Generalising AI model performance for Non Destructive Testing in railway systems

Georg Olm[1]
[1]Department of Civil System Engineering, Technische Universität Berlin
E-mail(s): georg.olm@tu-berlin.de

Abstract:

Maintaining railway systems is an essential task during operation of such public infrastructure, but the methods have remained unchanged in Germany for decades. Advancements in ML modeling, Simulation Techniques, and Data Management enable possibilities to accelerate the shift to more efficiently digitized evaluation processes. This research aims to analyze the capabilities of generalization of a previously identified Deep Learning model, trained on real-field ultrasonic data from regular inspection runs, and augmented by simulated defects. As this cannot be done by the usually applied summary statistics, it is necessary to analyze the model performance following the questions of how well the model learns data pattern explicitly, can interpolate between the trained parameters, and whether it did not overfit on specific simulation pattern. This is done by applying selective sampling strategies and an analysis of interesting results. The findings indicate a good fit for explicit patterns with an AUC of 0.964, while interpolation between parameters is successful only in specific use cases. Finally, we conclude that the model likely learns background patterns from simulations and may not necessarily apply simulated defects to real-world scenarios.

Keywords: Machine Learning, Railway Maintenance, Sensor Data, NDE

Erschienen in Tagungsband 35. Forum Bauinformatik 2024, Hamburg, Deutschland, DOI: 10.15480/882.13521

1 Introduction

Deep Learning-enabled AI models provide the potential to perform many tasks efficiently and can enhance the implementation of digitized processes. Rail inspection, essential for maintaining structural integrity and ensuring reliable infrastructure, has seen little change in methodology in recent decades. Evaluators still need to manually examine a lot of data collected on the maintenance train. Ultrasonic testing remains the primary data source for detecting internal rail defects. Evaluators must go through a long training period in order to learn the complex interactions between different signals and to sufficiently distinguish between normal expected indications and actual defects, which require actions. Once trained for the task, assessors can very quickly evaluate a single ultrasonic image for defects. Complex interactions in a large amount of data, which can be examined easily with experience, describe a prime use case for deep learning.

1.1 State of research

The field of ultrasonic-based railway maintenance, with its large volume of data requiring manual evaluation, presents significant opportunities for deep learning model applications. The research by Tang, De Donato, Besinović, *et al.* [1] and Chenariyan Nakhaee, Hiemstra, Stoelinga, *et al.* [2] provides an overview of how AI is used in the field of railway maintenance. Both emphasize the possibilities for AI methods in railway maintenance and see a growing number of published research in that field. They also see a lack of sufficient and meaningful data to train models suitable for the real world in the field spanning across most publications. Real defects, in particular, are scarce in general datasets. A common approach to overcome this problem is to generate synthetic data. Uhlig, Alkhasli, Schubert, *et al.* [3] conducted a review study on how synthetic and augmented ultrasonic data are used in multiple domains of NDE. They point out the huge potential of this approach as it enables meaningful training data sizes in the first place, but need to acknowledge that despite increasing interest in the last years its application remains low, compared to overall publications about AI in NDE. Using synthetic data for railway maintenance has not been explicitly researched by them, although it is generally considered part of the NDE field. Nevertheless, respective research has been conducted in the railway domain, but not in a way that reflects the applicability in the real world [4], [5].

This research focuses on utilizing large amounts of real field testing data from actual inspections, which contain all noise and ambiguous signals and allow real-world conclusions to be drawn. In addition, the lack of defect data in the field data will be overcome by introducing simulated defects into the training process. Both solve common gaps in this field of research.

Data collection followed the methods described by Armbruster, Heckel, and Fenger [6] and simulated according to Zhang and Heckel [7]. Neural networks are typically evaluated by looking at aggregate statistics that describe the model performance, e.g. precision, recall, area under the curve (AUC), or true positive and false positive rates (TPR, FPR). Before this research, a suitable model architecture has already been identified by assessing these indicators. The results can be seen in Table 1. It consists of six blocks of bidirectional LSTM layer, with 115 hidden units each including a residual step, which is followed by a multi-head attention layer following typical architecture from the literature [8], [9].

Table 1: Model performance overview of already identified model architecture. This model is used to during the research of this paper.

AUC	TPR-Defect	FPR-Defect	TPR-Artefact	FPR-Artefact
0.964	0.894	0.001	0.962	0.04

2 Research Method

Although these values look promising, considering proper hyperparameter tuning still offers capabilities for improvement, they need to be evaluated cautiously. Based on these summary metrics, it cannot be fully understood how well the trained patterns, especially through mutually exclusive division of defects and artifacts between simulated and field data, can be applied in the real world. Therefore, it is necessary to explore finer-grained parameters, to examine the model's capability to generalize

between the real world and the simulated pattern. This examination follows questions that are deriving from exploratory principles found in the basic literature [10], [11].

1. How well is the model classifying the learned pattern? – **Learning Explicit**
2. Can the model interpolate on data it has not explicitly been trained on? – **Learning Implicit**
3. Did the model prevent to learn undesired patterns? - **Not Learning**

The first question resembles the traditional deep learning questionnaire the most and can be answered through the summary statistics already assessed in Table 1 and a further confusion matrix in Figure 3.

Simulated defects are described by physical parameters (e.g., length, height, and angle) in greater detail than can be assessed through the evaluation of actual test data. Computational and time constraints prevent simulating all possible parameters. The model can only be trained on a finite set of defects and therefore needs to be able to interpolate between these values. Assessing this performance requires a very specific sampling strategy before training, which enables a statistical comparison of predictions within the sampled parameters and outside of them.

Finally, it is critical to differentiate classes between simulations and field data. It should be assumed that specific patterns are distinct in the simulated data, caused by the simulation process. Based on the currently available data, almost all artifact classes are from the field data and all defect classes are simulated. This differentiation is completely hidden in the summary statistics. Based on these assumptions, it is necessary to check whether the model is actually capable of classifying the defect data in reality or just detecting specific simulations. This must be taken into account in the preparation of the data and the artifact types sampled accordingly.

2.1 Data Preparation

Source data consists of 1,144 m of simulated data and 14,656 m of selected field data around indications, both of which will be converted into windowed sequences of size 115 mm with a stride of 10 mm. Depending on the actual question and task, it is necessary to define metrics that can be observed to assess the model's performance for not trained on testing data sets. **Explicit learning** on artifact detection and classification will be measured with ROC-AUC, combined with True Positive Rate (TPR) and False Positive Rate (FPR) for artifacts and defects. These fundamental results can already be seen in Table 1.

Testing for **implicit learning** results demands a more refined sampling strategy. Because the goal is to see out-of-training sample prediction on physical parameters, such as length, height, angle, it makes sense to only sample specific values for training and validation and keep a holdout set for testing. The amount of simulated data is currently still limited and the most variability is in the simulations length parameter. The sampling strategy is depicted in Figure 1. Variations of defects with drillings of lengths 3, 12, and 24 millimeters will be in the training and validation data, whereas lengths 6, 9, 15, 18, and 21 millimeters only occur in the testing dataset.

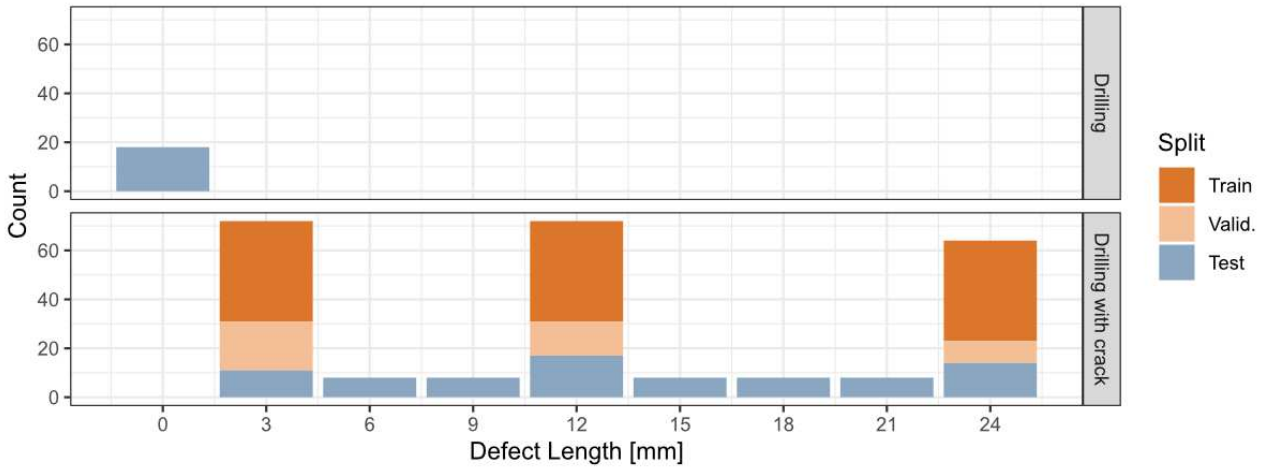

Figure 1: **Simulation Split** between training and testing the model. Data points with parameters used during training are further considered as *In Sampling* data and parameters, which are exclusively in the testing data will be considered as *Out of Sampling* data. This follows the question of the models implicit learning capabilities.

To assess whether the model learned an undesirable pattern, a special holdout simulation of a drilling without a crack has been created, and its sampling can also be seen in Figure 1. This artifact frequently occurs in the field data and its pattern should be well learned by the model. Consequently, misclassifications of simulations should closely align with those observed in field data.

3 Data and Results

3.1 Model Performance

The performance metrics of the model, based on Table 1 suggest a promising ability to learn the ultrasonic pattern, in general. An AUC of 0.964 confirms this, although a proper trade-off between TPR and FPR needs to be determined by domain experts before a final conclusion is possible. FPR and TPR are of the same categories. The values indicate a practical usability, but a deeper evaluation of why individual classifications happen or fail should be done before an assessment can be concluded. Looking at the FPR for artifacts of 0.04 demands special attention. Due to the very large number of normal rails in the inspection process, 4% false positive can lead to a large number of unnecessary signals for human evaluation. This value needs to be evaluated on a larger sample of measurements to fully assess the impact of defect and artifact detection on the overall performance of the model in detecting and classifying indications.

3.2 Parameter prediction

Figure 2 compares the difference in predicting in-sample and out-of-sample length parameters of rail indications, specifically drillings with crack under four different angles 0, 45, 90 and 135 degrees starting from a vertical crack on top of the drilling. Separation by angle is based on the underlying question, whether performance is dependent on the angle of crack expansion. The relative error is the highest for 0º angle, which is justifiable, as a vertical angle is hard to fully capture with the ultrasonic

angle beam probe. It also has a very apparent difference between the in- and out-of-sample values, which is much closer for the remaining categories. The relative error for all estimations seems very high, with typical relative errors leading up to 100% of their actual length.

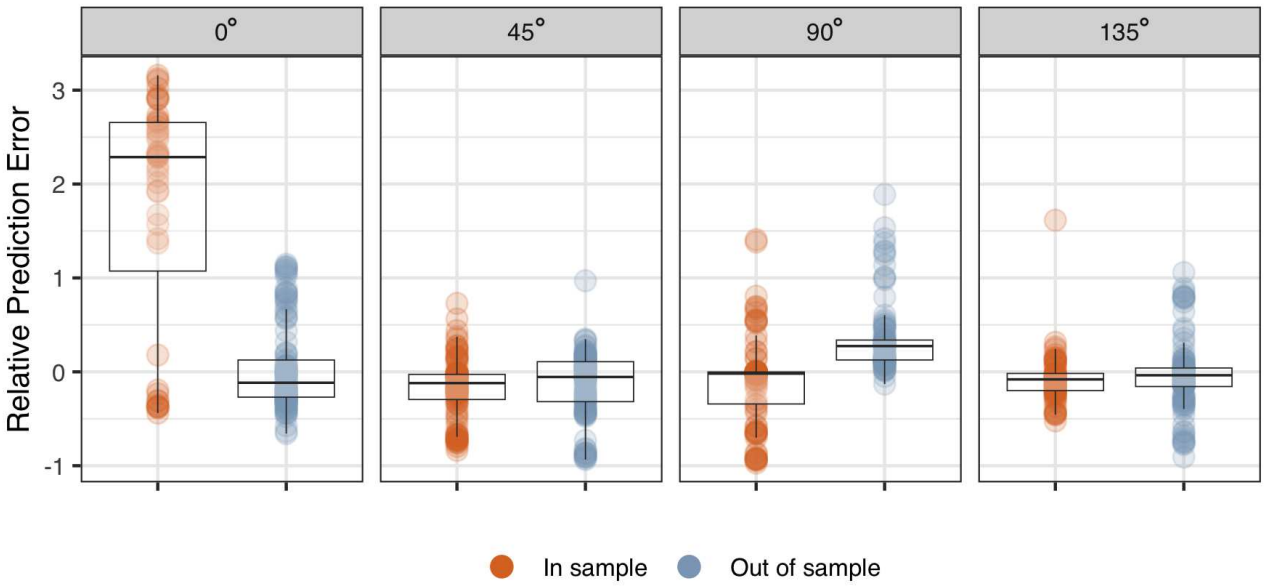

Figure 2: Comparative scatterplot of relative **Prediction Error Comparison**. Data grouping by angle of crack expansion. Color indicates whether the data defect parameter where also in training set or exclusively in test set. Data points augmented by boxplots to increase interpretability.

Table 2 describes the effect of sampling per angle based on Welch's t-test. For the 0º angle group we see the largest effect for out of sample estimation of around -1.701 and very low effects for 45º and 135º. The effects of sampling are very significant for 0º and 90º and not significant 45º and 135º.

Table 2: Influence of sampling strategy on error by angle

angle	effect term	estimate	conf.low	conf.high	std.error	p.value
0º	out of sample	-1.701	-1.974	-1.428	0.138	1.34e-24
45º	out of sample	0.036	-0.054	0.125	0.046	0.432
90º	out of sample	0.432	0.314	0.551	0.060	1.43e-11
135º	out of sample	0.041	-0.060	0.141	0.051	0.426

3.3 Effect of simulation pattern

Figure 3 provides more details for individual classifications. As can be expected from Table 1, most artifacts are classified correctly. Each failing classification should be closely examined to assess the actual model behavior, but the current research question is about the classification difference between simulations and field data. With regard to drillings with cracks, there is only one misclassification as a drilling. Considering cracks with a length of 3 mm are hardly visible in a data resolution of 3x3 mm, much more of these classifications would have been expected. Regarding the exclusively held out

simulated drillings, all are misclassified as drilling with crack. For simulated head defects, the results are similar. No misclassification of head defects as drillings or welds can be seen. There are only missed head defects. This can be expected because some head defects are simulated in a very close proximity to the UT entry shadow and are prone to be missed.

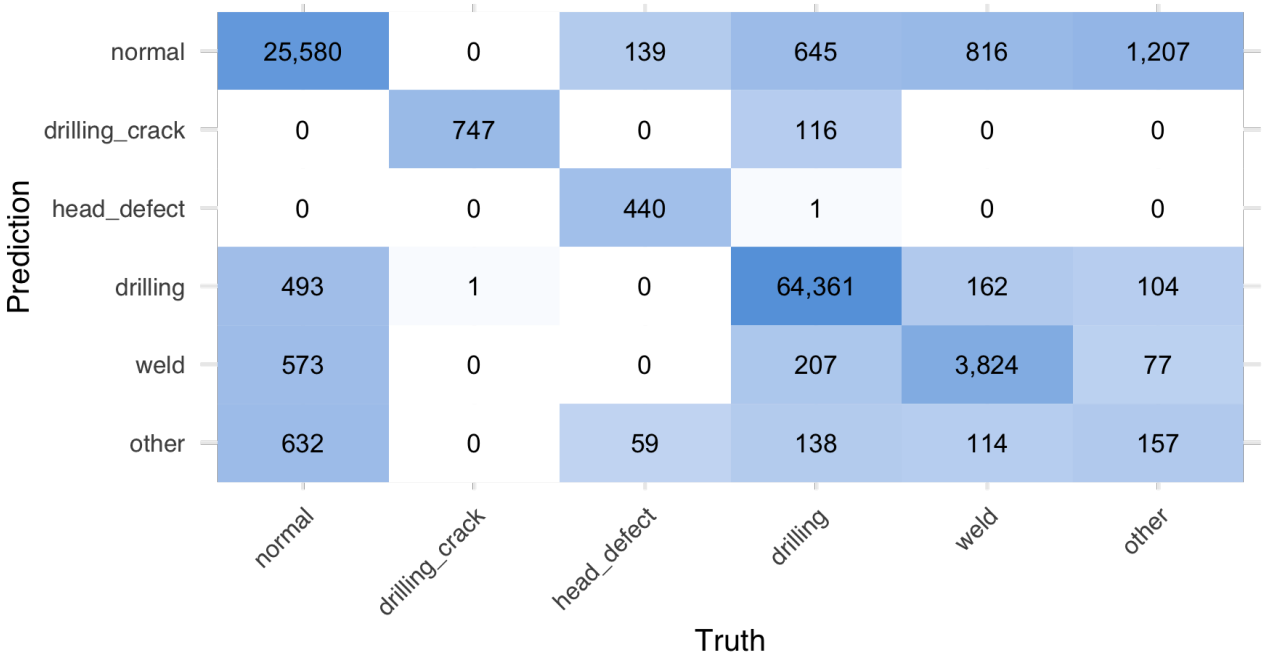

Figure 3: **Confusion Matrix** of all classified sequences in the validation set. Color scale is pseudo log transformed.

4 Discussion

The principal performance metrics appear to be good, but require more attention in the details. Besides a general drive to improve the results, it is necessary to perform closer examinations of failed classifications. It might not necessarily be the model that causes the problem, but the data which have certain properties that remained undefined. These can be indications that could not be properly labeled in the preparation phase or even new indications that were not considered even to be a problem. For example, certain noise levels can cause problems for the model and make the data unusable for defect detection. This realization might flow back into the inspection process, which could be adjusted to create cleaner data that are easier to classify.

The interpolation of the defect size for values that were specifically excluded in the training process shows mixed results. There are no significant differences in the predictions for 45 and 135 degree angles. The difference is significant for 90 degree and very large, as well as significant for 0 degree angle defects. In general, the error range is large, with error between -100% and 100% relative to the original size. The values have been normalized to generate a meaningful statistic with regard to difference between sampled and non-sampled values. Conclusions about the general quality of predictions cannot be drawn on the basis of these results. Different measures are necessary to do that, for example, the mean absolute error, which needs to be discussed in future research. From

these results, it is only possible to conclude that interpolation between learned defect sizes works for specific angles. A more thorough examination for predictions of 0º angle is necessary to understand how the models behave for this edge case.

Based on the confusion matrix in Figure 3, it must be assumed that the model learns to differentiate between the simulated and the real field data very well. Thereby a prioritization of data type over actual class is visible through the classification of simulated drillings as drillings with cracks and the lack of expected misclassifications for very small defects. In order to omit the theory that the model can generalize from simulation onto field defects, an explicit test for that case is needed but could not be completed due to a lack of data. Therefore, it is not possible to draw a final conclusion about this. There are examples in the literature where simulations can be generalized onto real data, such as Pyle, Bevan, Hughes, *et al.* [12], but the current data cannot reject doubts and even provides further evidence to assume that the model does not learn to classify real field data based on simulations.

However, interpretation of the results should be cautious and subject to certain limitations. In order to assess the models FPR a larger scale of normal rail data is needed and should be done with a special additional testing dataset to account for the overwhelming size of normal rail in reality. The data size for simulated defects is still relatively small. The work on further simulation continues and will be evaluated in further studies. In addition to creating more balance between field and simulation, equalizing different size parameter of defects will be key to this work. This research points out the possibility of failing generalization between simulations and real field data but cannot provide solutions to solve it. This needs to be done in future research, as possible approaches are still available for exploration. One solution can be a further assessment of the differences between simulations and field data. Knowledge about this can lead to an improved way to simulate defects and prevent deviating data pattern at the source. Another way of solving the problem might be an adjusted training method. Currently, training occurs naively with a combined set of simulations and field data. Training the model with a dedicated data set at different levels can yield more general results. Therefore, an approach of transfer learning should be implemented which first trains the model weights for field data and fine-tunes specific layers with simulated data afterwards.

5 Conclusion

This research provides valuable insights into the questions of how a deep learning model can generalize from trained data to new data. First, the model is capable of learning the pattern of defects and artifacts from ultrasonic testing data from the field and from simulations. Although further optimization and parameter tuning will be necessary to provide optimal value for the real world. Second, interpolating on different out-of-sample defect lengths works for specific angles of expansion and fails for difficult edge cases. For future research, it must be assessed whether the error in predicting the actual defect length is in an acceptable range to be applied in a productive system. Third, by observing the different classification of simulation and field data, it seems very likely that the model learned characteristic patterns, which are distinct from simulation and differentiate them from the actual field data. It is not clear whether this prevents a classification of real field defects, as this could not have been tested for now. However, a consequent assumption needs to be that the model did not learn to classify the

simulated pattern in the real field. Further work needs to specifically address this problem and how it might be solved by adjusting the training process or the simulations themselves.

6 Acknowledgements

This research is part of the project AIFRI, which is funded by the German Federal Ministry of Digital and Transport as part of the mFUND innovation initiative with a total of € 2.416 million.

References

[1] R. Tang, L. De Donato, N. Besinović, *et al.*, "A literature review of artificial intelligence applications in railway systems", *Transportation Research Part C: Emerging Technologies*, vol. 140, p. 103 679, 2022.

[2] M. Chenariyan Nakhaee, D. Hiemstra, M. Stoelinga, and M. van Noort, "The recent applications of machine learning in rail track maintenance: A survey", in *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification: Third International Conference, RSSRail 2019, Lille, France*, Springer, 2019, pp. 91–105.

[3] S. Uhlig, I. Alkhasli, F. Schubert, C. Tschöpe, and M. Wolff, "A review of synthetic and augmented training data for machine learning in ultrasonic non-destructive evaluation", *Ultrasonics*, vol. 134, p. 107 041, 2023.

[4] J. M. Ha, H. M. Seung, and W. Choi, "Autoencoder-based detection of near-surface defects in ultrasonic testing", *Ultrasonics*, vol. 119, p. 106 637, 2022.

[5] H. Mahajan and S. Banerjee, "A machine learning framework for guided wave-based damage detection of rail head using surface-bonded piezo-electric wafer transducers", *Machine Learning with Applications*, vol. 7, p. 100 216, 2022.

[6] R. Armbruster, T. Heckel, and S. Fenger, "Die gläserne schiene-fortgeschrittene ultraschallschienenprüfung", *ZfP-Zeitung*, vol. 121, pp. 65–68, 2010.

[7] T. Zhang and T. Heckel, "Virtuelle referenzschienen nach din en 16729-1 für die ultraschallprüfung von verlegten eisenbahnschienen", 2023.

[8] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification", *Neural Networks*, vol. 116, pp. 237–245, 2019.

[9] B. Sirisha, S. Naveena, G. Palanki, and P. Snehaa, "Multivariate time series sensor feature forecasting using deep bidirectional lstm", *Procedia Computer Science*, 2023.

[10] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[11] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] R. J. Pyle, R. L. Bevan, R. R. Hughes, R. K. Rachev, A. Ait Si Ali, and P. D. Wilcox, "Deep learning for ultrasonic crack characterization in nde", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.