AI for NDE 4.0 - How to get a Reliable and Trustworthy Result in Railway Based on the New Standards and Laws.

D. Kanzler[1,a], M. Selch[2,] G. Olm[3]

[1] Vrana Gmbh, Rimsting
[2] Deutsches Zentrum für Schienenverkehrsforschung, Dresden
[3] Technische Universität Berlin, Berlin
[a]*Email: KanzlerD@av-ndt.com*

Abstract

The potential of artificial intelligence (AI) in our modern society is virtually boundless. However, alongside this potential, we are witnessing an increase in challenges and risks within the field. In Europe, these concerns have spurred discussions leading to the development of the AI Act, a European law designed to harness the potential of AI technology while safeguarding personal rights and security.

This article will delve into the significance of AI in non-destructive evaluations (NDE) and (also) the necessary steps to establish reliable AI solutions. It's essential to note that this process should not be perceived solely as a regulatory requirement but as an opportunity to enhance value, ultimately enabling the creation of innovative maintenance concepts. As an illustrative example, we will explore the use of AI technologies in rail testing, a part of the ongoing AIFRI project in Germany.

Keywords: *NDT; Reliability; POD; AI; Railways; predictive maintenance*

1 INTRODUCTION

When embracing any new technology, the central question often revolves around on how to harness its full potential while mitigating its drawbacks. Achieving a perfect balance is admittedly impossible, but history has shown that humanity has consistently embraced significant developments such as the wheel, the letterpress, and various industrial revolutions. Artificial Intelligence (AI) stands as a transformative technology with the potential to reshape our world in profound ways. This very potential prompts people to consider AI in contexts like Industry 4.0 and Non-Destructive Evaluation 4.0 (NDE 4.0).

However, it's crucial to understand that NDE 4.0 represents more than just adopting a new technology; it signifies a revolution. It's about creating an environment that fosters seamless collaboration between humans and technology. AI plays a very important role in NDE 4.0 by revolutionizing the way the inspections are carried out for the integrity of materials, structures, and components. AI algorithms can be applied to various NDE methods, such as ultrasonic testing, radiography, and eddy current testing, to automate and enhance the inspection process. By analyzing vast amounts of data with remarkable speed and precision, AI can detect defects that might be challenging for human inspectors to identify. This leads to several benefits, including increased inspection efficiency, reduced human error, and the ability to detect certain defects which may be missed by manual inspections. Additionally, AI-driven NDE can facilitate in improving safety, and reducing downtime in industries like aerospace, manufacturing, and infrastructure.

Even though, AI presents significant advantages in NDE 4.0, there are potential drawbacks and dangers associated with its adoption. One major concern is that the excessive dependence on AI systems might lead to reducing the expertise of human inspectors and the critical evaluation needed for certain inspections. Moreover, the use of AI in NDE depends heavily on the quality and diversity of the data on which the AI is trained. Any discrepancy in the training data could influence the AI decision-making, potentially leading to false positive or false negative results. There is also the problem of security issues, in the event of these AI systems are hacked or manipulated. This leads to the loss in integrity in inspection processes and the safety of critical infrastructure.

This is precisely why discussing the regulation of AI in the realm of NDE is not only necessary but also highly valuable. It's about establishing the right framework to ensure that the incorporation of AI into NDE processes is not only regulated but optimized for the benefit of all.

1.1 Standardizations and regulation in Europe

The European AI-Act [1] is currently in the process of being reviewed. This act reflects the European approach to harness the potential of AI technologies while remaining in alignment with European values and basic rights. This broad development spans various sectors, from medical diagnostics to financial decision-making. While the field of NDE is a relatively small part of this extensive discussion, it is still valuable to examine the processes and incorporate beneficial approaches from other domains.

The utility of these approaches largely depends on the specific application area. Therefore, the initial discussion will focus on general aspects, such as the trustworthiness and applicability of the AI-Act to NDE, with a particular emphasis on railway testing. Despite the potential risks associated with AI misuse, the AI-Act provides a mechanism for risk analysis within its designated scope. It categorizes AI applications on a scale ranging from minimal risk to high risk as well as unacceptable risk.

From the perspective of NDE, rail inspection plays a crucial role in ensuring the integrity of critical railway infrastructure. As per the AI-Act under Annex III, railways are not classified as critical infrastructure, whereas streets and highways are considered highly critical components of our society. Nevertheless, the comparable nature of these sectors underscores the importance of careful consideration, especially when taking into account that the Act is still under development.

The Act also mandates a declaration of conformity, which is not new in the field of NDE. However, the implications of the AI-Act for users need to be understood. The Act delegates the responsibility of planning AI approaches, making decisions, and defining the various fields of application to the member states of the union. In Germany DIN, an independent platform for standardization in Germany, is concerned with this topic. From their perspective, the DIN Normungs roadmap [2] serves as a valuable tool for comprehending the requirements and needs in the implementation of AI. DIN provides a platform for experts to discuss AI's requirements within different sectors like medicine, automotive, and more.

While the NDE domain might not currently be featured in the roadmap, the fundamental concept of trustworthy AI has been defined, offering a useful framework for dealing with AI technologies. In addition, the roadmap is continually evolving, which suggests that AI for NDE may be included in the near future, as discussions progress and requirements become clearer.

1.2 Trustworthy AI

The concept of "Trustworthy AI" is at the core of NDE 4.0 and is closely related to established standards and it revolves around the fundamental question of how much we can rely on our NDE system. It's important to recognize that NDE itself doesn't change the reliability and functionality of a technical component or critical structure. Instead, it is often the primary means of obtaining information about their safe usage and performance. To achieve this, a functional quality management system as well as information on the trustworthiness of the NDE application, are essential. A failure in NDE doesn't directly lead to the failure of the component, but an unreliable testing system can result in unnecessary costs, and undetected potential threats may have to be addressed through design modifications or reductions in the component's lifetime.

For trustworthy NDE, the 2nd European American Workshop on the Reliability of NDT introduced the term "reliability." NDE reliability is defined as the extent to which an NDT system can effectively accomplish its objectives, regarding detection, characterization, and minimizing false alarms. This concept also led to the creation of the modular model (as shown in Fig. 1) of NDE, which includes intrinsic capability, application factors, human factors, organizational context, and the influential category of algorithms [3]. Within the subsection of algorithms, a variety of influences related to AI play a significant role, connecting reliable NDT to trustworthy AI.

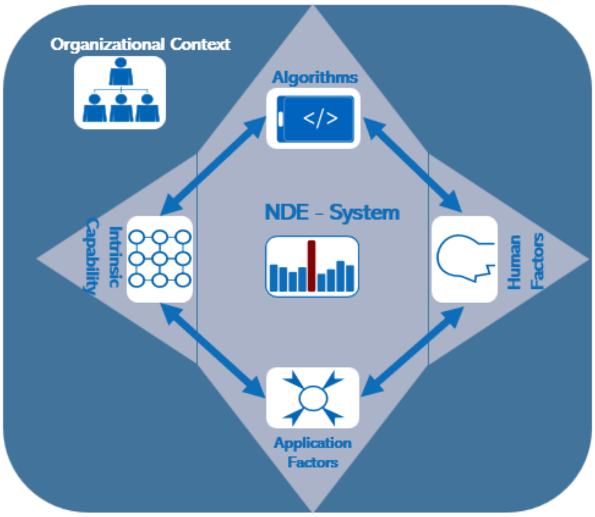

Fig.1 Modular Model for the Reliability assessment for NDE

Trustworthiness for AI is defined [4] by attributes such as fairness, autonomy and control, transparency, bias analysis, robustness, reliability [5], safety and security, as well as data protection. Here's how these attributes are relevant to NDE:

Fairness: Although fairness in the context of AI may not directly apply to NDE, it relates to the quality of data and coverage. Additionally, the collaboration with the human operator should be considered.

Level of Autonomy and Control: Current AI approaches aim to support human operators, who have the final say in testing system decisions. This Human-in-the-Loop approach enhances safety by requiring human confirmation, but it also places more emphasis on operator understanding of the method, raising concerns about issues like Automation Bias [6].

Transparency: NDE applications have an advantage in terms of transparency. Complex NDE applications often rely on the interaction between potential defects and fundamental physical testing knowledge to assess the testing system's capability. Metrics like the Probability of Detection offer transparent assessments of testing systems.

Bias Analysis: Evaluating the AI system for bias issues by examining how it performs across different demographic groups, materials, and defect types.

Robust Testing: Subjecting the AI system to a range of challenging conditions, including noisy data, variations in lighting or environmental factors, and difficult-to-detect defects, to assess its robustness.

Reliability: Evaluating the robustness of AI applications and estimating uncertainty, particularly the risk of making false decisions, is crucial for assessing reliability. In NDE, the Probability of Detection is a unique statistical metric used to assess reliability, especially in classification tasks where an incorrect prediction of the absence of a critical defect can have significant consequences, such as impacting railway service.

Safety and Security: While NDE AI applications don't directly cause physical harm to people, AI-based decisions can lead to unsafe conditions in the technical domain, akin to decision-making in medical diagnostics. The discussion of data security in NDE is unique and complex, often involving political considerations, due to the fact that the knowledge of testing results corporate secrets can be disclosed within NDE data.

In summary, ensuring trustworthy AI in NDE is a complex task, but it is essential to guarantee the reliability, safety, and security of critical infrastructure and technical components.

2 RELIABILITY ASSESSMENT FOR AI IN NDE

2.1 Reliability Assessment Concept for NDE

As previously mentioned, the typical metrics used to assess AI performance often involve Receiver Operating Characteristics (ROC), which include sensitivity and specificity, resulting in quantified characteristics like the Area under the Curve (AUC). Additional metrics such as the F1-Score are also commonly employed. However, all of these metrics overlook a significant difference between the use of AI in, for example, medicine and its use in NDE within maintenance programs.

In fields like medicine, even the slightest indication of a tumor is considered critical, but in context of the damage-tolerant concept in the technical field, certain defects, such as cracks, may not pose an immediate safety threat as long as they do not reach the critical crack size. Therefore, the size of potential critical defects is a crucial consideration in NDE. The question of when a testing system is capable enough to detect a signal resulting from the interaction of the defect, in relation to the defect's size, can be addressed through the Probability of Detection (POD).

The POD leverages the physical relationship of the testing method to describe its ability to detect defects. It takes potential interference into account which can introduce noise and data scattering, that may hinder defect detection. This relationship can be visually represented in the a (defect size) vs. â (signal response) graph (as shown in Fig. 2), which forms the basis for POD calculations, based on the decision threshold according to the distribution of the noise amplitudes, as depicted in Fig. 3.

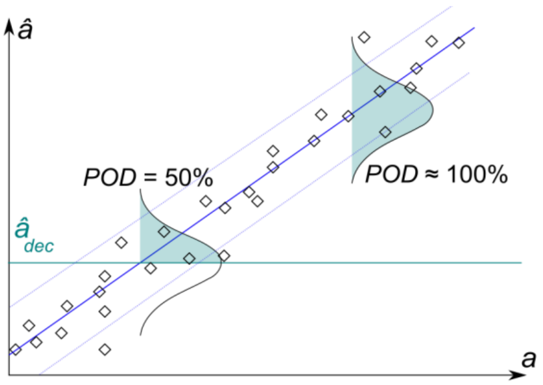

Fig.2 *â vs a graph for a POD evaluation*

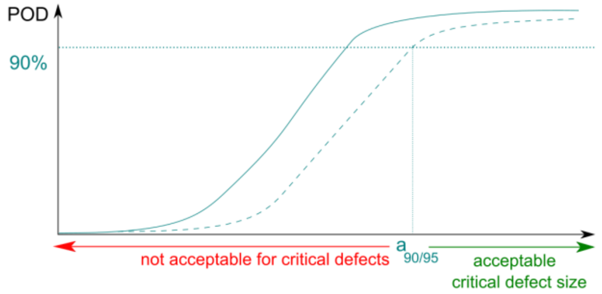

Fig.3 *POD graph an assessment of the NDT method*

The POD's ability to describe the physical relationship also contributes to the transparency aspect of AI performance assessment.

While the POD appears to be a suitable choice for evaluating AI in NDE, there are some aspects that warrant (a) discussion. As a statistical metric, the POD is effective under specific conditions and can be susceptible to misuse. Therefore, the demand for objectivity and potential third-party assessments, as stipulated in the AI-Act, is also applicable in this context. Organizations that offer objective assessments of NDE applications are limited. However, understanding how to establish and interpret PODs is essential, and fostering scientific discourse on these approaches within the community is necessary. Even though, POD assessments can be guided through various standards like MIL-HDBK 1823A, ASTM E3023/E2862, ENIQ Report no.41, etc. there are still many methods and techniques that are being used by several researchers. Many advanced methods like, multiparametric, model assisted POD methods, transfer function, etc. could not be easily implemented by individual organizations. Especially in the context of condition monitoring technologies, like structural health monitoring (SHM), NDE 4.0, etc., reliability assessments are extremely challenging. Hence, platforms for such discussions include the ICNDT Specialist International Group "NDT Reliability" and regular International Workshops on the Reliability of NDT/E.

2.2 Reliability Assessment Concept for AI

In assessing the specific AI process, it's essential to consider the well-established fact that the quality of data available for AI/Machine Learning (ML) algorithms, such as those utilized in the AIFRI project, is of paramount importance. As discussed in the preceding paragraphs, the Probability of Detection (POD) is a fitting metric for evaluating the use of AI in the field of NDE. As explained previous in Section 2.1, POD can be obtained based on the decision threshold applied to the signal response data. This decision threshold is, again, dependent on certain conditions on the distribution of noise amplitude data for a given false positive rate. Nonetheless, during the training phase of an AI system, it's crucial to remember that the unknown threshold established by the ML system, has a substantial impact on the system's detectability and false alarm rate.

For direct comparisons between different methods during the training phase, the general concepts of AI evaluation come into play. This means that the evaluation of an ML system for NDE is structured into two distinct phases. In the training phase, the ML system is assessed using Receiver Operating Characteristic (ROC) analysis to identify the optimal method based on factors like detectability and the false alarm rate (as shown in Fig. 4). In contrast, during the validation phase and the practical use (deployment) phase, the focus shifts towards evaluating the criticality of defects in terms of different metrics like size, area, volume, etc. As such, the ML system's performance is assessed using the POD, which aligns with its ability to accurately detect critical defects in real-world scenarios. This two-phase evaluation approach ensures that AI systems are both capable of detecting defects and reliable in practice.

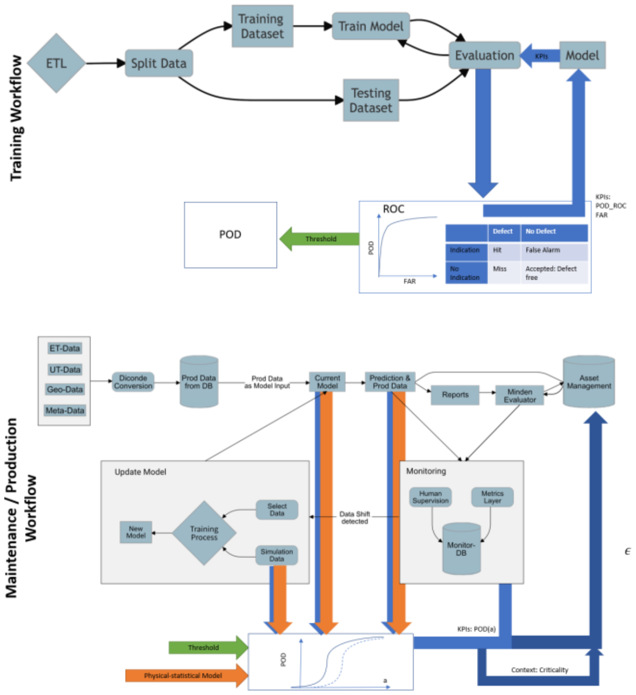

Fig.4 *Two-phase Data and Validation Concept with AIFRI: The Trainings Workflow and the Deployment Workflow*

3 MODEL DEVELOPMENT

3.1 Practical use of the AI evaluation in Railway

The state of railways in Germany is a critical topic, with implications for both sustainable travel and overall transportation quality. While it presents potential for eco-friendly travel across central Europe, the current quality of the transportation system is unsatisfactory. The high frequency of train delays, cancellations [7], and even derailments [8] indicates a need for significant improvement. It's worth noting that the costs associated with train derailments can be exorbitant, making it an imperative for infrastructure companies to ensure a safe and well-maintained rail network.

One of the complexities in the German rail system is its mixed use, with cargo, high-speed trains, and various other trains sharing a significant portion of the network. This variety in demands on rail applications needs a robust maintenance system to avoid unnecessary maintenance steps such as repairs or replacements. A key element in this context is the reliability knowledge derived from NDE techniques. The buzzword often heard in this context is "predictive maintenance." AI-driven NDE can facilitate predictive maintenance, extending the lifespan of critical assets, improving safety, and reducing downtime in railways due to maintenance activities. Apart from railways, the concept of predictive maintenance can multiply benefits to several industries, like aerospace, manufacturing and civil infrastructure.

Predictive maintenance is indeed an appealing concept, but its feasibility relies heavily on the reliability of the testing system. Without reliable data about probable defect size within the rails, without prioritizing and implementing predictive risk management strategies, the concept of predictive maintenance, becomes challenging.

One of the challenges stems from the fact that different stakeholders and engineering departments operate independently, causing a gap in efforts to combine all available information into a unified methodology. The AIFRI project endeavors to address this issue, aiming to gather information from various sources to tackle the future challenge of cost-effective rail maintenance and testing.

The testing process involves specialized testing trains that travel at high speeds over the rails, using Ultrasonic and Eddy current testing applications. The evaluation of the data is carried out separately, with decisions made about the need for further actions, such as detailed testing, speed restrictions, or section blockades. Within the evaluation phase, AI could potentially support human operators by expediting data analysis and improving its reliability.

Another issue is the limited number of testing trains responsible for the extensive rail grid in Germany. Currently, testing intervals are primarily determined by fixed time periods. However, there is a significant potential in shifting from fixed intervals to predictive variable time intervals. This transition must be highly accurate to ensure that critical rail sections are not left untested. Furthermore, due to outages of testing trains or the respective personnel, many scheduled test runs have to/had to be cancelled and must be rescheduled at short notice.

3.2 Evaluation results as the basis for planning test runs on railway tracks

The current scheme for scheduling rail test runs consists in fixed inspection intervals dependent on the maximum speed and ranges from four to 24 months. In this way, all track segments to be inspected, are classified into four different groups.

With the help of the results of the AI-based defect detection, a more individual and criticality-related approach can be derived. The POD analysis delivers an estimate of the size of small defects, which cannot be reliably detected by the testing system. The size corresponds to the $a_{90/95}$ threshold of the respective defect type, see Fig. 3.

Based on the information gathered in this project, an initial step involves classifying rail sections into different groups built on factors like the time since the last test, environmental influences, and testing train availability. Classification relies on mathematical approaches to plan the testing train routes and factors in the capability of the testing system in combination with the predicted behavior of potential defects within the rail.

For instance, considering a theoretical crack in the rail. Over time, this crack has the potential to grow due to the loads on the track and material behavior. The probabilistic nature of crack propagation behavior, combined with the probabilistic information about their detectability (POD), leads to a highly probabilistic situation. If the crack continues to grow without maintenance, it could lead to a catastrophic event before reaching the end of its life (EOL). Depending on the time and the potential for failure of a rail section, it's possible to calculate a metric that helps prioritize actions, a dimensionless quantity, in Fig. 5 called □. This metric □ considers material parameters, load conditions, and the capabilities of the testing system, making it an ideal tool for planning testing train operations.

The initial value of the metric depends on the applicable extent of potentially non-detected defects. While no further maintenance is conducted on a specific rail segment, the metric increases according to the growth of the assumed non-detected defects. After a new test run, the metric is set back to the initial value, since present larger defects would have been detected by the testing system, see Fig. 6.

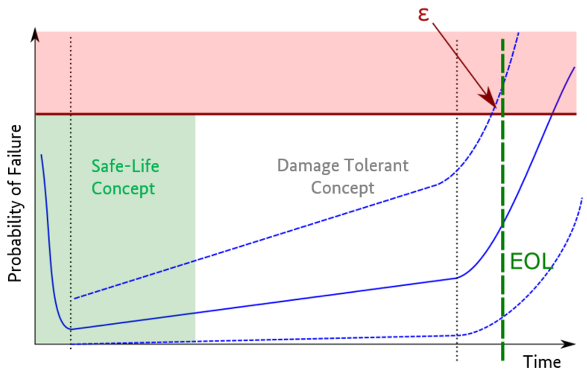

Fig.5 *Model for the Probability of Failure over time*

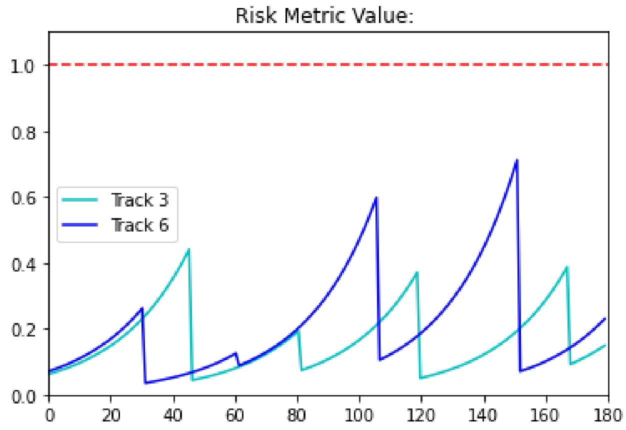

Fig.6 *Evolution of□□-values for two track segments over a horizon of 180 time periods with value set-backs at the times of track inspection*

In this way, an individual growth behavior of the criticality of each track, untested over a certain period of time, can be represented. These curve progressions, collected for all tracks of a network, input as objective parameters into a mathematical optimization model based on the models from literature [9,10]. While ensuring various planning constraints, combinatorial optimization can be used to address various objective criteria. Considering the □-metric, minimizing the aggregated risk of all tracks at each time during a planning period (security aspect) or aiming to keep the curve values under a threshold by performing as few test-runs as possible (economic aspect). At the same time, the schedule should form the most efficient and possible rotation of the rail testing trains, for which personnel availability must also be considered.

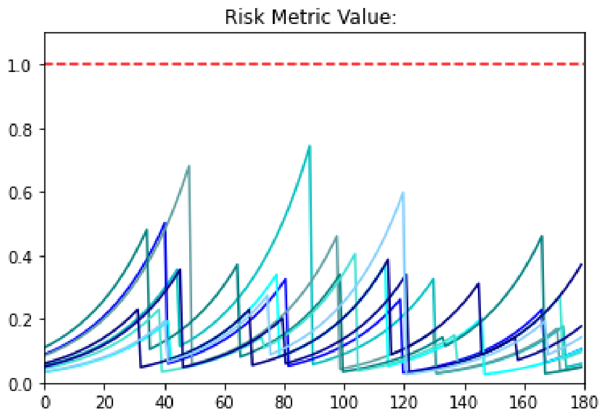

Fig. *7 Simulation of a regional network in eastern Germany with the objective of keeping the quadratic sum of the □ values of each track and time as low as possible for a planning window of 180 periods.*

4 RESULTS AND DISCUSSIONS

4.1 Reliability Framework

AI in NDE holds great promise for improving reliability, safety, and cost-effectiveness in various sectors. While AI has the potential to enhance NDE, it should be integrated thoughtfully and in conjunction with human expertise to mitigate these drawbacks and maintain the highest standards of safety and quality. Reliability assessment of AI plays an important role for improving safety and quality. Indeed, the Probability of Detection (POD) is a powerful and widely used tool with strong mathematical foundations for reliability assessments. It has been successfully customized for specific applications, even in high-reliability and safety-critical sectors, such as the nuclear industry and aviation. These approaches align well with the requirements set forth by the AI-Act and the Normungs roadmap.

4.2 Reliability data within AIFRI

A critical question remains for AI and NDE: What data do we need, and how much of it can we effectively use? While many AI applications have access to abundant data, even in the medical field, the technical component data, specifically those containing critical defects, can be quite scarce. This means that achieving trustworthy AI in NDE hinges on several factors, including the type of data, the extent of data coverage, and the overarching concept of "fairness" in data acquisition.

In the medical field, where vast amounts of data are available and some degree of comparability exists, NDE struggles to find relevant data on critical defects and achieve data comparability between components. However, in NDE, there is an opportunity to generate simulations or modeling tools as a data source for training and, to some extent, validation. Simulations allow for the creation of a substantial amount of data. Still, it is imperative to validate and benchmark this data to ensure that it accurately performs on real-world scenarios and is suitable for training AI systems effectively.

But first, the problem of imbalanced data in this research area has to be addressed. The majority of data are collected represents perfectly normal rail and simply taking all available measurements, could prevent the model from learning important patterns. The baseline for resampling training data is the number of artifacts that can be found in the measurements, which can be welds or drillings found in Ultrasonic data. From that information, it is possible to sample an appropriate amount of plain data to compose a meaningful data set suitable for training a model. The necessary amount of simulations is created with the same considerations in mind. All classification types need to be distributed evenly among each other to prevent the formation of biases.

In the future it is necessary for creating Deep Learning models with the help of simulation, the following questions has to be answered to prepare experiments and data for model validation and practical generalization. What should be learned that is in the data? With experts, artifacts and defects should be prioritized. Especially defects that are essential for the model to classify, e.g. cracks at drillings in 45° or 90° orientation, which will be simulated and serve as training data for the model. Hereby, we explicitly define and simulate which knowledge the model as to acquire. The next is designed to clarify the missing knowledge of the mentioned model. In reality, cracks are not necessarily limited to a 45° or 90° angle, but also might occur in between or outside of that range. In this case, cracks can be simulated with angles outside of the predefined range. They will not be part of the training data whatsoever and are solely considered during model evaluation. This way it is possible to examine how well the model can generalize and classify data outside of the explicit pattern it was trained on. Finally, it is to consider, whether the method introduced a possible bias towards simulated data and searched for patterns, that should not be mastered but that can occur in the data. Although the simulated data appears like truthful measurements, it cannot be ruled out that subtle differences are learned by the model. The models are not observable in the first place and cause a model bias towards simulated defects and fail with practically measured defects from the field. There is the need for specific artifacts, that are apparent in simulated data and field data alike. This enables to compare the model performance between both sources and evaluate further bias tendencies.

In summary, while data scarcity is a challenge in NDE, the ability to create simulations, offers a promising way to supplement this shortfall, but it also requires validation to ensure its applicability in real-world scenarios.

5 CONCLUSIONS

The AIFRI project is still ongoing, and it has achieved initial success in developing a model that interconnects various decision-making aspects within an objective decision-making process. The next phase of the project involves the following steps:

- **Development of the AI Technique:** The project will focus on further developing the AI technique, building upon the foundational work that has been accomplished.

- **Validating the AI Technique with Real Data:** To ensure the AI technique's effectiveness and reliability; it will be rigorously validated using real-world data. This step is crucial to demonstrate the AI model's practical applicability.

- **Development of a Train Planning Approach for a Specific Region in the German Rail Grid:** A specialized train planning approach will be created for a specific region within the extensive German Rail Grid. This region-specific approach will consider the unique requirements and characteristics of that area.

- **Discussion of Planning Results and Customization According to Non-Technical Requirements:** The project team will engage in discussions regarding the planning results and make necessary customizations to align the approach with non-technical requirements. These may include considerations related to regulations, policies, and other factors that influence decision-making beyond technical aspects.

These future steps will contribute to the project's ongoing efforts to enhance rail network maintenance and testing efficiency, ultimately improving the quality and safety of railway transportation in Germany.

ACKNOWLEDGMENT: The project about the real application is called AIFRI. In the collaboration project, funded by the research program mFUND IT, tools are used to improve the maintenance of rail-infrastructure. The project-partners are DZSF, DB Netz AG, Federal Institute of Materials Testing and Research (Bundesanstalt für Materialforschung und - Prüfung), TU Berlin, ZEDAS GmbH and Vrana GmbH. The project is funded by the Federal Ministry for Traffic and Digital Infrastructure. Thanks to all project partners for the real data (DB Netz), the support for the simulations (BAM) and the integration of the approaches (ZETAS). Thanks to Vamsi Krishna Rentala (Applied Validation of NDT) for the support in creating the article.

6 REFERENCES

[1] European Commission 2021 "Proposal for Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts" Document 52021PC0206 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[2] DIN/DKE 2022 "Deutsche Normungsroadmap Künstliche Intelligenz" Edition 2 https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen

[3] Kanzler, D., Rentala V.K. 2021 "Reliability evaluation of testing systems and their connection to NDE 4.0." Handbook of Nondestructive Evaluation 4.0

[4] Poretschkin, et al. 2021 "Guideline for Designing Trustworthy Artificial Intelligence – AI Assessment Catalog" Fraunhofer IAIS www.iais.fraunhofer.de/en/ai-assessment-catalog

[5] Rentala, V.K, Kanzler D., Fuchs P. 2022. "POD evaluation: the key performance indicator for NDE 4.0." Journal of Nondestructive Evaluation 41.1

[6] Bertovic, Marija 2016. "A human factors perspective on the use of automated aids in the evaluation of NDT data." AIP conference proceedings. Vol. 1706. No. 1. AIP Publishing.

[7] Spiegel 03.06.2023 “Verspätungen bei der Deutschen Bahn so häufig wie seit Monaten nicht“ https://www.spiegel.de/wirtschaft/unternehmen/verspaetungen-bei-der-deutschen-bahn-so-haeufig-wie-seit-monaten-nicht-a-c93ef48c-69ea-417c-bd97-8d4899f26ea4

[8] Dokus im Ersten 2023 „Sicher Bahnfahren!“ https://www.daserste.de/information/reportage-dokumentation/dokus/videos/sicher-bahnfahren-video-100.html

[9] Peter Nganga Muchiri, Liliane Pintelon, Harry Martin & Peter Chemweno (2013): Modelling maintenance effects on manufacturing equipment performance: results from simulation analysis, International Journal of Production Research, DOI: 10.1080/00207543.2013.870673

[10] Mohammad Rezaeimalek. Inspection planning based on preventive maintenance condition in a serial multistage manufacturing system under uncertainty. Ecole nationale supérieure d’arts et métiers - ENSAM; University of Teheran, 2019.