Proceedings of the 9th World Congress on Civil, Structural, and Environmental Engineering (CSEE 2024)
London, United Kingdom – April 14-16, 2024
Paper No. ICGRE 123
DOI: 10.11159/icgre24.123

MWD data Analysis for Risk Assessment and Process Optimization in Tunneling

Alla Sapronova[1], Thomas Marcher[1], Franziska Klein[1]
[1]Institute of Rock Mechanics and Tunnelling/Graz University of Technology
Rechbauerstraße 12, A8010, Graz, Austria
alla.sapronova@tugraz.at;
thomas.marcher@tugraz.at; franziska.klein@student.tugraz.at

Abstract **–**In drill and blast operations, the collected Measure While Drilling (MWD) data can be used for analysis. In this study we process MWD data with machine learning (ML) methods to demonstrate how real-time risk assessment can be conducted: by predicting the volume of over-excavation, one can aim in reducing risk, lowering project costs, and minimizing environmental impact. The complexity of MWD data makes it necessary to address several challenges before utilizing the data within ML frameworks, namely the three steps (data pre-processing, feature extraction, and normalization) shall take place before further modelling. Here, we discuss the data preparation process, showing how the output from the correlation analysis impacts ML models accuracy. We will show that information from the raw MWD data is preserved and, even, enriched in the correlation analysis. Such information enrichment allows ML models to discover patterns implicitly related to changes in the rock mass conditions in MWD data.
By including correlation analysis into the data preparation pipeline, combining it with encoding strategies, fine-tuning procedures, and careful selection of validation metrics, one can dramatically improve the accuracy of ML models in geotechnical applications.

Keywords**:** MWD data, machine learning, data science, predictive modelling

1. Introduction

The construction industry is a field where precision, safety, and efficiency are extremely important. Within this domain, tunneling projects are standing out by complexity of engineering challenges mixed with environmental and economic considerations.

Technological advancements and data analysis have already revolutionized numerous industries, and construction is no exception. In recent years, the field has seen the integration of digital tools, automation, and advanced analytical techniques that have dramatically increased productivity and accuracy in decision making.

One of transformative technologies in tunneling was an emerge of Measure While Drilling (MWD) technology that allowed to collect data as drilling operations proceed. By analyzing this data, engineers can gain real-time insights into the condition of the rock mass being drilled by predicting potential hazards and ensuring the integrity of the tunnel structure. This not only improves safety measures but also enhances the decision-making process, allowing for adjustments in drilling techniques that align with the encountered geological conditions. Now the MWD is a critical component of modern tunneling projects, including drill and blast tunneling.

The application of machine learning (ML) -a branch of artificial intelligence that thrives on the ability to learn from data - to MWD data in tunneling is a relatively new field, but it is already showed its great potential [1,2].

In this work we investigate how preparation of MWD data contributes to successful application of ML for predictive analysis and decision-making.

The primary objective of this study is to develop predictive ML model to forecast the overbreak quantity, based on MWD data. By using intelligent preparation of MWD data, we help ML to discover patterns that otherwise might be overlooked. This research also aims to bridge the gap between data collection and data utilization, showcasing the capabilities of relative simple ML models in yielding good results.

2. Data Preparation

Previous studies have demonstrated that ML-based predictive models outperform traditional mathematical approaches concerning accuracy and computational speed when applied to different aspects of blasting operations [3]. However, there

is still significant room for improving accuracy of models trained on MWD data. To tackle this challenge, we concentrated on two critical aspects: data preparation and predictive modeling.

With data preparation one aims to build representations of the data that improves the identification of underlying patterns during ML model training. The preparation process consists of three steps: data preprocessing, normalization, and feature engineering.

- **Preprocessing:** During the drill and blast operations, MWD systems continuously record a lot of parameters. This data is stored as instrumental logs of the real-time interaction between the drilling machinery and the rock mass in a plain text (ASCII) format. The preprocessing of MWD data is aimed at handling logs of drilling data and related information to enhance information required for data driven modeling. At first, all collected MWD data was converted to CSV format to facilitate subsequent processing stages. Then, raw MWD data was cleaned: this included filtering out outliers, handling missing values, and correcting any inconsistencies. The cleaning step is crucial to ensure that the ML models are not misled by erroneous data. At the next layer of the preprocessing step, we integrated various datasets to create a comprehensive view of the tunneling operation. This study does not rely solely on MWD data; it also encompasses an array of records including explosives data, tunnel geometry, and planned excavation geometry. Proper data integration involves aligning these differently structured data into a unified dataset where the impact of each parameter on the others can be examined. To achieve this integration, data from different sources were combined, ensuring that the records are synchronized based on timestamps or spatial coordinates. The integration process also involves checks on data consistency and completeness. In cases where data from different sources overlap, it was vital to resolve any discrepancies to maintain the integrity of the dataset.
- **Normalization:** Data normalization is essential for ensuring that the ML model does not become biased towards certain features due to the scale of the data. The variations in the drilling parameters shall be normalized to maintain their predictive value without being affected by scale differences.
- **Feature Engineering:** The feature engineering step focuses on selecting relevant MWD variables or creating new by combining existing ones, to improve ML model performance. The selection of the most relevant features usually based on domain knowledge and statistical analysis. For instance, the ratio of rotational speed to thrust might be a significant indicator of bit wear. The objective of this step is to develop a training dataset that provides a clearer signal to the predictive modeland therefore enhances the ability of ML model to detect complex patterns and relationships.

To support the preprocessing, normalization, and feature engineering steps, correlation analysis can be employed. Correlation analysis is a statistical tool that helps to quantify the relationships between MWD variables and allows to map the impact of various rock mass conditions with MWD data. It is important to note, that the correlation analysis not only aids in refining the feature set but also provides insights into the physics of rock drilling, which can guide future data collection and feature engineering efforts.

In the next section we aim to bridge the detailed examination of correlation analysis and the foundational aspects of dataset suitability for machine learning, to show that utilizing correlation coefficients, in conjunction or isolation with raw data, for ML training is advantageous.

2. Dataset Quality Assessment

Our analysis showed that datasets crafted from correlation indices are inherently richer in information. They exhibit increased variance and reduce self-correlation, properties that significantly enhance their utility for machine learning training. This section provides a theoretical explanation for our decision to use correlation coefficients as inputs to ML model training.

When preparing datasets for ML training, several key metrics are crucial: entropy, normalized variance, dimensionality, and average correlation. Each of these metrics provides insights into the characteristics and quality of the dataset. When the goal is to choose a dataset that not only encapsulates a rich, informative feature set but also maintains quality and manageability for effective ML modeling, the summary of these metrics can be used to determine which dataset is preferable for ML training.

In machine learning the informational diversity within a dataset can be described by entropy (H): a concept "borrowed" from information theory that quantifies the amount of information or uncertainty within a dataset. For a discrete random variable X with possible values $\{x_1, x_2, ..., x_n\}$ and probability function P(X), entropy is defined as:

$$H(X) = -\sum P(x_i) \log_2 P(x_i) \quad (1)$$

where b is the base of the logarithm, typically 2, making entropy units in bits. Higher entropy implies greater randomness and potentially more information content [4].

High entropy indicates diverse information, but it needs to be balanced against the risk of having too much noise. Measuring noise in a dataset is a complex task because noise can come in many forms and is not always quantifiable in a straightforward way. However, here are some general methods and considerations for estimating noise in a dataset: a statistical noise estimation and data quality analysis.

To estimate the statistical noise, we calculated the variance of the error and signal-to-noise ratio (SNR).

Calculation of the Variance of the Error: The dispersion of residuals—the divergence between observed values y and predicted values ŷ —serves as a noise indicator in a well-fitted model. The variance of residuals $\sigma^2_{residuals}$ can be expressed as:

$$\sigma^2_{residuals} = (\sum (y_i - \hat{y}_i)^2) / N \quad (2)$$

where N is the number of observations. A high residual variance implies considerable noise, potentially overshadowing the signal.

Calculation of the Signal-to-Noise Ratio (SNR): Originating from signal processing, SNR determines the strength of the signal, or the desired pattern, against background noise.

It is defined as:

$$SNR = \mu^2 / \sigma^2 \quad (3)$$

A higher SNR signifies a clearer distinction between the signal and the noise, indicating less noise within the data [4]. Interpreting results from statistical noise estimation involves a nuanced understanding of the dataset's context and the ML model's objectives. High residual variance and low SNR may necessitate data cleansing or further investigation. A significant number of outliers or inconsistencies might indicate poor data quality or collection issues, which could compromise ML model performance.

Variance in a dataset quantifies the spread of data points. Normalization of variance is essential, especially when comparing datasets with features on different scales. Normalized variance brings all features to a common scale, preventing features with larger numerical ranges from dominating the variance calculation. The formula for variance (σ^2) is:

$$\sigma^2 = (1/N) \sum (x_i - \mu)^2 \quad (4)$$

where x_i are the data points, μ is the mean, and N is the number of data points. For normalization, we used Min-Max scaling $x' = (x - \min(x)) / (\max(x) - \min(x))$.

Dimensionality of a dataset refers to the number of features or variables in the dataset. Higher dimensionality can lead to the 'curse of dimensionality', where the feature space becomes so large that the available data becomes sparse. This sparsity is problematic for ML models and can lead to overfitting [5].

Average correlation among features indicates the presence of linear relationships. In datasets with high multicollinearity (where features are highly correlated), models might struggle to distinguish between relevant features, leading to unstable coefficient estimates. It's calculated as the average of absolute values of correlation coefficients between different pairs of features.

3. Modelling and Results

2.1. Dataset Selection

As discussed in the selections above, selecting suitable dataset for ML training is a critical step to ensure high accuracy of ML model output and the selection process involves assessing several metrics: entropy, variance, dimensionality, and correlation. In our study, we performed a comprehensive analysis of three datasets:

- the original dataset (DF) made from the preprocessed and normalized MWD data,
- the statistical dataset (SDF) made from the preprocessed and normalized statistical values (per borehole min, max and mean values for every variable) of MWD data, and
- the derived dataset (DDF) made from the correlation coefficients between MWD data computed for 20sec sliding window.

The DDF, crafted from the DF, encapsulates the relational intricacies within the data, potentially unveiling hidden patterns beneficial for ML algorithms.

To evaluate the efficacy of both datasets for ML training, we first conducted a correlation analysis among features within DF, SDF and DDF to identify if the derived features in SDF and DDF introduce redundant information (shown as highly correlated features) that can hamper the performance of ML models.

Subsequently, we assessed the feature relevance through a feature importance analysis by using a Random Forest Regressor (RFR). This analysis provided a quantitative measure of the contribution of each feature in SDF and DDF compared to the original DF. We also evaluated the training and validation loss after first 100 epochs from two ML models: regression and transformers based (the details of ML models are described in section 4.2) trained on 80% of each dataset.

Next, we calculated variance of error and SNR to compare the noise content in data for each dataset. Furthermore, we evaluated the data quality of all datasets by analyzing the extent of missing data and the prevalence of duplicates. This step was crucial to ensure that the quality of SDF and DDF was not compromised in the process of feature engineering.

Lastly, we examined the dimensionality of all datasets. While a higher number of features in DDF could imply more comprehensive information, it was essential to balance it against the risks of overfitting and the curse of dimensionality, which can adversely affect ML models.

Comparison of datasets, shown in Table 1, allowed to identify the most suitable dataset for ML training, ensuring that the chosen dataset not only offers a breadth of information but also maintains the quality and relevance necessary for effective ML modeling.

The findings from this analysis can be taking as a guideline in the process of dataset selection for ML training.

Table 1: Comparison of datasets.

Dataset	DF	SDF	DDF
Length	72474	6525	6357
Entropy	3.27	3.27	3.32
Dimensionality	10	26	30
Correlation	0.28	0.32	0.12
Variance of Error	0.000004	0.000753	0.000042
SNR	74493.5	423.0	7700.5
Outliers Detected	5419	1233	97
Inconsistencies Detected	5010	741	97
Regression training/validation loss, after 100 epochs	0.0381 0.1086	0.0271 0.0888	0.0586 0.0565
Transformer training/validation loss, after 100 epochs	0.0427 0.1103	0.0374 0.0985	0.0340 0.0498

Comparing dataset's metrics from Table 1, we see that DDF characteristics make it the superior choice for developing robust ML models. First, it's important to note that the DDF has highest entropy value, indicating that the information was not lost as we transform the data from DF to the DDF. It shows the lowest average correlation among features, suggesting

that each feature brings unique information to the model. The SNR of the DDF is considerably high, demonstrating that the data contains useful signals compared to noise. Additionally, the DDF has the fewest outliers and inconsistencies, which eliminates the effort for the data cleaning process. The DDF, with its higher entropy, lower inter-feature correlation, and increased dimensionality, appears to offer a more complex and less redundant representation of the data. Figure 1 demonstrate how inter-correlation in a dataset decay from DF to DDF.

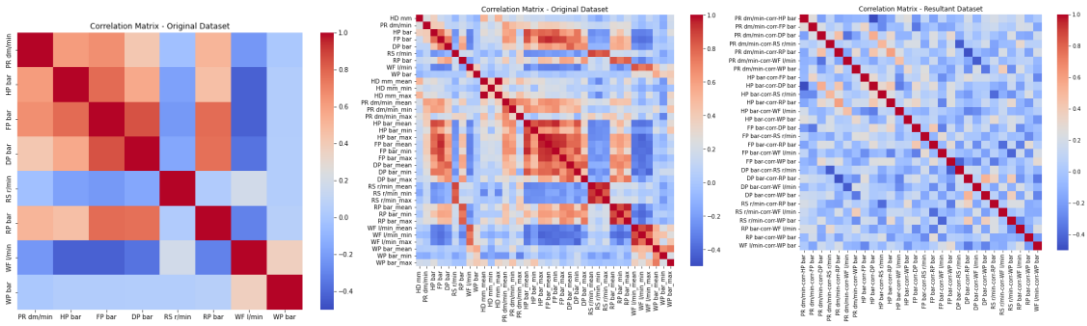

Fig. 1: Correlation matrices for DF (left), SDF (middle), and DDF (right).

Given these considerations, the DDF dataset was selected to use for ML model training.

Also, with approach used to construct DDF we implicitly apply data normalization and dimensionality reduction. The dimensionality reduction implies that instead of looking at MWD timeseries for each hole (or less informative average/min/max statistical description of these timeseries) we will end up with one row of correlation coefficients for each borehole.

2.2. Predicting Overbreak Quantity

Following the data preparation stage, we build and train predictive ML models to forecast the overbreak volume – an essential aspect concerning operational safety and cost-efficiency within tunnelling projects. Here, we employed two ML architecture to predict the volume of over excavation based on MWD data: RFR and Transformer. Transformer-based machine learning models is relatively new type of ML, but well known for their effectiveness. These models have revolutionized natural language processing, recently have being adapted for use with tabular data through transfer learning techniques [6].

Each ML model undergo training and then was tested and evaluated. The DDF serves as the training ground for all models. The models are trained using supervision method to predict volume of over/under-excavation [dm3]. Models' prediction accuracy is illustrated in Figure 2.

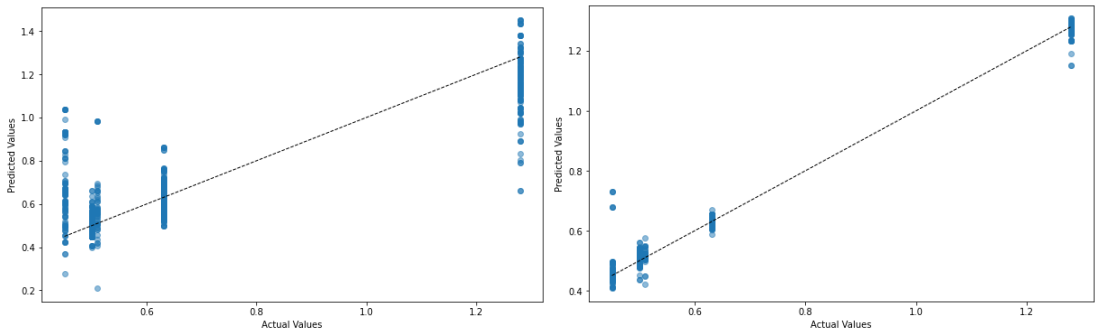

Fig. 2: Scatter plot of actual vs predicted under/over excavation volume for RFR (left) and transformer-based (right) ML models, both trained on DDF.

Transformer-based ML model demonstrates a higher degree of accuracy and able to resolve the volumes of over excavation with RMSE under 0.1 m3 for 98% of the cases.

4. Conclusion

This work presents a workflow for predicting the over excavation volumes with ML models trained on preprocessed MWD data. It demonstrates how the use of correlation analysis aids in preparing the training dataset that provides insights into the physics of rock drilling.

We also describe a process for dataset selection for ML training. The described methodology for evaluating the amount of information in the dataset can be applied as guideline to merely any dataset aimed for ML training. This allows to avoid commonly used time-consuming trial-and-error approach in feature- and dataset selection.

We demonstrate that predictive model with transformer-based architecture outperforms RFR model, even when trained on the same dataset. Transformer-based ML model showed a higher degree of accuracy in predicting overbreak quantity.

The proposed strategy for building ML models capable of accurately predicting the risk of over/under-excavation can contribute significantly to operational optimization:

- proactive approach to risk management can prevent incidents that might otherwise result in costly delays and endanger workers' safety.
- insights provided by the predictive models enable more efficient use of resources.
- optimizing drilling and blasting operations, the models help minimize waste and reduce the overall cost of tunnelling projects.

The study also considers the environmental aspect of tunnelling operations. The accurate predictive model can aid in reducing the environmental footprint by optimizing the use of materials and reducing unnecessary excavation, which in turn minimizes the disturbance to the surrounding ecosystem.

References

[1] X.Cheng, H.Tang, Z.Wu, D.Liang, Y.Xie, “BILSTM-Based Deep Neural Network for Rock-Mass Classification Prediction Using Depth-Sequence MWD Data: A Case Study of a Tunnel in Yunnan, China”, *Appl. Sci.* 2023, 13, 6050. [Online]. Available: https://doi.org/10.3390/app13106050

[2] F.Shan, X.He, H.Xu, D.J.Armaghani, D.Sheng, “Applications of Machine Learning in Mechanised Tunnel Construction: A Systematic Review.”, Eng 2023, 4, 1516-1535. [Online]. Available: https://doi.org/10.3390/eng4020087

[3] V.Isheyskiy, E.Martinyskin, S.Smirnov, A.Vasilyev, K.Knyazev, T.Fatyanov, "Specifics of MWD Data Collection and Verification during Formation of Training Datasets," *Minerals*, vol. 11, 798, 2021. [Online]. Available: https://doi.org/10.3390/min11080798

[4] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948. [Online]. Available: https://ieeexplore.ieee.org/document/6773024

[5] R.Bellman, *Dynamic Programming*, Princeton University Press, 1957. [Online]. Available: https://press.princeton.edu/books/hardcover/9780691653961/dynamic-programming

[6] H.Hotz, “Transformers for Tabular Data: How to use transformers with your own data” . *Towards Data Science* [Online]. Available: https://towardsdatascience.com/transformers-for-tabular-data-b3e196fab6f4