DOI: https://doi.org/10.68732/918bf63c090b5329

ISBN: 978-9989-2053-7-8

Overbreak/underbreak prediction in drill-and-blast tunnelling using machine learning: automated feature selection from MWD data

A. Sapronova & T. Marcher
Graz University of Technology, Graz, Austria
alla.sapronova@tugraz.at
0000-0003-2863-1536

Abstract

This overbreak and underbreak in drill-and-blast tunnelling affect excavation volume, support demand, construction time, and cost. This study investigates whether measurement-while-drilling (MWD) data can support the prediction of radial excavation deviation, ΔR, and which MWD-derived variables are most informative for this task.

Datasets from three tunnel projects were analysed. Features were computed at the contour-hole-pair scale and included raw MWD summary statistics, pairwise correlation coefficients between MWD channels, principal components of MWD statistics, and selected drilling-geometry variables. A multi-criteria feature-ranking workflow was applied using redundancy analysis, principal components. (PC) -based variance contribution, mutual information, Light Gradient Boosting Machine (LightGBM) importance, Random Forest importance, and Kolmogorov-Arnold Network (KAN)-based permutation importance.

The results show that feature relevance is strongly project-specific. Flushing-related correlation features were dominant in two projects, while penetration-rate/rotation-speed coupling was dominant in the third. Adding KAN-based permutation importance changed the top-ranked features in most tested project–representation combinations and identified features with negative permutation importance, meaning that their removal improved validation performance. Prediction benchmarking showed that selected feature subsets can reduce model size and improve or maintain predictive performance, although MWD-derived features alone were insufficient for robust prediction in some projects.

The study demonstrates that physically meaningful variables are not necessarily the most useful predictors in a given dataset. MWD-based feature selection should therefore be performed project-specifically and interpreted together with drilling geometry, blast design, and geological observations.

Keywords

Measurement While Drilling, Overbreak/underbreak, Feature selection, Permutation importance, Drill-and-blast tunnelling.

1 Introduction

In drill-and-blast tunneling, the excavated profile often deviates from the design contour, producing overbreak or underbreak. These deviations affect construction safety, support demand, excavation volume, schedule, and cost. Overbreak increases mucking and shotcrete consumption, while underbreak may require scaling, trimming, or re-blasting before support installation. Reliable prediction of profile deviation is a practical requirement for improving excavation control.

Measurement-while-drilling (MWD) technology provides continuous operational telemetry that records the drilling system's response to ground resistance and can therefore serve as an indirect source of information about geomechanical conditions. The physical basis for relating drilling response to rock resistance can be traced to the concept of specific drilling energy (Teale 1965), while subsequent work showed that drill-performance parameters can support rock-mass characterisation and RQD prediction (Schunnesson 1996; Schunnesson 1998). More recent studies have used MWD data for rock-mass assessment and support prediction (van Eldert et al. 2020), improved filtering and normalisation of tunnel MWD data (van Eldert et al. 2020), and detection of potential overbreak zones in tunnel blasting (Navarro et al. 2018).

Despite the increasing volume of available MWD data and the wider adoption of machine learning (ML), quantitative prediction of excavation-profile deviation remains difficult. The target variable (overbreak/underbreak) depends on rock-mass structure, drilling geometry, explosive charge distribution, drilling accuracy, excavation sequence, and local operational conditions. In addition, MWD datasets are usually sparse, project-specific, and affected by equipment-dependent control logic. Recent studies have applied ML models to overbreak prediction using geological, blasting, drilling, scanner, photogrammetry, or MWD-derived variables (Navarro et al. 2018; Zhao et al. 2023; Qiu et al. 2024; Liu et al. 2025; Im et al. 2025; Sakai et al. 2024). Variables selection for training ML models in this approach is commonly performed through isolated procedures, such as PCA-based reduction (Sakai et al. 2024; Im et al. 2025), tree-based importance (Zhao et al. 2023; Qiu et al. 2024; Sakai et al. 2024), SHAP-type or related interpretability explanations (Qiu et al. 2024; Zhang et al. 2025; Im et al. 2025), or manual engineering preselection of geological, blasting, drilling, or MWD variables (Navarro et al. 2018; Zhao et al. 2023; Sakai et al. 2024). This leaves unresolved the question of whether the same variable remains important when assessed simultaneously for redundancy, variance structure, statistical dependence on the target, model-based predictive contribution, and held-out generalization behavior. A single ranking criterion can favor redundant, model-specific, or split-specific predictors.

This motivates the use of a feature-ranking strategy that combines several complementary views of feature relevance. In the present study, redundancy analysis is used to penalise variables that duplicate information already carried by other predictors. PCA-based variance contribution identifies variables associated with the dominant structure in the input space; mutual information measures statistical dependence with the target variable; tree-based importance evaluates nonlinear predictive contribution in ensemble models; and permutation importance tests whether model accuracy decreases when one feature is randomly shuffled. The last component is especially useful because permutation importance can also be negative: the feature may have reduced the model's ability to generalise.

This workuses a feature-ranking workflow for MWD-based overbreak and underbreak prediction in drill-and-blast tunnelling. The workflow combines statistical, exploratory, and model-based criteria: redundancy analysis, PC-based variance contribution, mutual information, tree-based importance, and KAN-based permutation importance. Its purpose is to test whether a variable remains useful when evaluated from several complementary viewpoints. First, the work compares feature relevance across several definitions of importance and combines them into a single ranking. Second, it uses KANs as an additional diagnostic model for permutation importance. Third, it treats negative permutation importance as useful diagnostic information. This workflow helps separate variables that are physically meaningful from variables that are statistically useful for prediction in a given dataset.

2 Data and feature representations

All features are computed per contour-hole pair, which is the spatial unit of the target variable ΔR. Three MWD-derived representation types are used.

The first consists of raw MWD summary statistics (MWD). For each borehole, mean, maximum, and minimum values are computed from the depth-resolved MWD signals after excluding the initial 60 cm of each hole to reduce collaring effects. These borehole-level statistics are then averaged over the two adjacent contour holes that define the ΔR segment.

The second representation consists of pairwise correlation coefficients (CCs) between MWD channels. For each borehole, Pearson correlation coefficients are computed from the raw depth-resolved signals, before aggregation into summary statistics. These features capture within-borehole inter-channel coupling, for example how penetration rate co-varies with feed pressure or rotation speed along drilling depth. Correlation coefficients are bounded and dimensionless, which makes them less dependent on the absolute scale of rig-specific measurements than raw MWD magnitudes.

The third representation consists of principal components (PCA) fitted to standardised raw MWD statistics on the training partition only. Components explaining at least 90% of cumulative variance are retained. In this study, PCA is treated as an input representation in its own right, not only as a dimensionality-reduction step inside a downstream model.

Drillinggeometry variables (GEOM) are used during the benchmarking stage. GEOM includes the average drilling angle of the two adjacent contour holes and their spacing. GEOM features describe the geometric execution of the designed contour and treated separately from MWD-derived features.

The three MWD-derived representations are tested individually and in all combinations, giving seven input variants per project: MWD, CC, PCA, MWD+CC, MWD+PCA, CC+PCA, and MWD+CC+PCA. Geometry-augmented input variants, including GEOM, MWD+GEOM, CC+GEOM, PCA+GEOM, and MWD+CC+PCA+GEOM, are used to assess the extent to which drilling geometry contributes beyond MWD-derived information. The ranking is computed independently for each project-representation combination.

3 Methodology

The workflow consists of three steps: multi-criteria feature ranking, KAN-based permutation importance, and benchmarking of selected feature sets. All preprocessing, ranking, and model fitting are performed using the training set. Separate tunnel faces are used only for final evaluation and for permutation-importance testing. Each candidate feature receives a consensus rank derived from six scoring criteria (Table 1). The criteria cover redundancy, variance structure, statistical dependence with the target, and model-based predictive contribution. The consensus rank is the arithmetic mean of the six ranks, with lower values indicating more favourable features.

Table 1. Feature-scoring criteria used for consensus ranking.

Criterion	What it measures	Ranking direction
Redundancy	Mean absolute Pearson correlation between one feature and all other features	Lower is better
PCA loading contribution	Feature contribution to retained principal components, weighted by explained variance	Higher is better
Mutual information	Statistical dependence between the feature and ΔR	Higher is better
LightGBM importance	Gain-based contribution in a tree-boosting model	Higher is better
Random Forest importance	Mean decrease in impurity in Random Forest Regression	Higher is better
KAN permutation importance	Change in validation accuracy after randomly shuffling one feature	Higher is better; negative values retained

KANs are used as an additional diagnostic model for the permutation-importance test (Liu et al. 2025). In this study, their role is to provide an additional estimate of the usefulness of features. Positive permutation importance means that shuffling the feature worsens the prediction. Negative permutation importance means that shuffling the feature improves prediction, suggesting that the feature may reduce generalisation for the selected project split. Negative values are retained because they provide diagnostic information about potentially harmful predictors. For benchmarking, three feature sets are

defined for each project–representation combination. ALL contains all candidate features. TOP_K contains the top-ranked features at the elbow of the sorted consensus-rank curve, or K = 15 where no clear elbow is present. KAN selection contains only features with positive permutation importance. Random Forest Regression (RFR), LightGBM, and KAN models are trained on each feature set and evaluated on spatially separated validation faces using RMSE, R^2, and SMAPE. Workflow design (see Fig. 1) allows the effects of feature selection and model choice to be compared.

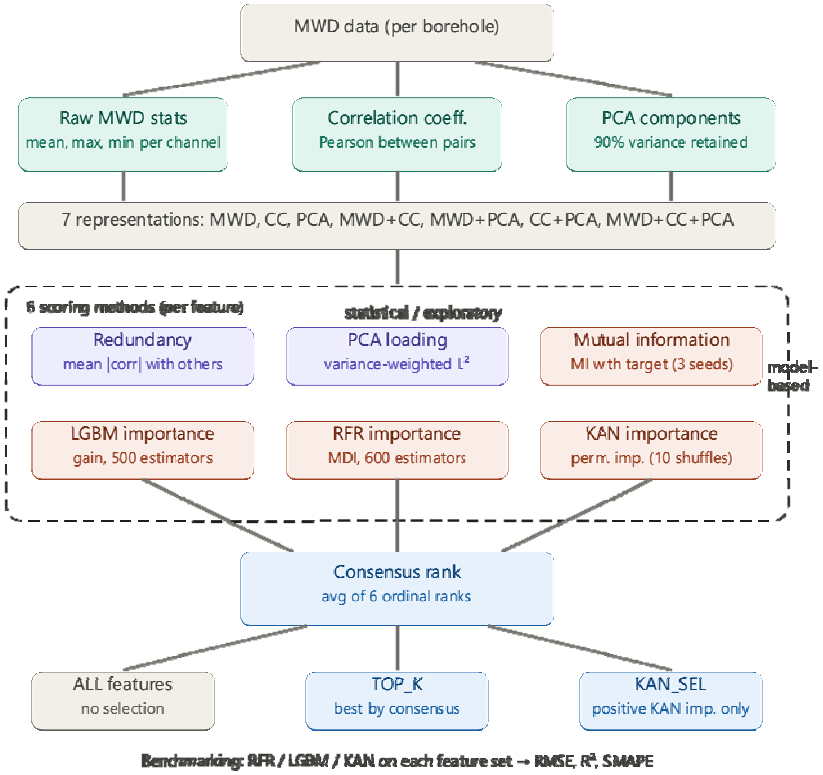


Fig. 1 Multi-criteria feature-ranking workflow: MWD data transformed into summary statistics, CC, and PCA; drilling geometry is added as a separate representation. Six scoring methods are applied: redundancy, PCA loading-based variance contribution, mutual information, LightGBM importance, Random Forest importance, and KAN permutation importance. The resulting ordinal ranks are averaged to obtain a single rank per feature. Three feature sets, ALL, TOP_K, and KAN_SEL, are then benchmarked using RFR, LightGBM, and KAN models.

4 Results and discussion

The consensus ranking was computed independently for each project and feature representation. The top-ranked correlation coefficient (CC) features differed across projects: Project A was dominated by flushing-flow/flushing-pressure couplings, Project B by penetration-rate/rotation-speed couplings, and Project C by flushing-related couplings. Adding KAN permutation importance changed the top-three consensus features in 16 of 21 project–representation combinations: 6 of 7 for Project A, 4 of 7 for Project B, and 6 of 7 for Project C (Table 2). Negative KAN permutation importance occurred frequently, especially in larger feature spaces. Under the full MWD+CC+PCA representation, 47% of Project A features, 18% of Project B features, and 48% of Project C features had negative importance. In the single-representation cases, 50% of Project A MWD features, 36% of Project C MWD features, and 56% of Project C PCA features had negative importance, while Project B had no negative-importance CC features. In the geometry-augmented ranking, contour-hole spacing ranked in the bottom quartile across all three projects.

Table 2. Condensed results from ranking and benchmarking.

Result	**Project A**	**Project B**	**Project C**	**Interpretation**
Top CC features	Flushing flow-pressure	PR-rotation speed	Flushing-related couplings	Best predictors are project-specific
Top-3 changed after adding KAN	6/7	4/7	6/7	KAN importance is not redundant with other scores
Negative importance examples	MWD 50%; all 47%	CC 0%; all 18%	MWD 36%; PCA 56%; all 48%	Larger input spaces contain more harmful predictors
Best consensus-ranked result	PCA: RMSE 129.46, R^2 −0.807	TOP_K CC: RMSE 97.43, R^2 −0.026	TOP_K CC: RMSE 59.17, R^2 0.161	MWD-only prediction is project-dependent

The results show that feature importance is strongly project-specific. The dominant CC features differed among the three tunnels: flushing-related couplings were most important in Projects A and C, whereas penetration-rate/rotation-speed coupling dominated in Project B. These patterns are physically explainable: flushing variables may reflect borehole hydraulic conditions, while penetration and rotation variables may reflect mechanical contrast along the borehole.

The effect of KAN-based permutation importance shows that feature selection depends on the model architecture used to estimate importance. Adding KAN changed the top-three consensus features in 16 of 21 project–representation combinations. This does not imply that KAN is the best predictive model. Its role in this study to provides an additional view of feature usefulness byshowing how prediction accuracy changes when a single feature is randomly shuffled. A negative permutation-importance value indicates that model accuracy improved after a feature was shuffled. Such a feature may carry noise or redundant information, thus, it does not generalise to the validation faces.

The results from GEOM clarify the engineering interpretation. Contour-hole spacing ranked low in the geometry-augmented ranking, although spacing is physically important for blast design. This shows the difference between physical relevance and statistical discriminative power. In the analysed datasets, spacing was likely controlled within a narrow range, so it carried limited information ML.

Prediction performance limits the operational interpretation. Project C showed the clearest benefit from feature selection: KAN trained on TOP_K CC features improved over KAN trained on all CC features. On Project A, removing features with negative KAN importance reduced RMSE. However, Projects A and B produced negative R^2 values for MWD-derived representations alone. This indicates that MWD-derived features, by themselves, were insufficient to predict ΔR better than the training-set mean under the tested validation split.

Several limitations remain. First, the validation is based on spatially separated faces, but alternative grouped splits were not tested. Second, the benchmarking does not exhaustively evaluate every project–representation–model combination. Third, the KAN architecture was fixed and used as a diagnostic instrument, rather than optimised as a final predictor. Fourth, the evaluation metrics treat overbreak and underbreak errors symmetrically, although their operational costs differ. Finally, all three projects involve percussive drilling in drill-and-blast tunnelling; the observed feature rankings may differ for other excavation methods, drilling systems, or geological environments.

5 Conclusions

This work introduced a multi-criteria consensus feature-ranking workflow for MWD-based overbreak and underbreak prediction in drill-and-blast tunnelling. The workflow combines redundancy analysis, PCA-based variance contribution, mutual information, tree-based importance, and KAN-based permutation importance.

The results show that no single MWD-derived feature is universally dominant. Feature relevance depends on the project, geological setting, rig configuration, available channels, and target distribution. Therefore, feature rankings should be recomputed for each project rather than transferred directly between tunnels.

KAN-based permutation importance changed the selected top features in most tested project–representation combinations and identified features with negative importance. This supports the use of architecturally diverse importance measures and shows that feature selection should distinguish between helpful, neutral, and harmful predictors.

The predictive results show that feature selection can reduce model size and improve validation performance. However, MWD-derived features alone were insufficient on some projects. Robust operational prediction of ΔR should integrate MWD data with drilling geometry, explosive charge information, drilling accuracy, and face-level geological mapping.

The proposed workflow provides a reproducible way to identify which signals are statistically useful in a given tunnel dataset and which variables may reduce generalisation.

References

Im H, Kurauchi T, Sato N, Kawamura Y, Jang H, Topal E (2025) Tunnel overbreak prediction: an integrated approach using 3D photogrammetry and machine learning. *Mining, Metallurgy & Exploration* 42:1441–1457. https://doi.org/10.1007/s42461-025-01278-1

Liu Y, Li A, Zhang H, Wang J, Li F, Chen R, Wang S, Yao J (2023) Minimization of overbreak in different tunnel sections through predictive modeling and optimization of blasting parameters. *Frontiers in Ecology and Evolution* 11:1255384. https://doi.org/10.3389/fevo.2023.1255384

Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, Hou TY, Tegmark M (2025) KAN: Kolmogorov-Arnold Networks. *Proceedings of the 13th International Conference on Learning Representations, ICLR 2025*. OpenReview. https://openreview.net/forum?id=Ozo7qJ5vZi

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

Qiu Y, Zhou J, He B, Armaghani DJ, Huang S, He X (2024) Evaluation and interpretation of blasting-induced tunnel overbreak: using heuristic-based ensemble learning and gene expression programming techniques. *Rock Mechanics and Rock Engineering* 57:7535–7563. https://doi.org/10.1007/s00603-024-03947-x

Sakai K, Miyanaga S, Yamagami M, Sapronova AV, Soliman A (2026) Study on machine learning algorithm for optimal tunnel blasting design. In: Deb D et al. (eds) *Advances in Rock Mechanics—Infrastructure Development, Volume 1*. Lecture Notes in Civil Engineering 783, Springer, Singapore, pp 245–254. https://doi.org/10.1007/978-981-95-4259-8_26

Sapronova A, Marcher T (2025) Improving data quality with advanced pre-processing of MWD data. *Geotechnics* 5(2):28. https://doi.org/10.3390/geotechnics5020028

Schunnesson H (1996) RQD predictions based on drill performance parameters. *Tunnelling and Underground Space Technology* 11(3):345–351. https://doi.org/10.1016/0886-7798(96)00024-7

Schunnesson H (1998) Rock characterisation using percussive drilling. *International Journal of Rock Mechanics and Mining Sciences* 35(6):711–725. https://doi.org/10.1016/S0148-9062(97)00332-X

Teale R (1965) The concept of specific energy in rock drilling. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* 2(1):57–73. https://doi.org/10.1016/0148-9062(65)90022-7

van Eldert J, Schunnesson H, Johansson D, Saiang D (2020) Application of measurement while drilling technology to predict rock mass quality and rock support for tunnelling. *Rock Mechanics and Rock Engineering* 53:1349–1358. https://doi.org/10.1007/s00603-019-01979-2

van Eldert J, Schunnesson H, Saiang D, Funehag J (2020) Improved filtering and normalizing of measurement-while-drilling data in tunnel excavation. *Tunnelling and Underground Space Technology* 103:103467. https://doi.org/10.1016/j.tust.2020.103467

Zhang Y, Zhou J, Li J, He B, Armaghani DJ, Huang S (2025) Advancing overbreak prediction in drilling and blasting tunnel using MVO, SSA and HHO-based SVM models with interpretability analysis. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources* 11:53. https://doi.org/10.1007/s40948-025-00963-1