2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) | 979-8-3503-8609-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEM62345.2024.10857013

Comparative Analysis of Machine Learning-Based Surrogate Modeling Approaches for Multi-Body Dynamic Simulation in Railway Digital Twin Platform

Shiyang Zhou[1]*, Artur Sogomonyan[1], Artur Ohanian[1], William Amminger[1], Yuxi Xia[2], Manfred Grafinger[1],

[1]Institute of Engineering Design and Product Development, Vienna University of Technology, Lehargasse 6/307, 1060 Vienna, Austria

[2]Faculty of Computer Science, University of Vienna, Vienna, Austria.

Abstract - **Machine learning (ML)-based surrogate models offer a promising alternative for Multibody Dynamics (MBD) Simulation of railway vehicle-track dynamics systems. A well-built ML model can accurately and quickly predict the dynamic responses to various track irregularities, significantly reducing computation time and unifying integration interfaces of different vehicle-track systems. However, training effective surrogate models is a complex process, influenced by the specific needs of the analysis. Different algorithms and training sets might be required for different surrogate models, making it essential to research the impact factors for building these models. This paper presents a comparative analysis of ML-based surrogate modeling approaches tailored for MBD Model within a railway digital twin platform. The primary focus is evaluating the performance of various ML-based surrogate modeling approaches, as well as the influence of neural networks, training parameters, and data sources. By leveraging extensive simulation and measurement data, we assess the ability of these surrogate models to predict key performance indicators under varying operational conditions. This analysis provides valuable insights for railway engineers in selecting appropriate surrogate modeling approaches, ultimately contributing to the advancement of predictive maintenance and optimization in railway operations.**

Keywords - **Systems Modeling and Simulation, Railway Vehicle-Track System, Intelligent Systems, Machine Learning, Multibody Dynamics Simulation**

I. INTRODUCTION

The developed surrogate model for replacing complex simulations has become a vital component in creating a digital twin (DT) platform for the holistic railway infrastructure system [1]. Zhou et al. [2] proposed utilizing a surrogate modeling approach to reduce computation time and unify the integration interfaces of different submodels of the railway system. They designed a machine learning (ML)-based surrogate modeling methodology for submodel integration in a holistic railway infrastructure DT platform and illustrated it through a case study. This study involved an ML-based surrogate model for multibody simulation of a railway vehicle-track dynamic system. The ML model accurately and quickly predicts vehicle-track system responses to various track irregularities. Finally, they proposed integrating the ML-based surrogate model into the DT platform using a standardized open-source Functional Mock-up Interface (FMI). However, training effective surrogate models is complicated. Various algorithms, data sources, and training sets may be necessary for different models, highlighting the importance of investigating the factors that influence their construction.

Network: Timeseries data prediction involves numerous well-established ML algorithms. An understanding of the temporal dependencies by “remembering” previous values has been proven a helpful ability by recurrent neural networks. Two types of recurrent neural networks were explored during development for the surrogate model: Long Short-Term Memory (LSTM) and Nonlinear Auto Regressive with exogenous inputs (NARX). They typically show promising results for time-series forecasting. Simon-Martin et al. [3] applied both LSTM and NARX models for forecasting electricity prices in the Spanish market. The results demonstrated that the LSTM model incorporating the exogenous variable exhibited superior predictive performance. This outcome can be attributed to the LSTM’s inherent ability to effectively model complex temporal dependencies within nonlinear system dynamics. Abbas et al. [4] explored networks to predict the expected battery life, this work handles multi-dimensional data. In this work, the high performance was attributed to the more complex model capturing the nuances of variability in electric power. Despite this, it is arguable how applicable these findings are to railway vehicle-track dynamics systems, as the behavior of electrical systems is quite different. An example of capturing physical phenomena was done by Wunsch et al. [5], using both LSTM and NARX models to predict the groundwater level in the Rhine region between France and Germany. Compared to other works, NARX performed best in this scenario, outperforming the LSTM in accuracy and computation time. It is pointed out, however, that the NARX model's performance was highly dependent on its initial conditions, highlighting again the stability of LSTMs. To summarize, simpler systems are captured more effectively by NARX, which should be considered for its quicker computation times, while more complex systems are suitable for LSTMs, which can provide more reliable results.

Training Parameter: Tuning specific hyperparameters, such as learning rate and batch size, had a greater impact on performance than modifying the model architecture, for ml-based surrogate models. R. Shankar et al. [6] proved that a model's performance significantly depends on tuning its parameters, specifically batch size and learning rate. While larger batch sizes degrade the model's performance, smaller batch sizes improve prediction accuracy by minimizing overfitting. Furthermore, the accuracy and efficiency of the model's training are greatly influenced by the learning rate. Higher learning rates provide a fair trade-off between accuracy and training time, while lower learning rates increase precision but slow convergence. Besides parameters, an essential aspect of training the model is data. M. Lindholm et al. [7] described utilizing LSTM neural networks for mortality forecasting with a focus on the efficient use of data. The research compared the baseline approach of withholding the last fraction of observations for validation, randomly sampling a fraction of observations in time, and splitting the population into two parts by sampling individual life histories. These methods enhance the model's predictive stability and efficiency by making better use of limited data. The boosted LSTM models show promising results, providing robust long-term predictions without compromising short-term accuracy. Another valuable point is that increasing the amount of training data can significantly lower overfitting in the LSTM network and enhance prediction accuracy. However, from the research of G. Ayzel et al. [8], it is also evident that the performance of the LSTM network cannot be significantly enhanced when the amount of training data exceeds a given value or range of values, and that the critical value varies for different basins. So, selecting a suitable quantity of training data can enhance the LSTM network's learning capacity.

Data Source and Quality: One of the most significant aspects of stable and robust ML model prediction, alongside model architecture suitability and correct hyperparameters, is data source and quality. In our domain, where we utilize track irregularity data from measurement trains for the ML-based surrogate model, which usually contains missing sections, outliers, duplicated sections, and random errors, proper data quality handling becomes quite essential. According to L. Budach et al. [9], unprocessed data usually leads to unstable and unreliable model performance. Through empirical research, they discovered that for regression tasks, the most critical dataset feature leading to poor deep learning model performance is data completeness and feature accuracy, i.e., the deviation of the input data from the ground truth values. The decrease in quality in one of these “data quality parameters” leads to a worse-than-linear decrease in algorithm performance. This finding underscores the importance of addressing missing data from measurement data in our case, which is inevitable when working with measurement data. In addition, Foroni et al. [10] try to quantify the influence of data quality issues on model performance. From the results obtained for regression tasks, a relatively low percentage of noise in the training data, such as 10%, increases the Mean Squared Log Error by 1.5 times. This underscores the importance of processing data to ensure that it contains less than 5% of problematic cases, thereby enabling the model to provide reasonable output.

This study seeks to address an existing gap by examining the efficacy and precision of various machine learning-based surrogate modeling techniques, including neural networks, training parameters, and data sources. Utilizing comprehensive simulation and measurement data, we evaluate these surrogate models' capability to forecast critical performance indicators across different operational scenarios. In Section 2, we detail the parameters for constructing the MBD model for railway vehicle-track dynamics, analyze various data sources, and outline the methods for developing and testing different machine learning-based surrogate modeling approaches for MBD simulations. Section 3 presents the results of a comparative analysis of these models, examining the impact of various factors. Finally, Section 4 concludes the study with a summary of our findings.

II. METHODOLOGY

We used an MBD model to simulate the railway vehicle-track dynamics, which is based on the Manchester Benchmarks Passenger Vehicle. The parameters of the MBD model are derived from realistic situations, as presented in Table 1.

Table I: Parameters of the MBD model

Name	Value	Unit
Track Length	5	km
Velocity	100	km/h
Vehicle Mass (with passengers)	31032	kg
Simulation Time	180	s
Sampling Rate	2000	Hz

Ensuring the robustness and accuracy of surrogate models is paramount. A proficient surrogate model must maintain the integrity of original data while providing reliable outcomes for optimization, prediction, and feasibility assessments. Consequently, validating trained ML models adequately becomes imperative. In our research, we utilize measurement data from the Austrian national railway track system to validate the fidelity of our ML-based surrogate model. However, due to measurement constraints, only vertical track irregularities are accessible. These irregularities comprise two dimensions: left-side track irregularities and right-side track irregularities. Corresponding to the measurement data, we generate two-dimensional input datasets. Extracted from SimPACK-Post, the vertical wheelset forces serve as the output data.

To evaluate the prediction quality, we employ the normalized Root-mean-square deviation (RMSE), a commonly used metric. Calculated by dividing RMSE by the range of simulated output data from the MBD model, normalized RMSE yields values between 0% and 100%, with lower values indicating better model fit. In this context, a normalized RMSE below 10% is deemed acceptable, aligning with industry expectations and signifying predictions within an acceptable deviation range.

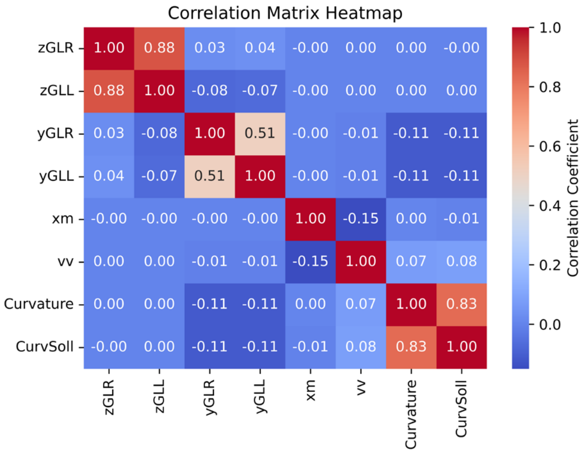

Fig. 1. Correlation matrix of measured track irregularities data

The correlation matrix shown above represents a table with pairwise correlation coefficients of the relevant input parameters, facilitating the discovery of patterns and trends between each input. In Figure 1, zGLR represents the vertical track irregularities of the right side of the track, while zGLL denotes the vertical track irregularities of the left side. yGLR represents the horizontal track irregularity of the right side of the track, while yGLL represents the left side. “xm” stands for kilometering, indicating the specific location of the track corresponding to each measurement. “vv” represents the speed during measurement, “Curvature” indicates the curvature of the measurement point, and “curvature soll” denotes the designed curvature of the respective measurement point. In this instance, irregularities in the track between the left and right sides have a high correlation in the vertical dimension, with a coefficient of 0.9. Additionally, irregularities in the horizontal dimension also show correlation, although not as strong as in the vertical dimension. Track irregularities measured along the y-axis (yGLR, yGLL) are also positively correlated, but the coefficient is not as high as for the vertical irregularities. Furthermore, the variable “curvature” is correlated with “CurvSoll,” representing the intended curvature value, making this strong correlation evident.

Effective parameter tuning is crucial for a surrogate model's generalization from training to unseen data. To ensure robustness and accuracy, parameters like batch size and learning rate must be carefully adjusted. Without proper tuning, models risk underfitting due to high bias or overfitting due to high variance, resulting in poor performance on new data. Figure 2 outlines the experimental process, starting with configuration and initialization, followed by a grid search through hyperparameters to identify optimal settings. The *DataManager* handles the slicing and processing of data, ensuring that training and test datasets are correctly generated with appropriate dimensions and sliding windows for sequential modeling. This processed data is

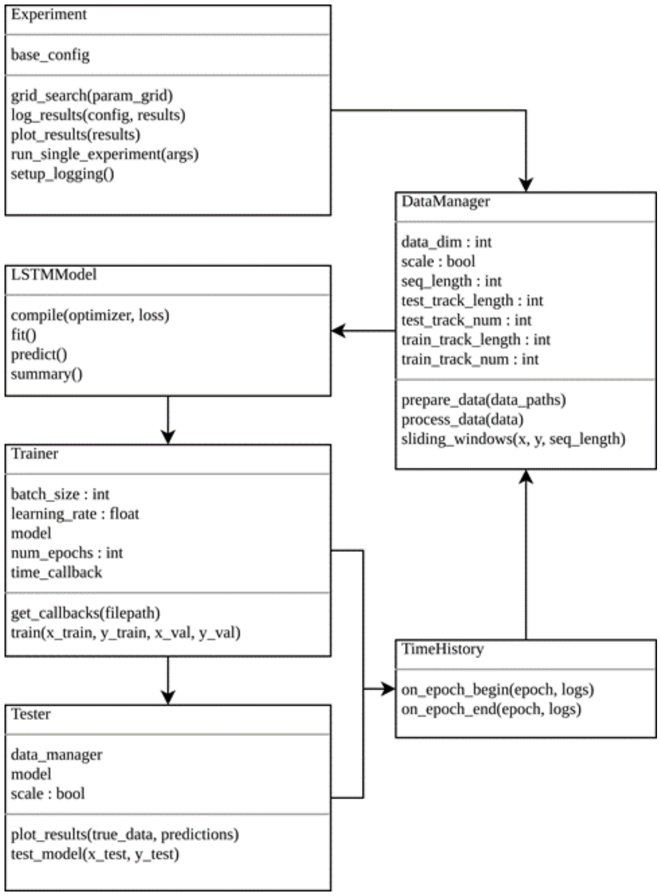

Fig. 2. Diagram classes of experiment process

fed into the *LSTMModel*, where the model is compiled with the chosen optimizer and loss function, then trained on the dataset. The *Trainer* block is responsible for executing the training process, where the learning rate, batch size, number of epochs, and callbacks (such as time tracking) are defined and managed. After training, the *Tester* evaluates the model’s performance on test data and generates predictions. Finally, *TimeHistory* tracks the epoch-wise training performance, ensuring detailed insights into the training dynamics.

III. RESULTS

The prediction results from various ML models are presented in Table II, categorized into four comparison groups: Group 1 includes Input Data Dimensions (Data No. 1, No. 5, and No. 7), which represent vertical, horizontal, and cross-level track irregularities; Group 2 covers Data Sources (Data No. 5 and No. 6), including Simpack-generated data and measurements from ÖBB; Group 3 focuses on Network Types (Data No. 1 and No. 2), specifically NARX and LSTM; and Group 4 pertains to Optimizer Types (Data No. 2 and No. 3), namely LM and Adam. Due to data protection requirements from our industrial partner, the RMSE for the 2-dimensional measurement data will be withheld.

Table II: Comparison of Prediction Results from Different ML-models

No.	Input Data	Data Source	Output Type	Network	Optimizer	RMSE	nRMSE
1	1-Dimensional	SimPack	Wheelset Force	NARX	lm	56.68	0.55%
2	1-Dimensional	SimPack	Wheelset Force	LSTM	lm	131.5	2.12%
3	1-Dimensional	SimPack	Wheelset Force	LSTM	adam	176.2	2.91%
4	1-Dimensional	SimPack	Acceleration in y axis	LSTM	lm	19.2	1.17%
5	2-Dimensional	SimPack	Wheelset Force	NARX	lm	-	1.01%
6	2-Dimensional	ÖBB	Wheelset Force	NARX	lm	-	3.52%
7	3-Dimensional	SimPack	Wheelset Force	NARX	lm	261.72	4.31%

Impact of Input Data Dimension: As we can see from Table II, the nRMSE of No.1 one-dimensional, No.5 two-dimensional and No.7 three-dimensional inputs, which are all generated from SimPACK and use the same Network NARX and Optimizer lm, are 0.55%, 1.01% and 4.31% respectively. We can clearly tell that the higher the input data dimension, the worse the machine learning predictions. Higher dimensions can lead to worse machine learning predictions due to the problem of the "curse of dimensionality". This refers to the fact that as the number of dimensions increases, the amount of data needed to effectively model the relationship between features and targets grows exponentially. As a result, there is a risk of overfitting the model to the limited data available, leading to poor generalization performance on unseen data. Additionally, high-dimensional data can also result in computational challenges and increased time and memory requirements.

Impact of Data Source: As we can see from Table II the nRMSE of No.5 two-dimensional inputs generated from SimPACK and the nRMSE of No.6 two-dimensional inputs collected from ÖBB are 1.01% and 3.52%, respectively, which indicates that the ML models have worse performance on measurement data compared to training data generated with SimPACK. The reasons might be 1. Data distribution shift: The distribution of the measurement data may differ from the distribution of the training data. This can result in a mismatch between the model's assumptions and the actual data, leading to reduced performance on measurement data. 2. Feature selection: The model may have been trained on a set of features that are not representative of the measurement data. This can result in poor performance if the model cannot effectively capture the relationships between the inputs and outputs in the measurement data. 3. Data quality: The quality of the measurement data may be lower than the quality of the training data. For example, the measurement data may contain missing values, outliers, or other anomalies the model was not designed to handle.

Impact of Network: As we can see from Table II the nRMSE of No.1 one-dimensional inputs trained with NARX and No.2 one-dimensional inputs trained with LSTM are 0.55% and 2.12% respectively, which indicates that the NARX Network has better performance compared to LSTM with one-dimensional input dataset. LSTM and NARX are both types of Recurrent Neural Networks (RNNs) used for processing sequential data. However, they differ in their architecture and the way they handle inputs. LSTM is a type of RNN designed to handle long data sequences and address the problem of vanishing gradients in traditional RNNs. They have memory cells and gates to control the flow of information, which allows them to store information from previous time steps and use it to make predictions about future time steps. This makes LSTMs well suited for tasks such as language modeling, speech recognition, and time series prediction. NARX, on the other hand, is a type of RNN that is designed to handle inputs that are not just dependent on the previous time step, but also on other inputs, called exogenous inputs. In NARX models, the relationship between the inputs and outputs is modeled as a nonlinear function, and this function is used to make predictions. NARX models are well suited for tasks such as time series prediction when there are additional inputs that need to be taken into account, which caters to our situation. As we can see from Table II the nRMSE of No.3 one-dimensional inputs trained with adam optimizer and in LSTM and No.2 one-dimensional inputs trained with lm optimizer in LSTM are 2.91% and 2.12%, respectively. It indicates, that lm optimizer has better performance compared to adam optimizer with one-dimensional input dataset using the LSTM Network. The reason might be as follows: In terms of convergence, Adam is often faster than lm because it does not require the computation of the second-order derivatives. However, lm is more efficient in terms of computation time because it only requires the computation of the Jacobian matrix, which is smaller than the Hessian matrix. lm is also better suited for nonlinear problems because it uses second-order information to estimate the local curvature of the loss surface. In general, adam is a good default choice for most machine learning problems, but lm can be a better option in specific cases where the problem requires a more customized optimization approach, such as nonlinear problems, which is similar to our usecase.

IV. CONCLUSION

In this study, we extensively analyzed machine learning models to forecast critical performance indicators in railway systems, examining the influence of various factors such as input data dimensionality, data sources, network architectures, and optimizer types on the efficacy and accuracy of ML-based surrogate models. Our exploration of the correlation matrix across different dimensions of input data revealed significant correlations among parameters, shedding light on vital patterns crucial for model development, particularly emphasizing the strong correlations between track irregularities in vertical and horizontal dimensions. Additionally, our investigation into the impact of input data dimensionality unveiled a trend wherein increased dimensionality led to diminished prediction performance, attributable to challenges like data sparsity, overfitting, and computational complexity, commonly referred to as the "curse of dimensionality." Moreover, our analysis highlighted the pivotal role of data quality and distribution congruence in model performance, showcasing the superiority of surrogate models trained on simulated data over those trained on real-world measurement data. Furthermore, our evaluation of different network architectures and optimizer types provided nuanced insights into their varying impacts on prediction accuracy, emphasizing the superiority of certain architectures and optimizers.

In conclusion, our study yields valuable insights for the refinement and optimization of ML-based surrogate models in railway systems, with future research poised to explore advanced techniques for data preprocessing, feature engineering, and model interpretation to enhance predictive capabilities in real-world applications. Collaborative efforts between academia and industry stakeholders are essential for translating these findings into actionable solutions, driving innovation and efficiency in railway infrastructure management and optimization. As the R4F Platform continues development, our focus will expand to applying the ML-based surrogate model methodology to other railway subsystems beyond the railway vehicle-track system. The automated parameter tuning methodology [11] will also be considered for application to identify the optimal combination of parameters for different deployed networks. Additionally, further analyses involving different algorithms and model integrations [12] will be pursued to validate the reliability and fidelity of the proposed integration methodology, marking a significant milestone in the holistic large-scale railway Digital Twin system's sustainable development.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of the COMET Program R4F (882504) within the COMET Competence Centers for Excellent Technologies from the Austrian Federal Ministry for Climate Action (BMK), the Austrian Federal Ministry for Digital and Economic Affairs (BMDW), the Vienna Business Agency and the Styrian Business Promotion Agency (SFG). The Austrian Research Promotion Agency (FFG) has been authorized for the program management, and the support from ÖBB Infrastructure and Siemens Mobility Austria GmbH.

REFERENCES

[1] S. Zhou et al., "A Conceptual Model-based Digital Twin Platform for Holistic Large-scale Railway Infrastructure Systems," Procedia CIRP, vol. 109, pp. 362-367, 2022.

[2] S. Zhou, A. Meierhofer, Ozan Kugu, Y. Xia, and Manfred Grafinger, "A Machine-Learning-based Surrogate Modeling Methodology for Submodel Integration in the Holistic Railway Digital Twin Platform," Procedia CIRP, vol. 119, pp. 345–350, Jan. 2023.

[3] Miguel de Simon-Martin, S. Bracco, E. Rosales-Asensio, G. Piazza, F. Delfino, and Pier Giuseppe Giribone, "Electricity Spot Prices Forecasting for MIBEL by using Deep Learning: a comparison between NAR, NARX and LSTM networks," Jun. 2020.

[4] G. Abbas, M. Nawaz, and F. Kamran, "Performance Comparison of NARX & RNN-LSTM Neural Networks for LiFePO4 Battery State of Charge Estimation," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Jan. 2019.

[5] A. Wunsch, T. Liesch, and S. Broda, "Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and nonlinear autoregressive networks with exogenous input (NARX)," Hydrology and Earth System Sciences, vol. 25, no. 3, pp. 1671–1687, Apr. 2021.

[6] R. Shankar, B. K. Sarojini, H. Mehraj, A. S. Kumar, R. Neware, and A. Singh Bist, "Impact of the learning rate and batch size on NOMA system using LSTM-based deep neural network," The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, vol. 20, no. 2, pp. 259–268, Oct. 2021.

[7] M. Lindholm and L. Palmborg, "Efficient use of data for LSTM mortality forecasting," European Actuarial Journal, vol. 12, no. 2, pp. 749–778, Apr. 2022

[8] G. Ayzel and M. Heistermann, "The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset," Computers & Geosciences, vol. 149, p. 104708, Apr. 2021

[9] L. Budach et al., "The Effects of Data Quality on Machine Learning Performance," arXiv:2207.14529 [cs], Nov. 2022.

[10] D. Foroni, Matteo Lissandrini, and Yannis Velegrakis, "Estimating the extent of the effects of Data Quality through Observations," Apr. 2021

[11] S. Zhou, Ozan Kugu, Y. Xia, and Manfred Grafinger, "A Reinforcement-Learning-based Parameter Tuning Methodology for Traction Control in the Holistic Railway Digital Twin System" Procedia CIRP 2024., in press.

[12] K., Ozan, S. Zhou, R. Nowak, G. Müller, S. H. Reiterer, A. Meierhofer, M.Grafinger. "An FMI-and SSP-based Model Integration Methodology for a Digital Twin Platform of a Holistic Railway Infrastructure System." Modelica Conferences, pp. 717-726. 2023.