Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-1149

Preprocessing of MWD data for data-driven modeling: advancing techniques for handling missing data

Franziska Klein, **Alla Sapronova**, Thomas Marcher
Institute of Rock Mechanics and Tunneling, TU Graz, Austria, alla.sapronova@tugraz.at

ABSTRACT: Measure While Drilling (MWD) data provides critical insights into subsurface conditions and drilling performance across industries such as tunneling, mining, and oil and gas exploration. However, MWD datasets often contain incomplete sections, specifically during unstable drilling phases, which can significantly impair their usability for data-driven modeling. Handling these missing data during data preprocessing is essential to ensure robust and accurate analyses for predictive and operational applications.Traditional methods for handling missing MWD data rely on statistical, domain-specific, and probabilistic techniques. Common approaches include mean, median, or mode imputation, linear or spline interpolation, and geotechnical knowledge-based rules, which often distort variance or rely heavily on expert judgment. Regression-based methods, such as simple and multiple imputations, account for relationships between variables but assume linearity and can be computationally intensive. Time-series techniques like forward/backward filling and decomposition assume stationarity, which may not hold for MWD parameters. Advanced methods like matrix factorization, expectation-maximization, and kriging consider patterns and spatial correlations, requiring specific assumptions about data distributions.While these traditional methods provide foundational solutions, they fail to account for the complex, sequential nature of MWD datasets. Recent advancements in machine learning (ML) offer more sophisticated alternatives and allow the discovery of temporal dependencies and long-range interactions within the data.This study proposes a new approach to reconstructing missing MWD data using advanced ML methods within deep learning architectures. By improving dataset completeness and reliability, more accurate predictive modeling and operational optimization in data-driven applications can be achieved. Furthermore, the proposed approach combines cutting-edge ML techniques with domain-specific preprocessing strategies to improve multi-parameter data reconstruction and ensure robust and scalable solutions for various analytical tasks.

KEYWORDS: drill and blast, tunneling, MWD, data preprocessing, machine learning.

1 INTRODUCTION

Measure-While-Drilling (MWD) data collected during drilling process contain information useful for both real-time and posteriori analysis. In drill-and-blast tunneling, drilling parameters such as penetration rate, rotation pressure, thrust, and flushing indicators are recorded at several seconds frequency to monitor the drilling process (van Oosterhout, 2016). These data can be used for various data-driven tasks, for example to identify rock mass class (Hansen et al., 2024) or predict rock mass strength (Komadja et al., 2025). They have also been used to detect rock support requirements (van Eldert et al., 2021) or detect potential overbreak zones (Navarro et al., 2018). The potential of MWD data for improving operational efficiency and safety is evident; however, realizing this potential requires that the data be of high quality and completeness. In practice, a significant challenge arises because MWD datasets often contain gaps or missing sections, which can severely compromise subsequent analysis and predictive modeling. These missing data typically occur during unstable drilling phases – for instance, at the start of each borehole in tunneling cycles when the drilling rig has not yet reached a steady regime and the rock at the face has been disturbed by the previous blast. As a result, the initial segment of every drill hole (around the first 0.5–0.6 m in the case study herein) is commonly unreliable and shall be routinely discarded. Discarding data from every borehole in this manner means losing roughly one-third to one-half of the information, which is highly detrimental to robust modeling. The core problem addressed in this paper is how to reconstruct these missing sections so that the resulting MWD dataset is sufficiently complete and of high fidelity for data-driven modeling.

Various traditional approaches, such as mean/mode imputation, linear interpolation, spline interpolation, kriging, regression imputation, are commonly used to handle missing data in sequential datasets, like MWD data (Zhou et al., 2024). The statistical imputation techniques replace missing readings with summary statistics (e.g. mean or median of available data), or carry forward the last known value. Such methods are easy to implement but these methods generally distort the original variance and inadequately represent the inherent sequential structure of datasets. As a result, imputed data produced by these methods often fail to accurately capture complex sequential variations, especially over larger gaps or non-linear sequences. The interpolation methods are filling gaps in data sequences by estimating intermediate values between known data points. Linear interpolation is a common choice that assumes a steady gradient between the end points of a gap. Spline interpolation allows a smooth curving fit, while kriging interpolation method predicts missing values by modeling the underlying correlation structure in the sequential dataset. However, each of these interpolation methods comes with limitations: a linear filling may lead to underfit if the true signal is nonlinear, spline interpolations can overshoot or generate unrealistic oscillations, and kriging requires careful selection and fitting of a correlation model (variogram), making it sensitive to modeling assumptions and computationally more demanding. Other techniques treat the problem in a probabilistic framework, such as expectation-maximization or Bayesian regression approaches, but these can be computationally intensive and assume a *certain* stationarity or distribution for the data. In general, the traditional imputation methods, though foundational, struggle to capture the sequential dynamics and multi-variable relationships in MWD data (Zhou et al., 2024).

Application of machine learning (ML) for handling missing data in sequential data shows promising results (Al-Fakih et al., 2025). Unlike static interpolation, data-driven models can learn complex patterns from the dataset itself, potentially exploiting temporal dependencies and correlations among multiple sensor readings. For example, recurrent neural networks such as Long Short-Term Memory (LSTM) networks have been employed in drilling applications to generate or infer missing log data (Osarogiagbon et al., 2020) demonstrated that deep learning can be used to synthesize missing gamma ray logs from other drilling parameters, achieving good accuracy. Similarly, Cheng and Zhang, (2020) introduced an ensemble LSTM (EnLSTM) architecture to generate well logs, showing

that ensembling multiple LSTM models can improve accuracy and better handle smaller datasets. These studies confirms that ML-driven imputation are more accurate in capturing non-linear relationships and long-range interactions than static methods. At the same time, more sophisticated models demand more training data and computational resources, and their performance gains over simpler methods are not always guaranteed if the dataset is limited or noisy. The aim of this study is to improve the preprocessing of MWD data by systematically evaluating the performance of traditional interpolation methods (linear, spline, kriging) against advanced ML models (LSTM, Transformer, Random Forest) for reconstructing missing data (Chen et al., 2021 and Caruso et al., 2024). Additionally, this research proposes a novel, hybrid deep-learning strategy that explicitly considers the drilling process characteristics and practical constraints of MWD datasets. Using MWD dataset obtained from drill-and-blast tunneling operations, this paper addressing the problem of reconstructing the initial missing segment of each borehole. The following sections describe the methodology, present the comparative results of different imputation techniques, discuss the implications for data-driven modeling, and suggest future steps to improve MWD data preprocessing.

2 METHODOLOGY

2.1 *Dataset*

The case study dataset comes from a drill-and-blast tunneling project in which MWD data were recorded for each advance length (round lenth). In the project, multiple boreholes (drill holes) of approximately 1.5 m in length, were drilled into each tunnel face for placing explosives. The MWD system logs various parameters along the depth of each borehole, including feed pressure, rotation pressure, penetration rate, percussion pressure, flushing pressure, and flushing flow at small depth intervals (centimeter-scale resolution). Due to machine instability and pre-conditioned rock near the start the first 0.60 m of every borehole's data are considered unreliable and are effectively "missing" for modeling purposes. The raw dataset contained several hundred boreholes collected over dozens of faces. To allow advanced analysis of sequential data, a data reorganization is performed: boreholes from consecutive faces were matched and concatenated to form longer continuous sequences along the tunnel alignment. The matching was done by pairing up holes in successive faces based on proximity, and lead to creating of extended MWD sequences spaning multiple advance rounds (in some cases, up to 10 faces, i.e. ~15m of drilling) to simulate a continuous series of boring data with the intervening missing segments (Fig. 1).

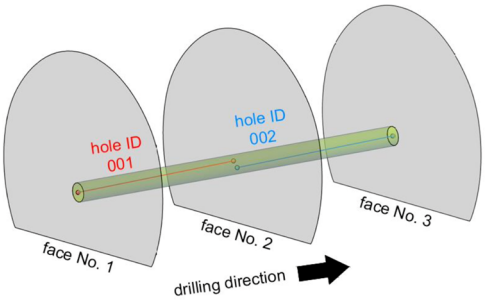

Figure 1. The borehole matching procedure: a borehole in one tunnel face is connected to the closest borehole from the next face. The illustration is from Klein (2025).

The outcome was a set of long composite borehole records that contain the missing data segments at each face junction.

Prior to imputation of missed data, a standard data preprocessing was applied as described by Sapronova and Marcher (2025). The erroneous readings and outliers removed also include boreholes flagged as "error" in the raw data and anomalous values caused by e.g. sensor faults. The first 0.60 m segment of each borehole (at each face break) was removed, so that these portions could later be reconstructed using the methods under study. For evaluation purposes we also created known data gaps of fixed length in the data where ground truth was known and could be compared to the reconstructed values.

2.2 *Approach*

Two strategies for reconstructing missing data in MWD data are compared: (1) a conventional strategy based on interpolation techniques, and (2) a data-driven strategy based on ML techniques.

Within the first strategy, three techniques, commonly used in engineering data, were investigated: linear interpolation, spline interpolation, and ordinary kriging. For spline interpolation the piecewise polynomial curves (cubic splines) were used to smoothly connect the boundary points of the gap. For kriging, the spatial correlation of each MWD parameter along the borehole depth were modeled using variograms, and then missing values were predicted as weighted averages of neighboring known points.

For the ML-based strategy, both deep- and ensemble learning models were considered. For deep learning, two cutting-edge deep learning architectures were designed: an LSTM-based recurrent neural network and a Transformer model. For ensemble model, a Random Forest regressor (RFR) was created. In this study, the RFR was used as a baseline machine-learning model because of its lower complexity and explainability. These three ML models were given known portion of MWD data sequences (either preceding a gap, following a gap, or both), and were trained to predict the readings in the missing segment. To make use of the multivariate nature of the data, the models were set up to consider multiple MWD parameters simultaneously – i.e. the input vector at each sequence step included all available MWD parameters, and the models learned to predict the missed values taking into account the collective behaviour of all parameters before or after the gap.

For training and validation purposes, the 0.60 m sequence selected of MWD readings were artificially removed (masked) to simulate the missing intervals. The models were then trained to reconstruct these missing values and the accuracy of the reconstructon was then calculated by comparing reconstructed data with masked true values.

Because deep learning models require substantial data for training, a staged training procedure was employed. First, the single-parameter models was trained and validated to gauge the feasibility. At this step, an ML model was trained to predict only one target parameter from the segment with missing data (feed pressure in our initial trial). This preliminary step revealed that the LSTM model could capture the sequencial patterns more effectively than the Transformer model with less training. The Transformer model did not show significantly better accuracy comparing to the LSTM despite its greater complexity. This is likely because the length of the sequences (and the size of the dataset) was too small to properly train the Transformer model. Transformer model also incurred a significantly higher computational cost for training. Therefore, in the final modeling phase we focused only on the LSTM as the deep learning method, and retained the Random Forest as a point of comparison.

Inspired by recent studies on bidirectional sequence modeling (Antariksa et al., 2022), two LSTM models were

trained for each MWD parameter: one forward model that predicts missing values in the forward (drilling) direction using preceding data, and one backward model that predicts in reverse using data from after the gap (this approach is conceptually similar to a bidirectional LSTM without coupling the two directions in one network). At inference time, the forward and backward predictions for the gap were combined (by averaging) to produce the final reconstructed sequence.

All model training and interpolation procedures were implemented in Python, utilizing libraries such as Pandas, SciPy (for interpolation routines), PyKrige (for kriging) and PyTorch for the neural networks.

3 RESULTS

3.1 *Borehole overlap method*

The initial idea of using overlapping borehole portions from consecutive faces provided an intuitive, domain-grounded way to fill missing segments. In cases where a borehole from face N+1 lays almost exactly in line with a borehole from face N, the latter's recorded data for depths 0.0 – 0.60 m beyond the common interface could, in theory, serve as a surrogate for the former's missing 0.0 – 0.60 m interval. This approach was implemented by identifying matching borehole pairs and substituting the preceding face's tail data into the subsequent face's head data. It was expected that application of this method would yield visually seamless join-ups. For some boreholes the transition was indeed smooth and the curves for MWD parameters aligned well across the boundary. However, in many cases, using the overlap substitution resulted in noticeable discontinuities ("jumps") at the junction, likely occurred due to differences in rock conditions or drilling parameters between consecutive boreholes.

In the plot of the combined borehole data some sequences exhibit a sharp kink at depth 0.6 m where the overlap data had been spliced. For example, of a mismatch in penetration rate and percussion pressure data is shown in Figure 2.

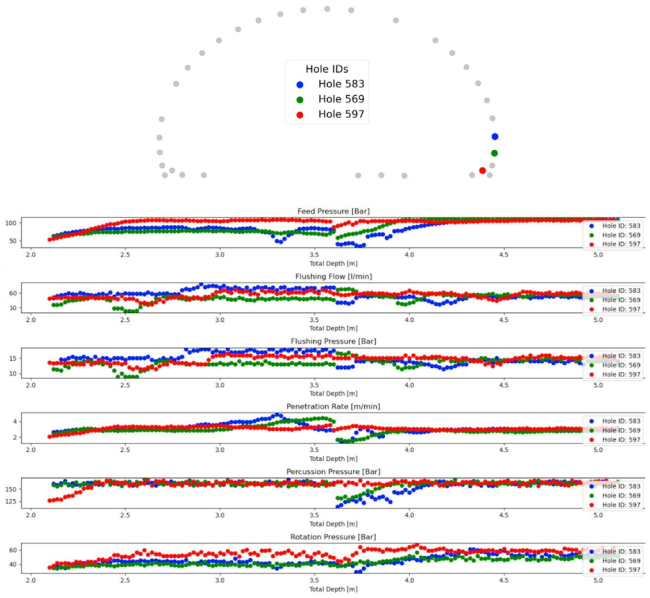

Figure 2. Schematic sketch of a tunnel face with boreholes' locations marked with colors (top), and MWD parameters along three boreholes of two consequent tunnel faces, reconstructed with borehole overlap method (bottom). The illustrations are from Klein (2025).

Presence of inconsistencies in the overlap-filling approach showed that a better strategy for data imputation is needed to accurately match the sequential data trends. For the scope of this study, the overlap method we set aside, and the focus was shifted to the interpolation and ML methods.

3.2 *Interpolation method*

All three interpolation methods (linear, spline, kriging) were applied and evaluated by comparing the filled-in values against the originally recorded data in artificially created data gaps segments. Several error metrics were computed per parameter, including the mean squared error (MSE), root mean square error (RMSE), coefficient of determination (R2), and a form of accuracy defined as the percentage of variance explained (ACC%). Table 1 summarizes the results of this evaluation for a representative subset of MWD parameters. Overall, the linear interpolation appears to be the most reliable method among the three. It produced the smallest errors in most parameters and the highest R2 scores. This finding was somewhat surprising but it can be explained by the fact that the missing intervals (0.60 m) are relatively short, and many MWD parameters are changing monotonically. Spline interpolation turned out to be overly flexible – in several instances, the spline fit oscillated or introduced implausible local extrema within the gap, especially where the endpoint values had a steep difference. As a result, spline interpolation approach turned to be unreliable as a general solution. Kriging interpolation performed comparably to linear interpolation in terms of error metrics on many parameters, and in a few cases it slightly outperformed linear interpolation on RMSE. However, the kriging method fell short on R2, and variance-explained measures compared to linear fills. The kriging method did not show a clear advantage likely because the short-range variogram had to be estimated from limited data, and the process essentially regressed toward a linear-like mean anyway.

Table 1. Results for interpolation methods

MWD Parameter	MSE	RMSE	ACC%	R2
Linear Interpolation				
Feed Pressure [Bar]	5.88	2.43	98.34	0.95
Flushing Flow [l/min]	1.46	1.21	98.48	0.99
Flushing Pressure [Bar]	0.61	0.78	97.10	0.92
Penetration Rate [m/min]	0.01	0.10	97.31	0.93
Percussion Pressure [Bar]	5.92	2.43	98.99	0.78
Rotation Pressure [Bar]	6.50	2.55	96.54	0.75
Spline Interpolation				
Feed Pressure [Bar]	5.93	2.44	98.41	0.96
Flushing Flow [l/min]	3.22	1.79	97.62	0.99
Flushing Pressure [Bar]	1.24	1.11	95.48	0.84
Penetration Rate [m/min]	0.01	0.12	96.75	0.91
Percussion Pressure [Bar]	22.49	4.74	97.98	0.18
Rotation Pressure [Bar]	38.90	6.24	91.40	-0.36
Kriging Interpolation				
Feed Pressure [Bar]	6.71	2.59	98.23	0.95
Flushing Flow [l/min]	2.41	1.55	98.06	0.90
Flushing Pressure [Bar]	1.04	1.02	96.26	0.87
Penetration Rate [m/min]	0.02	0.12	96.87	0.90
Percussion Pressure [Bar]	6.41	2.53	98.98	0.77
Rotation Pressure [Bar]	6.74	2.60	96.50	0.76

Considering also that kriging is more complex to implement and it is computationally expensive, it was concluded that a simple linear interpolation is the preferable conventional technique for this specific missing data problem. It offers ease of use and, in our results, the best fidelity to the actual measured values among the tested interpolators.

3.3 *ML method*

After establishing the baseline with linear interpolation, the ML-based approaches on the same gaps were evaluated. At first, the outcome of the single-parameter trials were explored. During the single-parameter trials ML models were trained to predict the missing data only for one parameter (using all others as inputs). In that experiment, the LSTM slightly outperformed the Transformer in prediction accuracy (by about 1% in our accuracy metric). The Random Forest model achieved error metrics comparable with the deep learning models for the single-parameter case. However, for the Random Forest model a crucial limitation emerged when extending to the autoregressive, multi-step prediction mode. The Random Forest (being a collection of decision trees) predicts each target point independently and has no inherent memory of sequence order, so to predict a sequence of 20 – 30 points forming the gap, it had to be applied iteratively: each predicted point shall be used as an input for the next step. This iterative prediction with Random Forest led to compounding errors and instability, as the prediction errors would propagate without any temporal corrective mechanism. In practical terms, while the Random Forest could predict the first few missing points reasonably, the latter part of the gap often diverged, yielding physically implausible spikes or drops. The LSTM, by contrast, is explicitly designed to handle sequence dependencies and can predict all points in the gap in one forward pass (or a few passes) while maintaining coherence. In the full multi-parameter gap-filling task, the LSTM model significantly outperformed the Random Forest, delivering smoother and more accurate reconstructions of the missing intervals. It was observed that the Random Forest's earlier advantage in static error metrics did not translate to the dynamic gap-filling scenario.

For the LSTM model, the bidirectional approach was used as described in the methodology section. One LSTM was trained on sequences moving forward (using data prior to the gap to predict into the gap) and another on reversed sequences (using data after the gap to predict backwards into the gap). Each model on its own performed well, but had certain biases: the forward model tended to slightly under-predict some parameters near the end of the gap (where it had not learned a "future" context), whereas the backward model sometimes misestimated the very beginning of the gap (lacking "past" context). By averaging the forward and backward predictions, a combined reconstruction was obtained using information from both sides. The summary of the evaluation of ML models is shown in Table 2.

Table 2. Results for ML methods

MWD Parameter	MSE	RMSE	ACC%	R2
Random Forest				
Feed Pressure [Bar]	12.22	3.5	65.99	0.92
Flushing Flow [l/min	2.23	1.49	85.14	0.99
Flushing Pressure [Bar	1.09	1.04	41.35	0.86
Penetration Rate [m/min]	0.02	0.15	39.08	0.84
Percussion Pressure [Bar]	10.98	3.31	-0.54	0.62
Rotation Pressure [Bar]	10.8	3.29	26.04	0.65
Bi-LSTM, averaged				
Feed Pressure [Bar]	5.05	2.65	98.82	0.97
Flushing Flow [l/min]	1.63	1.12	98.94	0.99
Flushing Pressure [Bar]	0.69	0.66	97.50	0.93
Penetration Rate [m/min]	0.02	0.10	97.04	0.95
Percussion Pressure [Bar]	4.26	2.46	98.99	0.85
Rotation Pressure [Bar]	6.09	2.01	96.91	0.80

While the accuracy of the LSTM model appears high, it is important to consider the errors relative to the parameter range. For instance, a maximum error of up to 50 bar was observed for feed pressure, which is significant compared to the overall parameter range of 211 bar. This suggests the presence of substantial prediction outliers. To manage this issue, a threshold-based method was introduced for each parameter. Specifically, if the difference between the predicted and actual values for the first two data points exceeded a predefined threshold—determined by the standard deviation of that parameter—the predictions were flagged as less reliable.

Across all test instances, the LSTM's reconstructions have captured the general trends of the missing data without introducing the abrupt artifacts as observed with some interpolation methods. In terms of quantitative performance, the LSTM-based approach achieved an average R2 in the range of 0.85 – 0.99 (varying by parameter), exceeding what linear interpolation attained (0.75 – 0.95 range in our tests). The improvements were most noticeble for parameters that have complex interdependencies – for example, the simultaneous rise in penetration rate and drop in feed pressure when the bit hits a softer layer is something the LSTM could learn to anticipate by looking at patterns in other channels, whereas linear interpolation would simply draw a straight line in each channel independently.

4 DISCUSSION

Obtained results provide a comparative insight into missing data handling for MWD datasets and carry several implications for practitioners and researchers in this domain. One immediate observation is the good results achieved by linear interpolation, a method often considered rudimentary.

In this research, the simplicity of linear interpolation turned out to be an advantage. Linear interpolation effectively captures first-order trends in data dynamics, yielding good accuracy in the data imputation procedure while avoiding modeling artifacts introduced by more sophisticated interpolation methods. This suggests that for short gaps in MWD records, where the true signal does not have time to vary dramatically, a linear assumption can be considered adequate. From a practical standpoint, this is an encouraging finding: it means that in absence of advanced tools, engineers can apply a quick linear fill to missing MWD sections and still retain a large portion of the data's informational value for subsequent analysis. However, the application of this method is limited: linear interpolation does not account for any higher-order pattern or cross-variable correlation. It treats each parameter independently and cannot predict, for example, a sudden change in penetration rate due to a different rock layer if that change is not reflected at the gap boundaries. In our controlled evaluation, linear interpolation performed well partly when the test data contained no extreme, unforeseen events. In scenarios where a geological discontinuity or anomaly lies in a missing section, more advanced methods have to be used to detect or infer the anomaly.

Employment of ML models, represented by the Random Forest and LSTM in our study, demonstrated a set of strengths. The LSTM model was able to recognize the sequential context and relationships between multiple parameters to produce a more informed estimation of missing values. For example, when an increase in penetration rate immediately precedes a data gap and a decrease in flushing pressure follows thereafter, an LSTM model could infer that this sequence is indicative of entry into a fracture zone and, consequently, predict a corresponding response—such as elevated vibration or reduced torque—within the gap, an association that a univariate linear

interpolation method would be unlikely to capture. Using such recognition, the model could reconstruct the missing data by replicating the behavior of other MWD parameters consistent with the pattern, such as a spike in vibration or a decrease in torque—features that a univariate linear interpolation would fail to capture. The fact that LSTM model was overperforming the Random Forest underscores the importance of accounting for sequence dependency when filling up missing MWD data. Tree-based models like Random Forest stumbled in autoregressive prediction due to lack an intrinsic notion of order or memory. In contrast, the LSTM's concept of internal state enables it to carry information across the entire missing interval, maintaining consistency and smoothing out predictions in a way that aligns with learned dynamics of drilling processes.

Another notable point is the failure of the Transformer model to show an advantage in this application. Transformers have revolutionized sequence modeling in fields like natural language processing, primarily due to their ability to capture long-range dependencies with attention mechanisms. In this research, however, the longest sequence of data for ML model training was at most on the order of tens of data points, and the training dataset was limited to few hundred instances. The Transformer's heavy architecture (with many parameters to train) likely did not get fully utilized, and its training overhead became a liability. This aligns with a broader understanding that for moderate sequence lengths and smaller data regimes, LSTMs type neural network can be more efficient and just as effective as Transformers. This likely means that when choosing an ML model for MWD data imputation, one should consider the volume of data and sequence length at hand – a simpler recurrent network might show sufficiently accurate results, whereas a more complex architecture could be overkill.

The experiment of using a bidirectional LSTM to capture both past and future context has precedents; for example, Cheng et al., (2023) applied a bidirectional LSTM to improve prediction of rock mass classes ahead of the tunnel face, demonstrating that considering information from both directions significantly improved prediction accuracy. In this work, the bidirectional LSTM concept for data imputation was effectively mirrored by combining forward and backward predictions. This approach allowed to eliminate the edge effects each single-direction model suffered from. This is particularly useful in ensuring that the reconstructed segment transitioned smoothly into the known data at both ends. The approach does introduce extra complexity (training two models instead of one) but given the small additional computational cost relative to the whole training process, it was well worth the effort in our results.

While this study focused on a specific MWD dataset and a specific kind of missing data (the start-of-hole sections), the methodology for filling the missed data can be transferred to other contexts. Many industrial data sequences encounter similar issues of missed data due to e.g. sensor outages, initialization periods of equipment, or data lost due to transmission problems. The methodology discussed in this work could be used as a guideline on how to tackle the problem: build a baseline model using simple and explainable method, compare a baseline to more complex approaches, and verify that added complexity worth an achieved benefit. In our case, the benefit of using a deep learning model over a linear interpolation approach mostly pronounced in scenarios where multi-parameter coupling was important.

For evaluation, an artificial gaps in data sequences strategy was created and the originally recorded data in the "missing" sections were used for benchmarking. This approach potentially raises a question about general quality of recorded MWD data.

In this work, methods, like linear interpolation, produced a smoothed version of missed data, which might be closer to the *true* underlying trend than the raw measurements were. The LSTM, trained on the raw data patterns, tends to reproduce whatever biases were present. This observation highlights an important point: success in filling missing data should not be judged solely by reproducing every wiggle of the original signal – especially if the original data is noisy. Instead, the accuracy of data imputation method should be judged by how well the filled data enables the end-goal analysis or model. In a follow-up study, the integration of the reconstructed data into a downstream task will be introduced to verify whether performance improves when using LSTM fills versus linear fills. That would ultimately validate the usefulness of the advanced preprocessing.

Based on the comparative results of this study, a preliminary guidance for similar MWD datasets can be offered: short, smooth gaps without abrupt geological transitions may be effectively reconstructed using linear interpolation, whereas gaps spanning potential parameter shifts or multi-variable dependencies benefit from advanced sequence-aware models such as LSTMs.

5 OUTLOOK

This also opens a possibility for future work: a hybrid approach might be beneficial where one uses simple methods as a first pass and then applies ML selectively in cases where those methods flag uncertainty or likely error. For example, one could use a rule-based system to detect when a linear interpolation might be misleading (perhaps by analyzing the difference gradient between predicted and measured data at either end of the gap or external indicators of geology) and only then invoke a trained ML model to fill the gap.

It is also insightful to compare the described MWD data imputation methodology with approaches used for Tunnel Boring Machine (TBM) data, as both involve large-scale sequential drilling data but in different contexts. Both MWD and TBM datasets are high-resolution time series of machine and drilling parameters, and both suffer from missing readings due to sensor dropouts or communication issues. Thus, many principles of handling missing data are shared. For example, our finding that simple linear interpolation works well for short, unremarkable gaps should likewise apply to TBM data – a brief loss of a TBM sensor signal can often be bridged by assuming a straight-line continuation if the machine's state doesn't change abruptly. Similarly, the advantage of using models that consider sequential context (like LSTMs) is conceptually applicable to TBM operations, where the state of the machine and the encountered ground conditions evolve continuously along the tunnel path.

6 CONCLUSIONS

Handling missing data in MWD datasets is an important data preprocessing step that significantly affects the success of subsequent data-driven modeling. This study compared traditional interpolation techniques with modern machine learning methods for reconstructing incomplete MWD records and introduced a hybrid deep learning approach for data imputation. The results show that a linear interpolation method, despite its simplicity, often performs well for short gap filling. At the same time, the LSTM-based model, applied in a forward–backward fashion, yields more accurate and realistic data reconstructions, capturing cross-parameter dynamics and the sequential nature of the problem that linear methods fail to address. The accuracy achieved by the ML approach demonstrates the benefit of employing deep learning for data

curation in drilling operations. By restoring the missing segments of each borehole's data, the completeness of the dataset is improved, enhancing its suitability for training predictive models — an important step for applications such as geological forecasting and operational optimization.

Through the evaluation of classical and advanced methods, preliminary guidance emerges: short, smooth gaps without abrupt geological transitions may be effectively reconstructed using linear interpolation, whereas gaps spanning potential parameter shifts or multi-variable dependencies benefit from advanced sequence-aware models such as LSTMs. As the field moves toward increasingly automated and intelligent analysis of drilling operations, the importance of reliable data preprocessing is increasing. The findings of this work contribute to the development of robust preprocessing pipelines in drill-and-blast tunneling projects, ultimately enabling more accurate predictions of subsurface conditions and more efficient, safer drilling practices.

It is also noted that the LSTM's advantage was not uniformly large for all parameters, and some signals were reconstructed more accurately with the linear model. This indicates that the choice of data imputation method should consider the specific parameter characteristics and acceptable error margins. Nonetheless, the LSTM consistently provided high-quality gap fills and demonstrated robustness across varying conditions and parameters present in the dataset.

These outcomes open the way for follow-up investigations where the imputation scheme shall be integrated into a complete predictive modeling pipeline, using reconstructed MWD data to train predictive models, for example for rock mass classification or prediction of overbreaks during the drill and blast tunneling projects. Comparative evaluation against models trained on datasets with gaps or simpler fills would provide a direct measure of the practical impact of advanced preprocessing. Further enhancement of the proposed method could be achieved through hybrid approaches, such as applying linear interpolation by default and switching to an ML model under specific conditions (e.g., when multiple parameters collectively indicate an anomaly or when cross-validation reveals high interpolation error). Such adaptive systems have the potential to combine efficiency with accuracy, ensuring complex models are used only when necessary.

7 REFERENCES

Al-Fakih, A., Koeshidayatullah, A., Mukerji, T. et al. (2025) 'Well log data generation and imputation using sequence based generative adversarial networks', Scientific Reports, 15, p. 11000. doi: 10.1038/s41598-025-95709-0.

Antariksa, G., Muammar, R., Nugraha, A. and Lee, J. (2022) 'Well log data imputation using deep learning method in West Natuna Basin, Indonesia', *SSRN Electronic Journal*. Available at: http://dx.doi.org/10.2139/ssrn.4279765 (Accessed: 4 August 2025).

Caruso, Camillo Maria, Paolo Soda, and Valerio Guarrasi. (2024) *Not Another Imputation Method: A Transformer-Based Model for Missing Values in Tabular Datasets*. arXiv, July 16, 2024, arXiv:2407.11540. https://doi.org/10.48550/arXiv.2407.11540.

Chen, Y. and Zhang, D. (2020) 'Well log generation via ensemble long short-term memory (EnLSTM) network', *Geophysical Research Letters*, 47(23), Article e2020GL087685. doi: 10.1029/2020GL087685.

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., Liu, Z. (2021) 'A transfer learning-based LSTM strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system', *Journal of Hydrology*, 602, 126573. doi: 10.1016/j.jhydrol.2021.126573.

Cheng, X., Tang, H., Wu, Z., Liang, D. and Xie, Y. (2023) 'BILSTM-based deep neural network for rock-mass classification prediction using depth-sequence MWD data: A case study of a tunnel in Yunnan, China', *Applied Sciences*, 13(10), p. 6050. doi: 10.3390/app13106050.

Hansen, T.F., Erharter, G.H., Liu, Z. and Torresen, J. (2024) 'A comparative study on machine learning approaches for rock mass classification using drilling data', *Applied Computing and Geosciences*, 24, p. 100199. Available at: https://arxiv.org/abs/2403.10404 (Accessed: 4 August 2025).

Klein, F. (2025) *Preprocessing of MWD Data for Data-Driven Modeling: Advancing Techniques for Handling Missing Data*. Master's thesis, Graz University of Technology.

Komadja, G.C., Westman, E., Rana, A. et al., (2025) 'Predicting rock mass strength from drilling data using synergistic unsupervised and supervised machine learning approaches', *Earth Science Informatics*, 18, p. 325. doi: 10.1007/s12145-025-01837-6.

Navarro, J., Sanchidrián, J.A., Segarra, P., Castedo, R., Costamagna, E. and López, L.M. (2018) 'Detection of potential overbreak zones in tunnel blasting from MWD data', *Tunnelling and Underground Space Technology*, 82, pp. 504–516. doi: 10.1016/j.tust.2018.08.060.

Osarogiagbon, U., Oloruntobi, O., Khan, F., Venkatesan, R. and Butt, S. (2020) 'Gamma ray log generation from drilling parameters using deep learning', *Journal of Petroleum Science and Engineering*, 195, p. 107906. doi: 10.1016/j.petrol.2020.107906.

Sapronova, A. and Marcher, T. (2025) 'Improving data quality with advanced pre-processing of MWD data', *Geotechnics*, 5(2), p. 28. doi: 10.3390/geotechnics5020028.

Stekhoven, D.J. and Bühlmann, P. (2012) 'MissForest—Non-parametric missing value imputation for mixed-type data', *Bioinformatics*, 28(1), pp. 112–118.

van Eldert, J., Funehag, J., Saiang, D. et al. (2021) 'Rock support prediction based on measurement while drilling technology', *Bulletin of Engineering Geology and the Environment*, 80, pp. 1449–1465. doi: 10.1007/s10064-020-01957-x.

Van Oosterhout, D. (2016) 'Use of MWD data for detecting discontinuities'. MSc Thesis. Delft University of Technology. Available at: https://doi.org/10.4233/uuid:7bc0f5b3-b9fc-4c7f-ae61-f922a29e88a7 (Accessed: 3 August 2025).

Zhou, Y., Aryal, S. and Bouadjenek, M.R. (2024) 'Review for handling missing data with special missing mechanism', *arXiv preprint* arXiv:2404.04905. Available at: https://arxiv.org/abs/2404.04905 (Accessed: 4 August 2025).