University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Data-driven hydrogeological mapping and its impact on the tunnel infrastructure

Thesis presented by

Zahra Hajighasemi

For the degree of

Research Master of Engineering Science

University College Cork

School of Engineering and Architecture

Supervisors:

Dr. Zili Li

Dr. Andrea Visentin

2025

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Zahra Hajighasemi

List of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
CERN	European Organization for Nuclear Research
LHC	Large Hadron Collider
SPS	Super Proton Synchrotron
ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
SVM	Support Vector Machine
SVR	Support Vector Regressor
KNN	K-Nearest Neighbors
CatBoost	Categorical Boosting
XGBoost	eXtreme Gradient Boosting
IC-XGBoost	Iterative Convolution eXtreme Gradient Boosting
MS-GAN	Multi-Scale Generative Adversarial Networks
RF	Random Forest
ANN	Artificial Neural Network
DNN	Deep Neural Network
CNN	Convolutional Neural Network
MPS	Multiple Point Statistics
MRF	Markov Random Field model
CMC	Coupled Markov Chain Simulation
DSI	Discrete Smooth Interpolation
LOOCV	Leave-One-Out Cross-Validation
M	Marl
S	Sandstone
M/C or M-C	Marl-Calcaire
Co	Conglomerate
C	Calcaire
GPR	Ground-Penetrating Radar
VR	Virtual Reality
GR	Groundwater Level
DCUM	Distance Cumulée

Abstract

Tunnel maintenance and predicting failures during operation have been heavily investigated. This investigation has led to new methods to improve prediction accuracy and extend the service life of tunnels. Among the many factors affecting tunnel performance, the hydrogeological profile of the tunnel route is widely considered one of the most important factors. Over time, it can cause various types of deterioration, such as deformation, material degradation, water leakage, cracks in the lining, and other subsurface issues. For these reasons, creating subsurface geological maps and studying tunnel infrastructure hydrology is essential.

Several traditional methods already exist for building geological maps along tunnels, including spatial interpolation, kriging, and Inverse Distance Weighting (IDW). However, these traditional approaches have notable limitations; they struggle to manage sparsely distributed data, scale poorly with large datasets, and have limited ability to capture complex and nonlinear relationships. As an alternative, Machine Learning (ML) methods have gained popularity because of their precise, fast performance and better handling of the aforementioned drawbacks. Supervised ML techniques can learn the correlations between the input features and predict the appropriate label for the output regarding the training set.

In this study, borehole spatial features and lithological information were manually extracted from paper-based reports from multiple sources, then structured, preprocessed, and integrated into a unified dataset. These data sets were used to train different ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), Categorical Boosting (CatBoost), and eXtreme Gradient Boosting (XGBoost); the performance of these models was compared. Each data point has the following attributes: latitude, longitude, altitude, ground level, borehole name, and type of soil/rock. In addition, the neighborhood aggregation was used to add new features to improve the models' accuracy. Among the implemented algorithms, the RF and CatBoost performed better than the others; both algorithms were used to develop a geological map for the Large Hadron Collider (LHC) tunnel in the European Organization for Nuclear Research (CERN) from point 3 to point 4, where leakage data was available. To complete the hydrogeological profile, the hydrological characteristics of the study area were assessed.

Comparing the hydrogeological profile and the tunnel's leakage data, no obvious relationship was noted between the geological characteristics and leakage. Nevertheless, a river near SPM9 (about DCUM 8700) contributes to the high leakage percentage in that section, whilst the terrain map is related to the leakage distribution in the longitudinal direction. Darcy's Law states that groundwater flow is directly proportional to the hydraulic conductivity, the cross-

sectional area, and the hydraulic gradient. As the difference in hydraulic head (i.e., the elevation difference between the ground surface and the tunnel level) increases, the discharge per unit of groundwater into the tunnel also increases. Therefore, the greater the tunnel depth, that is, the larger the difference between the altitude of the tunnel and ground level, the higher the groundwater table, and thus, more leakages are likely to occur.

Acknowledgement

First of all, I would like to thank God for giving me the strength and patience to complete this thesis.

I'm deeply grateful to my lovely husband for his endless support, encouragement, and understanding. His presence made this journey much easier and more meaningful.

I also want to thank my family for always believing in me. Their love and support have been with me every step of the way, even from far away.

Finally, I would like to sincerely thank my supervisors for their guidance, feedback, and support throughout this research. Their advice helped me grow and improve in many ways.

This research is supported by a grant from the European Organization for Nuclear Research (CERN), the Irish Centre for Applied Geoscience (iCRAG), and the Science Foundation Ireland (SFI) under Grant Number 13/RC/2092_P2. I gratefully acknowledge colleagues at CERN for their valuable contributions.

Chapter 1
Introduction

1 Introduction

1.1 Background

Large tunnel networks age over time and require repair and maintenance to ensure structural integrity and serviceability. Many studies are underway to find practical strategies for maintaining tunnels and preventing their deterioration in the long term. Natural and environmental parameters have a significant impact on these strategies (Wang, Xiao, et al., 2024)(Han et al., 2020)(Yuan et al., 2013). Various parameters can affect a tunnel's performance, such as the deformation of concrete lining, cracks, leakage, and seepage. The unknowns and uncertainties of the hydrogeological profile around subterranean infrastructure can make its condition more complex. Over time, the impact of soil/rock on tunnel functionality increases. The failures in the tunnel can cause catastrophic consequences, irreparable safety risks, and tunnel collapse (Baji et al., 2017).

To ensure serviceability, it is necessary to examine the tunnel conditions from different perspectives. One of the aspects that should be considered is the hydrogeological situation. Extracting meaningful information and useful parameters from geological data is challenging and essential. Traditional methods can hardly capture the complex relationships between various factors. For instance, spatial interpolation methods, including kriging, Inverse Distance Weighted (IDW), spline function, and Discrete Smooth Interpolation (DSI), are some traditional methods that can be used. Kriging and DSI offer higher precision geomodelling. Meanwhile, IDW and spline functions are generally faster and less computationally demanding (Sun et al., 2022). Moreover, traditional methods rely on discrete sampling points and expert interpretation, which may overlook the full complexity of subsurface conditions. However, modern approaches like Machine Learning (ML) algorithms can analyse variable interactions more quickly and accurately and find their relationships.

The ML algorithms require digitalized data while geological reports often contain borehole analyses and are presented in paper form. To address this obstacle, digitalization methods of reports should be deployed. Borehole reports typically contain geological details such as soil/rock composition, depth, permeability, and other subsurface characteristics. These reports are typically organized as unstructured tables written manually, which makes them hard to comprehend directly and need to be transformed and structured for analysis. By processing the structured datasets, hidden subsurface patterns and features can be obtained and used for tunnel construction and maintenance. Manual and automated techniques can provide structured reports ready for data extraction, revealing the relationships between the features and their impact on the tunnel. Integrating borehole data into comprehensive geological maps enables a clearer understanding of the underground condition and allows tunnel performance prediction over time.

Moreover, a major problem in civil engineering projects, particularly underground tunnels, is the high costs and complexity of geological data. Due to financial constraints, drilling many boreholes is not feasible, leading to a lack of knowledge about the site's geology. A practical solution for this challenge is to integrate borehole data collected from different areas and build a unified dataset. Integrating multi-source data and creating a single dataset allows engineers to access more information, extract patterns, and understand subsurface conditions. This approach not only helps to have higher-resolution underground geological maps but also can assess the impact of environmental factors on the tunnel more accurately.

Understanding the relationships between geological and hydrological parameters and tunnel defects can ensure these parameters are considered for future tunnel construction. This will also give more insights into designs and engineering practices, enhancing the lifespan and safety of large-scale tunnel infrastructure.

1.2 Research aims and objectives

The goal of this research is to address the impact of the hydrogeological profile on tunnel infrastructure and apply the data-driven methodologies to construct a geological map. Thus, the key objective of this study will be around the following:

1. *Developing a subsurface geological map providing insights into subsurface conditions:*
   One of the goals of this research is to explore the hydrogeological effects on the tunnel construction and performance. To achieve this goal, a detailed understanding of the geological features surrounding the tunnel is required. Therefore, a geological map was developed using the available information on the region.

2. *Manual structuring of borehole log information:*

   Different CERN engineers made the existing boreholes available over several decades. These records were scanned from handwritten documents dating back to the tunnel construction. Due to their limited quantity and various formats, it was not worth extracting tables automatically. So these reports were manually reviewed, and the valuable information was extracted and organized for future stages of the research.

3. *Extracting and integrating borehole information from multiple sources:*

   These boreholes were drilled for the construction of the CERN tunnels. In this study, the reports were collected from various sources. Some were obtained from CERN colleagues and engineers, while others were obtained from publicly available reports about the construction of the CERN tunnels through online browsing. These datasets differed in format, and the information needed to be structured, pre-processed, merged, and standardized.

4. *Applying ML algorithms to generate a geological map:*

   Different approaches exist for developing geological maps. One method that has recently gained popularity among researchers is the use of ML algorithms. These algorithms can learn available patterns with a limited amount of data and can make predictions for new entries and unsampled locations. These new, efficient, and accurate methods can update and improve their performance once more information is provided.

5. *Robustness of ML algorithm:*

   Cross-validation was used to split data sets and assess the model's performance, ensuring robustness despite limited data. At the first step, average accuracy and average balanced accuracy were used to evaluate the models' performance. Confusion matrices also provide details of the model's prediction by learning each data series.

6. *Linking the leakage data to the hydrogeological profile:*

   Leakage data were available for a specific section of the Large Hadron Collider (LHC) tunnel. This study attempted to investigate the correlations between the leakage data and hydrogeological profile, including the surrounding type of rock and soil, the soil cover depth above the tunnel to the terrain, the groundwater level, and the rivers nearby. The potential relationships can inform engineers for future risk assessment and tunnel maintenance.

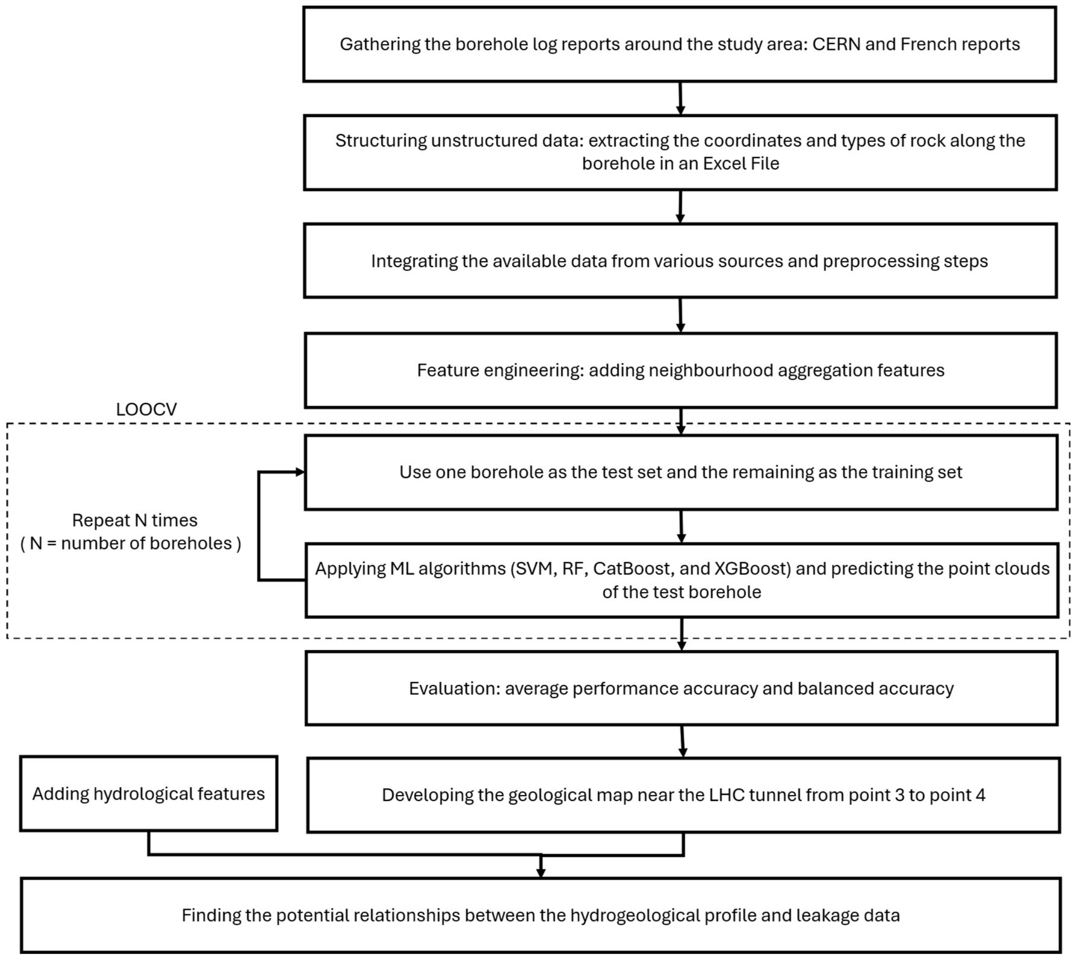


Figure 1:1 Proposed framework

Figure 1:1 illustrates the proposed framework in this research. The procedure begins with collecting and structuring the data, followed by updating the coordinates and integrating them into a unique origin. Then, the neighborhood aggregation technique was conducted to enhance the model's understanding by adding new spatial features. Using the Leave-One-Out Cross-Validation (LOOCV) approach, each borehole was used as a test set once, and the ML model was trained using the remaining boreholes. Comparing the performance of ML techniques, the selected ML algorithm was used to develop a geological map, and by incorporating the hydrological information, a comprehensive hydrogeological map was obtained. Finally, the potential relationship between the hydrogeological profile and leakage data was investigated.

1.3 Challenges and limitations

The current study encountered several challenges and limitations in achieving the aforementioned goals. This section outlines the key issues that have been identified during the research process.

Although data analysis methods are powerful tools for understanding the relationships between geological and environmental variables, extracting this type of information is still a challenge among researchers due to limited accessibility, the diversity of geological data, the conversion over time, and the interaction between various factors. Furthermore, despite the abundance of historical data from borehole logs and inspection reports, most of this data remains underutilized due to its complexity and format.

1.3.1 Data limitations

1.3.1.1 Data availability

The most critical issue encountered was the insufficient data. The limited number of boreholes made developing an accurate geological map difficult. Subsequently, it affected the definition of the sequence of soil/rock layers, making it more difficult to investigate the geological impact on the tunnel infrastructure failures.

1.3.1.2 Data format

A considerable portion of the available borehole data originated from scanned archives. Some of these records were incomplete; for example, the precise coordinates of the boreholes were not specified. Additionally, inconsistencies in the reference systems complicated the integration process. Thus, cross-referencing maps and shared boreholes in different sources were used to realign the coordinates to a unified reference frame.

1.3.1.3 Data sources

Moreover, the multi-source data with diverse formats caused data structuring challenges. As a result, meaningful information was extracted from them manually. However, this process can increase the risk of errors due to illegible content, formatting inconsistencies, and misinterpretation.

1.3.2 Complexity of the geological profile

From a geological perspective, the study aimed to construct a detailed geological map along the tunnel and consider the type of soil/rock instead of stratigraphic chronology. This approach causes more complex layers, requiring closer boreholes to characterize transitions accurately. Consequently, the sparsity and spatial distribution of boreholes posed an obstacle for the classification task, contributing to the reduced model accuracy and rise in analysis errors. In previous studies, researchers have examined the stratum for the far-away boreholes rather than lithology and material types to achieve better performance.

Furthermore, Figure 1:2 highlighted the physical condition of the boreholes after extraction and months of storage. Over time, the core borehole may undergo erosion or develop fractures. Since the available data are from old reports, there is a risk that the recorded data no longer reflect the subsurface conditions.

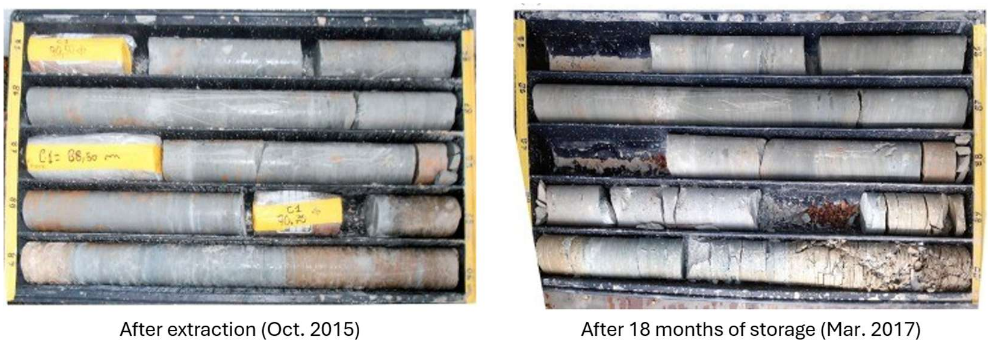


Figure 1:2 The core of borehole C1 at Point 1 at a depth of 85 m to 90 m changes over time(Fern et al., 2018)

The presence of overlapping composite rock types was addressed by selecting the most common groups and combinations. For instance, the three categories of marl and calcaire and the combination of these two(marl-calcaire) were applied as part of the final classification scheme.

1.3.3 Scope of study

The spatial limitation of leakage data made a comprehensive conclusion impossible. Most of the tunnel's segments are located in the molasse geological zone, whereas the leakage data are from a restricted different area with calcareous compositions. This difference in geological zones reduces the generalizability of the findings. As a result, more leakage data covering a wider range of tunnels is necessary for future studies.

1.4 Thesis layout

This thesis is structured into 6 chapters, and the organization is explained in the following paragraphs.

Chapter 2: Literature review

The research's two core aspects are geological mapping methods and the hydrogeological impact on the underground tunnels, which may lead to defects. The geological mapping section focuses on reviewing recent studies that apply ML techniques to the development of geological maps. The second part reviewed hydrogeological factors, including groundwater, rock characteristics, their reaction, and their long-term interactions with underground infrastructure. The chapter concludes by summarizing the main concepts, findings, and knowledge gaps.

Chapter 3: Methodology

In the methodology chapter, the study area, available data, and useful tools were described. Then, data collection, preprocessing steps, ML methods and implementation steps were outlined.

Chapter 4: Results and discussion

This chapter is divided into three main sections to discuss the outcome of the implemented ML algorithms: comparing the performance of the ML techniques across different datasets, generate a geological map in the selected area, and analysing the relationship between the hydrogeological profile and tunnel leakage.

Chapter 5: Conclusion

The conclusion chapter summarizes the key findings and highlights the main contributions, and it provides recommendations for future studies.

Chapter 2
Literature review

2 Literature review

The literature review section is a crucial component of initiating new research. It helps to identify knowledge gaps and explore new ways to address them. It also highlights the common methods and indicates which method is best suited for each. It's therefore easier to identify the research path and the gaps for further innovations. Given the purpose of this research, it is intended to answer the following questions:

- What are the state-of-the-art methods used for geological mapping?
- What are the common Machine Learning (ML) algorithms used to build the geological map?
- What input data was used to train these ML models?
- How was the model evaluation conducted?
- What are the knowledge gaps in geological mapping and the impact of hydrogeology on tunnel structure?
- How important are the effects of hydrogeological profiles on tunnel defects?

2.1 Search strategy

The filtration process of the papers is explained in this part. The keywords to find the related papers and previous works were "subsurface geological mapping", "machine learning", "deep learning", "neural networks", "borehole", "geotechnical engineering", "data-driven", "site characterization", "tunnel hazard", "tunnel deterioration", "lithological mapping", "hydrogeology", "hydrology", and "rock characteristics". Different combinations of these keywords were used to filter sources in various databases, including Google Scholar, ScienceDirect, and Scopus and no filters have been applied in terms of year and journal. In each series of results, the papers were filtered by reading the title first to identify the most relevant research objectives. The next step was to review the abstract and the shortlisted papers were compiled into an Excel file, and by reviewing the full text, useful information was extracted and analyzed.

2.2 Geological mapping

The geology of a region has a great impact on the construction and maintenance of underground infrastructure, including tunnels, subways, and pipelines. Most infrastructure was constructed decades ago and is now facing problems such as degradation, deformation, cracks, and shifts due to the surrounding geology. These failures can lead to serious safety hazards.

The geological profile affects the tunnel's stability. For instance, different combinations of soil and rocks (such as marl, sandstone, sand, and clay) react to changes over time differently. Geological mapping is a crucial step in maintaining underground infrastructure.

One of the most popular methods of geological mapping in recent years is ML-based methods. They are powerful tools for finding hidden, sophisticated, linear, or nonlinear relationships between parameters. Using the provided information, the model can be trained for further analysis. Before the widespread use of ML techniques, researchers used traditional methods, including spatial interpolation, kriging, and Inverse Distance Weighting (IDW). However, these traditional approaches have notable limitations: they struggle to manage sparsely distributed data, scale poorly with large datasets, and have limited ability to capture complex and nonlinear relationships. On the other hand, advanced monitoring solutions can meet these challenges. These state-of-the-art techniques are faster, more efficient, and more reliable in terms of their ability to monitor in real time.

Building a geological map can be considered a type of classification, as the goal is to comprehend and predict the subsurface layers. Each ML model consists of three essential components: input data, a modelling method, and an output.

In the first step, input data and gathering methods should be considered. In most cases, the input data needs to be pre-processed and cleaned before being fed to the model. There are various methods for gathering data and processing it to enhance knowledge of subsurface geology. In the present section, categorizing papers is based on the input data, which can be either well/borehole logs or spatial characteristics. In the first group, researchers used well or borehole logging data like coordinate, gamma ray test, calliper test, and density as the inputs to their model while in the second group, they used only spatial characteristics (coordinates) of boreholes or besides other geometric features which were extracted from coordinates using mathematical methods. According to the available data, the papers whose objective was to build a geological map only using spatial characteristics were mostly considered.

The next step is to find the optimized model or method. In this literature review, the main focus was on ML algorithms, and only a limited number of other methods, such as interpolation, were studied.

In terms of classification methods, papers used a variety of ML techniques regarding their approach. For instance, Radom Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Neural Networks are commonly used in previous studies. Neural Networks are more common when the size of the dataset is large enough and the relations between features are complicated. On the other hand, methods like SVM are appropriate for cases where there are decision boundaries between classes, and they aim to maximize the margin of the closest point. Although SVM is suitable for multiclass classification and high-dimensional spaces, its hyperplane complexity is less interpretable in higher dimensions. XGBoost is also a popular algorithm for complex classification. In this algorithm, each new tree is trained to correct the errors of the previous tree. Moreover, built-in parallel processing makes it a suitable choice for training large datasets quickly. RF is an ensemble method that builds multiple decision trees and improves accuracy by introducing randomness through bagging and feature selection at each split. This combination of diverse, decorrelated trees leads to a more robust and generalized model. RF, in most studies, had the best performance among all other algorithms. In RF, the risk of overfitting is reduced due to the combination of decision trees. In the section 3.6.1 the further details of the ML models are explained.

There are different approaches to the final goals of different papers. In some cases, researchers aim to classify the types of rocks in a specific area of their study, while others follow up with the next steps of the rock classification to build a 3D geological map. The model's output involves predicting the values for its voxels, and eventually, the results of predictions and categorized data should be evaluated and reported.

2.2.1 Methods and inputs

Existing methods for 3D modelling of subsurface stratigraphy may be divided into two categories: geological-interface-based modelling methods (L. Zhu et al., 2012)and point-cloud-based modelling methods (Shi et al., 2020), (Guo et al., 2019), (Lyu & Wang, 2024), (Zhang et al., 2020). In the first group, generated interfaces are followed by a surface-to-body approach for developing 3D models (Figure 2:2). For cases without extensive measurements, such as boreholes, interpolation methods like kriging can predict the interfaces. For sophisticated boundaries or interfaces, these methods may encounter difficulties, even with large datasets. On the other hand, point cloud-based models represent a 3D domain with distributed points, and each point stores the corresponding rock type (Figure 2:1). A variety of data-driven methods, including Multiple Point Statistics (MPS), Markov Random Field model (MRF), Coupled Markov Chain simulation (CMC), and Iterative Convolution eXtreme Gradient Boosting (IC-XGBoost), were compared for point-cloud-based modelling methods (Lyu et al., 2024).

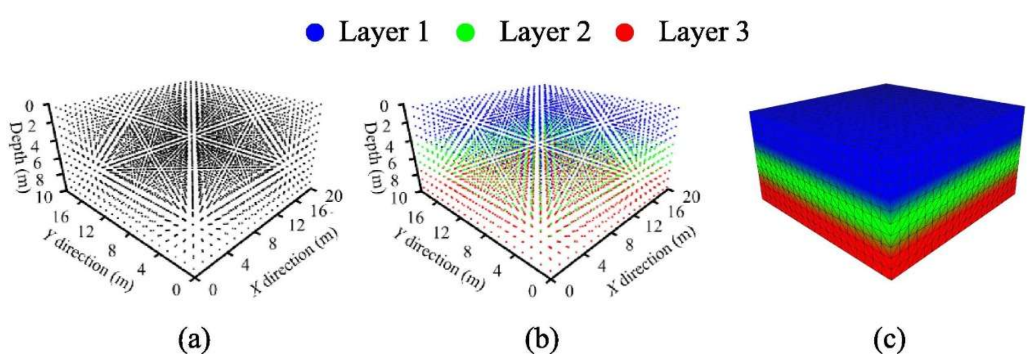


Figure 2:1 Illustrative example of a 3D model. (a) A point cloud without geological information, (b) a point cloud color-coded by geo-material types, and (c) voxelization of the point cloud.(Lyu & Wang, 2024)

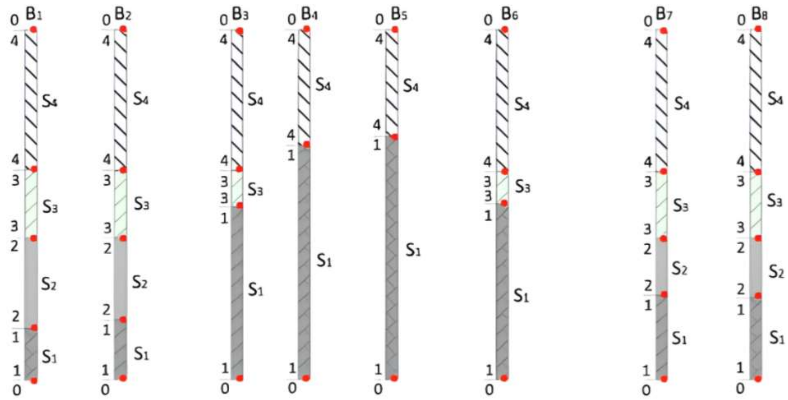


Figure 2:2 Discretization of borehole data and resulting scatter points for geological-interface-based modelling (L. Zhu et al., 2012)

The objective of this section is to investigate the methods that previous researchers have implemented to build a subsurface geological map using their available borehole data.

Table 2:1 provides an overview of previous works that utilized ML techniques to develop geological maps using various input data. This table compares the papers regarding their geological mapping methods, ML algorithms, input and output data, approaches, evaluation methods, and resolution of the Final maps. An attempt has been made to emphasize articles that used only borehole coordinates to train the ML model. There will be more papers to investigate where borehole log reports contain other test results, such as gamma tests and Ground-Penetrating Radar (GPR), which can be used as input for Artificial Neural Networks (ANN) and other methods.

In most of the studied cases, the resolution of the final map is less than 50 meters in each direction, while regarding the available borehole log data for the present study, which are farther away, achieving this resolution will be challenging. Comparing the evaluation methods and the number of boreholes, cross-validation can be an appropriate choice when the number of boreholes is limited.

Table 2:1 Summary of papers for geological mapping

Reference	Geological mapping method	ML method	Input	Output	Approach	Evaluation	Resolution
(C. Shi & Wang, 2022)	3D point cloud	ICXGBoost	2D cross-session images	Six group	Physics-informed and data-driven	Leave-one-out cross-validation.	x, y, z: 0.75,1,1 meter
(Lyu et al., 2024)	3D point cloud	MS-GAN	Limited boreholes and a 3D training image representing prior geological knowledge (real data: 12 borehole Site B + the 3D training image Site A)	Drente Formation (DR), Drachten Formation (DN), Peelo Formation (PE), and Urk Formation (UR)	Data-driven	3 boreholes for test, 9 boreholes for training	x, y, depth: 5,5,0.5 meter
(Guo et al., 2019)	3D point cloud	SVM, Bp Neural Network	Borehole coordinates (x, y, z)	Seven group	Data-driven	6 boreholes (10% of all) for validation	-
(L. Liu et al., 2024)	3D point cloud	Deep-learning-based method	Spatial coordinates (x, y, and z) and stratigraphic attributes	Holocene: 4; Upper Pleistocene: 3; Middle Pleistocene: 2; Lower Pleistocene: 1	Implicit modelling	10 sets of original data for each of the following drill depths: 30 m, 50 m, 70 m, and 90 m	depth sampling accuracies of 5 m and 1 m
(Hu et al., 2024)	3D point cloud (discrete grid points)	ANN, SVM, XGBoost, RF	Easting in meters, northing in meters, and elevation in meters + weighted summation of the true soil classifications in the neighborhood	E (estuarine clay), F1 (fluvial sand), F2 (fluvial clay), and M (marine clay)	Data-driven	165 borehole cross-validation, site1: 1D and 2D modelling, site2: 3D modelling	3D array of size 73 × 64 × 45 (vertical and horizontal resolutions: 0.5 meter)
(Kim & Ji, 2022)	Surface mapping	DNN	Starting depth of each soil, x, y coordinates	Landfill, sediment, weathered soil, weathered rock, soft rock (or engineering bedrock), medium-hard rock, and hard rock	Data-driven	Train set: 564,343, test set: 62,705	3D lattice network :10 m × 10 m × 2 m
(Cracknell & Reading, 2014)	Surface mapping	Naive Bayes, k-Nearest Neighbors, RF, SVM, and ANN	17 inputs (Digital Elevation Model, Total Magnetic Intensity, four Gamma-Ray Spectrometry channels, and Total Count channels)	13 lithological classes	Data-driven	Cross-validation	-
(Harris & Grunsky, 2015)	Surface/ lithological mapping	RF	Lake sediment geochemical (9 major and 7 minor elements), airborne low-resolution magnetic, and airborne gamma-ray spectrometer data	8 types of rock	-	-	-
(Lyu & Wang, 2024)	3D point cloud	Bayesian compressive sensing/sampling interpolation	-	-	-	-	voxel with an area of 10 m × 10 m × 5 m

In this section, the reviewed papers are divided into three groups. The first group includes papers that utilized ML algorithms and trained them using borehole spatial coordinates. In contrast, in the second group, other parameters (e.g., (Kumar et al., 2022), (Jia et al., 2021)) in the borehole log data are also used. The last group consists of papers that used other methods, such as mathematical interpolation.

2.2.1.1 Machine Learning - Input: borehole spatial coordinates

ML algorithms offer powerful tools for interpreting geological data. In the following paragraphs, papers utilizing ML algorithms for building geological maps, based solely on the spatial coordinates of boreholes as input, will be described.

(C. Shi & Wang, 2022) adopted a stratigraphic modelling and uncertainty quantification method called 3D iterative convolution eXtreme Gradient Boosting (IC-XGBoost3D) for automatically developing 3D subsurface geological domains from limited measurements. The resolutions in the x, y, and z directions are set to 0.75 m, 1.0 m, and 1.0 m, respectively, resulting in a total of 0.5 million voxels. Based on the geology revealed by boreholes, the site primarily comprises six distinct soil types. Cross-section images are used as training and test sets, and the model can learn the features of a pair of perpendicular 2D training images. The algorithm is modified from the Convolutional Neural Network (CNN) and utilizes readily available tools in deep learning for feature extraction and classification. Moreover, they considered the impact of the number of measurements on the model's accuracy. There is a slight reduction when decreasing the number from 36 to 18, but after that, the decrement increases as the measurement data number is further reduced.

(Lyu et al., 2024) proposed a generative ML method called Multi-Scale Generative Adversarial Networks (MS-GAN) for developing 3D subsurface geological models from limited boreholes and a 3D training image representing prior geological knowledge. A 3D training image, represented by a 3D point cloud and stored with prior stratigraphic information, was used by the model as input for training. Next, at the first scale of the iterative process, the boreholes, regularly distributed at the specific site, are considered as input data, and GAN models are used to predict values at several prescribed locations only. This process was repeated, using the prediction result as input for the next scale. This iterative generation process continued until the 3D map was obtained entirely. Two real case studies illustrated the model.

(Guo et al., 2019) demonstrated an implicit three-dimensional geological modelling method based on ML that converts the three-dimensional modelling problem of strata into the attribute classification problem of underground space grid units. Borehole data was converted to voxel points with coordinates and labels, and, after normalization, fed to train two different models: an SVM and a Backpropagation neural network. The radial basis function had the best performance among the kernel functions of SVM in handling nonlinear problems. Finally, 6 random boreholes (10% of the data) were used as a test set, and the model's accuracy was calculated.

(L. Liu et al., 2024) introduced a Deep Learning (DL)-based method for 3D geological modelling. Raw drilling data comprises planar coordinates, elevation, and stratigraphic depth of exploration points. The data are categorized into depth ranges as follows: less than 30 m (25,269 drills), 30 to 50 m (8207 drills), 50 to 70 m (7260 drills), and greater than 90 m (4335 drills). The input samples were partitioned into subnetworks based on geological age, geological origin, and geotechnical type. The processed sample data was subsequently fed into a fully connected deep neural network with N hidden layers and a K-dimensional configuration for each layer. The three subnetworks operate independently and are later amalgamated to establish a geomathematical model. At last, a 3D geological model with depth sampling accuracies of 5 m and 1 m was developed using PyVista.

(Hu et al., 2024) proposed an improved data-driven ML framework boosted with the neighborhood aggregation technique for modelling three-dimensional (3D) subsurface soil stratigraphy. The classical inputs were easting, nothing, and elevation. In this study, the target soil categories included E (estuarine clay), F1 (fluvial sand), F2 (fluvial clay), and M (marine clay). Neighborhood aggregation added new features corresponding to the weighting parameter of the occurrence of different soil categories (FILL, E, F1, F2, M) for any borehole, from its neighbor borehole, to improve the prediction results of classical ML models. 165 boreholes were utilized for four selected ML models: XGBoost, RF, ANN, and SVM. The results suggested that additional new features developed through neighborhood aggregation are beneficial to the XGBoost and RF models, with an increase in accuracy of approximately 5% (Figure 2:3).

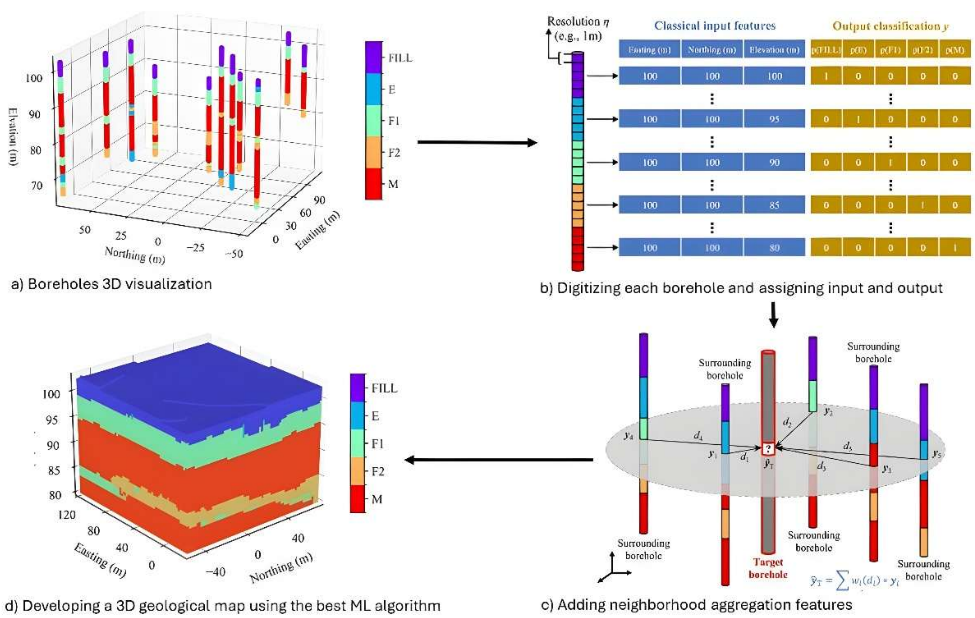


Figure 2:3 The implemented process for subsurface geological mapping by (Hu et al., 2024)

(Kim & Ji, 2022) presented the design of an optimum DNN-based learning model using a large-scale borehole database from geotechnical-layer depths and 3D spatial coordinates of boreholes. The geotechnical layers were divided into seven groups: landfill, sediment, weathered soil, weathered rock, soft rock (or engineering bedrock), medium-hard rock, and hard rock. They used k-means for outlier detection. In this study, researchers developed a DNN modelling architecture that incorporates data preprocessing, multi-layer perceptron modelling, 3D data processing, and display functionality using different tools.

2.2.1.2 Machine Learning - Input: borehole logs

The following section introduces papers that utilized a variety of parameters extracted from borehole logs, in addition to spatial characteristics, as input for ML algorithms.

(Cracknell & Reading, 2014) implemented a performance comparison between 5 ML algorithms: Naive Bayes, k-Nearest Neighbors (KNN), RF, SVM, and ANN for a supervised lithology classification. The airborne geophysical data used in this study included a Digital Elevation Model, Total Magnetic Intensity, and four Gamma-Ray Spectrometry channels, comprising Potassium, Thorium, Uranium, and Total Count channels. They expressed that the X coordinates were the most important variable in training. RF performed well across several evaluated aspects, such as stability, ease of use, processing time, and prediction accuracy. However, the computational cost of KNN was lower than that of the others. Overall, they concluded that RF is the best choice for remote sensing lithology classification.

(Harris & Grunsky, 2015) used RF for lithological mapping with two different training strategies, which were the extraction of rock type in each sample station from the legacy geology map and reconnaissance field mapping. Their ultimate goal was to predict eight types of rock in the study area. Five RF classification experiments were conducted using a training strategy based on legacy geology and geochemical sample points (both raw and corrected) and legacy field samples. Three data types were used to create the predictive lithology maps: lake sediment geochemical data (comprising 9 major and 7 minor elements), airborne low-resolution magnetic data, and airborne gamma-ray spectrometer data. They mentioned in their results that the correction of raw data had no significant effect on the accuracy of the model.

2.2.1.3 Other methods

(Lyu & Wang, 2024) proposed a framework for the visualization of 3D subsurface ground models in geo-applications using Virtual Reality (VR) technology (Figure 2:4). The subsurface map was represented as a 3D point cloud with spatial coordinates and an attribute as stratigraphic type. Therefore, each point had x, y, z, and C (corresponding to the class type). In this research, 65 sparse boreholes were used to extract ground surface elevation information, and 61 boreholes were used for rockhead elevation with the Bayesian compressive sensing/sampling interpolation method. Based on these two surfaces, a 3D point cloud was constructed with dimensions of 52 × 92 × 53. To develop the 3D subsurface ground model, each point of the 3D point cloud was converted to a voxel with an area of 10 m × 10 m × 5 m.

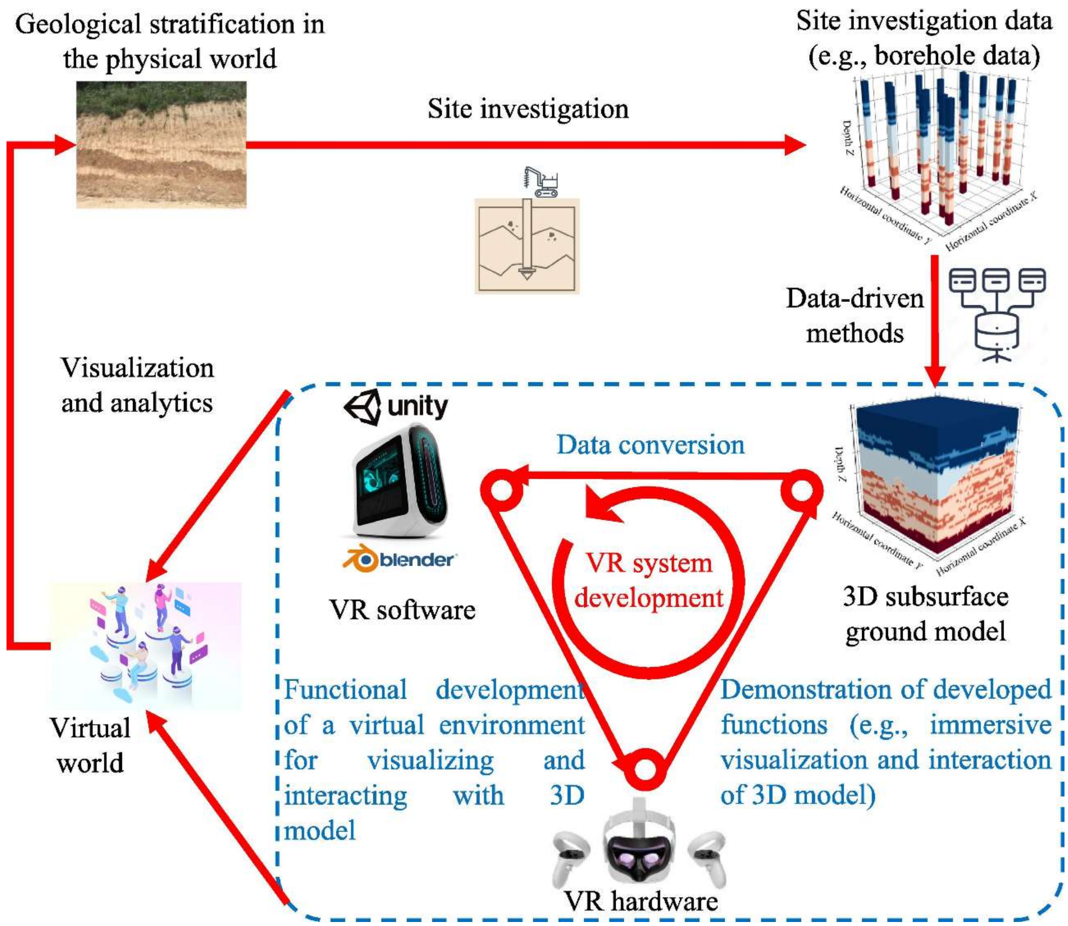


Figure 2:4 Proposed framework of developing a VR system for immersive visualization of 3D subsurface ground models in a virtual environment (Lyu & Wang, 2024)

(L. Zhu et al., 2012) described a method to 3D solid modelling of sedimentary system from borehole data, called the Borehole–Surface–Solid method. The most prominent feature of this method is its comprehensive treatment of three types of missing strata: non-deposition, erosion, and the combination of non-deposition and erosion. The process consists of 11 steps, like defining the boundary of models according to boreholes, determining the integrity of strata, interpolating the elevation of surfaces, understanding the type of missing strata and sorting them, and building a 3D model for visualization. A voxel-based solid model, filled with triangular prism meshes, can also be established by applying 3D discretization methods. Finally, two case studies were investigated to evaluate the model.

(J. Guo et al., 2021) presented an automatic implicit 3D geological modeling and visualization method that can be applied to urban geotechnical drilling data. First, by analyzing the general characteristics of these types of data, the implicit Hermite radial basis function surface was optimized, followed by the marching tetrahedra visualization method, for use in the field of 3D urban drilling modeling.

(Yang et al., 2023) Markov Chain Monte Carlo-based Bayesian method for the simulation of geological formations utilizing available boreholes from a specific site. They just used 7 boreholes. The researchers suggested a log-normal distribution when the borehole data is limited.

(Yan et al., 2023) proposed a rigorous random field-based framework to model the 3D stratigraphic uncertainty distribution with improved physical interpretation. They suggested that ML and statistical methods could be employed in future studies.

(Zhang et al., 2020) designed a method to extract paper-based borehole log information automatically with ML algorithms. The structured borehole data and geological knowledge were integrated into the process of building a 3D geological model. After the borehole data was loaded, point cloud data at different locations in the same stratum could be obtained. Furthermore, they focused on the processing of pinch-out, lens, and faults to improve their map.

2.2.2 Knowledge gap and future studies

Different methods of geological modelling have their advantages and disadvantages. In traditional methods like MRF and CMC, a transition probability matrix is needed, but they are widely used in 2D geological modeling, and building a 3D geological map using them is rare (Lyu et al., 2024). On the other hand, ML performed well in geological mapping with accurate results. It has gained popularity in recent years, and many studies have benefited from AI methods.

AI can process a wide range of data sources, handle misclassification by learning from large datasets, and also consider external influences. Despite the potential of AI and ML in geological mapping and rock classification, some challenges should be addressed. ML algorithms can provide accurate results with large datasets, although obtaining high-quality and complex geological data is a challenging task. With inadequate or biased data, the AI model may fail to perform properly. Therefore, validating the ML models with real-world observations is crucial. A combination of ML approaches and empirical methods during geotechnical studies is suggested (Niu et al., 2024).

From another perspective, lithological observations, such as boreholes, are limited and sparse, and indirect observations can be used to supplement the information about the subsurface structure. Therefore, DL-based methods are suitable for handling and analysing a huge amount of data and extracting geological information. Additionally, DL methods can accommodate various inputs, including images and multimodal data. Meanwhile, there are some limitations, including the limited availability of training samples, the lack of public databases for data integration, and the model's dependency on the quantity and quality of the data for accuracy and efficiency. Furthermore, there are factors like network architecture and loss function that can impact the accuracy of the DL model (Zhan et al., 2023).

As a result, the following can be considered for future studies:

- Expanding datasets with more diverse and high-quality geological conditions
- Development of benchmark databases
- Real-time monitoring and using ML models to update data
- Hybrid models, data-driven and empirical methods, to benefit from both methodologies
- Coupling the DL-based and traditional methods for subsurface identification
- Collaborating with other fields to find out the latest techniques
- ML applications to focus on anomalies and rare geological events detection
- Novelty in DL-based architecture and loss function
- Transfer learning application for DL-based models

2.3 Tunnel defects and Geological hazards

Tunneling in complex geological settings may face difficulties related to geotechnical and hydrogeological factors. Ground deformation, rock bursts, water inrushes, and structural failures are some common challenges that can occur during the construction or operation of a tunnel. For instance, Figure 2:5 shows some recorded defects in CERN tunnels. There can be a variety of reasons for these challenges, depending on factors such as groundwater interaction, rock type, and chemical reactions. These factors are interdependent. In this section, previous studies will provide a clearer understanding of why it is essential to consider the hydrogeological profile for underground infrastructures and the hazards that can lead to tunnel defects or even collapses.

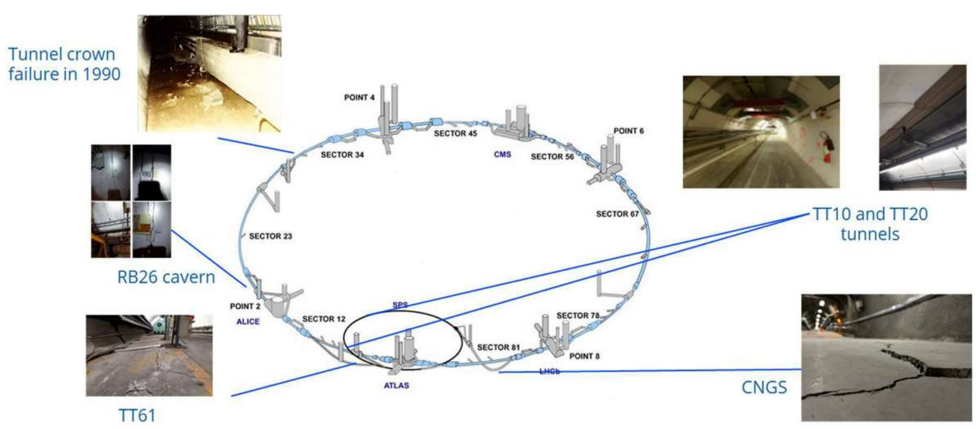


Figure 2:5 Recorded defects in CERN tunnels (Ouyang et al., 2023)

Recently, (L. Li et al., 2024) reviewed 155 cases of typical tunnel collapse during their operation and categorized the causes of collapse into three major groups: geological, lining, and accidental. The collapse was a serious accident resulting in significant damage to tunnels or parts of them, such as walls or ceilings, during the operating period. They also considered the hidden dangers in design, construction, and the operational period that led to collapse cases. From the statistical analysis, the geological factors had the most significant share, as Figure 2:6 indicates.

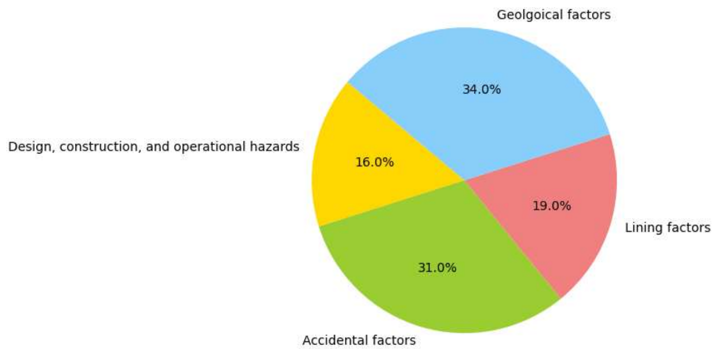


Figure 2:6 Pie chart of causes of tunnel collapse cases during operation (L. Li et al., 2024)

The geological causes of tunnel collapse include cavities behind the lining, the stability of the surrounding rocks, poor geological formation, groundwater, and high geostress. In the aforementioned review paper, post-lining cavities were identified as the primary cause, followed by poor geological conditions and groundwater as the next factors. Incidental causes such as fire, combustion and explosion, earthquakes, heavy rains, and mudslides also played a significant role in tunnel collapses. Among the lining causes, lining crack expansion was identified as a common problem that may lead to tunnel serviceability non-compliance, with lining spalling and aging as secondary and tertiary contributions (Figure 2:7).

Some other studies reviewed different causes of tunnel failure or collapse, such as the impact of excavation methods (Frenelus et al., 2021), seismic behavior of tunnels (Roy & Sarkar, 2017), and design errors in support strength (Qin et al., 2023). (L. Li et al., 2024) indicated that tunnel collapses during operation can result from multiple factors rather than a single cause.

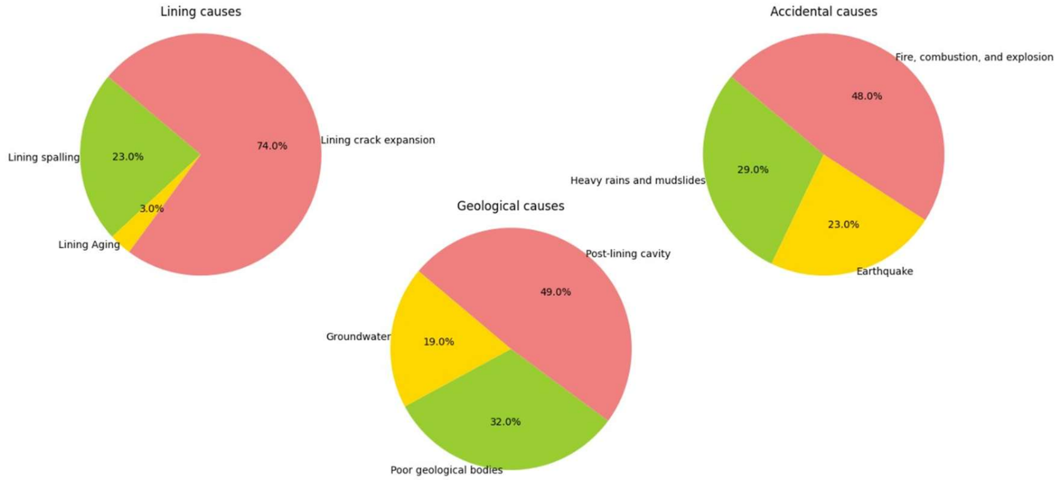


Figure 2:7 Pie charts for different subcategories of tunnel collapse causes during operation *(L. Li et al., 2024)*

In the following, the impact of the hydrogeological profile on the underground tunnel will be considered from three different aspects. First, the influence of hydrology factors, including groundwater profiles, hydraulic permeability, and hydraulic deterioration, was investigated. Next, the effects of rock characteristics and their interaction with groundwater were studied. Finally, papers with multiple geological impacts were reviewed.

2.3.1 Hydraulic influence

Several factors, including groundwater level, the hydraulic permeability of the ground, tunnel linings, water leakage, and tunnel deterioration, can significantly impact the long-term performance of the tunnel. (Wang, Friedman, et al., 2024) reviewed studies about hydraulic influences on the tunnel infrastructure in the long term, and a summary is presented in this literature review. In the mentioned review, the hydraulic factors are considered in three main categories: groundwater profile, hydraulic permeability, and hydraulic deterioration. Table 2:2 presents some examples of papers assessing the groundwater effect on tunnel infrastructure.

Table 2:2 Example papers for groundwater impact on tunnel infrastructure

Reference	Cause	Hazard	Goal	Method	Result
(Colombo et al., 2018)	Rise of groundwater	Flooding hazards	The hydrogeological hazard for the underground infrastructures caused by the rise of the groundwater level in Milan	A stochastic model to assess the hydrogeological hazard (in probabilistic terms) for the underground infrastructures for different scenarios of the aquifer system	A significant increase in water level by 5 m in the next 15 years, depending on the scenarios taken into account
(Gattinoni & Scesi, 2017)	Rise of groundwater	Changes in the flow path and an increase in the velocity of water	The hydrogeological hazard increases for the existing underground structures and infrastructure due to the rising trend of the water table	3D numerical model of the groundwater flow system and several scenarios of the aquifer system evolution	Local deformation and changes in flow path and danger of underground infrastructure
(Soga et al., 2017)	Ground permeability, compressibility, and anisotropy	Consolidation-induced settlement for tunnels located in low-permeability London clay	Focusing on ground surface movements and tunnel lining deformation in the interest of engineering concerns	A series of 3D finite element analyses and interpretation of limited field case studies to explain mechanisms of long-term ground movements and tunnel behaviour after tunnelling in clayey soil	Groundwater seepage evaporation in low-permeability clay, negligible effect of soil permeability anisotropy for impermeable tunnel
(Zhou et al., 2020)	Ground permeability	Tunnel settlement	Find the cause of the settlement and improve	Establishing a settlement prediction model to analyse the potential influencing factors based on the Deep Belief Network (DBN) and calculating the correlation degree between influencing factors and settlement through sensitivity analysis	The main factors affecting settlement: the permeability coefficients of the layer and the cutter head torque
(Yi et al., n.d.)	Leakage through concrete cracks	Crack growth and enabling interconnection of the water flow routes, and an increase in permeability	Water permeability and how it is affected by hydraulic pressure and crack widths in cracked concrete	Using the splitting and reuniting method to manufacture concrete specimens with controlled crack widths	An Increase in water transport as the crack width and hydraulic pressure increased

2.3.1.1 Effect of groundwater profiles

Groundwater is one of the most impactful factors contributing to tunnel failures. Water inrush and groundwater seepage can lead to leakages, ground weakening, and structural instability. The deterioration of tunnels caused by groundwater can be divided into three aspects: changes in water pressure of the tunnel lining due to the rise and fall of groundwater, water-rock interaction, and deterioration of the concrete material by groundwater (R. Li et al., 2020).

For example, groundwater may cause erosion of rocks since it contains dissolved substances. Then, in cold regions, groundwater seeps into cracks, and when the temperature drops, the freezing groundwater in the rock leads to volume expansion and rock crumbling (Niu et al., 2024).

In the aforementioned review paper, groundwater level (GL) is also a critical factor for tunnel maintenance, as it can cause engineering issues such as leakage, cracks, and lining deformation. The related papers were divided into the following categories:

1. *Maintained phreatic surface:*

   In many previous studies, a constant water table was assumed for the simplification of problems. The results showed that the justification for studies assuming a constant GL typically falls into four categories:

   a. Water table unaffected by tunnel leakage
   b. Sufficient recharge balancing, negligible tunnel leakage
   c. Ease of numerical and analytical solutions
   d. Complexities in representing cyclic and transient hydro-geological environments

   Although most researchers accept this assumption, it does not accurately represent the variability of realistic groundwater and should not be ignored. Additional information about accurate changes in the water table can help to understand further tunnel deformation and other defects in tunnels. Thus, investigating groundwater conditions and the tunnel's response to changes is crucial.

2. *Groundwater drawdown:*

   Several objectives, including dewatering, insufficient recharge, pumping, seasonal climate change, and construction activities, can lead to groundwater drawdown in different ways. This is more common when the construction ground is highly permeable. Similarly, using hydrostatic groundwater levels to represent the drawdown and failing to consider the long-term evolution of groundwater changes will result in an inaccurate estimate of tunnel functionality.

3. *Rising groundwater table:*

   Sea level rise, prolonged heavy precipitation, and human-induced underground construction can increase groundwater levels, potentially causing tunnel uplift and surface structure failures. The groundwater level rise is attributed to:

   a. Restoration of groundwater level due to deindustrialization/de-abstraction
   b. Climate and weather-induced GL rise
   c. Reduced water inflow due to drainage deterioration

   Linear cracks, radial deformation, pore water pressure on the tunnel lining, and tunnel uplift are risks associated with the rise of the groundwater level that can affect the long-term serviceability and integrity of the tunnel.

   For instance, (Colombo et al., 2018) suggested a stochastic model to assess the hydrogeological hazards in different aquifer system scenarios. In their research, the underground infrastructure caused changes in the water table. The flooding of underground structures and infrastructures, the corrosion of foundations, and increased groundwater pollution will be some of the consequences of the rising water table. Additionally, other studies have highlighted the impact of increased groundwater levels in various cities (Gattinoni & Scesi, 2017), (Vazquez-Sune et al., 1997), (Wilkinson, 1985) (Wilkinson, 1985).

2.3.1.2 Effect of hydraulic permeability

The next impactful factors in long-term tunnel performance are ground and liner permeability. Ground permeability correlates with the rate of groundwater flow. Lining permeability controls the water flow regime around the tunnel drainage boundary(Wang, Friedman, et al., 2024). From 3 aspects, these two factors were compared:

1. *Magnitude:*
   Ground permeability influences the tunnel's performance and stability by pore water pressure distribution on lining, ground deformation, and tunnel deformation and loading. Generally, high permeability causes the rapid dissipation of pore water pressure, reduces the hydrostatic pressure on the tunnel lining, facilitates quick drainage, and promotes fast consolidation. Therefore, ground permeability can lead to some changes in lining internal forces and deformation. In previous studies, the magnitude of the effect was determined by assuming homogeneous ground layers, constant ground permeability, and subjective permeability variations; however, these simplifications cannot accurately represent the realistic ground permeability effect during tunnel operation.
2. *Anisotropy:*
   Permeability anisotropy is the ratio of horizontal permeability to vertical permeability, and different ground types have different values. It contributes to the groundwater transmission ability in each direction, subsequently affecting the tunnel's performance throughout its lifetime. The results confirmed that ground permeability anisotropy is a key component that can cause variations in water inflow patterns and rates, pore water pressure distributions behind the lining, ground consolidation, and ground settlement characteristics. The soil permeability anisotropy ratio and water inflow are inversely related, meaning that as one increases, the other decreases.
3. *Uniformity:*
   Lining hydraulic permeability is linked to water infiltration into tunnels and varies over time in response to changes in the hydrogeological profile and the tunnel. Studies with the hypothesis of uniform permeability concluded that a reduction in permeability results in a rise in pore water pressure within the tunnel. This assumption works properly in specific situations, for example, when there is no defect in the tunnel's hydraulic system. On the other hand, many studies prefer non-uniform lining permeability for tunnel joints. Understanding tunnel functionality will be more realistic when the effect of lining permeability non-uniformity is considered, but accurately quantifying non-uniform circumferential or longitudinal lining permeability is challenging in practice.

In a study to investigate the soil-tunnel consolidation interaction, (Soga et al., 2017) discussed the ground surface movements and tunnel lining deformation. A series of 3D finite element analyses was conducted to describe the various mechanisms of long-term ground movement and tunnel response. The researchers reviewed the mechanism of tunneling-induced soil consolidation for single tunnels, twin tunnels, and cross-passage tunnels.

In another study, (Zhou et al., 2020) found the cause of the settlement and a novel Deep Belief Network to improve it. They established a settlement prediction model to analyze the potential influencing factors and calculated the correlation degree between them. The researchers concluded that the permeability coefficients of the layer and cutter head torque are the main factors affecting settlement.

2.3.1.3 Effect of hydraulic deterioration

Hydraulic deterioration typically occurs in the form of water leakage, tunnel drainage system blockage, and tunnel lining hydraulic degradation.

1. *Water leakage:*

   Water leakage is a common defect along the tunnel and can be divided into: Uniform tunnel leakage and Localized tunnel leakage.

   Some studies focused on estimating water inflow and preferred a uniform tunnel leakage analysis over an examination of uniform leakage on a tunnel's performance. However, uniform water leakage is rare in tunnel infrastructure. The hydrogeological profile, including rock permeability, fault zones, and heterogeneous groundwater flow patterns, impacts the water inflow. Consequently, the leakage locations can vary regarding the location's characteristics. Three main results were achieved from previous studies. First, the goal of most researchers is to concentrate on localized leakage through circumferential joints. Second, a hypothetical leakage distribution was assumed through segmental joints in tunnels. Lastly, most previous works have ignored the realistic permeability of partial leakage areas.

2. *Drainage system blockage:*

   The unwanted groundwater is evacuated from the tunnel by the drainage system, which consists of two main components: drainage layers and drainage channels. The faults in the drainage system can lead to pore water pressure acting on the tunnel lining and additional loads. A lack of quantitative assessment of the tunnel's drainage system was reported.

3. *Lining hydraulic deterioration:*
   Hydraulic deterioration of lining concrete, lining joint, and lining corrosion are different types of tunnel hydraulic deterioration. There is an interaction between lining cracking and water leakage in joints. Additionally, the corrosion of the lining increases the pressure and load, leading to cracking.

Crack development reduces the lining's bearing capacity, and as a result, the tunnel's strength becomes unreliable.

To wrap up, this review paper concluded that the main challenges are the simplifications and a lack of enough data to analyze. It suggested a realistic groundwater level, detailed permeability features, and a quantitative assessment of deterioration for future studies. Therefore, gathering more data and finding the relationships between them will help us find more precise and accurate models to evaluate the tunnel's long-term performance.

2.3.2 Rock mass characteristics

Another factor that affects tunnel stability significantly is the rock characteristics. Tunnel excavation can lead to degradation of the surrounding rocks' characteristics. The unloading process affects the primary features of the rock and can cause instability. In the long-term assessment of tunnel performance, considering the deterioration of rock mechanical properties is essential because rock strength may be reduced (Frenelus et al., 2021).

Weak rocks in tunnels are a threat to the stability of the underground excavations. Various factors, such as humidity, temperature, and changes in water content throughout each season and time, lead to degradation of the rock mass properties. Consequently, this impacts the tunnel, potentially threatening the stability of the structure. (Narimani Dehnavi & Sadeghi, 2019) studied the failure of the specific section of the Golab tunnel, and it was deduced that the temporary support system that was applied was not appropriate. This research highlighted the importance of the long-term behavior of rocks for the design and application of structures.

Time-dependent deformations, such as creeping, squeezing, and swelling, are remarkably attributable to rock rheology. Additionally, the weathering of rocks surrounding the infrastructure can impact the stability of the tunnel. The physical and mechanical characteristics of rocks have the most significant impact on the weathering process (Showkati et al., 2021).

(Maleki et al., 2020) identified the probable hazards affecting the alignment of the Gelas water conveyance tunnel and the magnitude of the risk in some sections. Several potential hazards, including squeezing and groundwater inflow, were assessed, and geotechnical characteristics of rock masses were evaluated. In another study, (Khanlari et al., 2012a) assessed the rock mass of part of the Karaj-Tehran tunnel and they concluded that in the crushed zones and fractured zones, the squeezing potential is higher than in other areas.

The safety of the tunnel during the operational period is primarily influenced by the deterioration of the rock mass, which depends on its characteristics. It plays a crucial role in tunnel maintenance. The deterioration of the rock mass can lead to a reduction in the mechanical parameters due to weathering, aging, and other factors (Kong et al., 2022).

From the long-term aspect, chemical reactions between groundwater and rock lead to erosion and problems in the tunnel structure. The stream of groundwater, which contains mineral material of rock, may be dissolved with a concrete lining or other materials. Subsequently, the deterioration of support measures impacts the functionality of the tunnel. Over time, the failure of support and collapse of the excavation will result.

(Fathi Salmi et al., 2019) investigated the hydrogeological factors that can be important in the deterioration of the segmental concrete linings at the Zagros water transfer tunnel. Eventually, they proposed some techniques to prevent damage to the segmental linings due to future chemical and physical deteriorations.

The same study by (J. X. Shi et al., 2020) was done. The researchers developed a mathematical model based on the physical and chemical reactions of surrounding rocks to analyze the relationship between leakage and the interaction between groundwater, rock, and concrete lining. Ultimately, the results showed that the erosion of rocks by groundwater, the channels created by groundwater in the rock, and the probability of leakage are positively correlated.

The future problems of the corrosiveness of concrete due to groundwater chemistry along the Tabriz metro tunnel, investigated by (Ghobadi et al., 2016). The formations had the greatest impact on the deterioration of groundwater quality, and they predicted problems with the lining for the near future.

Another group (Su et al., 2025) classified the water-rock interaction into physical, chemical, and mechanical interactions. The dynamic change in groundwater can lead to the deterioration of stress conditions and rock degradation. This reflects the interaction between water and rock in tunnel engineering.

All the reviewed papers in this section are summarized in Table 2:3.

Table 2:3 Summary of papers considered rock characteristics as a cause of hazard in a tunnel

Reference	Cause	Hazard	Goal	Method	Result
(Khanlari et al., 2012b)	Rock characteristics	Squeezing	The assessment of squeezing potential along the second part of the water supply Karaj to Tehran tunnel	Lab and field assessment of geotechnical properties of rock, and using empirical and semi-empirical equations to evaluate the tunnel	Fair squeezing potential in crushed and fractured zones along the tunnel route
(Ghobadi et al., 2016)	Rock characteristics and groundwater	Concrete lining deterioration	Investigation of the future problems of the corrosiveness of concrete due to groundwater chemistry	Using geological maps to understand the impact of geological formations on groundwater quality and chemical parameters' relation with concrete lining	The groundwater in the east part is highly aggressive, and an expectation of deterioration of the concrete lining
(Narimani Dehnavi & Sadeghi, 2019)	Rock characteristics	Failure in a specific section	Describe the case of the Golab access tunnel, where it is subject to failure in a specific section containing meta-shales	Point load test to find fault zones in rocks, estimating the effect of the deterioration on strength parameters and finding the correlation between point load and SDI	Another support system suggested to prevent hazards
(Fathi Salmi et al., 2019)	Rock characteristics and groundwater	Deterioration of supports and collapse	Investigation of the hydrogeological factors resulting in the deterioration of the segmental concrete linings	Field inspections at different tunnel locations to assess the deterioration of the concrete linings, Photographic analysis based on Microscopy Analysis, MA, and XRD technique to investigate the mechanisms of deterioration in micro-scale level.	Deterioration of the concrete lining because of the sulphate attack
(Maleki et al., 2020)	Rock characteristics	Groundwater in-rush and squeezing	Identifying the probable hazards affecting the alignment of the Gelas water conveyance tunnel in Iran and the magnitude of risk in various sections along the tunnel route	Evaluating the geotechnical characteristics of the rock masses based on lab investigation, selecting sections of the tunnel, and applying analytical methods to classify the route from groundwater hazards	High susceptibility to groundwater in-rush and squeezing for areas with faults and crush zones
(J. X. Shi et al., 2020)	Water-rock interaction	Leakage	Analysing the mechanism of the interaction between the groundwater and the surrounding rock of a tunnel	Analysing the interaction between the chemical components in the groundwater, rocks, and concrete lining by a mathematical model based on reactions	Erosion of rock by ground water leaded to leakage
(Showkati et al., 2021)	Weathering and degradation of rock, deterioration of primary support	Corrosion in rock bolts, degradation of lining	Predicting long-term stability of tunnels that accounts for weathering of rock and deterioration of primary support	Semi-empirical logarithmic equation for model semi-empirical logarithmic equations for modelling of the diminution of rock mass characteristics and deriving the equations for quantifying damages to their mechanical properties with time	Increase in the overall ground pressure and internal forces and recommendation of tunnel rehabilitation after 60 years
(Kong et al., 2022)	Deterioration of rock mass and shotcrete lining	Changes in support pressure of second lining	Long-term deterioration of rock mass numerical simulation	Simulating the long-term behaviour of rock mass and shotcrete in tunnels by using the three-dimensional finite difference method. Simulating the degradation of rock mass by weakening the mechanical parameters and the deterioration of shotcrete by decreasing the thickness	The deterioration of rock mass and shotcrete increased the rock pressure on the secondary lining
(W. Liu et al., 2022)	Rock characteristics	Deformation during construction, risk of crack	Long-term, real-time, and continuous monitoring of the structural stress of the Tunnel	Regression analysis of the structural stress by constructing four typical stress-time mathematical models and evaluating the long-term safety and in-service durability	Periodically fluctuation of the structural stress, risk of local cracking and damage to concrete, meet the safety factors for early stage
(Su et al., 2025)	Water-rock interaction	Leakage, freezing, waterproofing, and drainage system	Classifying the water-rock interaction into physical, chemical, and mechanical and its influence on tunnel stability	Introducing the classification and impact of temperature and humidity on them and analysing the impact on the tunnel	Different deterioration trends for different rocks and probable crack growth, reduction of rock strength, and pore water pressure lead to leakage and freezing

2.3.3 Multiple influences

In some studies that considered the long-term performance of a tunnel, multiple factors were found to have an impact on the tunnel's functionality. A combination of different parameters like rock characteristics, groundwater, in situ stress, fracture zones, lining material, etc. The interaction between these parameters can cause hazards to tunnel stability, deformation, corrosion, or structural support failure. However, finding the correlation between different parameters is a challenging task, and a lack of data may exacerbate the difficulties in identifying these relationships. Summary of the reviewed papers for this section is shown in Table 2:4.

To investigate the potential hazard sources,(Z. Chen et al., 2020) applied 5 sets of conditions- crossing a fault zone, shallow tunnels in soft rock, crossing a landslide deposit, and deep tunnels in hard and soft rock formations to the case study tunnel. The major potential hazard sources for the studied location were soft surrounding rock, high geo-stress, high water pressure, fault fracture zones, high seismic intensity, and geological disasters. Furthermore, the combination of the surrounding rock strength and the in-situ conditions can affect the squeezing of the surrounding rock and fractured zone, and water problems could reduce their integrity.

Table 2:4 Summary of papers considered multiple hydrogeological factors that caused hazards for the tunnel

Reference	Cause	Hazard	Goal	Method	Result
(Lipponen et al., 2005)	Weathering, swelling clays, erosion, and groundwater change	Block falls, fracture zones	Study the effect of water and geological factors on tunnel	Study possible causal relationships of block-falls and several geological parameters and analyse data for a specific part of the tunnel	Groundwater inflow and many factors caused block falls, need for planning regarding the fracture zones
(Cao et al., 2018)	In-situ stress, soft weak rock, groundwater seepage, weak support, delay closure of primary support	Squeezing and deformation	Lessons from runnel construction in weak rock mass under high in situ stress and controlling tunnel base deformation	Adjusting support type and excavation method dynamically and improvement of support stiffness, additional support, etc.	High horizontal in situ stress, soft-weak rock, groundwater seepage, weak support, and delay closure of primary support resulted in the occurrence of large deformation and failure of the tunnel during construction.
(Chen et al., 2020)	Soft rock, in-situ stress, fracture zone, water pressure, and seismic intensity	Structural safety hazard	Investigation of the potential hazard sources for tunnels constructed in complex and difficult mountainous areas	5 sets of conditions: crossing a fault zone, shallow tunnels in soft rock, crossing a landslide deposit, and deep tunnels in hard and soft rock formations. A program of field tests was performed to study the changes in rock pressure and the performance of support structures in different geological conditions	Squeezing of surrounding rock determined by rock strength and in-situ stress, high impact of time-dependent deformation, and geological hazards on structural safety

For instance, the damages observed in a tunnel in Finland were attributed to weathering, swelling clays, erosion, and changes in groundwater or pressure conditions. The main factor that triggered these changes was the water that flowed both along the tunnel and into the fractures of the surrounding bedrock. The damage concentrated on zones of existing weakness and sections that have not been lined with shotcrete.(Lipponen et al., 2005)

The other case study was a railway tunnel in China. The large deformations and failures at the initial stages necessitated further analysis. There have been many factors that result in the occurrence of squeezing failure of the tunnel during construction, such as high in situ stress, soft-weak rock, groundwater seepage, weak support, delayed closure of primary support, etc. (Cao et al., 2018)

To summarize, several potential challenges, including soft rock, high in-situ stress, high water pressure, fault fracture zones, and geological hazards, can significantly contribute to tunnel defects. However, the research confirms that tunnel problems are the result of multiple factors, and there is not only one exact reason for tunnel failures and collapses. (Z. Chen et al., 2020), (Cao et al., 2018)

Overall, the hydrogeological factors play a crucial role in the tunnel maintenance and in predicting the faults in long-term operation time. The novelty of this work lies in the development of a geological map around the tunnel using ML techniques, considering the long-term impact of detailed geology and groundwater levels on a large underground tunnel.

2.3.4 Knowledge gaps and future studies

Despite advancements in underground infrastructure maintenance, many aspects of the impact of the hydrogeology profile on tunnels remain unknown. Most studies assumed a static condition of the groundwater and rock situation, whereas a dynamic behavior is more realistic.

The previous researcher primarily focused on numerical studies, employing mathematical simplifications that overlooked fluctuations and changes in the hydrogeological profile. These assumptions were about GL, water-rock interactions, and rock characteristics. There is also a lack of discussion on structural performance, like deformation. Therefore, a more realistic assessment of the GL, permeability, and deterioration characteristics is suggested.(Wang, Friedman, et al., 2024)

In addition, for geological problems and disasters such as high geostress, geothermal stress, and hydraulic pressure, there is a need for the automatic extraction of vast amounts of geological information using digital methods. Hence, DL and ML algorithms can analyse big data of surrounding rocks and hydrogeological profiles for safe construction and maintenance.(H. Zhu et al., 2019)

2.4 Conclusion

Understanding the geological conditions near the tunnel is crucial for maintaining its safety. There are different ways to gather information about the surrounding rocks and groundwater around the tunnel, and this information can be analysed from different aspects. In this literature review, two main subjects were considered. Primarily, the geological mapping methods with a focus on ML methods were studied. Next, the impact of the hydrogeological profile on the underground tunnel is investigated.

In the first section, different input data and algorithms were considered. Some research emphasized the use of borehole logs, while others relied solely on the coordinates of the boreholes to construct a geological map. Different algorithms, including RF, SVM, and XGBoost, were implemented to train the model. In cases involving big data and complex correlations between features, ANN and DNN can perform efficiently.

The other section focused on explaining the impact of the geological profile on the tunnel infrastructure's performance. Hydraulic influences, including groundwater, permeability, and deterioration factors, were also studied in addition to the rock characteristics. These factors can lead to irreparable events, such as crack expansion, uncommon leaks, and even tunnel collapse.

To better understand and anticipate subsurface conditions, future research should focus on developing data integration and utilizing ML models. Model generalizability and reliability will be improved by creating standardized benchmark databases and expanding datasets to incorporate more varied and high-quality geological situations. Continuous updates and adaptive learning can be made possible by integrating ML models with real-time monitoring systems. It may be possible to better represent subsurface behaviors by combining data-driven and empirical methods through hybrid modeling and by combining DL techniques with conventional methods. To integrate the most recent developments in sensing, computation, and modelling, interdisciplinary cooperation is crucial.

Chapter 3
Methodology

3 Methodology

This chapter describes the systematic methodology employed, from data collection and preprocessing to the construction of the geological map. The methodology begins with an overview of the study area, followed by gathering boreholes from multiple sources. Next, the available data is stored in a structured data frame, ensuring compatibility with Machine Learning (ML) algorithms.

3.1 Study area description

CERN, the European Organization for Nuclear Research, was founded in 1952 on the Franco-Swiss border in Geneva, Switzerland. Its focus is on exploring the fundamental structure of the universe. The first particle accelerator was built at this particle physics laboratory.

CERN hosts an extensive underground infrastructure, including particle accelerators, detectors, and experimental facilities. Two key components of underground infrastructure are the Large Hadron Collider (LHC) and the Super Proton Synchrotron (SPS).

The LHC is the most powerful particle accelerator ever built. It pushes protons or ions close to the speed of light and runs beneath Switzerland and France. Due to geological considerations, the accelerator is situated at an average depth of 100 meters below the surface and a slight gradient of 1.4% depth, varying between 175 m beneath the Jura mountains and 50 m towards Lake Geneva (Figure 3:1).[1] It is embedded within a 27-kilometer (26659-meter) circular tunnel of superconducting magnets with several accelerating structures that boost the energy of the particles along the way. Along its path, the tunnel traverses a range of geological formations.[2]

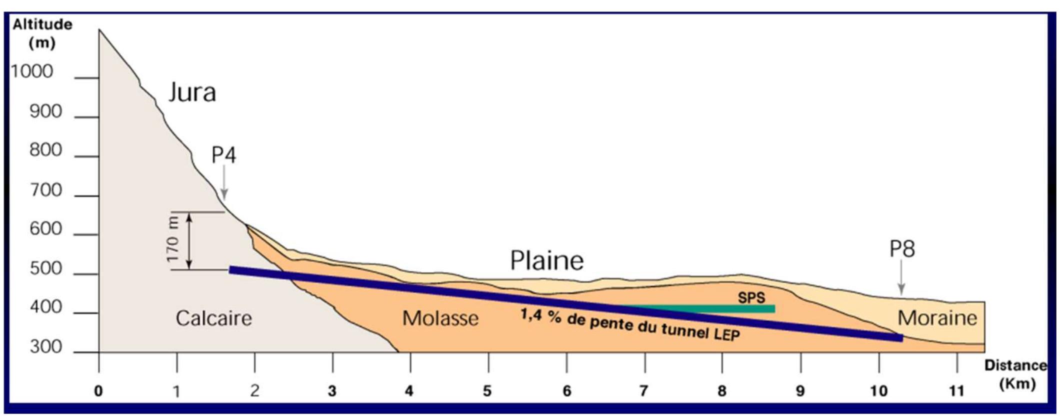


*Figure 3:1 2D visualization of the study area and the tunnel positioning in geological layers (*Large Colliders:Civil Engineering and Siting*, n.d.)*

[1] https://cds.cern.ch/record/2255762/files/CERN-Brochure-2017-002-Eng.pdf
[2] https://home.cern/resources/faqs/facts-and-figures-about-lhc

Seven experiments are installed at the LHC as follows [1]:

- A Large Ion Collider Experiment (ALICE)
    - Location: Segny (access from St. Genis Pouilly), France
    - Point2
- A Toroidal LHC ApparatuS (ATLAS)
    - Location: Meyrin, Switzerland
    - Point1
- The Compact Muon Solenoid (CMS)
    - Location: Cessy, France
    - Point5
- The Large Hadron Collider beauty (LHCb) experiment
    - Location: Ferney-Voltaire, France
    - Point8
- The Large Hadron Collider forward (LHCf) experiment
    - Location: Meyrin, Switzerland (near ATLAS).
    - Point1
- The TOTal Elastic and diffractive cross section Measurement (TOTEM) experiment
    - Location: Cessy, France (near CMS)
    - Point 5
- Monopole and Exotics Detector at the LHC(MoEDAL)
    - Location: Ferney-Voltaire, France (near LHCb)
    - Point 8

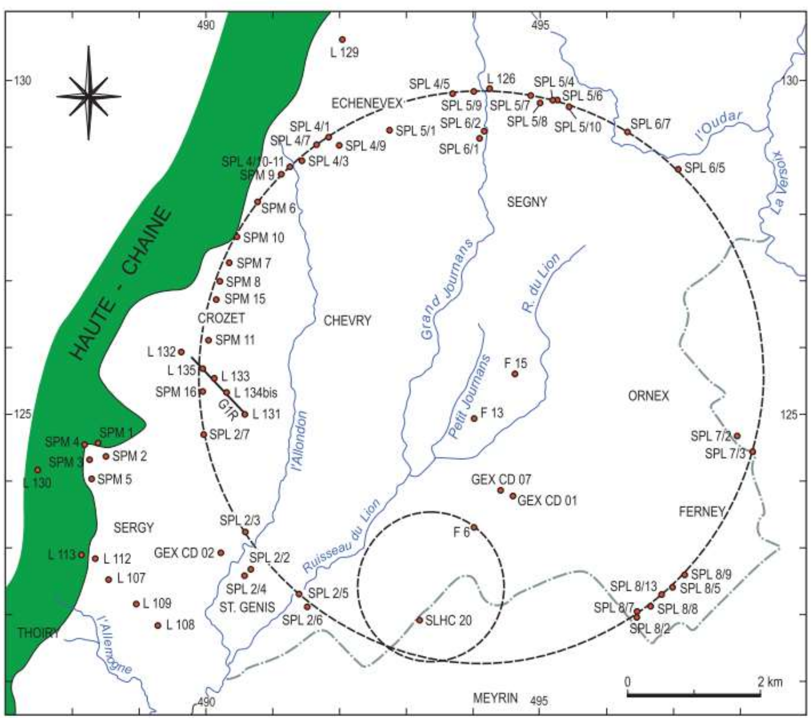


Figure 3:2 Location of the boreholes in the French report (Charollais et al., 2007)

Figure 3:2 shows the location of boreholes and cities near them and Figure 3:4 presents the underground map of the tunnels and the locations of the explained points.

A variety of excavation methods were used in different sections of the LHC tunnel. For instance, the TT10 tunnel was constructed using an alpine road-header machine with an excavation span of around 1.2 m length (Murro & College, 2019). In another section, to construct the TI8 tunnel, a road-header was used towards SPS; however, along the LHC, a Tunnel Boring Machine (TBM) was used (Fielder, 2000).

As shown in Figure 3:3, for the tunnel sections were constructed through molasse, the typical cross-section consists of a 100 mm primary segmented concrete lining, followed by a 150 mm secondary cast-in-situ lining, separated by an intermediate waterproofing layer (Abada et al., 2019).

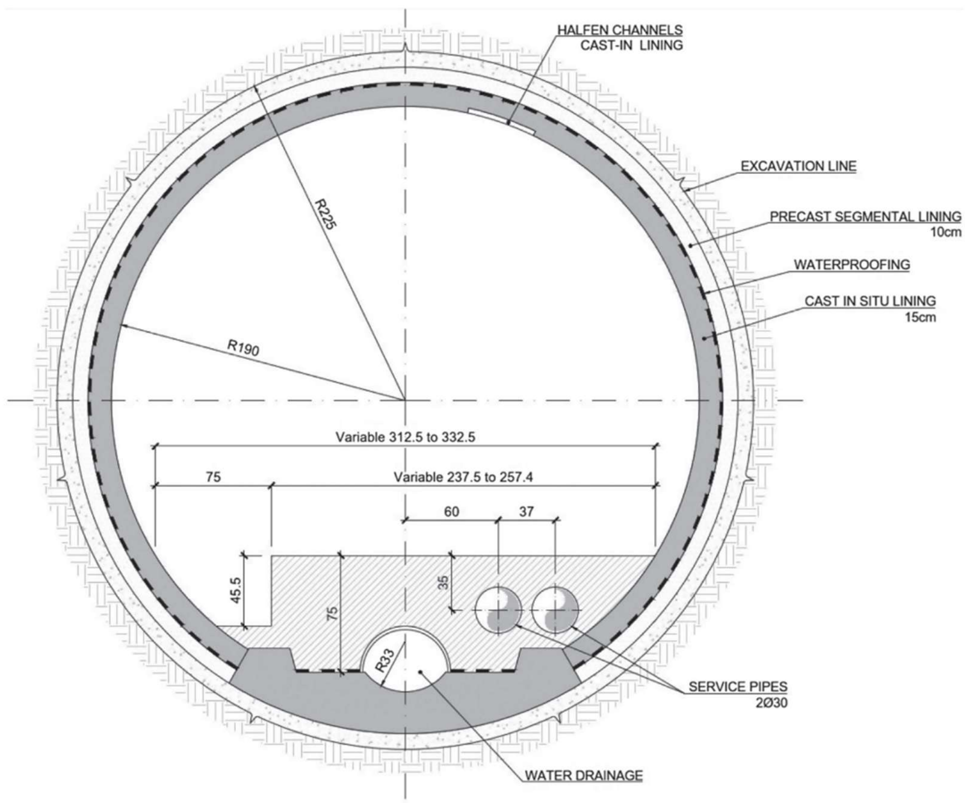


Figure 3:3 Typical cross-section of LHC tunnel (Abada et al., 2019)

The SPS is the second-largest accelerator in CERN's complex. The SPS tunnel has a circumference of 6910 meters and a cross section of 4 meters in inner diameter. It is situated at an elevation of 400 meters above sea level, with a depth below the surface ranging from 23 to 65 meters. The tunnel walls are reinforced with a concrete lining approximately 30 centimetres thick.

The summary of the tunnel specification is represented in Table 3:1.

Table 3:1 LHC and SPS tunnels specifications

Tunnel	Circumference	Altitude	Depth
SPS	7km	400 m	23-65 m
LHC	27km	450 m	50-175 m

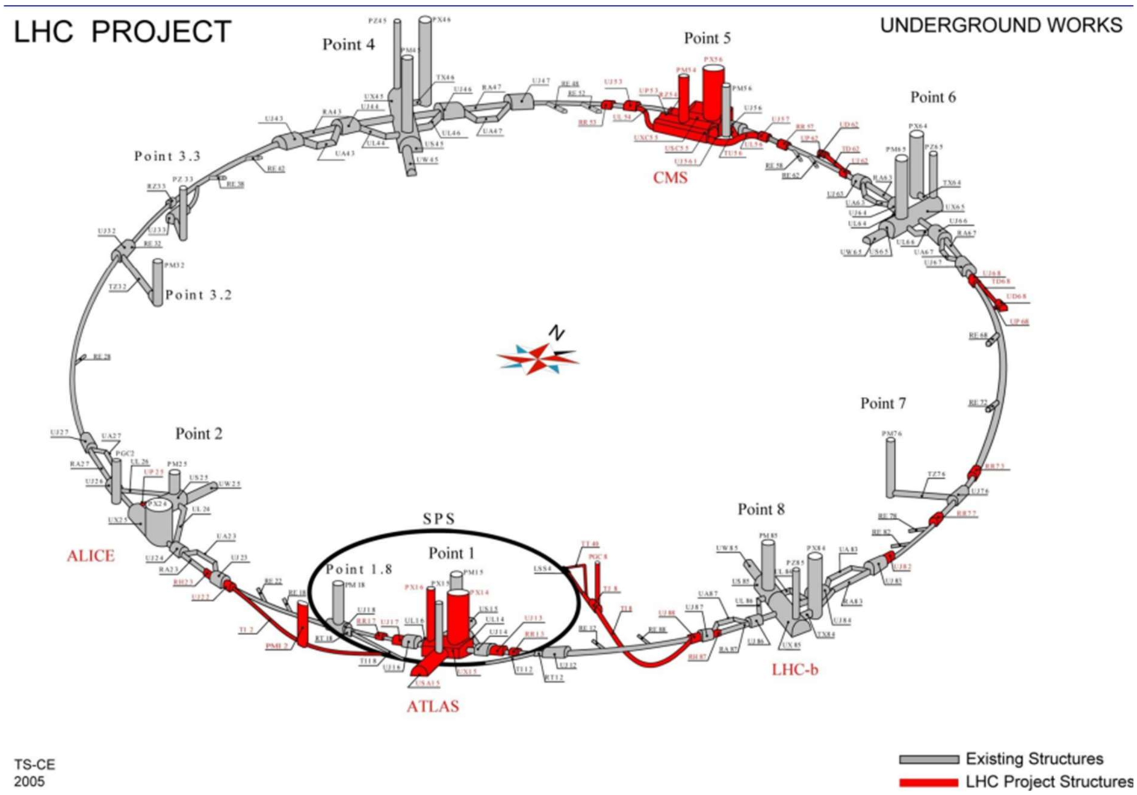


*Figure 3:4 Underground map of the LHC and SPS tunnel and location of tunnel points (*Large Colliders:Civil Engineering and Siting*, n.d.)*

Geologically, Switzerland consists of three distinct regions: the Alps, the Jura, and the molassic plateau. The Alps are predominantly composed of strong sedimentary and crystalline rocks; the Jura is composed of medium-strong limestones; the Swiss plateau is comprised of molasse, a weak to medium-strong sedimentary rock.

(Fern et al., 2018) focuses on a specific type of molasse (the red molasse), which is a weak sedimentary formation. This rock mass comprises two rock units – marl and sandstone – and three sub-units (very weak, weak, and medium-weak for marl and weak, medium-strong, and strong for sandstone) exist for each rock unit. Marl, a calcium carbonate-rich mudstone, and sandstone, composed of compacted sand-sized particles, are particularly significant for this study due to their varying permeability and mechanical properties. In the bedrock, these units occur as lenses that grade smoothly into one another, which makes them hard to predict. The rock mass is generally massive and can be considered a continuous solid. However, local fractured zones are observed, mostly in very weak layers of marl, and can be treated as a soil-like rock unit.

3.2 Software and tools

This research utilizes various software applications and computational tools. A brief introduction and application are provided below.

Python:

Python is one of the most popular programming languages, widely used in data science and ML. In this study, Python 3.11.11 was selected. The following libraries were used:

- Pandas: a data manipulation library for handling, reading, and cleaning structured data
- NumPy: numerical computing library, providing arrays, matrices, and scientific computing
- Matplotlib: a visualization library
- Scikit-learn: ML library, tools for preprocessing, classification, regression, evaluation metrics, etc
- XGBoost: an optimized distributed gradient boosting library
- CatBoost: a gradient boosting library that can handle categorical features

Google Colab:

Google Colab is a cloud-based Jupyter notebook environment. Users can run Python codes with free GPU support, and it can handle large-scale computations for ML models.

Excel:

Excel is a Microsoft software that can help organize and visualize tabular data. Microsoft® Excel® for Microsoft 365 MSO (Version 250) 64-bit was used in this study.

GetData Graph Digitizer:

GetData is a data extraction tool that can retrieve numerical values from graphical images like plots or scanned images. This study employed GetData (version 2.24) to estimate unknown borehole locations based on known reference points. In this software, the unknown values are retrievable after defining the axis. Therefore, the axis will be determined based on two known points, and then, by clicking anywhere on the screen, the new coordinates will be stored in a new table.

3.3 Data collection and sources

This borehole data was collected from various sources. Some borehole reports were obtained through collaboration with CERN, such as documents of SPS boreholes, SHLC boreholes, and some LHC boreholes for points 3 to 4 (L135, SPM 11, SPM10, SPM11, SPM7, SPL4.10, SPL4.1, SPM9, SPM6, SPM8, SPL4.3, R43.2, R43). Each report contains detailed information such as coordinates, altitude, depth, and a description of the stratigraphic layers. The results section of the report describes the lithological compositions in detail, specifying the actual type of rock/soil materials—marl, sandstone, calcaire, or others. Moreover, the reports contained maps identifying the location of each borehole. These maps were subsequently used to reference ambiguous boreholes that were not clearly identified.

Figure 3:5, Figure 3:6, and Figure 3:7 are some examples of the stratigraphy obtained from borehole reports.

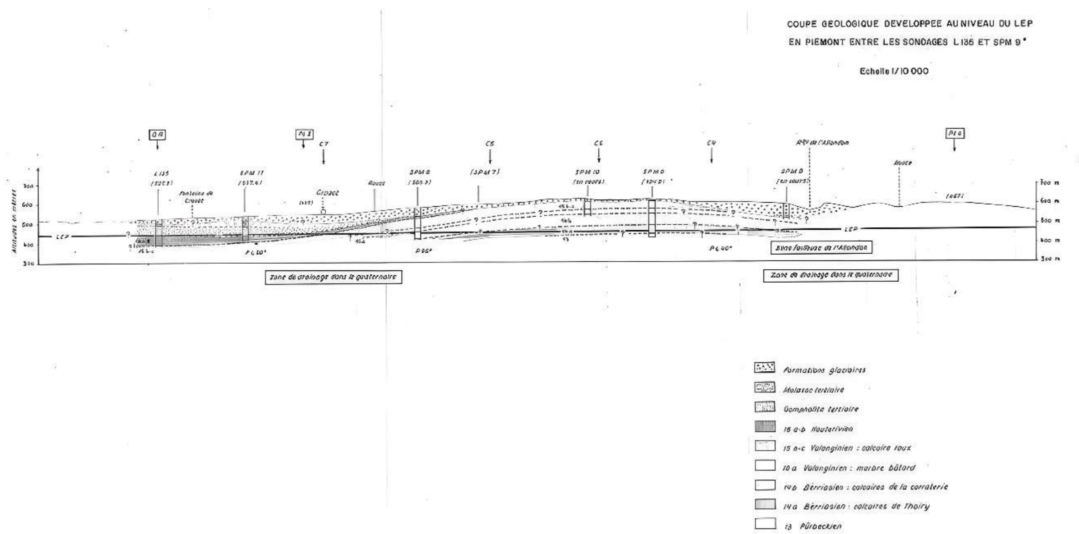


Figure 3:5 LHC point 3 to point 4, 2D borehole visualization along the tunnel from the CERN reports

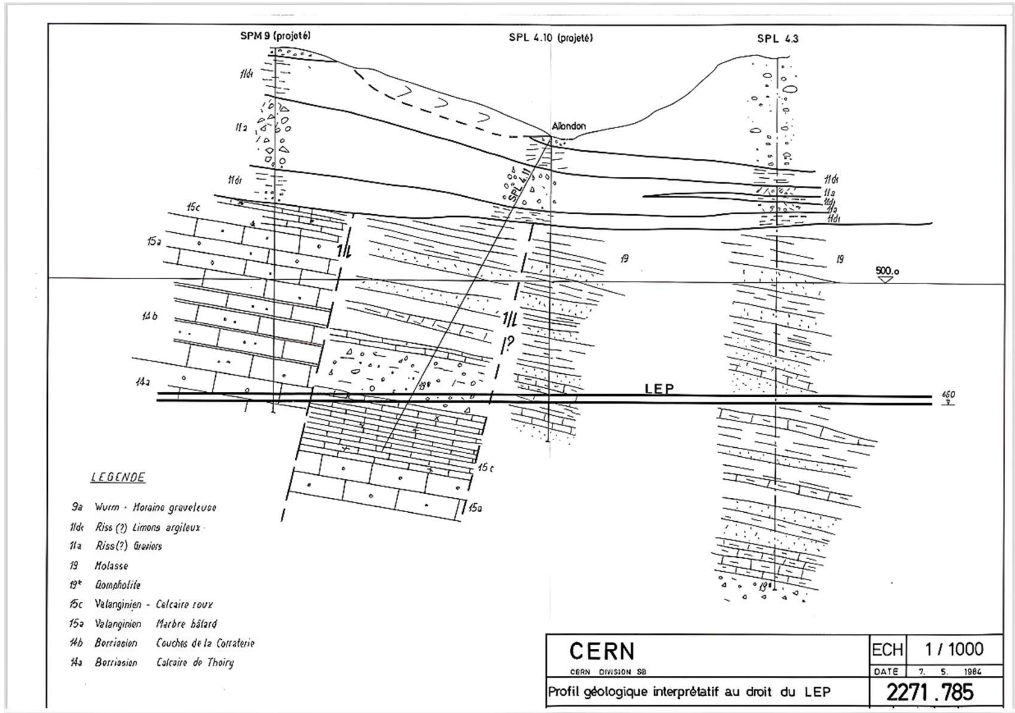


Figure 3:6 Details of SPM9, SPL4.10, SPL4.11, SPL4.3 boreholes from CERN reports

Additional borehole data were extracted from a French study on the area's geology titled "Archives des Sciences, volume 60, fascicules 2-3, November 2007".[3] The publication discusses the Molasse formation in the Greater Geneva area and its underlying Mesozoic substrate, based on the boreholes investigations and engineering surveys conducted across the region. The information from this source provided valuable insights into the area's lithology and complemented the data collected directly from CERN (Charollais et al., 2007).

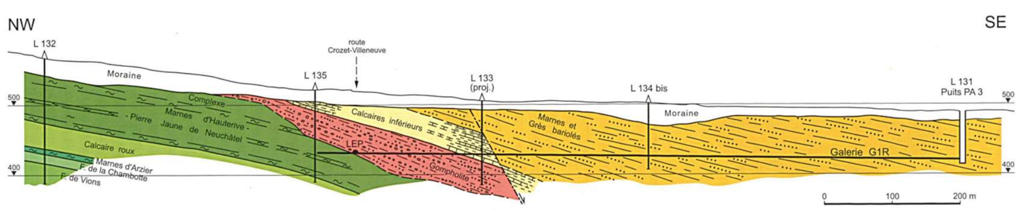


Figure 3:7 The geology and location of L132, L132, L133, L134, L131 boreholes (Charollais et al., 2007)

Each data set contained information such as coordinates (X, Y, Z), the type of rock or soil at the desired altitude, and stratum. The Excel files were then converted to data frames using the Pandas library for further analyses.

To structure the consecutive lithological layers, the altitude of each data point is represented as the lower boundary of that layer. Also, it corresponds to the upper boundary of the layer below. For the first layer of each borehole, two data rows

[3] https://www.unige.ch/sphn/Publications/ArchivesSciences/Volumes.php

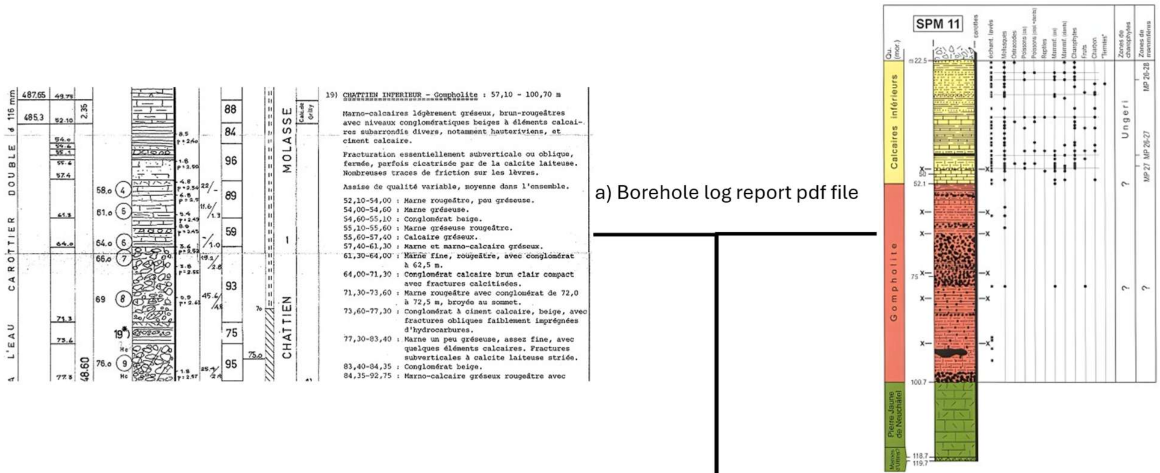
with similar characteristics are added, and only the altitude differs between these rows. Two data entries with identical characteristics were created for the first layer in each borehole, differing only in altitude to represent the layer's start and end points (Figure 3:13). Also, a new column was added to the structured data indicating the sequential layer number within each specific borehole. This can preserve the order of layers and ensure their maintenance throughout the preprocessing steps.

3.4 Data preprocessing

3.4.1 Structuring unstructured data

Unstructured data lacks a predefined schema or tabular arrangement and typically consists of text, images, and other non-uniform formats. Due to its heterogeneous nature, the analysis and extraction of meaningful information is complex. Examples of this ambiguous data include PDF or Word files, images, videos, or graphs. In contrast, structured data is usually found in databases that can efficiently categorize information and perform the necessary analysis. Researchers can selectively retrieve columns and rows, allowing for efficient data manipulation. (Verma et al., 2020)

One of the methods of converting unstructured data into structured data is to use image processing techniques. For example, (Zhang et al., 2020) accessed several paper borehole log data sets and applied image processing, table line recognition, cell positioning, text recognition, and text correction, enabling the automatic conversion of existing paper data into structured data in the form of a file. Since the current study has a limited amount of borehole data, converting data from unstructured to structured was performed manually. Figure 3:8 shows the process of extracting data from PDF files, storing them in Excel files, and then importing them as data frames using the Python library.



b) Structured Excel file

x	y	z	depth	altitude	survay	material	number
49004	12610	537.4	52.1	485.3	SPM11	calcaire	273
49004	12610	537.4	54.6	482.8	SPM11	marl	272
49004	12610	537.4	55.1	482.3	SPM11	conglomei	271
49004	12610	537.4	55.6	481.8	SPM11	marl	270
49004	12610	537.4	57.4	480	SPM11	calcaire	269
49004	12610	537.4	64	473.4	SPM11	marl	268
49004	12610	537.4	71.3	466.1	SPM11	conglomei	267
49004	12610	537.4	73.6	463.8	SPM11	marl	266
49004	12610	537.4	77.3	460.1	SPM11	conglomei	265
49004	12610	537.4	83.4	454	SPM11	marl	264
49004	12610	537.4	84.35	453.05	SPM11	conglomei	263
49004	12610	537.4	92.75	444.65	SPM11	marl/calca	262
49004	12610	537.4	94	443.4	SPM11	calcaire	261

c) Data frame

	x	y	altitude	survay	material	number
987	49004	12610	487.0	SPM11	calcaire	273
988	49004	12610	486.0	SPM11	calcaire	273
989	49004	12610	485.0	SPM11	calcaire	273
990	49004	12610	484.0	SPM11	marl	272
991	49004	12610	483.0	SPM11	marl	272
992	49004	12610	482.0	SPM11	conglomerat	271
993	49004	12610	481.0	SPM11	calcaire	269
994	49004	12610	480.0	SPM11	calcaire	269
995	49004	12610	479.0	SPM11	marl	268
996	49004	12610	478.0	SPM11	marl	268

Figure 3:8 Structuring data procedure

In the current research, initial data were obtained through PDF files, which included the reports for boreholes drilled during the tunnel construction in the 1980s. This information was for the SPS tunnel and parts of the LHC tunnel. By importing the relevant data from these borehole logs into Excel spreadsheets in a categorized format, the converted structured data was suitable for data analysis and ML model training.

Four different series of datasets were used in this study: SPS tunnel data, SLHC data near point 5 in the LHC tunnel, data from points 3 to 4 in the LHC tunnel, and distributed data along the LHC tunnel. The first three series were obtained from CERN archive files, while the fourth was obtained from the French report (Charollais et al., 2007). Comparing these two data sources, six boreholes were common. From each document, the following parameters were extracted and categorized in a table:

- Borehole coordinates (X, Y)
- Borehole name
- Altitude
- Stratigraphy
- Lithology or type of soil/rock

Table 3:2 Borehole distribution in datasets

	SPS	SLHC	LHC 3_4	LHC French	all
Number of boreholes	36	7	12	24	73
Source	CERN	CERN	CERN	(Charollais et al., 2007)	-
Document year	1970-1971	1996	1982-1986	2007	-
Type of soil/rock	M/S	M/S	M/S/C/M_C /CO	M/S/C/M_C /CO	M/S/C/M_C /CO

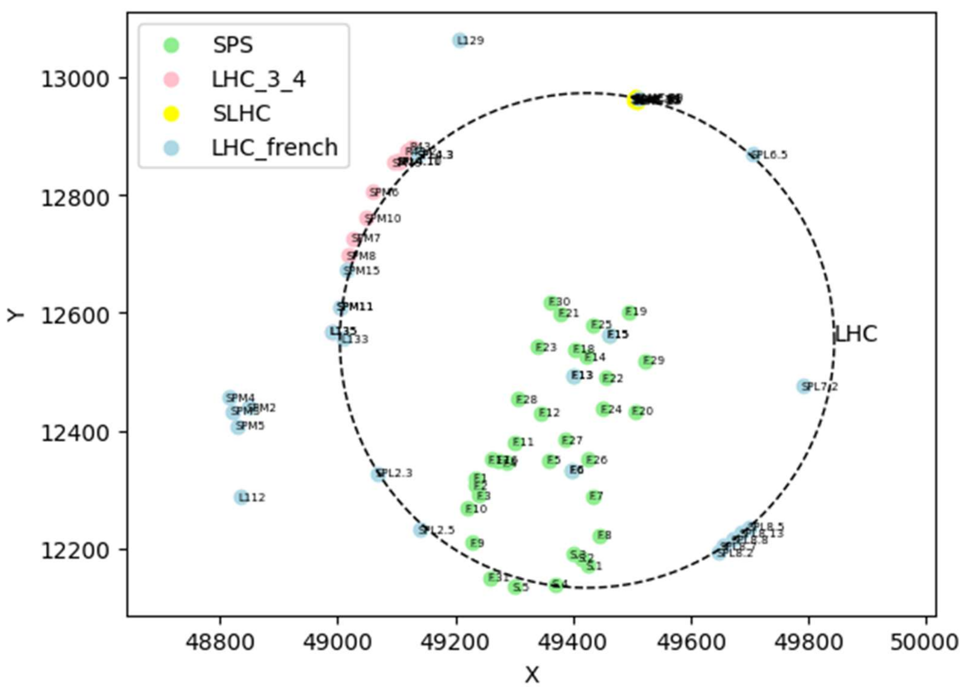


Figure 3:9 The 2D visualization of the borehole location according to the X and Y coordinates

The data frames from each Excel file were named as "df_sps" for the SPS dataset, "df_slhc" for the SLHC section of the LHC, "df_lhc_3_4" for the LHC point3 to point4, and "df_lhc" for the LHC data from the French source. Table 3:2 indicates the final number of boreholes in each dataset. Figure 3:9 shows the 2D spatial distribution of the boreholes along the tunnels, with different colours representing their respective dataset assignments, while Figure 3:11 provides the 3D visualization of the boreholes and the corresponding colours to describe the lithology. In Figure 3:10 the overview of all proposed steps from data collection to model implementation is shown.

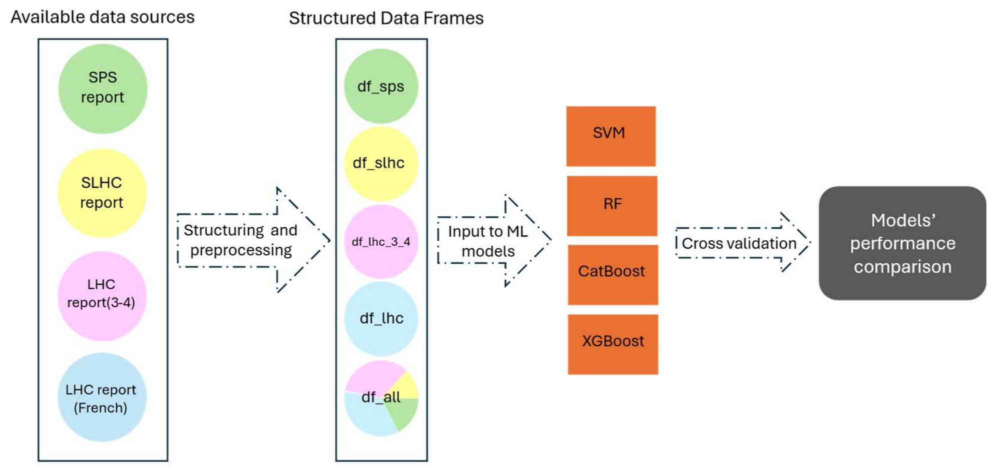


Figure 3:10 Flow chart of the implemented steps

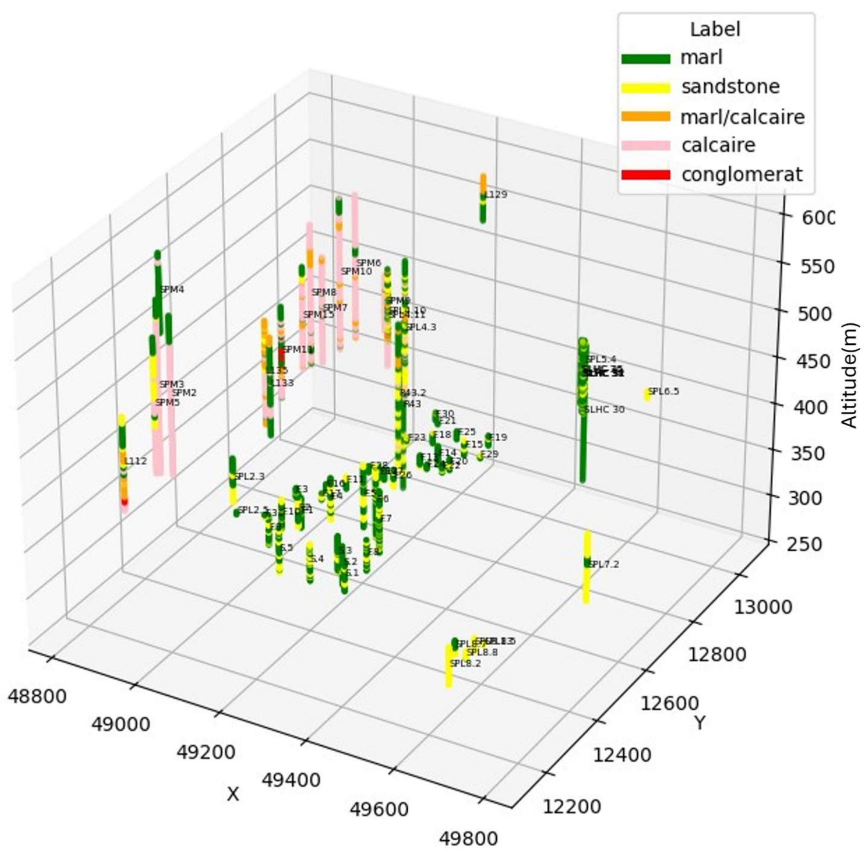


Figure 3:11 The detailed 3D visualization of boreholes according to their coordinates

3.4.2 Data integration

Due to the difference in origin and coordinate systems across the various reports, it was necessary to standardize the spatial references of all the boreholes. By cross-referencing common borehole (specifically F.6, F.13, F.15 for the SPS tunnel, and SPL4.3, L135, SPM11 for the LHC tunnel) found in both CERN and French documents, and further utilizing borehole location maps provided within the CERN documentation, the unified spatial coordinates of all boreholes associated with the SPS and LHC tunnel systems were retrieved through the application of Get Data software.

For the SPS tunnel, matched boreholes - F6 and F15 - were used to establish a reference coordinate system to recalculate the location of other boreholes. Similarly, SPM15 and SPM11 defined new axes for the LHC tunnel, and the locations of the remaining boreholes were updated according to the obtained coordinate system. Then, by combining existing maps and selecting an updated origin for boreholes from different documents, the coordinates of all boreholes were unified using the Get Data software, ensuring compatibility for use in the subsequent ML analysis.

Figure 3:12 presents a screenshot from the Get Data software for updating the boreholes' locations in the LHC tunnel between points 3 and 4. The SPM 11 and SPM 15 were used to calibrate the origin and axis, and the other borehole locations were retrieved. The new coordinates are available on the table (the right side of the screen).

The standardized and combined dataset from all sources was consolidated into a single integrated data frame called "df_all".

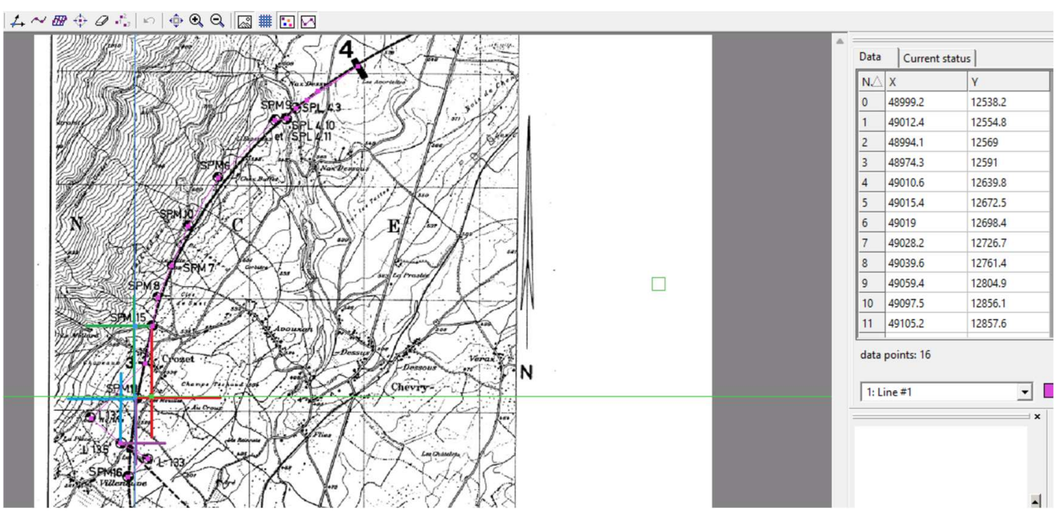


Figure 3:12 The screenshot from the Get Data software for updating the location of boreholes

3.4.3 Preprocessing

Data preprocessing is a crucial step and depends on the type of data and the ML algorithm. Previous research indicates that applying valid preprocessing can improve the model's accuracy significantly (Alam & Yao, 2019).

Image resizing, cropping, normalization, image augmentation, noise reduction, and color conversion are common preprocessing steps for image classification tasks. However, text data needs tokenization, lowercasing, removing stop words, lemmatization, stemming, and encoding to be cleaned and organized as input for an ML model. In this case study, the data format is tabular. Preprocessing approaches encompass minimizing the noise by distinguishing the optimized number of features, handling missing values and outlier data, text-to-number conversion, and data normalization (Albahra et al., 2023).

3.4.3.1 Adjusting altitude

To facilitate data analysis, the information stored in the Excel files was converted to data frames using the Pandas library in Python. Regarding the prior sections, the altitude column in the data frame represents the current geological layer's endpoint and the next layer's starting point. The raw altitudes for boreholes from each report had different degrees of measurement accuracy, often exceeding the required resolution. Accordingly, the altitude values were rounded up if their decimal number was greater than 0.5 meters and otherwise were rounded down.

3.4.3.2 Filtering type of soil/rock

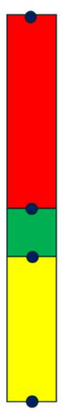
In each borehole, there were specific kinds of soil/rocks concerning the lithology of the region. For example, within the molassic region—home to the SPS tunnel and most of the LHC tunnel—marl and sandstone are the predominant soil types. However, a considerable part of the LHC tunnel is in the calcaire area (point 3 to point 4), and a combination of marl, sandstone, calcaire, and conglomerate layers coexist. Regarding the borehole log reports, some layers were distinguished as a mixture of marl and calcaire, indicating that these two were not separable in those intervals. Therefore, the five types of soil/rocks were selected —marl, sandstone, marl-calcaire, calcaire, and conglomerate— because they were the layers most frequently encountered in the examined boreholes.

To prevent data loss and preserve data integrity, the "strip" function removed whitespaces for material (type of soil/rock) and survey ("name of borehole") columns.

3.4.3.3 Adding new points

In the available data, only the top and bottom boundaries of each layer within a borehole are known. However, the model cannot recognize the internal points as part of that layer and requires the characteristics throughout the entire thickness of the layer. To address this, it is necessary to know the rock type at each depth within a layer. Considering the resolution in the Z-direction, which is assumed to be one meter, the layers are converted to 1-meter sections. If the distance between the top and bottom of each layer is one meter, there is no need to have a new point between them, although if the thickness of the layer is more than one meter, it has to be divided into one-meter segments, and new points will be added to the data. Subsequently, the top of the first 1-meter segment corresponds to the bottom of the overlying layer, while the bottom of the last segment matches the original bottom altitude of the layer. For all these blocks, the layer "number" column in that borehole is the same to prevent missing the order of layers.

Figure 3:13 illustrates a sample borehole to clarify this approach. The dark blue points are known from the sources, while the light blue points are added to the data set for the layers with a more than one meter thickness. (On the right-side borehole, each block is equal to 1 meter; for example, for the yellow layer, which is 3 meters, two new points should be added to the data set.)



X	Y	Altitude	Survey	Material
49398	12333	464	F.6	Marl
49398	12333	460	F.6	Marl
49398	12333	459	F.6	Sandstone
49398	12333	456	F.6	Calcaire

X	Y	Altitude	Survey	Material
49398	12333	464	F.6	Marl
49398	12333	463	F.6	Marl
49398	12333	462	F.6	Marl
49398	12333	461	F.6	Marl
49398	12333	460	F.6	Marl
49398	12333	459	F.6	Sandstone
49398	12333	458	F.6	Calcaire
49398	12333	457	F.6	Calcaire
49398	12333	456	F.6	Calcaire

Figure 3:13 Part of a sample borehole with three layers (known points: dark blue, unknown points: light blue)

3.4.3.4 Removing duplicate points

After rounding the altitudes, duplicate rows containing identical values across all columns appeared in the dataset. For instance, consider two data points with altitudes of 135.7 m (marl) and 135.8 m (sandstone). After rounding to the nearest integer, both values become 136 m. Consequently, the dataset contains two records with identical coordinates and altitude but different labels. Although some information may be lost if only one of those points is kept in the dataset, retaining equal rows with the same features can result in biases or misinterpretations by the ML model. The ML algorithm may assign disproportionate weights to the repeated entries, potentially leading to overfitting. Eventually, duplicate points were removed, and only the first occurrence among them was retained in the data set.

3.4.3.5 One-hot encoding

One-hot encoding is a widely used technique for encoding categorical variables. It compares the variables to a fixed reference by transforming a single variable into binary categories. In each observation, one of the categories will be 1 (indicating the presence of that category), and the others will be set to 0.(Hancock & Khoshgoftaar, 2020)

In this study, new labels using the one-hot encoding method were added because the output consists of more than two soil/rock groups, and there is no ordinal superiority for this categorical data.

3.4.4 Summary

For a clearer view of the available data, the number of data points in each dataset and the total number of data points are shown in Table 3:3. In the SPS and SLHC data, there are marl and sandstone, while the LHC has other groups as well (Table 3:3).

Table 3:3 The number of data points for different categories in each dataset

soil/rock \ Data	df_sps	df_slhc	df_lhc_3_4	df_lhc	df_all
marl	594	479	370	528	1857
sandstone	252	103	226	355	838
calcaire	0	0	713	620	1210
marl-calcaire	0	0	149	93	225
conglomerate	0	0	17	44	23
Total	846	582	1475	1640	4153

Table 3:3 shows that the distribution of the number of points from each group is imbalanced, with marl and calcaire comprising the majority, followed by a smaller proportion of sandstone. Given the spatial distribution of boreholes and the lithology along them, it is obvious that marls and sandstones are not easily distinguishable. In fact, they frequently occur together, forming the molasses layer.

In addition, Table 3:3 illustrates that along the SPS tunnel and SLHC section, approximately 70 % or more of the recorded material is marl. Based on this, the model should achieve an accuracy exceeding 70% to reflect learning from the dataset.

Furthermore, the number of conglomerate data points in the French report is 44, while in the integrated data, it decreased to 23. This discrepancy arises because of the duplicate boreholes, where a slight difference exists between the French data report and the CERN data report. For the borehole L135 in the French dataset, there are 23 conglomerate data points, while from the CERN borehole log report, the number is 0. In addition, in SPM 11, these numbers are 7 for the

French dataset and 13 for the CERN dataset. Thus, when both datasets covered the same borehole, the CERN measurements—considered more precise and reliable—took precedence.

3.5 Split train and test data

There are several ways to split data into required data sets. The amount of data will determine the percentage of each dataset. The most straightforward approach is to allocate 70 - 80 percent to the training and 30 - 20 percent to the testing. However, without a validation set, hyperparameter tuning can cause leakage. Using the “train_test_split” method in the scikit learn library, splitting arrays or matrices into random train and test subsets will be done. The data can be divided into three subsets to address the limitations of train-test splitting: training, validation, and test sets. Therefore, the validation set will be used for the hyperparameter tuning and model selection. The common ratio for these subsets is 60%:20%:20%.

Sufficient data should be available for both training and testing the model. In addition, the training and testing sets must be drawn from similar distributions. Otherwise, the evaluation cannot be reliable.

When data availability is limited, Cross-Validation methods can be an effective alternative to a simple train-test split. One of the most popular types is k-fold cross-validation, which divides the dataset into k subsets (folds). Then, the k-1 folds are used as the training set, and the remaining fold is used as the test set. This process repeats k times, allowing every borehole to be used as the test set once. Finally, the average accuracy across all iterations evaluates the model's performance.

Leave-One-Out Cross-Validation (LOOCV) is a specific variant of k-fold cross-validation where the total number of observations is n, and in each iteration, one observation is held out as the test set, and the model is trained on the remaining n–1 observations. This method maximizes the data for training in each iteration, while the computation cost and variance will be higher.

In the present study, all points were grouped by their survey’s name (borehole name), and then LOOCV was applied according to a limited number of boreholes to train and evaluate the models. In each iteration, one of the boreholes from the dataset was held out as the test set, while the remaining boreholes were the training set. This grouping ensures that the model is evaluated on unseen spatial data and is aligned with real-world scenarios.

3.6 ML Models and Evaluation

This section will explain the common ML algorithms and their theoretical foundations. Then, it will introduce the evaluation methods and their functionality in understanding model performance.

3.6.1 Model description

The following ML models were selected and compared. In the next step, the best models regarding accuracy were retrained with new features and investigated more accurately by comparing the balanced accuracies and confusion matrices.

3.6.1.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification problems. It can be an appropriate alternative for algorithms like maximum likelihood. The SVM method aims to fit an optimal hyperplane for classification. In SVM classification, only samples lying at the edge of the class distribution in feature space are needed to build the decision surface. It was designed for binary classification; however, it had been extended to multi-class scenarios.(Mathur & Foody, 2008)

Mathematically, the training sample features are x_i and the training sample labels are y_i which can be -1 or +1 for a binary classification. The goal is to find a hyperplane, which is defined as $\omega.x + b = 0$ where x is a point lying on the hyperplane, ω is normal to the hyperplane, and b is the bias. For two classes, the equation will be defined as follows:

$$y_i(\omega.x_i + b) - 1 \geq 0 \tag{3-1}$$

The support vectors (parallel to the optimized hyperplanes) are $\omega.x_i + b = \pm 1$. In the next step, the margin between these two planes should be maximized, and the equation below has to be solved under the above inequality constraint.

$$\min \{\frac{1}{2} \|\omega\|^2\} \tag{3-2}$$

3.6.1.2 Random Forest (RF)

RF is an ensemble method that combines decision trees and, by aggregating their predictions, improves model accuracy and reduces overfitting. Each decision tree node is based on features chosen randomly, and the leaf represents the output label or value. In RF, random subsets of the dataset are used to train each tree. The final output of classification is the majority of predictions by the trees.

Summarized steps of the RF by (He et al., 2016) are as follows and are shown in Figure 3:14:

1. Bootstrap sampling: sampling training data randomly with replacement to create subsets.
2. Model training: building trees by selecting a random feature and finding the best split based on the criteria
3. Model forecasting: the predicted class from each tree
4. Aggregation: the most voted class as the final prediction.

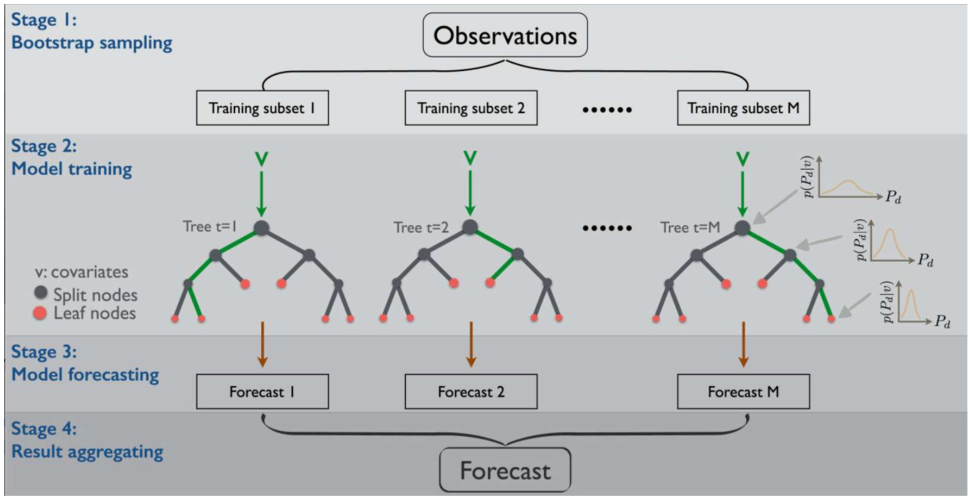


Figure 3:14 Random Forest steps (He et al., 2016)

3.6.1.3 Categorical Boosting (CatBoost)

CatBoost is based on gradient-boosted decision trees that were developed to handle categorical features efficiently. In the training process, a set of decision trees is built consecutively, and after comparing the reduced loss to the previous tree, the successive tree is built. Therefore, each successive tree attempts to correct the errors made by the previous ones. For each feature, quantization is performed to find the preliminary calculation of splits. Then, depending on the format of features, the conversion to numerical ones will be implemented. One of CatBoost's key advantages is its native support for categorical features, which it handles without requiring extensive preprocessing or manual encoding. Finally, using a greedy method, features are selected regarding the preliminary calculation of splits and transformations of features, and the tree structure will be created. Before building a new tree, a random permutation is performed to increase the algorithm's performance.[4]

Figure 3:15 shows the steps of CatBoost algorithm,

[4] https://catboost.ai/docs/en/

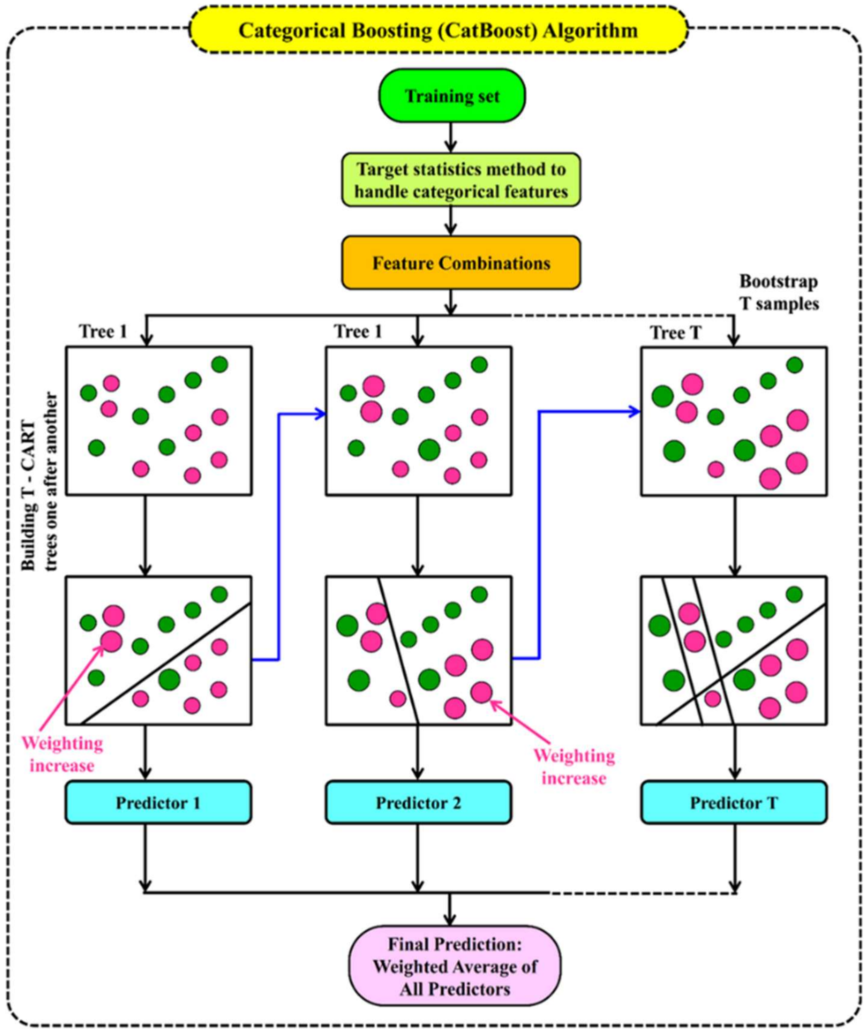


Figure 3:15 Flow chart of the CatBoost algorithm (Pandey et al., 2023)

3.6.1.4 eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting is a supervised learning algorithm based on ensemble trees. It can build a new decision tree in each new iteration to reduce the error of the previous tree. It will continue until the expected accuracy, or the maximum number of trees, is achieved. This technique aims to optimize the Loss function and the regularization term.

$$L(\varphi) = \sum_{i} l(\widehat{y_\iota}, y_i) + \sum \gamma \, T + \frac{1}{2} \, \lambda \, \|\omega\|^2 \qquad (3\text{-}3)$$

In the above equation, the first part, $l(\widehat{y_\iota}, y_i)$, is a differentiable convex loss function to measure the difference between the prediction $(\widehat{y_\iota})$ and the actual value (y_i). The second term aims to regularize and smooth the output to avoid overfitting. This model's output class will have the highest probability across trees.

The most important factor of the XGBoost is its scalability in all scenarios because of the system and algorithmic optimization. Some benefits of this algorithm include parallel learning, using a greedy algorithm, and sparsity awareness (T. Chen & Guestrin, n.d.).

Figure 3:16 presents the flowchart of XGBoost implementation.

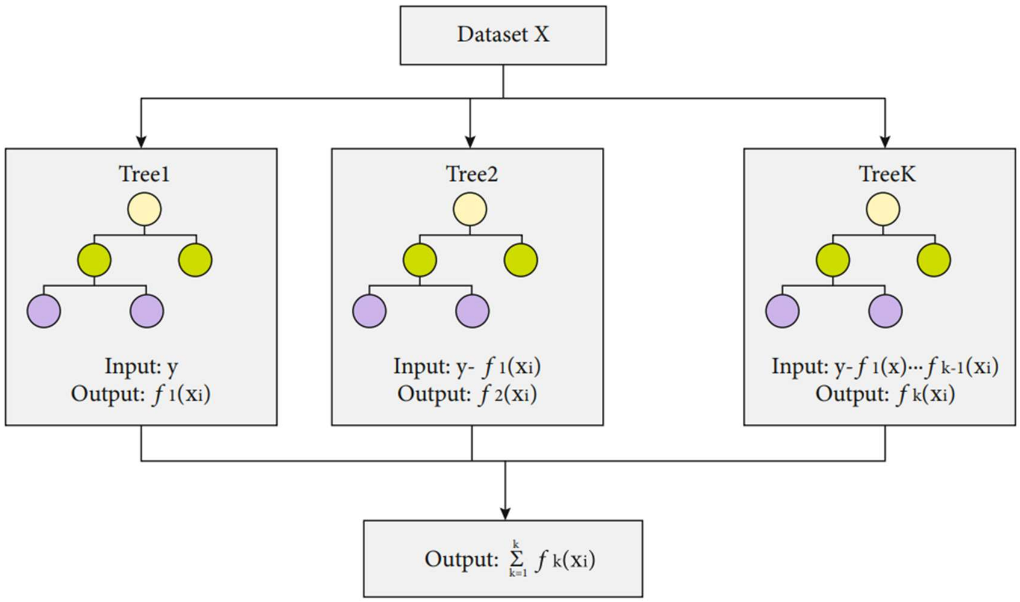


Figure 3:16 Flow chart of XGBoost algorithm (J. J. Liu & Liu, 2022)

3.6.2 Neighborhood aggregation

(Hu et al., 2024) used the neighborhood aggregation to add new features as inputs to the ML model for the soil/rock classification task. It is a technique as a feature engineering where the information from some near neighborhoods is combined and becomes an additional feature for improving the prediction. In this method, for the digitized grid point of the target borehole, y_T, neighboring points at the same elevation level will be aggregated, and using the equation below, the value will be assigned as a new feature for that digitized grid point.

$$y_T = \sum_{i=1}^{n} w_i(d_i) \cdot y_i \tag{3-4}$$

$$w_i(d_i) = \frac{{}^{1}/_{d_i}}{\sum_{i=1}^{n} {}^{1}/_{d_i}} \tag{3-5}$$

In this equation, the y_i is the true classification of the ith neighbor at the same altitude, $w_i(*)$ is the weighting coefficient based on d_i which is the distance between two points (the target point and the ith neighbor - Figure 3:17).

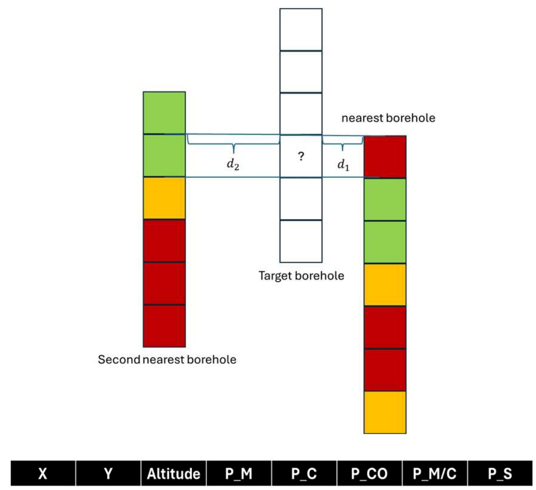


Figure 3:17 Neighborhood aggregation

Finally, the number of additional features introduced corresponds to the number of classes, and each one represents the weight associated with that specific class based on the proximity of neighbors. This study considered the two nearest neighbors because the boreholes were positioned along a circular alignment, resulting in limited directional coverage.

To implement this idea, it was assumed that each pair of boreholes being compared had distinct minimum and maximum altitudes. Therefore, the overlapping depth will vary according to their positioning. Figure 3:18 shows all the possible states. No overlapping altitude exists in the “a” and “e” cases; meanwhile, the overlapping altitude range can be identified and utilized to add the new features in the other three.

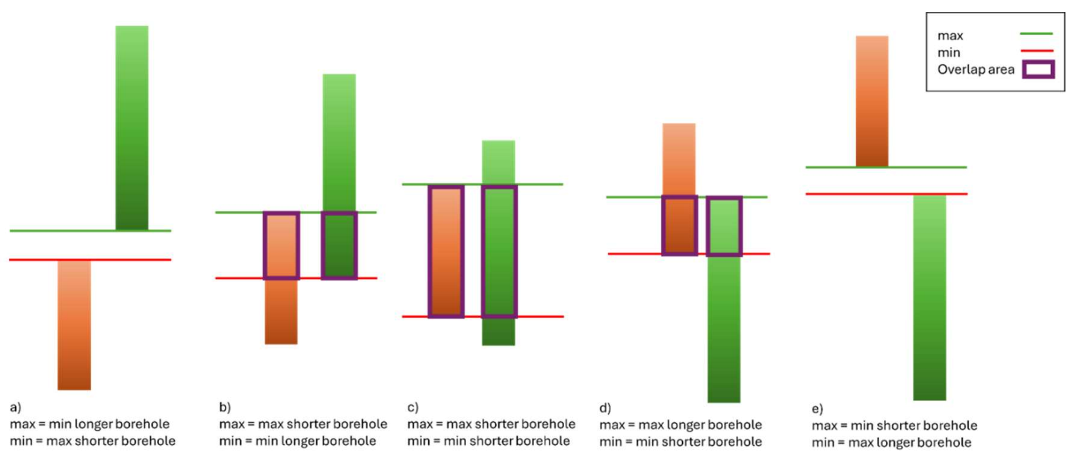


Figure 3:18 Borehole states to find the overlap area

In the final step, the additional features derived from the overlapping areas with the nearest borehole in each dataset were incorporated to train and test the previously described algorithms.

3.6.3 Evaluation methods

Different metrics are used to evaluate the ML algorithms' performance for a classification task, and each one extracts a specific aspect of the results.

3.6.3.1 Confusion matrix

The confusion matrix is a table illustrating the model's performance in a classification task. Each matrix cell records the number of occurrences between two raters and the actual and predicted classification. For binary classification, the matrix is 2*2, and it can be shown as Table 3:4.

Table 3:4 The confusion matrix for binary classification

Actual \ predicted	Positive (1)	Negative (0)
Positive (1)	True Positive (TP)	False Negative (FN)
Negative (0)	False Positive (FP)	True Negative (TN)

The matrix is n*n for the classification with n labels. This matrix compares the predicted and actual labels. The classes are listed in the same order, and the main diagonal shows the correctly predicted labels. (Grandini et al., 2020)

Table 3:5 shows that in the multi-class confusion matrix, E_{XY} are the samples whose actual class was X, but they were misclassified as class Y. Thus, FP_X will be the sum of the E_{YiX}, while FN_X will be the sum of the E_{XYi}. (Tharwat, 2018)

Table 3:5 The confusion matrix for multi-class classification

Actual \ Predicted	Class A	Class B	Class C	Class D
Class A	TP_A	E_{AB}	E_{AC}	E_{AD}
Class B	E_{BA}	TP_B	E_{BC}	E_{BD}
Class C	E_{CA}	E_{CB}	TP_C	E_{CD}
Class D	E_{DA}	E_{DB}	E_{DC}	TP_D

3.6.3.2 Accuracy

Accuracy is a parameter that shows the overall performance of the model. It is the proportion of correctly predicted labels and the number of samples. If the $\hat{y}_i$ is the predicted label and y_i is the actual label, the following function returns the accuracy.

$$accuracy(\hat{y}_i, y_i) = \frac{1}{n_{sample}} \sum_{i=0}^{n_{sample}-1} \mathbb{1}(\hat{y}_i = y_i) \quad (3\text{-}6)$$

Using the confusion matrix value, the accuracy will be as follows:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (3\text{-}7)$$

In the accuracy calculation, all data points' weights are equal. The final accuracy number will be between 0 and 1, where 0 means the worst performance and 1 means that the model can correctly predict all the labels.

The "accuracy_score" from the scikit-learn library calculates the accuracy.

3.6.3.3 Balanced accuracy

The accuracy score alone may not provide sufficient insight into model performance when dealing with imbalanced datasets. For example, in a dataset with 100 data points, 30 points belong to class A, and 70 points to class B. Therefore, if a model naively predicts all instances as class B, it will still achieve an accuracy of 70%. However, no learning will happen in this model because it cannot distinguish between the classes.

The balanced accuracy score computes the accuracy where each sample is weighted according to the inverse prevalence of its actual class. For binary classification, the below relation calculates the balanced accuracy and is the mean of sensitivity (TP rate) and specificity (TN rate).

$$balanced\ accuracy = \frac{1}{2}\ (\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \quad (3\text{-}8)$$

If the model's performance for all classes is similar, the accuracy and balanced accuracy will be equal. On the other hand, in scenarios where the model's predictions are no better than random guessing, the balanced accuracy can drop to $\frac{1}{n_{class}}$. For instance, if there are 5 output classes and the balanced accuracy is near 20 % ($\frac{1}{5}$), it means that the model is not learning from the input features, and the predictions are random.

The formula for balanced accuracy of multi-class classification is an average of the recall obtained on each class and can be calculated from the equation below (Grandini et al., 2020):

$$balanced\ accuracy = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} \frac{TP_i}{TP_i + FN_i} \quad (3\text{-}9)$$

The "balanced_accuracy_score" method was used to compute this score for the desired data and implemented in the scikit-learn library.[5]

3.6.3.4 Precision, Recall, and F1-score

The precision score indicates the model's performance for correctly predicting positive cases. It is useful when the aim is to minimize the FP cases, and the following relation demonstrates how it can be computed.

$$precision = \frac{TP}{TP + FP} \qquad (3\text{-}10)$$

Additionally, the recall score shows the classifier's ability to find positive cases. The equation below illustrates how this score can be calculated.

$$recall = \frac{TP}{TP + FN} \qquad (3\text{-}11)$$

The difference between these two scores is in their goal. The precision answers the questions about how many points were truly predicted as positive from all the positive instances. Meanwhile, the recall evaluates how the model can correctly identify the most positive cases. Therefore, low recall means the model is missing many positive cases, while low precision is a sign that the model is incorrectly predicting the output in the positive group.

The F1-score, the harmonic mean of precision and recall, can be used where both situations are essential.

$$F1\ score = 2 \times \frac{precision\ \times\ recall}{precision + recall} \qquad (3\text{-}12)$$

3.7 Implementation and summary

The present research aims to develop a geological map near CERN tunnels and investigate the impact of the hydrogeological profile on tunnel defects.

The process began with the borehole log data collection and structuring. Next, the preprocessing step was done to prepare the inputs and outputs of the models. After that, with the LOOCV, the data has been split according to the borehole names. The specific borehole was designated as a test set in each iteration within a data set, and the remaining boreholes were used to train and fit the selected model. The RF and SVM were implemented using the built-in functions from the scikit-learn library, and the default values for

[5] https://scikit-learn.org/stable/modules/model_evaluation.html

hyperparameters were used.[6][7] Moreover, XGBoost and CatBoost were implemented using their specific libraries.[8][9]

For model training, the input features were x, y, and altitude for each data point, and the output was the type of soil/rock.

Then, using neighborhood aggregation, additional features were added as input, and each dataset was trained by different algorithms and tested. Given the circular arrangement of the boreholes around the tunnel, only the two nearest neighbors were considered for calculating these new features.

To build the geological map near the LHC tunnel from point 3 to point 4, using the x coordinates and z value - the ground level - for the x values in this area, the z values were predicted by Support Vector Regression (SVR) which is an extension of SVM adapted for predicting continuous numerical values instead of class labels. Then, the y values were calculated using the equation of a circle, based on the tunnel's known points. The estimated circle equation had been extracted from the coordinates of three points on the LHC tunnel, and was expressed as follows:

$$(x - x_{center})^2 + (y - y_{center})^2 = radius^2 \quad (3\text{-}13)$$
$$x_{center} = 49424$$
$$y_{center} = 12553$$
$$radius = 424$$

Artificial boreholes were generated at the interpolated coordinates, and the type of soil/rock along each borehole was predicted using all available data and a trained model. Therefore, the 3D geological map near the LHC tunnel from Point 3 to 4 was developed.

The 3D map along the tunnel was converted to a 2D plane, and the spatial coordinates were converted to cumulative distances to facilitate comparison between the geological map and leakage data. To build this 2D map along the cumulative distances, the Euclidean distance between two boreholes next to each other was calculated and added to a constant number, which is the location of the starting point from point 3 to point 4.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[8] https://xgboost.readthedocs.io/en/stable/parameter.html
[9] https://catboost.ai/docs/en/concepts/python-reference_catboostclassifier

Chapter 4
Result and discussion

4 Results and discussion

This chapter first presents the Machine Learning (ML) model results and compares their performance across the different datasets. It then introduces the resulting geological map. Finally, it discusses how the inferred hydrogeology affects tunnel leakage.

4.1 Result of ML algorithms

The performance of SVM, RF, CatBoost, and XGBoost on 5 different data sets using various evaluation metrics was compared. The average accuracy and average balanced accuracy of the cross-validation iterations were gathered in Table 4:1. The results indicated that, despite reducing the number of data points for training after the neighborhood aggregation, the average accuracy and average balanced accuracy were maintained when the integrated dataset was used. After applying the neighborhood aggregation, the number of boreholes used changed because of their limitations regarding the overlapped altitude limitation.

In the SPS and SLHC datasets, the geological composition is limited to two types: “marl” and “sandstone”. In the SPS dataset, 70 % of the data points are labelled as “marl” and 30 % are classified as “sandstone”, while in the SLHC dataset, the ratio is 82 % “marl” to 18 % “sandstone”. As Table 3:3 shows, for the LHC point 3 to point 4 and the LHC_French datasets, there are 5 output categories. Most of the labels are “marl” and “calcaire”, while there is a moderate proportion of “sandstone”. However, the data from “marl-calcaire” and “conglomerate” classes is notably limited.

Table 4:1 indicates that the average balanced accuracy for all models using SPS and SLHC datasets is between 0.53 and 0.58. This range suggests that the models performed only marginally better than random chance, because with two output classes, the balanced accuracy of a random model will be 0.5. For the SPS dataset with additional features, the performance of the RF improved to almost 0.6. However, in the SLHC dataset, neighborhood aggregation did not change the balanced accuracy because of the limited number of boreholes.

Within the LHC point 3 to 4 dataset, five soil/rock types comprised the target labels, and the average balanced accuracy varied between 0.25 and 0.30. Therefore, the model's performance was slightly better than that of a random classifier, as predicting the labels randomly would result in a balanced accuracy of 0.2. In addition, there was no significant enhancement when additional features were considered as input for this dataset.

Finally, in the LHC_French data set and the integrated dataset using all datapoints, the average balanced accuracy markedly exceeded the random classification threshold of 0.20, in the range of 0.48 to 0.55. However, the SVM algorithm was exceptional when the LHC_French dataset was used, which presented an outlier performance. A slight improvement happened in some cases when the neighborhood aggregation weights were added to the features. When the ML algorithms were trained by all the available datapoints (df_all), the average accuracy and average balanced accuracy attained peak performance, ranging from 0.52 to 0.6 and 0.48 to 0.51, respectively.

In this study, only 2 nearest neighbors were considered to calculate the weights. In addition, the variation in the ground level caused differences in the borehole depths. The upper sections of the boreholes were mainly labeled as categories distinct from the five chosen for classification. To mitigate the risk of predicting incorrect labels for the top section of the boreholes, overlapping points of the designated test borehole with the nearest neighboring borehole were considered to train the models with additional features.

Table 4:1 Comparing the performance of the models with different inputs

dataset	input	output classes	number of boreholes	metric	ML model			
					SVM	RF	CatBoost	XGBoost
SPS	x,y,altitude	M,S	36	avg_acc	0.75	0.66	0.72	0.66
				avg_balanced_acc	0.58	0.53	0.57	0.54
	x,y,altitude, w_M, w_S		23	avg_acc	0.68	0.67	0.66	0.62
				avg_balanced_acc	0.6	0.61	0.59	0.58
SLHC	x,y,altitude	M,S	7	avg_acc	0.81	0.77	0.82	0.77
				avg_balanced_acc	0.5	0.58	0.54	0.58
	x,y,altitude, w_M, w_S			avg_acc	0.8	0.79	0.79	0.76
				avg_balanced_acc	0.5	0.55	0.53	0.56
LHC_3_4	x,y,altitude	M,S,C,CO,M_C	12	avg_acc	0.5	0.41	0.45	0.46
				avg_balanced_acc	0.3	0.25	0.25	0.28
	x,y,altitude, w_M, w_S,w_C,w_CO,w_M/C			avg_acc	0.49	0.46	0.52	0.43
				avg_balanced_acc	0.25	0.3	0.29	0.29
LHC_French	x,y,altitude	M,S,C,CO,M_C	24	avg_acc	0.12	0.55	0.53	0.46
				avg_balanced_acc	0.13	0.52	0.48	0.49
	x,y,altitude, w_M, w_S,w_C,w_CO,w_M/C		20	avg_acc	0.15	0.4	0.46	0.4
				avg_balanced_acc	0.15	0.55	0.45	0.55
all	x,y,altitude	M,S,C,CO,M_C	73	avg_acc	0.57	0.6	0.65	0.59
				avg_balanced_acc	0.48	0.51	0.51	0.5
	x,y,altitude, w_M, w_S,w_C,w_CO,w_M/C		59	avg_acc	0.54	0.6	0.63	0.5
				avg_balanced_acc	0.49	0.53	0.52	0.48
abbreviations :								
w_M:	weight for marl	M:	marl	avg_acc:	average accuracy			
w_S:	weight for sandstone	S:	sandstone	avg_balanced_acc:	average balanced accuracy			
w_C:	weight for calcaire	C:	calcaire					
w_CO:	weight for conglomerat	CO:	conglomerat					
w_M_C:	weight for marl_calcaire	M_C:	marl_calcaire					

Comparing the algorithms in Table 4:1, raising the number of boreholes and considering a wider area of study increased both average accuracy and average balanced accuracy. For the RF and CatBoost models trained on df_all, the average accuracy is nearly 0.6, and the average balanced accuracy is about 0.52 when neighborhood-aggregation features are integrated in the inputs.

Next, Table 4:2 compares each class's average accuracy, average balanced accuracy, precision, recall, and F1 score. The input features of the models include X, Y, altitude, and additional features from neighborhood aggregation. The “-” values indicate the absence of “calcaire”, “conglomerate”, and “marl_calcaire” labels. Regarding the average accuracy score in cross-validation, the CatBoost performed better than the others across all datasets. Meanwhile, evaluating an imbalanced dataset using average accuracy can be misleading, as explained in the previous chapter. Therefore, the average balanced accuracy can assess the more robust performance of algorithms, compared to the average accuracy.

Table 4:2 Comparison of ML models' performance on datasets with neighborhood aggregation features

	SPS				SLHC				LHC_3_4				LHC_French				all			
	SVM	RF	CatBoost	XGBoost	SVM	RF	CatBoost	XGBoost	SVM	RF	CatBoost	XGBoost	SVM	RF	CatBoost	XGBoost	SVM	RF	CatBoost	XGBoost
avg_acc	0.68	0.67	0.66	0.62	0.8	0.79	0.79	0.76	0.49	0.46	0.52	0.43	0.15	0.4	0.46	0.4	0.54	0.6	0.63	0.5
avg_balanced_acc	0.61	0.61	0.59	0.58	0.5	0.55	0.53	0.55	0.25	0.3	0.29	0.29	0.14	0.55	0.45	0.55	0.49	0.53	0.53	0.48
precision_C	-	-	-	-	-	-	-	-	0.59	0.8	0.74	0.75	0.31	0.33	0.41	0.48	0	0.73	0.66	0.64
recall_C	-	-	-	-	-	-	-	-	1	0.79	0.79	0.74	0.62	0.31	0.55	0.34	0	0.68	0.77	0.58
F1_score_C	-	-	-	-	-	-	-	-	0.74	0.8	0.76	0.75	0.41	0.32	0.47	0.4	0	0.7	0.71	0.61
precision_CO	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0.56	0	0	0	0	0
recall_CO	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0.13	0	0	0	0	0
F1_score_CO	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0.21	0	0	0	0	0
precision_M	0.65	0.66	0.65	0.63	0.8	0.82	0.81	0.82	0	0.37	0.39	0.32	0.17	0.36	0.47	0.39	0.42	0.56	0.63	0.52
recall_M	1	0.75	0.78	0.68	1	0.94	0.96	0.89	0	0.61	0.45	0.55	0.1	0.67	0.49	0.75	1	0.82	0.73	0.74
F1_score_M	0.79	0.7	0.71	0.65	0.89	0.88	0.88	0.85	0	0.46	0.42	0.41	0.12	0.47	0.48	0.51	0.59	0.67	0.68	0.61
precision_M/C	-	-	-	-	-	-	-	-	0	0.33	0.27	0.08	0	0	0	0	0	0.29	0.08	0
recall_M/C	-	-	-	-	-	-	-	-	0	0.04	0.1	0.02	0	0	0	0	0	0.03	0.083	0
F1_score_M/C	-	-	-	-	-	-	-	-	0	0.07	0.15	0.03	0	0	0	0	0	0.05	0.04	0
precision_S	0	0.36	0.33	0.29	0	0.45	0.44	0.35	0	0.18	0.21	0.18	0	0.75	0.21	0.46	0	0.39	0.3	0.34
recall_S	0	0.26	0.21	0.25	0	0.19	0.12	0.23	0	0.12	0.19	0.08	0	0.11	0.13	0.19	0	0.15	0.16	0.17
F1_score_S	0	0.3	0.26	0.27	0	0.27	0.19	0.28	0	0.14	0.2	0.11	0	0.19	0.16	0.27	0	0.22	0.21	0.22

From the perspective of models' comparison, as the values shown in Table 4:2 , the recall for "marl" class in the df_sps, df_slhc, and df_all, and the recall for "calcaire" class for df_lhc_3_4 and df_lhc is equal to 1 when SVM was implemented, although precisions and recalls for some other categories are equal to 0. Therefore, SVM frequently degenerates into a majority-classifier. This happened because of the SVM's inherent objective of finding hyperplanes to separate the points of different classes. Although the kernel function is a radial basis, the classes cannot be separated in the studied dataset.

The "conglomerate" prediction proved challenging for models due to the limited data points, resulting in precisions and recalls of 0 in most cases. In addition, only a few data points were classified as "marl-calcaire" correctly for the same reason. The recall value for "marl" and "calcaire" is higher than the others, indicating that the model can identify most of these samples. If the aim was to improve predictions for the minority classes like "conglomerate" or "marl-calcaire," it could be beneficial to apply techniques such as class weighting or oversampling to address the imbalance in data set.

As a result, training df_all for the RF and CatBoost performed better than the SVM and XGBoost. The F1 score for "marl" and "calcaire" is about 0.7, suggesting the model learned how to predict these two classes well.

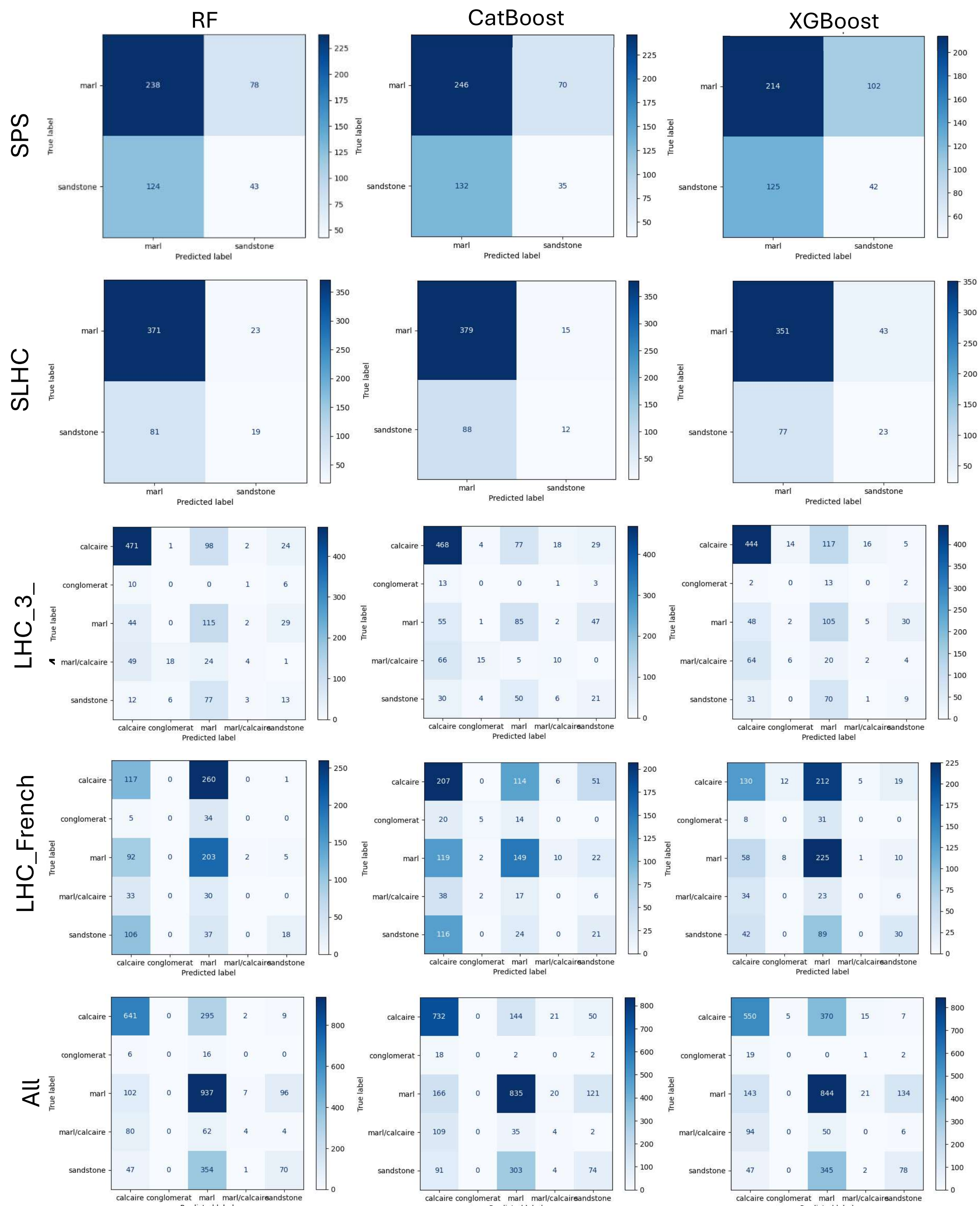


Figure 4:1 Confusion matrices for RF, CatBoost, and XGBoost for each dataset

The confusion matrices in Figure 4:1 confirm that RF and CatBoost can achieve relatively good predictions for the “marl” and “calcaire”. Meanwhile, there are misclassifications for the “sandstone” group, which can be attributed to the complexity of the layers and the limited boreholes in the molassic area where the layers of marl and sandstone are thin and hard to discrete. Figure 4:1 illustrates that the models struggle to differentiate between “marl” and “sandstone”, leading to frequent misclassification.

In conclusion, RF and CatBoost performed more effectively than SVM and XGBoost: by leveraging the larger training dataset, they produced more accurate label predictions. It is worth noting that RF is considerably faster than CatBoost, especially when working with large datasets and multiple iterations on cross-validation.

4.1.1 Address class imbalance

In the available dataset, the class distribution is imbalanced. There are different methods to handle this issue.

One implemented solution for addressing imbalances in the dataset was the use of an Oversampling method, the Synthetic Minority Oversampling Technique (SMOTE), which aims to balance the class distribution by generating new synthetic minority instances between existing ones, rather than simply duplicating samples. In the first step, the k-nearest neighbors of each minority-class sample are identified using the Euclidean distance within the same class. Then, based on the specified sampling rate, new synthetic points are generated along the line segments between the sample and its neighbors.

To apply SMOTE, in each iteration of LOOCV, synthetic samples of “marl-calcaire” and “conglomerate” were generated until their class sizes matched that of “sandstone”. Figure 4:2 shows the confusion matrix when all the real and synthetic boreholes were used as a test set once, and RF was used to train the model. The average accuracy was 0.6.

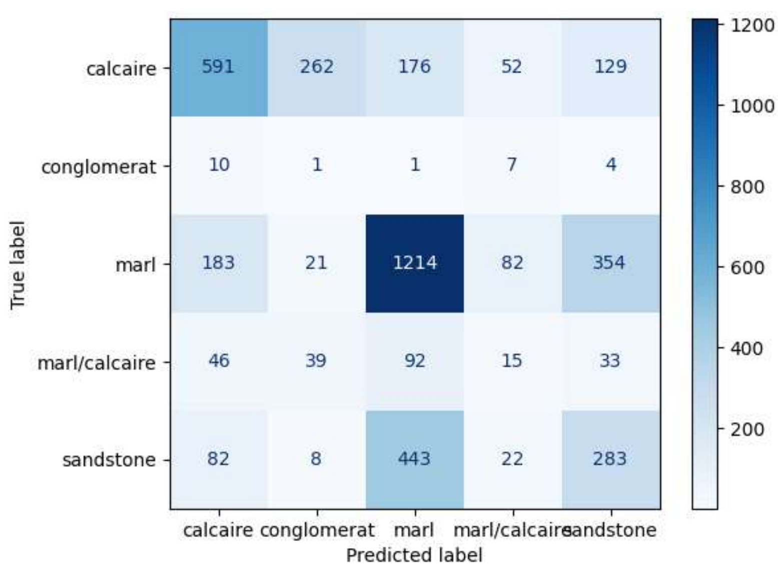


Figure 4:2 The confusion matrix after applying SMOTE on the integrated dataset and LOOCV using the RF

Another solution to address imbalanced classes is using the parameter “class_weight” with a value of “balanced” in the RF algorithm. This approach assigned higher weights to minority classes based on their frequency in the dataset, penalizing misclassification without modifying the original data.

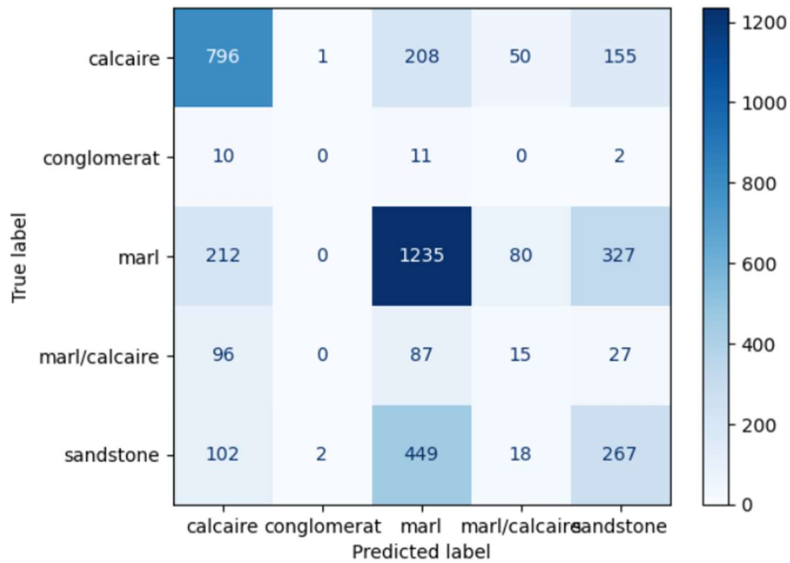


Figure 4:3 The confusion matrix of LOOCV and "RandomForestClassifier(class_weight ='balanced')" on the integrated dataset

Figure 4:3 illustrates the confusion matrix, when RF was called with the "class_weight" parameter, and the average accuracy of the LOOCV was 0.63.

In the above solution, the neighborhood aggregation features were not used as the input features.

Despite these approaches, the overall model accuracy did not improve. Both SMOTE and class weighting increased the representation of minority classes, but their effect on classification performance was limited. The dataset suffers from class imbalance, and due to the limited number of samples, neither assigning higher weights to minority classes nor generating synthetic samples (e.g., using SMOTE) enables the classifier to learn patterns that generalize well to unseen data.

4.2 Developed geological map

As the literature review stated, point cloud-based mapping has been used in many studies. In this research, a 3D geological map from point 3 to point 4 of the LHC tunnel was developed using the point cloud technique. This section was selected regarding the available leakage data.

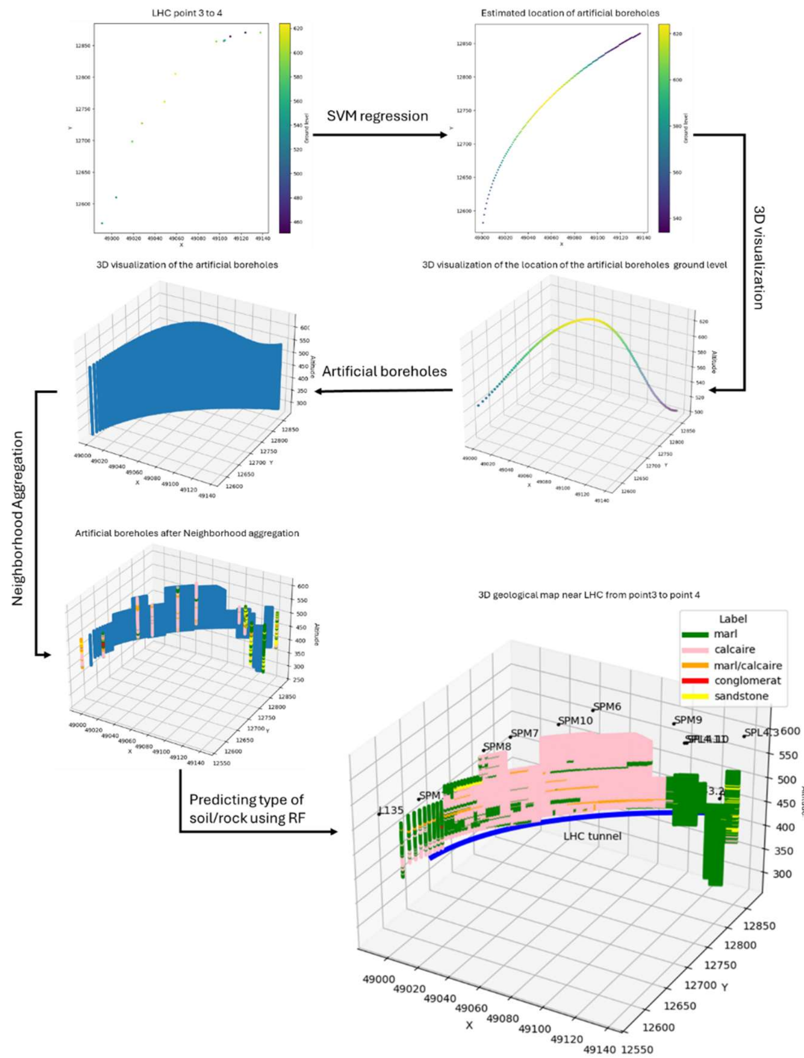
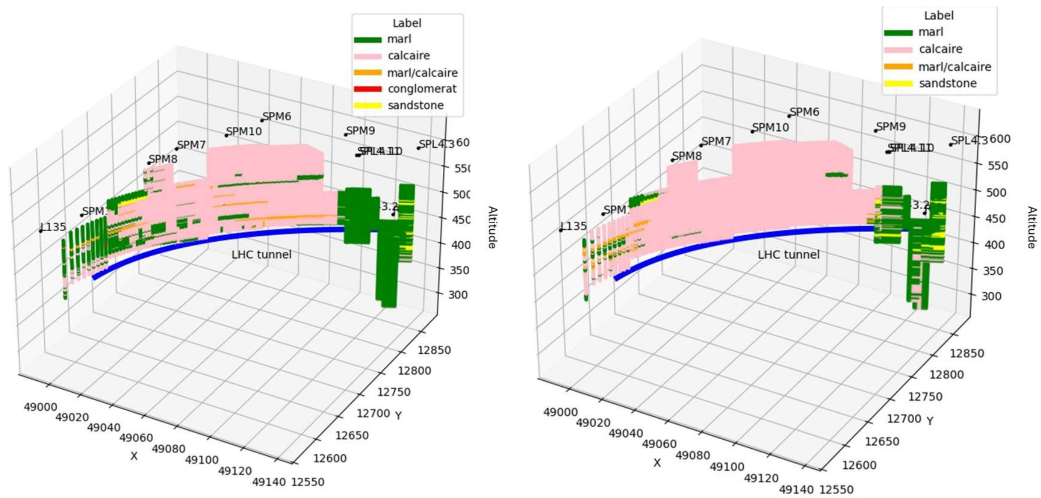


Figure 4:4 Geological mapping process

As Figure 4:4 shows, first, the SVR was employed to predict the ground level in each area. From the df_lhc_3_4 dataset, the 'x' values were the input features, and the 'z' values (ground level) served as the output. Next, using the circle formula obtained from the three specific boreholes, the 'y' values were also derived. This allowed for the positioning of artificial boreholes, which were then used to construct the geological map.

Once the location of the artificial boreholes was determined and data points filtered using the overlapped area with their neighbors, the type of soil/rock for each point cloud remained unknown and had to be predicted using the previously discussed algorithms. Therefore, RF and CatBoost were selected to predict the type of soil/rock along these artificial boreholes, and neighborhood aggregation was incorporated into the input features using the df_all points regarding the accuracy and balanced accuracy of this dataset on these algorithms.

The empty parts at the top of the artificial boreholes correspond to the topsoil, gravel, silt, and other types of soil/rock, which are irrelevant to the scope of this study. By filtering the rock type and considering the 2 nearest real borehole features, these two final maps were presented in Figure 4:5.



(a) (b)

Figure 4:5 Final geological map using (a) the RF, and (b) the CatBoost

As the developed maps show in Figure 4:5, the geology near this part of the tunnel does not change significantly. In most areas in the middle of this section, calcaire is the most appeared rock. Near points 3 and 4 on the sides, the ground level decreases, and the tunnel enters the molassic area.

In the final map using the CatBoost, all the middle area was predicted as “calcaire”. However, RF predicts some inner points as “marl”. The maps and confusion matrices indicate that when both “calcaire” and “marl” are plausible, CatBoost tends to assign the class to “calcaire”, whereas RF tends toward “marl”. This is due to the foundations of these algorithms. RF builds the decision trees randomly, while CatBoost attempts to correct the previous errors. Hence, when predicting most of the “calcaire” labels is the objective, CatBoost can be a better choice, and when the “marl” class is more important to be predicted correctly, RF’s performance can be better.

4.3 Tunnel defects

In the previous studies, researchers investigated the variety of defects along the CERN tunnels. (Brien et al., 2023) studied deep learning methods used to detect cracks along the tunnel automatically. Researchers illustrated an innovative image-based crack detection method accompanied by crack classification using Convolution Neural Network (CNN) specifically for the underground infrastructure environment, where one of the challenges was disparity of illumination combined with the congested image data caused by pipes, steel mesh, wires, and other tunnel facilities. Moreover, (Ouyang et al., 2023) described a remote and automated system for crack monitoring at a pixel-level scale using robot-mounted imaging technology. The system can collect and stitch the pictures to build a panorama image.

The goal of the present study was to develop a geological map in the vicinity of the CERN tunnels. Specifically, because the leakage data from points 3 to 4 was available, the geological map for this section was built. For a more straightforward comparison of the results, the 3D map was converted to a 2D visualization along the tunnel section.

To do this, the first borehole location was estimated and converted to a cumulative distance, Distance Cumulée (DCUM). According to this, the DCUM was calculated for the remaining boreholes. The Euclidean distance between boreholes was calculated, with the first borehole serving as the reference point for a distance of zero. Starting with the first borehole the Euclidean distance to each subsequent borehole was added to the cumulative distance of the previous borehole for all remaining boreholes. The resulting values were scaled by a factor of 12 and offset by 5700, thereby placing the estimated DCUM for the starting point of the borehole chain on the same scale as the leakage data. The scaling factor (12) and the offset value (5700) were determined by fitting the estimated data to the reference data provided by CERN so that the elevation above ground level at each DCUM matched between the reference data and studied dataset.

In Figure 4:6, the results indicate that leakage is less pronounced near points 3 and 4 along the tunnel's flanks—where marl and sandstone dominate and the alignment lies within the molassic zone—than in the central section, which is characterized by more calcareous rock. This can be attributed to the inherently more fractured nature of the calcaire and its higher permeability compared to marl. Nevertheless, there is no other strong relation between the geology of the surrounding area and the leakage data for this section of the tunnel because most parts near the tunnel surface are composed of "calcaire" between DCUM 7500 and 9000, where more leakage was observed. Therefore, the hydrology of the area was also investigated.

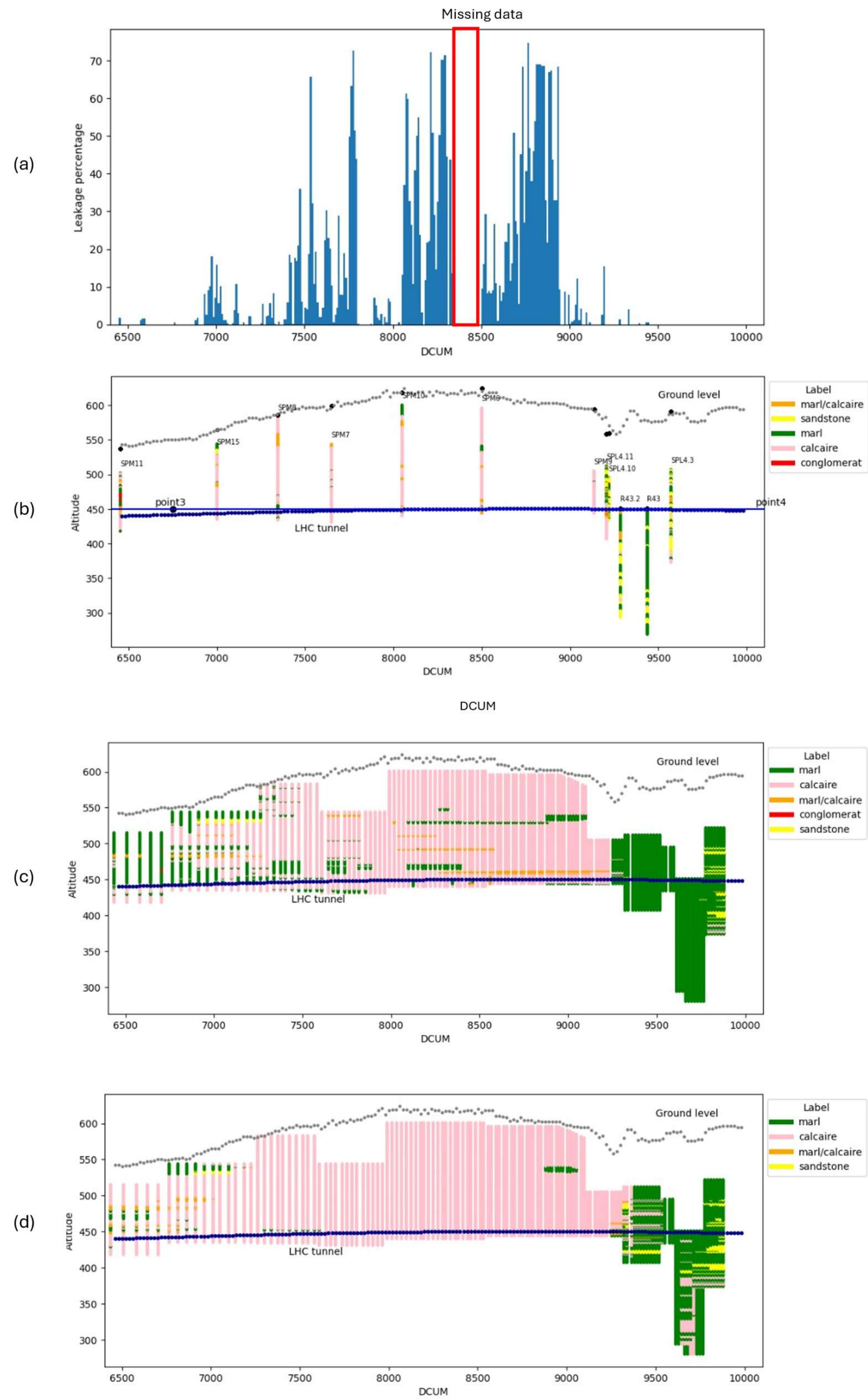


Figure 4:6 The LHC tunnel from point 3 to point 4: (a) leakage graph, (b) 2D map of boreholes in LHC_3_4 dataset, (c) final 2D map with RF, (d) final 2d map with CatBoost

Figure 4:6 (a) shows leakage data investigated by CERN on-site surveys, covering a 3 km section of the LHC tunnel, from Point 3 to Point 4. It is important to note that data are missing between DCUM 8360 and 8490. They reported that the tunnel section with DCUM 8750 to 8940 encountered the most severe leakage. They also mentioned that the drainage system was not functioning properly. In Figure 4:6 (b), black points represent the ground level of the boreholes, and in (c) and (d), the grey points indicate the ground level. The dark blue line indicates the location of the LHC tunnel.

To investigate the region's hydrology, the available maps in the French report show that a river is near SPM9 (Figure 4:7). In addition, the location of the boreholes was estimated on Google Maps (Figure 4:8), which confirms the river's existence.

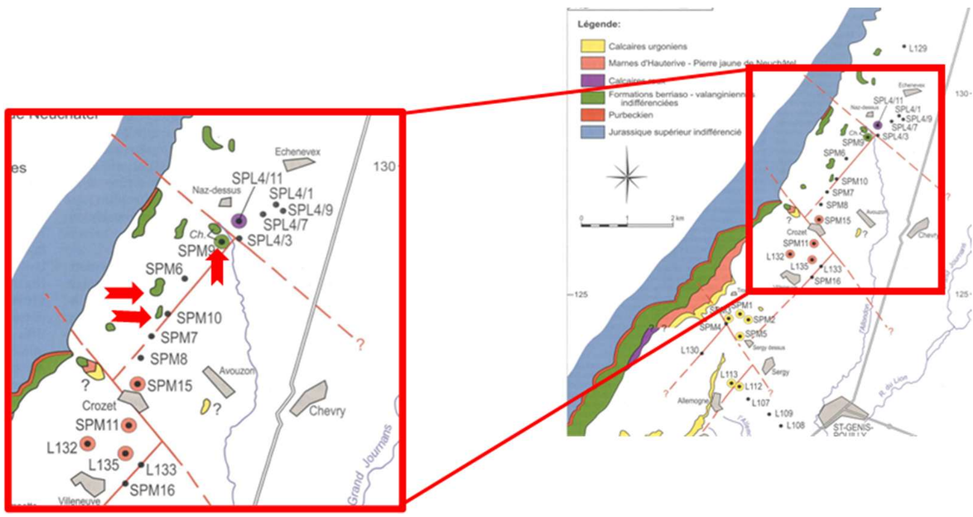


Figure 4:7 The SPM9, SPM6, and SPM10 location in French report maps (Charollais et al., 2007)

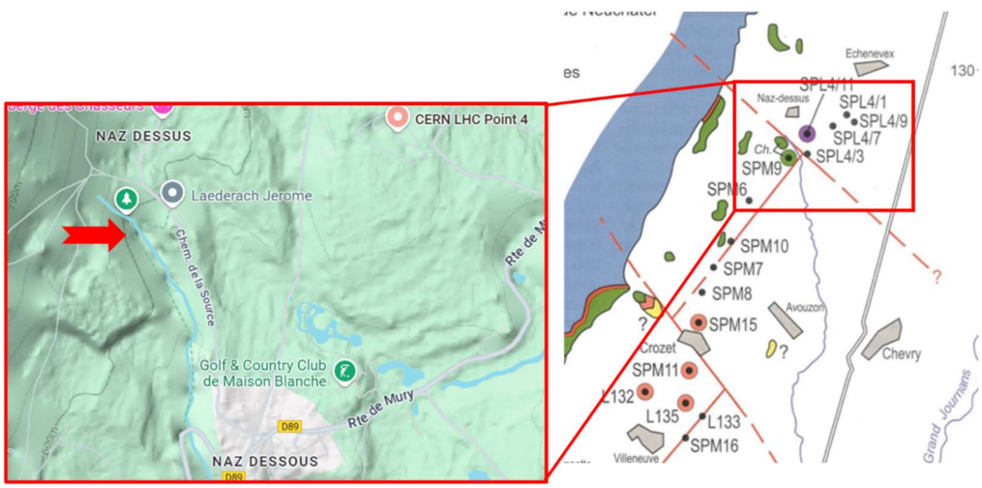


Figure 4:8 The river location on Google Maps [10]

Another hydrogeological parameter is the ground level and the difference in altitude between the tunnel and the ground surface. As Figure 4:6 illustrates, the ground level is about 535 m near SPM 11 and rises to roughly 625 m near SPM6,

[10] https://maps.app.goo.gl/TH1mAX7th5ryjRws9

and the highest altitudes occur between DCUM 7000 and 9000, where most of the leakages were observed.

(Huang et al., 2019) implemented an analytical solution of tunnel leakage regarding Darcy's law. The leakage or discharge per unit width can be given by:

$$q_1 = K_1 \times \frac{h_u - h_c}{d_1} \times 2\pi R \tag{4-1}$$

$$q_2 = K_2 \times \frac{h_c - h}{d_2} \times 2\pi(R + d1) \tag{4-2}$$

where the q_1 and q_2 are the leakages per unit width with a concrete lining and the grouting zone in the surrounding rock, K_1 and K_2 are the hydraulic conductivity of the two grouting zones, d_1 and d_2 are the thicknesses of the two grouting zones, and finally, the h_c is the water table above the lining and h_u refers to the upper storage reservoir water level (Figure 4:9). If we simplify the formula with the assumption of $q_1 = q_2$ and $K_1 = K_2$ the equation can be written as:

$$q = K \times \frac{h_u - h}{d_1 + d_2} \times 2\pi R \tag{4-3}$$

In conclusion, an increase in hydraulic head, h, leads to a corresponding rise in total discharge.

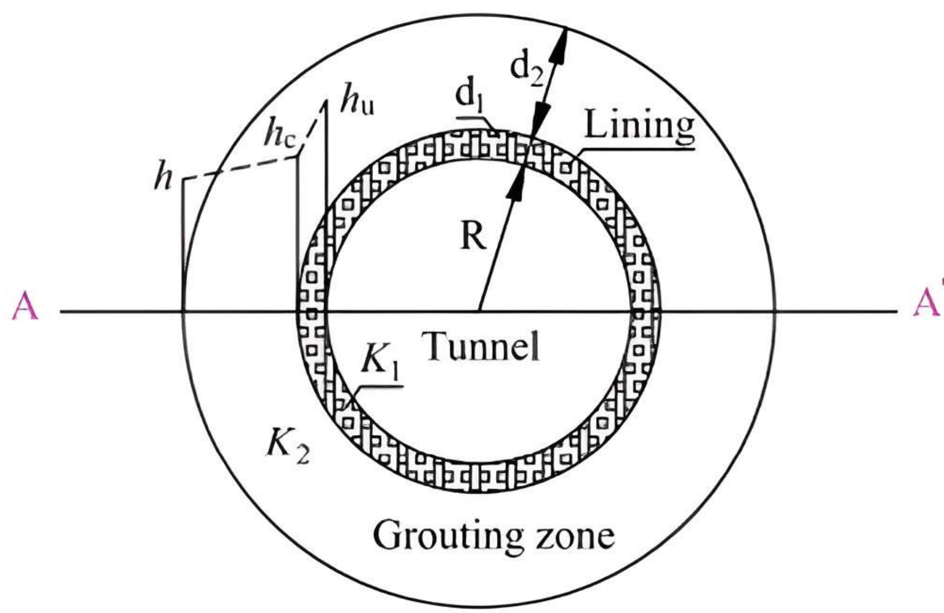


Figure 4:9 A schematic diagram of the tunnel section(Huang et al., 2019)

To summarize, comparing the leakage data and the developed geological map reveals that leakage decreases as the tunnel enters the molassic from the calcareous area. Although the type of soil or rock can contribute to tunnel defects, this study found no strong correlation between lithology and leakage data. Furthermore, the hydrology of the region demonstrated the presence of a river near the leaking section, which may contribute to inflow. However, the most significant relation was found with the hydraulic head, where the spatial trend almost mirrors the leakage data.

Chapter 5
Conclusion

5 Conclusion

To enhance the functionality of tunnel infrastructure and gain deeper insights into the hydrogeological processes behind tunnel deterioration, this study employed machine learning (ML) algorithms to create a geological map of the Large Hadron Collider (LHC) tunnel section between points 3 and 4. This study began with collecting borehole reports from multiple sources. As these reports were unstructured, extensive preprocessing was required: the data were first tabulated in Excel and then imported as data frames in Python.

Each dataset included key spatial and geological attributes: x and y coordinates, ground level, borehole name, rock type, and the lower boundary of each lithological layer. During the preprocessing steps, the rock types were filtered, and internal points were added at one-meter intervals to represent the rock type along the borehole. The duplicate points were compacted, and one-hot encoding was used to encode the output classes.

The coordinates of the available boreholes were originally in different reference systems and required transformation into a unified coordinate system with the exact origin. The common boreholes in the datasets from various sources were used to retrieve all new coordinates and integrate all datasets.

Neighborhood aggregation was applied to generate new features as new inputs. In this process, the overlapped altitude within the nearest boreholes was found, and the new features were computed using the distances to the neighboring boreholes and their corresponding rock type at the same altitude.

Various supervised Machine Learning (ML) methods were trained using spatial features derived from borehole data. The results show:

- Among the tested models, Random Forest (RF) and Categorical Boosting (CatBoost) demonstrated superior performance compared to Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost). Due to its nature, SVM attempts to fit a hyperplane to separate the classes; however, in the available datasets, it failed to distinguish correct categories and classified nearly all data points into a single group. While XGBoost performed considerably better than SVM, its performance was still inferior to RF and CatBoost.
- In this dataset, the boreholes were spaced relatively far apart, and the resolution in the z-direction was 1 meter. Predicting the correct label at each altitude was challenging under these conditions. Although some information was lost during the neighborhood aggregation, the model's accuracy remained unchanged. This outcome highlights the importance of feature engineering in improving model robustness and performance.
- In the central section of the generated map, the RF predicted some thin layers of "marl", while CatBoost predominantly classified the points in

this area as “calcaire”. This discrepancy arises from how the two algorithms construct decision boundaries. The RF builds trees by randomly selecting a subset of features, which can potentially lead to more localized predictions. In contrast, CatBoost builds the trees sequentially, with each new tree aiming to correct the error of the previous one, resulting in a more generalized classification.

- Furthermore, the integrated dataset with more data points achieved higher average accuracy and average balanced accuracy. In conclusion, as the dataset size increases and more boreholes are incorporated to train the model, the map's accuracy improves, enabling the generation of higher-resolution maps.
- In addition, when the tunnel enters the molassic area and the surrounding rocks transition from calcaire to marl, the leakage decreases. This is likely because the calcaire is more fractured and permeable compared to the marl.
- Although no strong correlation was observed between geological characteristics and leakage pattern, hydrological factors appeared to have a more substantial influence. In particular, the proximity of a river near borehole SPM9 and variations in terrain altitude relative to the tunnel were found to correlate with increased leakage rates. These findings suggest that hydrological conditions are a critical driver of water ingress into the tunnel.

Overall, this study highlights the potential of ML models for developing geological maps. The main limitation of this work was the limited number of boreholes. Specifically, the available boreholes were far apart (approximately 300 meters distance between each of the two near boreholes along the LHC), while the vertical resolution of the data was only 1 meter. The complex lithology of the geological profile and imbalanced datasets were other challenges addressed through simplification, integration of the datasets, and the use of the balanced accuracy metric to evaluate the models.

Future work could involve expanding the datasets, creating a geological map encompassing all tunnel sections, and examining other defects, such as cracks and deformations, along the tunnel to identify possible correlations and develop a more realistic model. The ML models can then forecast the tunnel's future failures, taking into account the hidden correlation between the hydrogeological profile and current problems along the tunnel from all aspects.

Moreover, given the importance of layer orders along a borehole, sequence-based models such as Long Short-Term Memories (LSTM) are suggested to be trained using complete boreholes (H. C. Liu et al., 2025).

Chapter 6

References

6 References

Abada, A., Zhong, Y.-M. ;, & Zimmermann, F. ; (2019). He-LHC: The high-Engery Large Hardon Collider - Future Circular Collider Conceptual Design Report Volume 4. *European Physical Journal: Special Topics*, *4*(5), 1109–1382. https://doi.org/10.1140/epjst/e2019-900088-6

Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*(3), 319–335. https://doi.org/10.1007/S10588-018-9266-8/FIGURES/5

Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S. V. S., Ockunzzi, S., Lallo, D., Hu, B., & Rashidi, H. H. (2023). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Seminars in Diagnostic Pathology*, *40*(2), 71–87. https://doi.org/10.1053/J.SEMDP.2023.02.002

Baji, H., Li, C. Q., Scicluna, S., & Dauth, J. (2017). Risk-cost optimised maintenance strategy for tunnel structures. *Tunnelling and Underground Space Technology*, *69*, 72–84. https://doi.org/10.1016/J.TUST.2017.06.008

Brien, D. O., Andrew Osborne, J., Perez-Duenas, E., Cunningham, R., & Li, Z. (2023). Automated crack classification for the CERN underground tunnel infrastructure using deep learning. *Tunnelling and Underground Space Technology*, *131*, 104668. https://doi.org/10.1016/J.TUST.2022.104668

Cao, C., Shi, C., Lei, M., Yang, W., & Liu, J. (2018). Squeezing failure of tunnels: A case study. *Tunnelling and Underground Space Technology*, *77*, 188–203. https://doi.org/10.1016/J.TUST.2018.04.007

Charollais, J., Weidmann, M., Berger, J.-P., Engesser, B., Hotellier, J.-F., Gorin, G., Reichenbacher, B., & Schäfer, P. (2007). *La Molasse du bassin franco-genevois et son substratum*.

Chen, T., & Guestrin, C. (n.d.). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Chen, Z., He, C., Yang, W., Guo, W., Li, Z., & Xu, G. (2020). Impacts of geological conditions on instability causes and mechanical behavior of large-scale tunnels: a case study from the Sichuan–Tibet highway, China. *Bulletin of Engineering Geology and the Environment*, *79*(7), 3667–3688. https://doi.org/10.1007/S10064-020-01796-W/FIGURES/27

Colombo, L., Gattinoni, P., & Scesi, L. (2018). Stochastic modelling of groundwater flow for hazard assessment along the underground infrastructures in Milan (northern Italy). *Tunnelling and Underground Space Technology*, *79*, 110–120. https://doi.org/10.1016/J.TUST.2018.05.007

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, *63*, 22–33. https://doi.org/10.1016/J.CAGEO.2013.10.008

Fathi Salmi, E., Soltani Asadi, Z., Bayati, M., & Sharifzadeh, M. (2019). Assessing the Hydrogeological Conditions Leading to the Corrosion and Deterioration of Pre-cast Segmental Concrete Linings (Case of Zagros Tunnel). *Geotechnical and Geological Engineering*, *37*(5), 3961–3983. https://doi.org/10.1007/S10706-019-00886-1/FIGURES/16

Fern, E. J., Di Murro, V., Soga, K., Li, Z., Scibile, L., & Osborne, J. A. (2018). Geotechnical characterisation of a weak sedimentary rock mass at CERN, Geneva. *Tunnelling and Underground Space Technology*, *77*, 249–260. https://doi.org/10.1016/J.TUST.2018.04.003

Fielder, R. (2000). *EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH ORGANISATION EUROPÉENE POUR LA RECHERCHE NUCLÉAIRE CERN-ST Division VARIOUS TUNNEL EXCAVATION METHODS USED ON THE LHC PROJECT*.

Frenelus, W., Peng, H., & Zhang, J. (2021). Long-term degradation, damage and fracture in deep rock tunnels: A review on the effect of excavation methods. *Frattura Ed Integrita Strutturale*, *15*(58), 128–150. https://doi.org/10.3221/IGF-ESIS.58.10

Gattinoni, P., & Scesi, L. (2017). The groundwater rise in the urban area of Milan (Italy) and its interactions with underground structures and infrastructures. *Tunnelling and Underground Space Technology*, *62*, 103–114. https://doi.org/10.1016/J.TUST.2016.12.001

Ghobadi, M. H., Firuzi, M., & Asghari-Kaljahi, E. (2016). Relationships between geological formations and groundwater chemistry and their effects on the concrete lining of tunnels (case study: Tabriz metro line 2). *Environmental Earth Sciences*, *75*(12), 1–14. https://doi.org/10.1007/S12665-016-5785-0/FIGURES/14

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. https://arxiv.org/abs/2008.05756v1

Guo, J. T., Liu, Y. H., Han, Y. F., & Wang, X. L. (2019). Implicit 3D Geological Modeling Method for Borehole Data Based on Machine Learning. *Dongbei Daxue Xuebao/Journal of Northeastern University*, *40*(9). https://doi.org/10.12068/j.issn.1005-3026.2019.09.021

Han, X., Xia, Y., Ye, F., & Wang, Y. (2020). Ageing models and maintenance strategy for road tunnels. *Structure and Infrastructure Engineering*, *16*(5), 831–846. https://doi.org/10.1080/15732479.2019.1670680

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 1–41. https://doi.org/10.1186/S40537-020-00305-W/FIGURES/4

Harris, J. R., & Grunsky, E. C. (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers & Geosciences*, *80*, 9–25. https://doi.org/10.1016/J.CAGEO.2015.03.013

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, *52*(10), 8217–8237. https://doi.org/10.1002/2016WR019034

Hu, Y., Wang, Z. Z., Guo, X., Kek, H. Y., Ku, T., Goh, S. H., Leung, C. F., Tan, E., & Zhang, Y. (2024). Three-dimensional reconstruction of subsurface stratigraphy using machine learning with neighborhood aggregation. *Engineering Geology*, *337*, 107588. https://doi.org/10.1016/J.ENGGEO.2024.107588

Huang, Y., Qiu, H., Dou, Z., Fu, Z., & Zhou, Z. (2019). An analytical method for estimating leakage from a hydraulic pressure tunnel. *Hydrogeology Journal*, *27*(1), 73–86. https://doi.org/10.1007/S10040-018-1853-8

Jia, R., Lv, Y., Wang, G., Carranza, E. J. M., Chen, Y., Wei, C., & Zhang, Z. (2021). A stacking methodology of machine learning for 3D geological modeling with geological-geophysical datasets, Laochang Sn camp, Gejiu (China). *Computers & Geosciences*, *151*, 104754. https://doi.org/10.1016/J.CAGEO.2021.104754

Khanlari, G., meybodi, R. G., & Mokhtari, E. (2012a). Engineering geological study of the second part of water supply Karaj to Tehran tunnel with emphasis on squeezing problems. *Engineering Geology*, *145–146*, 9–17. https://doi.org/10.1016/J.ENGGEO.2012.06.001

Khanlari, G., meybodi, R. G., & Mokhtari, E. (2012b). Engineering geological study of the second part of water supply Karaj to Tehran tunnel with emphasis on squeezing problems. *Engineering Geology*, *145–146*, 9–17. https://doi.org/10.1016/J.ENGGEO.2012.06.001

Kim, H. S., & Ji, Y. (2022). Three-dimensional geotechnical-layer mapping in Seoul using borehole database and deep neural network-based model. *Engineering Geology*, *297*, 106489. https://doi.org/10.1016/J.ENGGEO.2021.106489

Kumar, T., Seelam, N. K., & Rao, G. S. (2022). Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. *Journal of Applied Geophysics*, *199*, 104605. https://doi.org/10.1016/J.JAPPGEO.2022.104605

Large Colliders:Civil engineering and Siting. (n.d.).

Li, L., Wang, G., Chen, D., Wang, X., Bai, J., Lei, B., Ding, W., Huang, X., & Zhang, Q. (2024). Analysis of tunnel collapse disasters during operation and exploration of disaster damage mechanisms. *IOP Conference Series: Earth and Environmental Science*, *1333*(1), 012049. https://doi.org/10.1088/1755-1315/1333/1/012049

Li, R., Gao, X., & Kong, C. (2020). Study on the Aging Deterioration Curve of Tunnel Lining in Water Rich Area. *IOP Conference Series: Materials Science and Engineering*, *741*(1), 012021. https://doi.org/10.1088/1757-899X/741/1/012021

Lipponen, A., Manninen, S., Niini, H., & Rönkä, E. (2005). Effect of water and geological factors on the long-term stability of fracture zones in the Päijänne Tunnel, Finland: a case study. *International Journal of Rock Mechanics and Mining Sciences*, *42*(1), 3–12. https://doi.org/10.1016/J.IJRMMS.2004.05.006

Liu, H. C., Zhang, N., Yin, Z. Y., & Wang, Y. (2025). Interpretation of subsurface stratigraphic variations from limited boreholes using Dual Bi-LSTM. *Canadian Geotechnical Journal*, *62*. https://doi.org/10.1139/CGJ-2024-0455;REQUESTEDJOURNAL:JOURNAL:CGJ;PAGEGROUP:STRING:PUBLICATION

Liu, J. J., & Liu, J. C. (2022). Permeability Predictions for Tight Sandstone Reservoir Using Explainable Machine Learning and Particle Swarm Optimization. *Geofluids*, *2022*. https://doi.org/10.1155/2022/2263329

Liu, L., Li, T., & Ma, C. (2024). Research on 3D Geological Modeling Method Based on Deep Neural Networks for Drilling Data. *Applied Sciences*, *14*(1). https://doi.org/10.3390/app14010423

Liu, W., Chen, J., Luo, Y., Chen, L., Zhang, L., He, C., Shi, Z., Xu, Z., Zhu, H., Hu, T., & Dong, F. (2022). Long-term stress monitoring and in-service durability evaluation of a large-span tunnel in squeezing rock. *Tunnelling and Underground Space Technology*, *127*, 104611. https://doi.org/10.1016/J.TUST.2022.104611

Lyu, B., & Wang, Y. (2024). Immersive visualization of 3D subsurface ground model developed from sparse boreholes using virtual reality (VR). *Underground Space*, *17*, 188–206. https://doi.org/10.1016/J.UNDSP.2023.11.004

Lyu, B., Wang, Y., & Shi, C. (2024). Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge. *Computers and Geotechnics*, *170*, 106336. https://doi.org/10.1016/J.COMPGEO.2024.106336

Maleki, Z., Farhadian, H., & Nikvar-Hassani, A. (2020). Geological hazards in tunnelling: The example of gelas water conveyance tunnel, iran. *Quarterly Journal of Engineering Geology and Hydrogeology*, *54*(1). https://doi.org/10.1144/QJEGH2019-114

Mathur, A., & Foody, G. M. (2008). Multiclass and Binary SVM Classification: Implications for Training and Classification Users. *IEEE Geoscience and Remote Sensing Letters*, *5*(2), 241–245. https://doi.org/10.1109/LGRS.2008.915597

Murro, V. Di, & College, M. E. (2019). *Long-Term Performance of a concrete-lined Tunnel at CERN*.

Narimani Dehnavi, R., & Sadeghi, M. (2019). Deterioration of weak rocks over time and its effect on designing tunnel support systems. *Bulletin of Engineering Geology and the Environment*, *78*(2), 1045–1056. https://doi.org/10.1007/S10064-017-1154-9/FIGURES/13

Niu, G., He, X., Xu, H., & Dai, S. (2024). Development of Rock Classification Systems: A Comprehensive Review with Emphasis on Artificial Intelligence Techniques. *Eng 2024, Vol. 5, Pages 217-245*, *5*(1), 217–245. https://doi.org/10.3390/ENG5010012

Ouyang, A., Di Murro, V., Cull, M., Cunningham, R., Osborne, J. A., & Li, Z. (2023). Automated pixel-level crack monitoring system for large-scale underground infrastructure – A case study at CERN. *Tunnelling and Underground Space Technology*, *140*, 105310. https://doi.org/10.1016/J.TUST.2023.105310

Pandey, M., Karbasi, M., Jamei, M., Malik, A., & Pu, J. H. (2023). A Comprehensive Experimental and Computational Investigation on Estimation of Scour Depth at Bridge Abutment: Emerging Ensemble Intelligent Systems. *Water Resources Management*, *37*(9), 3745–3767. https://doi.org/10.1007/S11269-023-03525-W

Qin, Y., Shang, C., Li, X., Lai, J., Shi, X., & Liu, T. (2023). Failure mechanism and countermeasures of rainfall-induced collapsed shallow loess tunnels under bad terrain: A case study. *Engineering Failure Analysis*, *152*, 107477. https://doi.org/10.1016/J.ENGFAILANAL.2023.107477

Roy, N., & Sarkar, R. (2017). A Review of Seismic Damage of Mountain Tunnels and Probable Failure Mechanisms. *Geotechnical and Geological*

Engineering, *35*(1), 1–28. https://doi.org/10.1007/S10706-016-0091-X/TABLES/14

Shi, C., & Wang, Y. (2022). Machine learning of three-dimensional subsurface geological model for a reclamation site in Hong Kong. *Bulletin of Engineering Geology and the Environment*, *81*(12), 1–18. https://doi.org/10.1007/S10064-022-03009-Y/FIGURES/12

Shi, C., Wang, Y., & Shi, C. (2020). Nonparametric and data-driven interpolation of subsurface soil stratigraphy from limited data using multiple point statistics. *Https://Doi.Org/10.1139/Cgj-2019-0843*, *58*(2), 261–280. https://doi.org/10.1139/CGJ-2019-0843

Shi, J. X., Li, Z. H., He, Y. Q., Zhang, G. M., & Wei, M. Q. (2020). The Influence of Physical and Chemical Reactions on Water Leakage in a Multiarch Tunnel. *Advances in Civil Engineering*, *2020*(1), 8830739. https://doi.org/10.1155/2020/8830739

Showkati, A., Salari-rad, H., & Hazrati Aghchai, M. (2021). Predicting long-term stability of tunnels considering rock mass weathering and deterioration of primary support. *Tunnelling and Underground Space Technology*, *107*, 103670. https://doi.org/10.1016/J.TUST.2020.103670

Soga, K., Laver, R. G., & Li, Z. (2017). Long-term tunnel behaviour and ground movements after tunnelling in clayey soils. *Underground Space*, *2*(3), 149–167. https://doi.org/10.1016/J.UNDSP.2017.08.001

Su, X., Zhang, C., Zou, Z., Wang, Y., Lai, J., & Liu, T. (2025). Influence of Water Rock Interaction on Stability of Tunnel Engineering. *Pol. J. Environ. Stud*, *34*(1), 535–548. https://doi.org/10.15244/pjoes/186612

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/J.ACI.2018.08.003

Vazquez-Sune, E., Sanchez-Vila, X., Carrera, J., Marizza, M., Arandes, R., & Gutierrez, L. A. (1997). Rising groundwater levels in Barcelona : Evolution and effects on urban structures. *Groundwater in the Urban Environment - Vol I: Problems, Processes and Management*, 267–271.

Verma, S., Jain, K., & Prakash, C. (2020). An Unstructured to Structured Data Conversion using Machine Learning Algorithm in Internet of Things (IoT). *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3563389

Wang, C., Friedman, M., Wu, W., Zhang, D., & Li, Z. (2024). Hydraulic influences on the long-term performance of tunnels: A review. *Transportation Geotechnics*, *48*, 101329. https://doi.org/10.1016/J.TRGEO.2024.101329

Wang, C., Xiao, Z., Di Murro, V., Osborne, J., Friedman, M., & Li, Z. (2024). The impact of thermal–hydraulic variation on tunnel long-term performance: a tale of two tunnels. *Geotechnique*. https://doi.org/10.1680/JGEOT.23.00074

Wilkinson, B. (1985). Rising groundwater levels in London and possible effects on engineering structures. *Rising Groundwater Levels in London and Possible Effects on Engineering Structures*, *18*(4), 145–157.

Yan, W., Shen, P., Zhou, W. H., & Ma, G. (2023). A rigorous random field-based framework for 3D stratigraphic uncertainty modelling. *Engineering Geology*, *323*, 107235. https://doi.org/10.1016/J.ENGGEO.2023.107235

Yang, H. Q., Chu, J., Qi, X., Wu, S., & Chiam, K. (2023). Stochastic simulation of geological cross-sections from boreholes: A random field approach with Markov Chain Monte Carlo method. *Engineering Geology*, *327*, 107356. https://doi.org/10.1016/J.ENGGEO.2023.107356

Yi, S., Hyun, T., Materials, J. K.-C. and B., & 2011, undefined. (n.d.). The effects of hydraulic pressure and crack width on water permeability of penetration crack-induced concrete. *ElsevierST Yi, TY Hyun, JK KimConstruction and Building Materials, 2011•Elsevier*. Retrieved April 9, 2025, from https://www.sciencedirect.com/science/article/pii/S0950061810006380

Yuan, Y., Jiang, X., & Liu, X. (2013). Predictive maintenance of shield tunnels. *Tunnelling and Underground Space Technology*, *38*, 69–86. https://doi.org/10.1016/J.TUST.2013.05.004

Zhan, C., Dai, Z., Yang, Z., Zhang, X., Ma, Z., Thanh, H. V., & Soltanian, M. R. (2023). Subsurface sedimentary structure identification using deep learning: A review. *Earth-Science Reviews*, *239*, 104370. https://doi.org/10.1016/J.EARSCIREV.2023.104370

Zhang, X., Zhang, J., Tian, Y., Li, Z., Zhang, Y., Xu, L., & Wang, S. (2020). Urban Geological 3D Modeling Based on Papery Borehole Log. *ISPRS International Journal of Geo-Information 2020, Vol. 9, Page 389*, *9*(6), 389. https://doi.org/10.3390/IJGI9060389

Zhou, C., Gao, W., Cui, S., Zhong, X., Hu, C., & Chen, X. (2020). Ground Settlement of High-Permeability Sand Layer Induced by Shield Tunneling: A Case Study under the Guidance of DBN. *Geofluids*, *2020*(1), 6617468. https://doi.org/10.1155/2020/6617468

Zhu, H., Yan, J., & Liang, W. (2019). Challenges and Development Prospects of Ultra-Long and Ultra-Deep Mountain Tunnels. *Engineering*, *5*(3), 384–392. https://doi.org/10.1016/J.ENG.2019.04.009

Zhu, L., Zhang, C., Li, M., Pan, X., & Sun, J. (2012). Building 3D solid models of sedimentary stratigraphic systems from borehole data: An automatic method and case studies. *Engineering Geology*, *127*, 1–13. https://doi.org/10.1016/J.ENGGEO.2011.12.001