International Society for Rock Mechanics and Rock Engineering
Norwegian Group for Rock Mechanics

ISRM International Symposium
Eurock 2025 – Expanding the Underground Space
Trondheim, Norway, 16-20 June
ISBN 978-82-8208-079-8

Using Machine Learning for Predicting Collapse extending in Abandoned Underground Mines

L-M. Guayacán-Carrillo
Laboratoire Navier, Ecole nationale des ponts et chaussées, Institut Polytechnique de Paris, Université Gustave Eiffel, CNRS, 77455 Marne la Vallée, France.

N. Conil
Institut National de l'Environnement Industriel et des Risques (Ineris), Nancy, France
nathalie.conil@ineris.fr

A. Kadri
Laboratoire Navier, Ecole nationale des ponts et chaussées, Institut Polytechnique de Paris, Université Gustave Eiffel, CNRS, 77455 Marne la Vallée, France.
Faculté de sciences et technologies de Nancy, Université de Lorraine

Abstract

The main objective of this work is to explore the advantages of applying machine learning tools to the analysis of collapse related to abandoned underground mines. To this end, the present work focuses on the analysis of data recorded by Ineris for 143 collapses cases, with particular attention to the extension of the collapse to the surface. This work presents the procedure followed from the creation of a clean database, through the selection of features, to the proposal of a tool capable of estimating the extension of the collapse to the surface. This exploratory work confirms the significant potential of using machine learning on geotechnical data.

Keywords

Abandoned underground mines; machine learning; synthetic data generation; random forest; smote.

ISRM

NBG

1 Introduction

Because of its geology and industrial history, France has an enormous number of underground cavities. Their number is estimated at several hundred thousand, many of which are not located. These underground mines deteriorate over time, which can lead to land movements such as subsidence or local or widespread collapse. Therefore, particular interest should be paid to abandoned mines. With climate change, we can expect an increase in the frequency of exceptional hydroclimatic events (intense droughts, heavy rainfall or floods), which could lead to significant changes in groundwater levels or watercourses or cause heavy water infiltration. All these phenomena are likely to increase the risk of instability in these water-sensitive underground mines. An accidentology study of collapses associated with underground mines of water-sensitive rocks (gypsum, chalk, limestone) was initiated by Ineris in 2022. 550 collapse-type ground movements were identified (Conil & Gombert, 2024). Initial analysis of the data showed that only a few percent of the cases studied could be linked to "extreme rainfall". It confirms that this factor alone is of course not sufficient to explain the risk of collapse of a water-sensitive underground structure and that other factors related to the geometric and geomechanical characteristics of the mines must also be considered (Conil et al., 2024, Conil et Gombert 2024).

In order to improve the ability to analyse complex behaviour observed in situ, a number of methodologies have been developed to integrate various classical numerical techniques with machine learning tools. Indeed, over the last decade, machine learning (ML) techniques have experienced significant growth in geotechnical engineering. An overview of ML applications in geotechnical engineering can be found in several recent papers that present a detailed state of the art, mainly with applications to ground-structures interaction and to constitutive model identification (e.g. Morgenroth et al. 2019, Jong et al. 2021, Gao 2021, Baghbani et al. 2022). These methods offer significant advantages due to their superior computational capabilities and applicability to complex, high-dimensional problems. However, there are still concerns about their effectiveness and reliability when dealing with incomplete, noisy and limited datasets.

The current study focuses on examining data compiled by Ineris for 143 collapse incidents, paying particular attention to the extent of collapse reaching the surface. The main objective of this work is to determine how effectively machine learning tools can examine this inventory and derive valuable insights into the mechanisms responsible for the collapses. This paper provides a brief overview of the abandoned mines in France, followed by an explanation of the procedure used to create a clean database. It then discusses the soft computing approach used to estimate the extent of collapses and examines the effect of using synthetic data generation to improve the training dataset for machine learning algorithms, thereby optimising the results.

2 Abandoned mines collapse in France

Since 2021, Ineris has been compiling an inventory of abandoned mine collapses in France. The aim of this inventory is to carry out analyses in order to understand the circumstances of the collapses. This involves the collection of various types of information (dates, descriptions, geographical and geological parameters, etc.) selected on the basis of hypotheses about the possible causes of the events that have occurred. In order to be robust, the analyses must therefore be based on complementary quantitative data (e.g. distance from a watercourse) and qualitative data (e.g. type of material used) that make it possible to propose a scenario of events linked to predisposing or aggravating factors. Prior to this, it is necessary to identify the relevant information that needs to be collected in order to understand the ageing kinetics of these structures that led to the collapses.

To date, the list of collapsed mines covers approximately 150 communes in 24 departments. The location with the highest representation are Gironde, Indre-et-Loire, and Île-de-France, corresponding to the most commonly extracted materials: limestone, chalk, and gypsum. It should be noted that the initial aim was to obtain representative samples (several dozen) from underground mines in homogeneous geological and hydroclimatic contexts. However, this list is not exhaustive, as information on mines is still lacking in these and many other departments, particularly in the south of France, where most of the intense hydroclimatic events are concentrated. The distribution over the country of the cases in the inventory for which the location is known is shown in Figure 1.

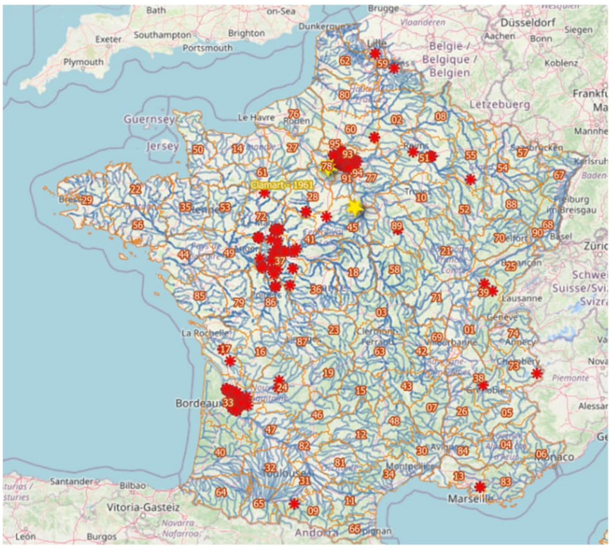

Fig. 1 Collapsed mines listed in the Ineris inventory at 31/12/2023 are represented by red dots (Conil & Gombert 2024).

3 Dataset preparation

3.1 Setting up the database

In first stage, a detailed analysis of the inventory was carried out to develop a clean database. Only surface collapses with a diameter of more than 5 metres were included. An initial pre-processing and standardisation phase of the variables (both quantitative and qualitative) was performed. It should also be noted that the inventory has a large number of missing values, which required careful consideration of their handling, both in terms of database representation and imputation.

The process of creating the dataset is divided into two stages:

- An initial selection of relevant variables was made according to two criteria: (1) variables with a significant amount of missing data and a low completion rate were excluded; (2) variables considered by engineering experts to be potentially influential in the occurrence of collapses were included.
- The data are statistically analysed and visualised to confirm the selection of key variables associated with collapse and to identify any correlations between these variables. The diameter of the collapse was used as a target variable to indicate the severity of the collapse.

In the end, a dataset consisting of six key variables was retained for analysis. A description of the six variables is given in Table 1. As a result of this selection, the rate of missing values in the final dataset was only 14%.

Table 1 Database description

Variable	Descriptive	Category
Material extracted	Type of material extracted. Three classes are identified: limestone, chalk, gypsum	Exploitation
Extraction ratio	The average rate of exploitation of the mine	Exploitation
Geomorphology 1 (see Figure 2)	Describes the position of the mine in relation to the valley mine. Five classes are identified: valley, lower slope, middle slope, upper slope and plateau.	Mine environment
Geomorphology 2 (see Figure 2)	Describes the geomorphological features of the terrain around a mine. 4 classes are identified: Cornice, spur, crest, plateau.	Mine environment
Distance to valley	Distance to watercourse	Mine environment
Equivalent surface diameter	Diameter of the collapse at the surface (in metres)	Collapse

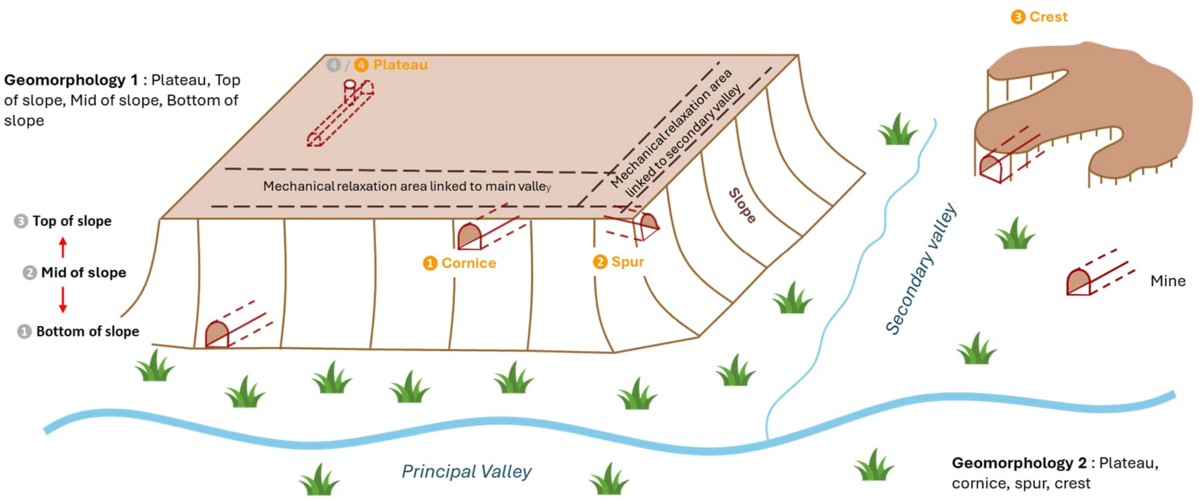
Geomorphology 1 : Plateau, Top of slope, Mid of slope, Bottom of slope
4 / 4 Plateau
3 Crest
Mechanical relaxation area linked to main valley
Mechanical relaxation area linked to secondary valley
Slope
3 Top of slope
2 Mid of slope
1 Bottom of slope
1 Cornice
2 Spur
Secondary valley
Mine
Principal Valley
Geomorphology 2 : Plateau, cornice, spur, crest

Fig. 2 Geomorphology conditions

Therefore, the database account with five variables that will serve as input parameters for the training algorithm, related to two main categories: mine exploitation (material extracted and extraction ratio) and mine environment (distance to valley, geomorphology 1 and geomorphology 2). Finally, a variable related to collapse information (equivalent surface diameter) serves as the basis for the generation of the output parameter, as explained in section 3.2.

3.2 Output: Extension of collapse

In a first step, a classification process is proposed in order to obtain the optimal gain from the use of machine learning tools from the constituted database with only 143 cases and with variables that are quantitative and qualitative. For this purpose, the variable 'equivalent surface diameter', which is a quantitative variable, has to be divided into classes. In order to construct the new target 'collapse extension', three collapse classes were defined: (1) small collapse, (2) moderate collapse and (3) large collapse. These classes were designed to be balanced in order to ensure the robustness of the future model. The strategy followed for this construction is divided into three stages:

- Analysis of the distribution of collapse diameters.
- Construction of an "intermediate" variable. In fact, for each diameter value, a specific modality is assigned to the following three identified classes: small collapse (from 5 to 20 metres in diameter), medium collapse (from 20 to 50 metres in diameter) and large collapse (from 50 to the maximum diameter recorded, in this specific case 451.5 m).
- Coding of the intermediate variable. This last step consists in transforming the variable into an ordinal variable that will be used to train the algorithm: 0 - small collapse; 1 - medium collapse; 2 - large collapse.

Figure 3 shows the dispersion of the equivalent surface diameter of the cases studied and the final distribution per class defining the "collapse extension" output.

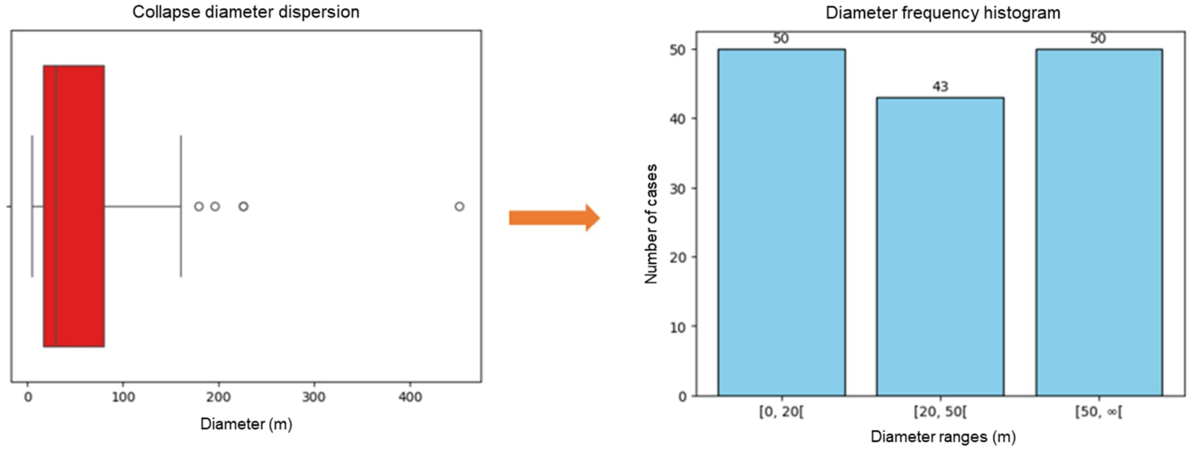

Fig. 3 Output variable distribution.

4 Soft computing approach

Small data sets can cause learning models to overfit, resulting in poor predictions. Combining multiple models to produce an overall and optimised prediction is a practical solution to this problem. Previous studies (e.g. Richa et al. 2024, Tristani et al. 2024) have concluded that 'ensemble methods' work best and provide the best prediction accuracy when a limited dataset is used to train the model. To this end, the use of an ensemble method such as Random Forest, which is well known for its high performance, will be tested below.

4.1 Random forest

Random forest (RF) is an ensemble machine learning technique developed by Breiman (2001). It can be applied to both regression and classification problems. As explained by Hastie et al. (2010), RF is a modification of bagging (a technique for reducing the variance of an estimated prediction function) that builds a large collection of decorrelated trees and then averages them. This technique involves randomly selecting subsets of the data with replacement, introducing diversity among the individual trees. In classification tasks, the final prediction is determined by the majority vote of the individual trees' predicted classes. This increases the robustness of the model to overfitting and reduces variance. RF is a popular technique for soil-structure interaction applications (e.g., Zhou et al. 2017, Xie et al. 2019) due to its ease of understanding and good predictive accuracy.

4.2 Applicability for estimating collapse extension

The main challenge of this work is related to the sufficiency of the data (the quality and quantity of the data available for the study). In fact, as explained above, this study deals with a small data set of only 143 cases and with about 14% of missing values. In order to optimise the use of the RF algorithm for predicting collapse extension, the use of imputation methods to complete the dataset is explored.

4.2.1 Imputation of missing data

As explained above, the final dataset contains 14% missing values. It is therefore necessary to use imputation methods to complete the dataset. Particular attention should be paid to the accuracy of imputing missing values in a dataset containing both qualitative and quantitative variables. One solution is a sequential imputation strategy (imputing the quantitative variables first, then the qualitative variables), but this approach has a major drawback: it ignores the possible relationships between the different types of variables, which may lead to less accurate imputations. Therefore, given the interdependence of the qualitative and quantitative variables in the dataset, mixed imputation methods that take these interactions into account are appropriate. After a thorough analysis of the available options, the MissForest method (Stekhoven & Bühlmann 2011) was used to impute missing data. This is a random forest based completion method known for its high performance in mixed imputation.

4.2.2 Training

Once the dataset is complete, it is divided into three groups for training. First, 10 cases were reserved for the training data to test the model's ability to generalise results to unknown cases. Then, the training set consisting of 133 cases is divided into two sub-sets: (i) a training set with 70% of the data and (ii) a test set with the remaining 30%. Noted that, as the aim was to select the best predicting model, to ensure a reliable assessment of model performance, we used the cross-validation technique. Given the relatively small size of dataset, a cross-validation with k = 3 folds is performed. This means that the dataset was subdivided into three equal parts: at each iteration, one subset was used as a test set, while the other two were used to train the model.

4.2.3 Results

Figure 4 shows the prediction results obtained from the 10 cases selected for validation, compared with the real values. Despite accuracies of 0.5 on the validation data and 0.525 on the test data, the performance of the Random Forest model remains well below our expectations, especially for a critical phenomenon such as collapse.

In fact, the model showed difficulty in correctly discriminating moderate collapse cases (class 1), especially when they were close to or similar to cases in the small collapse and large collapse classes. These results can be explained by the small size of the model training set, which is insufficient to capture the nuances of the different types of collapse. Given this limitation, it seems important to explore techniques to increase the size of the test training, such as the use of synthetic data generation methods, which will be discussed in the next section.

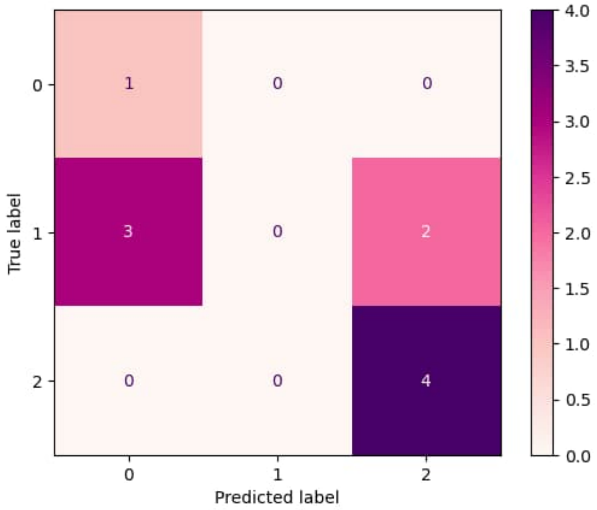

Fig. 4 Confusion matrix for the validation data obtaining after training a model with an initial imputation of data.

4.3 Discussion

As explained above, the weak performance of the model is mainly due to the limited data available. Therefore, to overcome this weakness, the Synthetic Minority Over-sampling Technique (SMOTE, Fernández et al. 2018) was used to generate synthetic data. This is a commonly used method for rebalancing unbalanced datasets by artificially creating new instances of the minority class. The algorithm creates these new instances by taking an existing piece of minority class data and adding a small variation based on one of its nearest neighbours. One of the strengths of this technique is that it creates unique instances, helping to improve the performance of machine learning models on unbalanced datasets. There are two main objectives: (1) First, to generate synthetic data for unbalanced data, as is generally the case in geotechnical projects (in this specific case, after retrieving the 10 cases randomly selected for validation, we obtain data from three classes that are unbalanced: 49 cases for small collapse, 38 cases for medium collapse and 46 cases for large collapse); (2) Considering the ability of this technique to generate new synthetic data based on nearest neighbours rather than duplicating the existing ones, the ensemble data set will also be expanded to explore the possibility of increasing the performance of the models with small data sets by generating synthetic data.

The results obtained show that this method had a positive impact on the performance of the model, both in the test phase and in the validation phase. Figure 5 shows the prediction results obtained for the 10 cases selected for validation. The accuracy is indeed improved, obtaining a value of 0.7 on the validation data, instead of 0.5 obtained when training on the original dataset. It should be noted that different tests were carried out by increasing the training set of different sizes, from 5 to 100 times the original data set. It can be observed that, for the dataset used, an increase of only 10 times the initial size allows to obtain a good accuracy (there is no significant difference with a larger dataset). In fact, the performance of these methods strongly depends on the sufficiency of the initial dataset used. In this particular case, it seems that even increasing the amount of synthetic data (for more than 10 times), since this increase depends on the nearest neighbour base, the model does not learn more new information from the data. It should be noted that this test has been performed with different sets of 10 cases of validation data and the results remain similar. This issue will be investigated further and other oversampling techniques will be tested and compared.

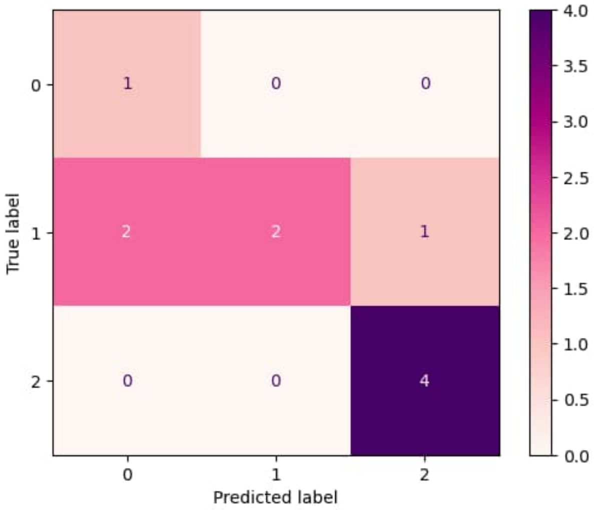

Fig. 5 Confusion matrix for the validation data obtained after training a model with synthetic data generation.

5 Conclusions

The main objective of this work was to evaluate the ability of machine learning methods to explore, analyse, and understand the existing database of 143 collapsed mines, which was created and is managed by Ineris in France. Then, to develop an initial model based on machine learning approach capable of estimating the extension of collapses linked to abandoned underground mines.

A step-by-step methodology was followed. First, a data analysis is performed to identify the factors that have a significant impact on collapses, depending on the data available. This was followed by the creation of a clean dataset in order to prepare the data used in the training algorithm. The final database account with five input variables, related to two main categories: mine exploitation (material extracted and extraction ratio) and mine environment (distance to valley, geomorphology 1 and geomorphology 2). And the output parameter 'Extension of collapse' was therefore created on the basis of a variable related to collapse information (equivalent surface diameter).

The training of a Random Forest (RF) algorithm was performed after a process of imputing missing data to complete the dataset. RF is one of the most popular ensemble methods in machine learning because of its high performance on small datasets. It was observed that imputing missing values helps to improve data quality and model performance. Finally, a first test of using techniques as Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic data and rebalance the classes, showed a high improvement in the performance of the predictive model. The accuracy of the model and its ability to generalise to new cases were tested using a validation dataset consisting of 10 new cases.

Based on the positive feedback of this exploratory work, it is concluded that a promising avenue of research is open, using an interdisciplinary approach, combining the knowledge acquired in geotechnic with that of artificial intelligence, mainly the field of machine learning. The present work has laid the foundations for a solid predictive model using imputation and oversampling techniques. The database is currently being completed and improved, and with the new data the model can be improved. In particular by including other key features (variables) well known to have a significant influence on the occurrence of collapses (pillar slenderness, extraction rate, geomechanical properties, etc.). So, in the future, it will be important to continue to expand the database to include as many variables as possible.

In addition, this work will be complemented by comparisons with other machine learning methods, including other ensemble methods such as XGBoost and CatBoost, which are also known for their performance on small datasets. Once a complete database is obtained, approaches such as symbolic regression will also be tested. This is indeed a valuable option, as it provides a mathematical expression linking input and output data, allowing for fast and reliable estimates for engineering applications.

Finally, another interesting point that will be further explored is to propose a predictive tool to help local authorities make the right decisions about abandoned, non-collapsed mines.

References

Baghbani, A., T. Choudhury, S. Costa et J. Reiner (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. Earth-Science Reviews 228. March, p. 103991. DOI: 10.1016/j.earscirev.2022.103991.

Breiman L (2001). Random forests *Mach. Learn.* 45 pp 5–32.

Conil N, Gombert P (2024). Amélioration de la connaissance des mécanismes d'instabilité et de l'impact du changement climatique sur les carrières souterraines abandonnées en France : premiers résultats Ineris. https://www.ineris.fr/fr/amelioration-connaissance-mecanismes-instabilite-impact-changement-climatique-carrieres. Accessed at 15/11/24.

Conil N., Hauquin T., Gombert P., Al Heib M., Maghami C., Rétro-analyse d'effondrements de carrières souterraines abandonnées (France). 12. Journées Nationales de Géotechnique et de Géologie de l'Ingénieur (JNGG), Jun 2024, Poitiers, France. ⟨ineris-04676407⟩

Fernández, Alberto, Garcia, Salvador, Herrera, Francisco, and Chawla, Nitesh V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61 :863–905, 2018.

Gao W (2018). A comprehensive review on identification of the geomaterial constitutive model using the computational intelligence method. Adv. Eng. Inform. 38 pp 420–440.

Hastie T, Tibshinari R, Friedman J (2010). The Elements of Statistical Learning - Data Mining, Inference, and Prediction Springer 2nd Edition New York.

Jong S.C, Ong D.E.L, Oh E (2021). State-of-the-art review of geotechnical-driven artificial intelligence techniques in underground soil-structure interaction Tunn. Undergr. Space Technol. 113.

Morgenroth J, Khan UT, and Perras MA. (2019). An Overview of Opportunities for Machine Learning Methods in Underground Rock Engineering Design. Geosciences 2019;9:504.

Richa T, Pereira JM, Guayacán-Carrillo LM, Gilles C & Lanquette F. Accuracy of Machine Learning techniques in forecasting tunnelling-induced soil settlements with limited data. In: Geotechnical Engineering challenges to meet current and emerging needs society. Lisbon, 2024.

Stekhoven D. J & Bühlmann P (2011). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Tristani A, Guayacán-Carrillo LM & Sulem J (2024). Data-driven tools to evaluate support pressure, radial displacements and face extrusion for tunnels excavated in elastoplastic grounds. International Journal for Numerical and Analytical Methods in Geomechanics. DOI: 10.1002/nag.3889.

Xie Q, Peng K (2019). Space-time distribution laws of tunnel excavation damaged zones (EDZs) in deep mines and EDZ prediction modeling by random forest regression. Adv. Civ. Eng. 2019.

Zhou J, Sh X, Du K, Qiu X, Li X, Mitri H.S. (2017). Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel. *Int. J. Geomech.* 17, pp 1–12.