking sufficient training data and thereby adapting a LLMs output syntax. Literature review showed, the method of using a fusion of RAG and FSL to improve a LLM’s capability to solve unknown machine learning tasks has not been addressed to date.

### 3 Methodology

#### 3.1 Dataset

The experiments conducted in this study aim to evaluate the performance of the proposed approaches for the classification of phishing emails. For this purpose, a dataset containing both phishing and legitimate emails was created by concatenating two publicly available datasets. The *CSDMC Spam Corpus*[12] includes 2,949 so-called “ham emails”, legitimate messages that do not fall into the categories of phishing or “spam”. It has already been used in similar studies as [10]. The phishing emails were sampled from the *Phishing Pot* [13] dataset and are real emails collected from August 2022 to July 2024. In contrast to [14] this approach do not include synthetic phishing samples or emails collected well into the past, as in [15]. By choosing an up-to-date source dataset newer phishing techniques are also represented in our final dataset. From each source dataset 2,900 emails were randomly sampled to build a new set with a total of 5,800 emails, balanced between the two classes *phishing* and *no phishing*. Samples with an email body of less than 50 characters or more than 420,000 characters were not considered as valid samples and were discarded in the selection step. In a subsequent data-cleaning process, all non-ASCII characters in the messages were removed. Each message sample consists of the concatenation of the email’s subject and its body. If the message body was available in text format and the HTML format, this approach prioritized the HTML part and converted to plain text by removing all HTML-related fragments. This study does not address the role of email attachments as an attack vector, all attachments included in the samples were removed.



### 3.2 Model Selection

The experiments were evaluated for a variety of Large Language Models that represent the current state of the art and are published under an open license. The approach deliberately refrained from the use of commercial models such as GPT4 (OpenAI). The selected AI models are OpenChat 7B [16], Mixtral 8x7B [17], Mistral 7B [18], Gemma2 9B and 27B (Google Deep Mind) [19], Llama3.2 1B and 3B [20], Mistral-Small 22B [18], Command-R 35B [21], as well as Llama3.1 8B and Llama3.1 70B (Meta AI) [20]. All models were pre-trained by their respective authors on different datasets and differ in their architecture and the number of parameters. While models with a larger number of parameters generally have a greater ability to understand complex patterns and relationships, they may tend to show over-fitting behavior and be less applicable to new and unseen data.

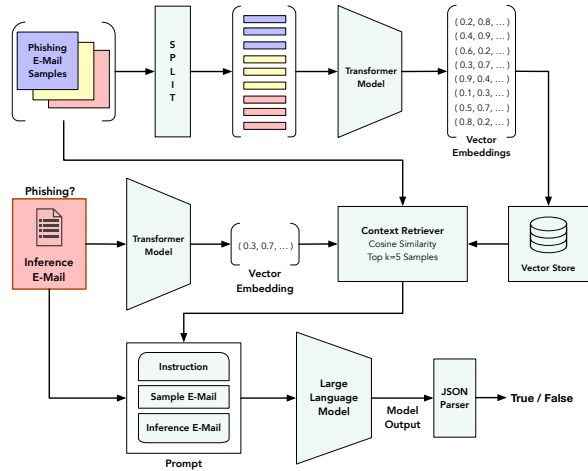


Fig. 1. Fusion of Few-Shot Learning (FSL) and Retrieval-Augmented Generation (RAG).

### 3.3 Detection of Phishing Emails with Large Language Models

Following the creation of the dataset, two different prompts<sup>4</sup> problem-specific prompts were developed to guide the LLMs to perform the classification task and produce the desired output. The same prompts were used across all AI models to allow for a consistent evaluation process. A fundamental prompt is provided for the evaluation of an email (**Prompt1**):

```
You are an expert in detecting phishing emails.
Your task is to determine whether it is a phishing email or not.
You are not supposed to justify or explain your decision.
{format_instructions}
E-mail:
'''{email}'''
```

<sup>4</sup> <https://github.com/n-vent/llm-phishing-detection-paper>

In the prompt, the LLM is given a fictitious role and the specific task to perform. The structure follows the *Persona Pattern*, a commonly used instruction scheme [22] that is independent of the area of application and the choice of model. **Prompt 2** extends the first prompt with a list of characteristics that may proof a phishing attempt. These include characteristics such as an impersonal generic greeting, urgent calls for action or demanding personal information such as the victim’s bank details.

The evaluated models are trained to generate textual output in natural language. The models are guided to generate structured output by providing additional formatting instructions as JSON schema [23] in the prompt. At the time of inference, the passed sequence of prompt, preprocessed email and formatting instructions is concatenated and used as model input. The email is inputted directly into the language model, without the need for any feature extraction. A subsequent parser extracts the result of the classification from the model’s text output as JSON, holding a boolean variable. The simple architecture of this approach does not include any components other than the described input construction, the respective language model and the JSON parser.

### 3.4 Context Augmentation through Few-Shot Learning and Retrieval-Augmented Generation for Improved Phishing Detection

This paper presents an improved approach for the classification of phishing emails by augmenting the knowledge of an already trained Large Language Model in-context and at the time of inference. The approach combines the methods of *Few-Shot Learning* (FSL) [24] and the *Retrieval Augmented Generation* (RAG) [25]. With FSL, also referred to as in-context learning, the model receives task demonstrations in natural language for a considered problem as part of the passed prompt. This allows the pre-trained models to address unknown tasks without a comprehensive training process or fine-tuning. FSL has shown to extend a language model’s capability outside of the data it has been trained on [11].

Instead of prompting prepared and static FSL examples of phishing emails, it is proposed to dynamically select a relevant set of examples at the time of inference based on the input email. Using the technique of Retrieval Augmented Generation, examples from a knowledge base are selected and integrated into the prompt before generation. The LLMs gain access to domain-specific information that was not present in their training data. The model does not persistently store the augmented information and its parameters remain unchanged.

Figure 1 shows the architecture of the proposed RAG FSL fusion approach. A collection of examples of phishing emails serve as the RAG knowledge source and are individually split into blocks with a maximum length of 200 characters. By leveraging a transformer model, vector embeddings are obtained for each block as numerical vectors that represent semantic relationships. The pre-trained embedding model *Sentence-BERT* (all-MiniLM-L6-v2) was selected as the transformer, which maps natural language sentences into a 384-dimensional vectorspace and is commonly used for semantic search [26]. A vector database is populated with the generated embeddings for the 200-character-blocks. For the classification of a suspicious email, it is encoded into its vector representation using the same transformer model in order to find similar phishing examples to use in FSL. A set of relevant examples is selected by the maximum possible semantic similarity via a vector search. The cosine similarity [27] between the email embedding and each vector of blocks in the database is considered as a similarity measure (context retriever). The  $k$  most similar samples from the phishing dataset with the lowest semantic distance are selected as relevant FSL examples.

The prompt provided to the model is a concatenation of the instruction for the prediction task and the output schema,  $k = 5$  positive phishing examples for in-context augmentation and the email message to be classified (**RAG FSL**):

```
You are an expert for detection of phishing emails.
For example, the following emails are phishing emails:
  Example 1 is a phishing email: {sample-email-1}
  [...]
  Example k is a phishing email: {sample-email-k}
Your task is to scan the following email, to decide whether or not
it is a phishing email and to use the provided JSON schema
for answering the question.
{format-instructions}
Question: Is the following email a phishing mail?
'''{email}'''
```

The result is extracted from the model output in a structured form using a JSON parser in the same way as the first approach (see section 3.3).

## 4 Experiments and Results

The 5,800 emails in the constructed dataset were systematically shown to the language models and the result of the classification evaluated for each sample. Each email was processed with **Prompt 1**, the extended **Prompt 2** as well as the proposed approach **RAG FSL**. Each prompt was evaluated across all of the selected eleven models. The model temperature parameter was set to 0.0, determining whether the output is more creative and random or more predictable. Other hyper-parameters than temperature were not changed. A total of 191,400 classification were run in this study. Model inference was performed on a NVIDIA A100 GPU with 80 GB of memory. In the evaluation, the performance of the approaches is assessed and various models using the quantitative metrics of precision, recall, F1 score, specificity and classification accuracy. If no valid JSON-data could be parsed from the model output by, the result was discarded in the evaluation. This could lead to an unbalanced number of positive and negative classes, which is met by calculating the metrics weighted by the number of samples as defined in [28]. The 2,900 emails of the positive class in the phishing dataset serve as knowledge source for RAG FSL. To guarantee the validity of the evaluation results and prevent target leakage, it was verified that the RAG phishing sample were not equal to the email test candidate at prediction time.

Table 1 shows the results of the conducted experiment to evaluate the presented approaches for the phishing email classification problem. The variance in results of the individual LLMs across the three different experiments shows the influence of the different model architectures and their training datasets on the performance in this classification task. The size of the models, measured by the number of active parameters, can be seen as an important but not decisive factor in the accuracy of the prediction. While Llama3.1 70B with 70 billion parameters always clearly outperforms the smaller 8B variant, the model Mixtral 8x7B with 13 billion active parameters already achieves an accuracy of 94.71% with prompt 1 compared to Llama3 with 92.82%.

The LLMs with less than 10 billion active parameters show a mostly higher performance using prompt 1 or prompt 2. When using RAG FSL on smaller models, the quality of the prediction seems to decline. Gemma2 9B can be noted as an exception, since the

**Table 1.** Performance of different Large Language Models for phishing email classification for *Prompt 1*, the extended *Prompt 2* and the proposed context-augmenting *RAG FSL* fusion approach. Models are in ascending order by their number of active parameters.

Model	Experiment	Precision	Recall	F1	Specificity	Accuracy
Llama3.2 1B	Prompt 1	41,94 %	50,39 %	33,86 %	<b>99,86 %</b>	50,39 %
	Prompt 2	51,09 %	<b>51,08 %</b>	<b>51,07 %</b>	50,27 %	<b>51,08 %</b>
	RAG FSL	<b>54,91 %</b>	<b>51,08 %</b>	38,28 %	96,33 %	<b>51,08 %</b>
Llama3.2 3B	Prompt 1	68,32 %	<b>53,77 %</b>	<b>41,76 %</b>	98,58 %	<b>53,77 %</b>
	Prompt 2	<b>74,02 %</b>	51,40 %	35,71 %	<b>99,96 %</b>	51,40 %
	RAG FSL	63,92 %	52,09 %	37,47 %	99,14 %	52,09 %
OpenChat 7B	Prompt 1	91,19 %	<b>91,16 %</b>	<b>91,16 %</b>	89,92 %	<b>91,16 %</b>
	Prompt 2	87,53 %	84,08 %	83,72 %	69,10 %	84,08 %
	RAG FSL	<b>91,43 %</b>	90,43 %	90,37 %	<b>98,15 %</b>	90,43 %
Mistral 7B	Prompt 1	87,43 %	85,99 %	85,84 %	95,87 %	85,99 %
	Prompt 2	89,52 %	<b>89,38 %</b>	<b>89,37 %</b>	92,39 %	<b>89,38 %</b>
	RAG FSL	<b>89,91 %</b>	88,13 %	88,01 %	<b>98,65 %</b>	88,13 %
Llama3.1 8B	Prompt 1	<b>87,82 %</b>	<b>87,70 %</b>	<b>87,69 %</b>	90,68 %	<b>87,70 %</b>
	Prompt 2	83,88 %	77,76 %	76,57 %	99,18 %	77,76 %
	RAG FSL	78,91 %	66,24 %	61,85 %	<b>99,44 %</b>	66,24 %
Gemma2 9B	Prompt 1	93,43 %	92,86 %	92,84 %	87,27 %	92,86 %
	Prompt 2	94,44 %	94,30 %	94,29 %	91,54 %	94,30 %
	RAG FSL	<b>95,16 %</b>	<b>95,00 %</b>	<b>95,00 %</b>	<b>92,01 %</b>	<b>95,00 %</b>
Mistral-small 22B	Prompt 1	94,97 %	94,54 %	94,53 %	99,43 %	94,54 %
	Prompt 2	93,64 %	92,85 %	92,81 %	<b>99,61 %</b>	92,85 %
	RAG FSL	<b>95,79 %</b>	<b>95,66 %</b>	<b>95,66 %</b>	98,23 %	<b>95,66 %</b>
Gemma2 27B	Prompt 1	95,55 %	95,49 %	95,48 %	93,62 %	95,49 %
	Prompt 2	95,97 %	95,97 %	95,97 %	95,36 %	95,97 %
	RAG FSL	<b>96,15 %</b>	<b>96,12 %</b>	<b>96,12 %</b>	<b>97,28 %</b>	<b>96,12 %</b>
Command R 35B	Prompt 1	90,88 %	90,36 %	90,33 %	96,06 %	90,36 %
	Prompt 2	91,84 %	91,71 %	91,71 %	94,50 %	91,71 %
	RAG FSL	<b>93,91 %</b>	<b>93,45 %</b>	<b>93,43 %</b>	<b>98,58 %</b>	<b>93,45 %</b>
Mixtral 8x7B	Prompt 1	<b>94,85 %</b>	<b>94,71 %</b>	<b>94,70 %</b>	97,54 %	<b>94,71 %</b>
	Prompt 2	92,26 %	92,02 %	92,01 %	88,32 %	92,02 %
	RAG FSL	94,33 %	93,88 %	93,87 %	<b>98,88 %</b>	93,88 %
Llama3.1 70B	Prompt 1	93,54 %	92,82 %	92,79 %	99,26 %	92,82 %
	Prompt 2	92,42 %	91,27 %	91,21 %	<b>99,54 %</b>	91,27 %
	RAG FSL	<b>96,22 %</b>	<b>96,18 %</b>	<b>96,18 %</b>	97,69 %	<b>96,18 %</b>

RAG FSL approach provides the best results. According to the data produced during the experiments Llama3.2 1B as well as 3B show behavior of randomly guessing and an overall lack of capability for the given task. Event tough the model size is small in parameters, the RAG FSL combination leads to an accuracy of 95% using Gemma2 9B which performs remarkably well in comparison to the results of larger models.

Prompt 2 is not capable of improving the performance of the models noticeably. On the contrary, it produces more inaccurate classification results than the more fundamental prompt 1. Only Mistral 7B shows improvements using prompt 2, outperforming other approaches.

The larger models consistently perform better using RAG FSL. The only exception is Mistral 8x7B which achieves its best classification results using prompt 1. An remarkable conspicuity is that the larger models while using the RAG FSL only marginally outperform Gemma2 9B by an accuracy delta of 1.18%. This leads to the conclusion that after a certain point model size doesn't improve the results very much, but leads to increased resource consumption. Some of the larger models even perform worse than Gemma2 9B when looking at the F1-score, the ones performing better do this only by a very small percentage.

The proposed approach for context reinforcement using a fusion of FSL and RAG outperforms the results of prompt 2 in the experiments for most of the larger models with the exception of Mixtral 8x7B which performs best using prompt 1.

A maximum accuracy of 96.18% is achieved with the Llama3.1 70B model. The performance of Llama3.1 70B increased the most, from 92.82% and 91.27% to 96.18% accuracy, with a reduced false negative rate. Also the performance of the much smaller Gemma2 9B improves from 92.86% and 94.30% to 95.00% accuracy.

The results show that choosing the right model and methodology is crucial for the effectiveness of phishing detection. It can be concluded, that most smaller models lack the capability of using the RAG effectively in the context of phishing detection.

## 5 Conclusion and Future Work

This work evaluates how well LLMs are able to distinguish legitimate emails from phishing emails. The paper presents an approach that improves the effectiveness of detection by combining the methods of Few-Shot Learning and RAG for contextual reinforcement. The knowledge of the language model is dynamically enhanced at the time of inference by in-context and problem-specific learning without the need of computationally intensive adjustments to the actual AI model and its parameters. Experiments on a generated test dataset have shown that our approach significantly increases the recognition rate of models with fewer parameters and lower resource requirements, and outperforms previous approaches using open LLMs. This approach achieves an accuracy of 96.18% for the classification of phishing emails.

The results of this work raise further questions for future research on the detection of phishing emails with LLMs. In a next step, it should be investigated how a fusion of the RAG information source with additional data sets affects the detection accuracy. A promising approach could be the generation of phishing examples by an LLM itself, as already used by attackers. In addition, the use of other embedding models and different semantic search methods should be evaluated. It would also be useful to consider email metadata and file attachments. Furthermore, agent approaches that extend the capabilities of LLMs with functional tools, e.g. for retrieving API interfaces, could be investigated.

## References

1. Cloudflare: Bericht zu Phishing-Bedrohungen 2023. Technical report, München (2024)
2. OpenAI, Achiam, J., et al., S.A.: Gpt-4 technical report (2024)
3. Nahmias, D., Engelberg, G., et al., D.K.: Prompted contextual vectors for spear-phishing detection. (2024)
4. Crawford, M., Khoshgoftaar, T., Prusa, J.e.a.: Survey of review spam detection using machine learning techniques. *Journal of Big Data* **2** (10 2015) 23
5. Thakur, K., Ali, M.L., Obaidat, M.A.e.a.: A systematic review on deep-learning-based phishing email detection. *Electronics* **12**(21) (2023)
6. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Next-generation spam filtering: Comparative fine-tuning of llms, nlps, and cnn models for email spam classification. *Electronics* **13**(11) (2024)
7. Rosa, S., Gringoli, F., Bellicini, G.: Hey chatgpt, is this message phishing? (06 2024) 1–10
8. Heiding, F., Schneier, B., et al., A.V.: Devising and detecting phishing: Large language models vs. smaller human models (2023)
9. Hou, X., Zhao, Y., et al, Y.L.: Large language models for software engineering: A systematic literature review (2024)
10. Koide, T., Fukushi, N., et al., H.N.: Chatspamdetector: Leveraging large language models for effective phishing email detection. (2024)
11. Baumann, N., Diaz, J.S., Michael, J.e.a.: Combining retrieval-augmented generation and few-shot learning for model synthesis of uncommon dsls. *Modellierung 2024 Satellite Events* (2024)
12. Zhang, R.: Csdmc2010 spam corpus. International Conference on Neural Information Processing. <https://github.com/zrz1996/Spam-Email-Classifier-DataSet> (2010)
13. Anonymous: Phishing pot dataset. <https://github.com/rf-peixoto/phishing-pot>. (2024)
14. Jamal, S., Wimmer, H., Sarker, I.H.: An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *SECURITY AND PRIVACY n/a(n/a)* (2024) e402
15. Patel, H., Rehman, U., Iqbal, F.: Large language models spot phishing emails with surprising accuracy: A comparative analysis of performance. (2024)
16. Wang, G., Cheng, S., Zhan, X.e.a.: Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235* (2023)
17. Jiang, A.Q., Sablayrolles, A., et al., A.R.: Mixtral of experts. (2024)
18. Jiang, A., Sablayrolles, A., et al., A.M.: Mistral 7b. (2023)
19. Gemma Team, G.D.: Gemma 2: Improving open language models at a practical size. (2024)
20. Touvron, H., Lavril, T., et al., G.I.: Llama: Open and efficient foundation language models. (2023)
21. Gomez, A.: Command r: Retrieval-augmented generation at production scale. (2024)
22. White, J., Fu, Q., Hays, S.e.a.: A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv e-prints* (February 2023) [arXiv:2302.11382](https://arxiv.org/abs/2302.11382)
23. Pezoa, F., Reutter, J.L., Suarez, F.e.a.: Foundations of json schema. In: *Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee* (2016) 263–273
24. Brown, T.B., Mann, B., Ryder, N.e.a.: Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20* (2020)
25. Lewis, P., Perez, E., Piktus, A.e.a.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20* (2020)
26. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. (2019)
27. Singhal, A.: Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* **24** (2001) 35–43
28. Pedregosa, F., Varoquaux, G., Gramfort, A.e.a.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12** (2011) 2825–2830