



School of Science and Technology

Engineering Project Management

Postgraduate Dissertation

“Analysis and Assessment of Geotechnical Risks Using Machine
Learning Techniques”

Konstantinos Symeonidis

Supervisor: Denise-Penelope Kontoni

Patras, Greece, June 2026

Theses / Dissertations remain the intellectual property of students (“authors/creators”), but in the context of open access policy they grant to the HOU a non-exclusive license to use the right of reproduction, customisation, public lending, presentation to an audience and digital dissemination thereof internationally, in electronic form and by any means for teaching and research purposes, for no fee and throughout the duration of intellectual property rights. Free access to the full text for studying and reading does not in any way mean that the author/creator shall allocate his/her intellectual property rights, nor shall he/she allow the reproduction, republication, copy, storage, sale, commercial use, transmission, distribution, publication, execution, downloading, uploading, translating, modifying in any way, of any part or summary of the dissertation, without the explicit prior written consent of the author/creator. Creators retain all their moral and property rights.



“Analysis and Assessment of Geotechnical Risks Using Machine Learning Techniques”

Konstantinos Symeonidis

Supervising Committee

Supervisor:

Denise-Penelope Kontoni

Associate Professor, Department of Civil
Engineering, University of the
Peloponnese

Co-Supervisor:

Niki Beskou

Collaborating Educational Staff
Hellenic Open University

Patras, Greece, June 2026

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Associate Professor Dr. Denise-Penelope Kontoni, for her valuable guidance, support, and encouragement throughout the preparation of this postgraduate dissertation.

Abstract

This thesis investigates how machine learning can support the preliminary assessment of strength-related geotechnical risk indicators through the systematic analysis of rock laboratory data from Doha, Qatar. The study is motivated by the need to extract interpretable engineering value from heterogeneous geotechnical datasets while recognising that model outputs cannot replace geological logging, laboratory testing, source-record review, or professional judgement. In this thesis, geotechnical risk is therefore treated in a screening-level sense, rather than as a full quantitative risk assessment based on probability, consequence, exposure, or formal risk ranking.

Doha's shallow subsurface is characterised by variable carbonate and shale-bearing rock units, including Simsima Limestone, Midra Shale, Dukhan Limestone, and the Rus Formation, where strength and stiffness may be affected by lithology, depth, weathering, dolomitisation, karstification, discontinuities, moisture, and groundwater-related conditions. The thesis develops a staged and reproducible workflow covering exploratory data inventory, category-specific cleaning, parameter summary and outlier review, feature selection, model development, performance evaluation, error diagnosis, and engineering interpretation. The available rock laboratory dataset contained 5,638 records and 67 columns. Unconfined compressive strength (UCS) was selected as the main modelling target because of its engineering relevance and relatively suitable data availability. The final UCS model used bulk density, depth, lithology, and moisture content as predictors, while a supplementary point load index Is_{50} workflow used depth and lithology because richer paired data were not sufficiently available.

Ridge Regression and Random Forest Regression were compared. For UCS, Random Forest Regression performed better on the held-out test set, with moderate predictive capability, but the result was accompanied by possible overfitting and persistent underprediction of high-UCS samples. The Is_{50} workflow showed weaker to modest performance, reflecting the sparse predictor set. Overall, the findings indicate that machine learning can help identify broad strength trends, data gaps, influential variables, and unusual cases within cleaned Doha laboratory data. However, the models remain screening-level decision-support tools and require validation against geological context, source records, and engineering judgement before practical use.

Keywords

machine learning; geotechnical risk assessment; unconfined compressive strength; lithology; Doha, Qatar; carbonate rock

Περίληψη

“Ανάλυση και Αξιολόγηση Γεωτεχνικών Κινδύνων με Τεχνικές Μηχανικής Μάθησης”

Η παρούσα διατριβή εξετάζει πώς η μηχανική μάθηση μπορεί να υποστηρίξει την προκαταρκτική αξιολόγηση δεικτών γεωτεχνικού κινδύνου που σχετίζονται με την αντοχή, μέσω της συστηματικής ανάλυσης εργαστηριακών δεδομένων πετρωμάτων από τη Ντόχα του Κατάρ. Η μελέτη βασίζεται στην ανάγκη εξαγωγής ερμηνεύσιμης γεωτεχνικής αξίας από ετερογενή σύνολα γεωτεχνικών δεδομένων, αναγνωρίζοντας παράλληλα ότι τα αποτελέσματα των μοντέλων δεν μπορούν να αντικαταστήσουν τη γεωλογική καταγραφή, τις εργαστηριακές δοκιμές, την ανασκόπηση των αρχικών τεχνικών αρχείων ή την επαγγελματική γεωτεχνική κρίση. Στο πλαίσιο της διατριβής, ο όρος γεωτεχνικός κίνδυνος χρησιμοποιείται σε επίπεδο προκαταρκτικού ελέγχου και όχι ως πλήρης ποσοτική αξιολόγηση κινδύνου με βάση την πιθανότητα, τη συνέπεια, την έκθεση ή την τυπική κατάταξη κινδύνου.

Το ρηχό υπέδαφος της Ντόχα χαρακτηρίζεται από μεταβλητές ανθρακικές και σχιστολιθικές ενότητες, όπως ο Simsima Limestone, ο Midra Shale, ο Dukhan Limestone και ο Rus Formation, όπου η αντοχή και η δυσκαμψία μπορεί να επηρεάζονται από τη λιθολογία, το βάθος, την αποσάθρωση, τη δολομιτώση, την καρστικοποίηση, τις ασυνέχειες, την υγρασία και τις συνθήκες που σχετίζονται με το υπόγειο νερό. Η διατριβή αναπτύσσει μια σταδιακή και αναπαραγωγική ροή εργασίας, η οποία περιλαμβάνει διερευνητική απογραφή δεδομένων, καθαρισμό ανά κατηγορία παραμέτρων, στατιστική περίληψη παραμέτρων και έλεγχο ακραίων τιμών, επιλογή χαρακτηριστικών, ανάπτυξη μοντέλων, αξιολόγηση απόδοσης, διαγνωστική ανάλυση σφαλμάτων και τελική γεωτεχνική ερμηνεία. Το διαθέσιμο σύνολο εργαστηριακών δεδομένων πετρωμάτων περιλάμβανε 5.638 εγγραφές και 67 στήλες. Η μονοαξονική θλιπτική αντοχή (UCS) επιλέχθηκε ως ο κύριος στόχος μοντελοποίησης λόγω της γεωτεχνικής της σημασίας και της σχετικά κατάλληλης διαθεσιμότητας δεδομένων. Το τελικό μοντέλο UCS χρησιμοποίησε ως μεταβλητές πρόβλεψης τη φαινόμενη πυκνότητα, το βάθος, τη λιθολογία και την περιεκτικότητα σε υγρασία, ενώ η συμπληρωματική ροή εργασίας για τον δείκτη σημειακής φόρτισης Is50 χρησιμοποίησε το βάθος και τη λιθολογία, επειδή δεν ήταν επαρκώς διαθέσιμα πλουσιότερα συζευγμένα δεδομένα.

Συγκρίθηκαν τα μοντέλα Ridge Regression και Random Forest Regression. Για τον στόχο UCS, το Random Forest Regression παρουσίασε καλύτερη απόδοση στο ανεξάρτητο σύνολο ελέγχου, με μέτρια προβλεπτική ικανότητα, αλλά το αποτέλεσμα συνοδεύτηκε από ενδείξεις πιθανής υπερπροσαρμογής και επίμονης υποεκτίμησης των δειγμάτων με υψηλές τιμές UCS.

Η ροή εργασίας για τον Is50 παρουσίασε ασθενέστερη έως μέτρια απόδοση, αντανακλώντας το περιορισμένο σύνολο μεταβλητών πρόβλεψης. Συνολικά, τα ευρήματα δείχνουν ότι η μηχανική μάθηση μπορεί να συμβάλει στην αναγνώριση γενικών τάσεων αντοχής, κενών δεδομένων, σημαντικών μεταβλητών και ασυνήθιστων περιπτώσεων σε καθαρισμένα εργαστηριακά δεδομένα της Ντόχα. Ωστόσο, τα μοντέλα παραμένουν εργαλεία υποστήριξης αποφάσεων σε επίπεδο προκαταρκτικού ελέγχου και απαιτούν επικύρωση με βάση το γεωλογικό πλαίσιο, τα αρχικά τεχνικά αρχεία και την επαγγελματική γεωτεχνική κρίση πριν από πρακτική εφαρμογή.

Λέξεις – Κλειδιά

Μηχανική μάθηση· αξιολόγηση γεωτεχνικού κινδύνου· μονοαξονική θλιπτική αντοχή· λιθολογία· Ντόχα, Κατάρ· ανθρακικά πετρώματα

Table of Contents

Abstract	5
Περίληψη	7
Table of Contents	9
List of Figures	12
List of Tables	13
List of Abbreviations & Acronyms	14
1. Introduction	20
1.1 Background	20
1.2 Problem Statement	22
1.3 Aim of the Thesis	23
1.4 Thesis Structure	24
2. Literature Review	25
2.1 Geotechnical Engineering and Risk Assessment	25
2.2 Machine Learning in Geotechnical Engineering	26
2.3 Review of ML Applications in Geotechnical Engineering	27
2.4 Types of geotechnical data used in ML	29
2.5 Common machine learning methods	32
2.6 Applications in classification and prediction	34
2.6.1 Tunnelling and settlement prediction	34
2.6.2 Slope stability, landslide assessment and subsidence	37
2.6.3 Foundations, bearing capacity and pile capacity	39
2.6.4 Liquefaction and seismic-related applications	39
2.6.5 Soil and rock property prediction	40
2.6.6 Classification and site characterisation	42
2.7 Summary of review findings	43
2.8 Model evaluation and interpretability	47
2.9 Limitations and research gap	48
3. Theoretical Background of Machine Learning	51
3.1 Basic Concepts of Machine Learning	51
3.2 Geotechnical engineering data	55
3.3 Geotechnical parameters	56

3.4	Model training and evaluation	57
3.5	Relevance to geotechnical engineering	59
3.6	Classification of machine learning methods	61
3.6.1	Supervised learning	61
3.6.2	Unsupervised learning	61
3.6.3	Deep learning	62
3.7	Commonly used machine learning methods in geotechnical engineering	63
3.7.1	Linear regression	63
3.7.2	Decision trees	64
3.7.3	Random forests	64
3.7.4	Support vector machines	65
3.7.5	Artificial neural networks	66
3.7.6	Gradient boosting methods	66
3.7.7	Clustering methods	66
3.8	Summary	67
4.	Data and Methodology	68
4.1	Exploratory data inventory	68
4.2	Category-specific cleaning of data	69
4.3	Parameter summary and data suitability assessment	70
4.4	Feature selection and target definition	71
4.5	Machine learning model development and training	72
4.6	Model performance evaluation and comparison	73
4.7	Error analysis and diagnostic interpretation	74
4.8	Engineering interpretation and workflow limitations	75
4.9	Reproducibility and data storage	76
5.	Machine Learning Workflow	78
5.1	Workflow overview	80
5.2	Stage 1: Exploratory data inventory	81
5.3	Stage 2: Category-specific cleaning	83
5.4	Stage 3: Parameter summary	84
5.5	Stage 4: Feature selection	90
5.6	Stage 5: Machine learning model development and training	91

5.7	Stage 6: Model performance evaluation and comparison	94
5.8	Stage 7: Error analysis and diagnostic interpretation	96
5.9	Stage 8: Engineering interpretation and workflow limitations	107
5.10	Chapter summary	112
6.	Results	113
6.1	Exploratory data analysis results	113
6.2	Feature selection results	114
6.3	Model 1 Ridge Regression results	114
6.4	Model 2 Random Forest Regression results	115
6.5	Comparison of model performance and diagnostic interpretation	115
6.6	Summary	116
7.	Discussion	118
7.1	Interpretation of results	118
7.2	Comparison with traditional geotechnical interpretation	119
7.3	Practical relevance	120
7.4	Limitations	120
7.5	Engineering judgement	121
7.6	Summary	122
8.	Conclusions and Recommendations	123
8.1	Conclusions	123
8.2	Main limitations	124
8.3	Recommendations for future work	125
8.4	Final statement	125
	References	127

List of Figures

Figure 1.1 Bar graph of search results for the number of SCOPUS publications per year that include the key terms "Machine Learning" & "Civil Engineering". (Source: SCOPUS, Literature search in December 2025).....	21
Figure 1.2 Bar graph of search results for the number of SCOPUS publications per year that include the terms "Machine Learning" & "Geotechnical Engineering". (Source: SCOPUS, Literature search in December 2025.).....	22
Figure 5.1 Bulk density versus depth grouped by lithology.	86
Figure 5.2 Dry density versus depth grouped by lithology.....	86
Figure 5.3 Moisture content versus depth grouped by lithology.	87
Figure 5.4 Porosity versus depth grouped by lithology.	87
Figure 5.5 Point load index Is_{50} versus depth grouped by lithology.	88
Figure 5.6 Unconfined compressive strength versus depth grouped by lithology.....	88
Figure 5.7 Young's modulus versus depth grouped by lithology.....	89
Figure 5.8 Poisson's ratio versus depth grouped by lithology.....	89
Figure 5.9 Measured versus predicted UCS for WD1.	98
Figure 5.10 Residual distribution for WD1.	98
Figure 5.11 Residuals versus predicted UCS for WD1.	99
Figure 5.12 Largest absolute errors for WD1.	99
Figure 5.13 Mean absolute error by lithology for WD1.	100
Figure 5.14 Measured versus predicted Is_{50} for WD2.	103
Figure 5.15 Residuals versus predicted Is_{50} for WD2.	103
Figure 5.16 Residuals versus depth for WD2.	104
Figure 5.17 Residual distribution for WD2.	104
Figure 5.18 Ridge Regression residuals by lithology for WD2.....	105

List of Tables

Table 2.1 Summary of machine learning applications in geotechnical engineering	45
Table 5.1 Summarises the main lithological units considered in the Chapter 5 dataset interpretation.....	79
Table 5.2 Summarises the dataset, source data, prepared dataset, and target parameter used in each workflow application.....	83
Table 5.3 Summarises the input dataset, and output dataset for Stage 2 in each workflow application.....	84
Table 5.4 Summarises the input dataset, output dataset, summary outputs, and target parameter used for Stage 3 in each workflow application.....	85
Table 5.5 Summarises the input dataset, output summaries, target parameter, and selected features used in Stage 4.....	91
Table 5.6 Summarises the input dataset and selected features, model input records, target parameter, and Stage 5 output for WD1 and WD2.	93
Table 5.7 Summarises in table sections (a) and (b) the Stage 6 performance evaluation metrics, outputs, workflow script, and target parameter used for each workflow.....	95
Table 5.8 Summarises the input diagnostic files, models, diagnostic outputs, and target parameter used in Stage 7.....	97
Table 5.9 Supplementary WD1 test-set error metrics by lithology.	101
Table 5.10 Supplementary WD2 test-set error metrics by lithology.	105
Table 5.11 - Summarises the Stage 8 engineering interpretation and workflow limitations for WD1 and WD2.....	110

List of Abbreviations & Acronyms

Abbreviation / Acronym	Full Term	Explanation
AdaBoost	Adaptive Boosting	AdaBoost is an ensemble method that improves prediction by training several weak models sequentially, with later models giving more attention to earlier errors.
ADTree	Alternating Decision Tree	ADTree is a tree-based classification method used in the reviewed geohazard susceptibility literature.
AI	Artificial Intelligence	Artificial intelligence refers to computational methods that perform tasks associated with reasoning, learning, perception, or decision-making.
ANN	Artificial Neural Network	An artificial neural network is a machine-learning approach made up of linked processing elements that can capture complex, nonlinear patterns in geotechnical datasets.
AOA	Archimedes Optimisation Algorithm	AOA is an optimisation algorithm used in the reviewed literature to tune XGBoost for rock strength prediction.
AUC	Area Under the Curve	AUC is a classification evaluation measure commonly associated with receiver operating characteristic analysis.
AutoML	Automated Machine Learning	AutoML refers to automated procedures for selecting, tuning, and comparing machine learning models.
BLR	Bayesian Logistic Regression	Bayesian logistic regression is a probabilistic classification method used in the reviewed susceptibility mapping literature.
BPLDA	Bundling with Penalised Linear Discriminant Analysis	BPLDA is an ensemble classification-tree method used in the reviewed landslide susceptibility modelling literature.
BPNN	Back-Propagation Neural Network	A back-propagation neural network is trained by updating connection weights through error feedback so that the difference between those values predicted and those observed is reduced.
CART	Classification and Regression Tree	CART is a decision-tree method used for classification or regression tasks.

CatBoost	Categorical Boosting	CatBoost is a gradient-boosting algorithm that is particularly effective for datasets containing categorical variables.
CBR	California Bearing Ratio	California Bearing Ratio is a pavement and embankment engineering index used to describe subgrade or material strength.
CEEMDAN-LSTM	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-Long Short-Term Memory	CEEMDAN-LSTM combines signal decomposition with a long short-term memory network and has been used for forecasting tunnel settlement behaviour.
CHAID	Chi-square Automatic Interaction Detection	CHAID is a decision-tree method that uses chi-square tests to identify appropriate classification splits.
CNN	Convolutional Neural Network	A convolutional neural network is a deep-learning architecture often applied to image, spatial, or grid-format data.
CPT	Cone Penetration Test	The cone penetration test is an in situ geotechnical test used to obtain cone resistance and related indicators of soil behaviour.
CSR	Cyclic Stress Ratio	Cyclic stress ratio is a seismic loading parameter used in liquefaction assessment.
DL	Deep Learning	Deep learning is a branch of machine learning that uses neural networks with several layers to learn complex data patterns.
DT	Decision Tree	A decision tree predicts an outcome by repeatedly splitting the dataset into groups based on selected input variables.
ELM	Extreme Learning Machine	An extreme learning machine is a feedforward neural network approach used in some forecasting and prediction studies.
F1-score	F1 score	The F1-score summarises classification performance by combining precision and recall into one balanced measure.
FNN	Feedforward Neural Network	A feedforward neural network transfers information from the input layer to the output layer without recurrent feedback connections.

GAM	Generalised Additive Model	A generalised additive model is a statistical modelling approach that represents nonlinear behaviour through additive smooth functions.
GAN	Generative Adversarial Network	A generative adversarial network is a deep-learning framework that can generate synthetic data or support image-based modelling tasks.
GANFIS	Genetic Algorithm - Adaptive Network-Based Fuzzy Inference System	GANFIS combines a genetic algorithm with an adaptive neuro-fuzzy inference system and has been applied to soft-soil strength prediction.
GBM	Gradient Boosting Machine	A gradient boosting machine is an ensemble method that improves prediction by adding weak learners sequentially to reduce model error.
GBRT	Gradient Boosting Regression Tree	A gradient boosting regression tree is a boosting model that uses regression trees as weak learners.
GIS	Geographic Information System	A geographic information system manages and analyses spatial data used in hazard and susceptibility assessment.
GLM	Generalised Linear Model	A generalised linear model extends linear regression to response variables with different statistical distributions.
GSI	Geological Strength Index	Geological Strength Index is a rock mass classification index used to describe rock mass quality.
KNN	k-Nearest Neighbours	K-nearest neighbours predict a new case by comparing it with the most similar examples in the training dataset.
LightGBM	Light Gradient Boosting Machine	LightGBM is an efficient gradient boosting framework commonly applied to structured or tabular datasets.
LMT	Logistic Model Tree	A logistic model tree combines a decision-tree structure with logistic regression models fitted at the terminal nodes.
LR	Logistic Regression	Logistic regression is a statistical classification method used in applications such as liquefaction and susceptibility assessment.
LSTM	Long Short-Term Memory	Long short-term memory networks are recurrent neural networks developed to learn patterns in sequential or time-dependent data.

MAE	Mean Error	Absolute	Mean absolute error expresses the average magnitude of prediction errors without accounting for whether the errors are positive or negative.
MAPE	Mean Percentage Error	Absolute	Mean absolute percentage error presents prediction error as an average percentage of the observed value.
MARE	Mean Relative Error	Absolute	Mean absolute relative error measures prediction error relative to observed or reference values.
MARS	Multivariate Adaptive Regression Splines		MARS is a regression technique that represents nonlinear relationships using piecewise basis functions.
ML	Machine Learning		Machine learning describes computational methods that learn patterns from data to support prediction, classification, or decision-making.
MLP	Multilayer Perceptron		A multi-layer perceptron is a feed-forward neural network containing one or more hidden layers.
MORF	Multi-Objective Random Forest		MORF is a random forest approach used for multi-output formation percentage prediction in shield tunnelling.
MSE	Mean Squared Error		Mean squared error is a regression metric calculated from the average of squared differences between observed and predicted values.
NATM	New Tunnelling Method	Austrian	NATM is a tunnelling method referenced in the reviewed tunnel-face rock mass quality literature.
NGBoost	Natural Boosting	Gradient	NGBoost is a probabilistic boosting method used to estimate predictive distributions as well as point predictions.
NN	Neural Network		A neural network is a machine-learning model inspired by interconnected layers of computational nodes.
NSE	Nash-Sutcliffe Efficiency		Nash-Sutcliffe efficiency is a model performance metric used to assess predictive skill relative to observed variation.
N-XGBoost	Natural Boosting XGBoost Learners	Gradient with Base	N-XGBoost is a probabilistic hybrid model that uses XGBoost as the base learner within an NGBoost framework for rockhead prediction.

PANFIS	Particle Swarm Optimisation - Adaptive Network- Based Fuzzy Inference System	PANFIS combines particle swarm optimisation with an adaptive neuro-fuzzy inference system and has been used for soft-soil strength prediction.
PSO-SVR	Particle Swarm Optimised Support Vector Regression	PSO-SVR uses particle swarm optimisation to tune support vector regression for improved regression performance.
Q-system	Rock Tunnelling Quality Index System	The Q-system is a rock-mass classification method used to evaluate rock quality for tunnelling and support design.
R ²	Coefficient of Determination	The coefficient of determination shows the proportion of variation in the target variable explained by the regression model.
RF	Random Forest	Random forest combines multiple decision trees and can be used for classification, regression, and assessment of variable importance.
RMI	Rock Mass Index	Rock Mass Index is a rock-mass classification system used to describe the quality of rock masses.
RMR	Rock Mass Rating	Rock Mass Rating is a geotechnical rock-mass classification method used in engineering and tunnelling assessment.
RMSE	Root Mean Squared Error	Root mean squared error measures prediction error while placing greater emphasis on larger deviations than MAE.
ROC	Receiver Operating Characteristic	Receiver operating characteristic analysis evaluates classification performance across decision thresholds.
RPM	Revolutions Per Minute	Revolutions per minute describes rotational speed, such as a TBM cutterhead or related tunnelling parameter.
RQD	Rock Quality Designation	Rock Quality Designation is an index of rock core quality based on the proportion of sound core pieces.
SGD	Stochastic Gradient Descent	Stochastic gradient descent is an optimisation method that updates model parameters using small batches or individual samples.
SHAP	SHapley Additive exPlanations	SHAP is a model-interpretation method that estimates the contribution of each feature to a prediction.

SMOBN	Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise	SMOBN is a resampling method used to address imbalanced regression targets, including rare extreme values.
SPT	Standard Penetration Test	The standard penetration test is an in situ geotechnical test that provides blow-count data for soil interpretation.
SVM	Support Vector Machine	A support vector machine is a supervised learning method suitable for classification and regression problems.
SVR	Support Vector Regression	Support vector regression adapts the support vector machine framework for continuous prediction problems.
TBM	Tunnel Boring Machine	A tunnel boring machine is mechanised tunnelling equipment whose operational data can support ground condition and performance prediction.
TPE-GBRT	Tree-structured Parzen Estimator - Gradient Boosting Regression Tree	TPE-GBRT combines tree-structured Parzen estimator hyperparameter optimisation with gradient-boosted regression trees.
TPE-RF	Tree-structured Parzen Estimator - Random Forest	TPE-RF combines tree-structured Parzen estimator hyperparameter tuning with random forest modelling.
UCS	Uniaxial Compressive Strength	Uniaxial compressive strength is a rock-strength parameter commonly used in rock mechanics and tunnelling assessment.
VAF	Variance Accounted For	Variance accounted for is a model performance metric used to describe how much observed variation is captured by predictions.
WOE	Weights of Evidence	Weights of evidence is a statistical method used in susceptibility and spatial prediction studies.
XGBoost	Extreme Gradient Boosting	XGBoost is a gradient boosting algorithm widely used for structured datasets and predictive modelling tasks.

1. Introduction

1.1 Background

Geotechnical engineering is central to the planning, design, construction, operation, and maintenance of civil infrastructure. Foundations, slopes, retaining structures, embankments, excavations, tunnels, transport corridors, and other ground-related assets depend on the interaction between engineered works and naturally variable soil and rock masses. Because ground conditions cannot be observed completely before construction or operation, geotechnical risk assessment is an important part of infrastructure decision-making. Transport infrastructure standards, for example, treat geotechnical risk assessment and hazard management as processes that support the safety of operations, infrastructure, adjacent people, and nearby property (Transport for NSW, 2014).

Geotechnical interpretation is therefore concerned not only with calculating design values, but also with understanding uncertainty. Site investigation data, field observations, laboratory test results, geological interpretation, empirical correlations, and monitoring records each provide partial evidence about the ground. These sources may be sparse, variable, biased by investigation location, affected by measurement error, or dependent on engineering judgement. Phoon and Kulhawy (1999) describe geotechnical variability as arising from various sources of uncertainty, such as the inherent variability, or the sources of measurement error, and even the transformation uncertainty. This uncertainty is the main reason why geotechnical engineering has traditionally relied on conservative design assumptions, factors of safety, observational judgement, and experience-based interpretation.

Traditional geotechnical methods remain essential, but they have limitations. Empirical correlations may be derived from datasets, soil types, test procedures, or geological settings that do not fully match the project being assessed. Deterministic analyses may present a single factor of safety, settlement estimate, or capacity value even when the input parameters are uncertain. Duncan (2000) notes that conventional geotechnical practice often applies the same factor of safety to a given class of problem without explicitly reflecting the degree of uncertainty or the consequences of unsatisfactory performance. Duncan (2000) further argues that reliability measures should supplement, rather than replace, factors of safety, because deterministic and probabilistic measures each provide useful but incomplete information.

Machine learning (ML) provides one possible way to support geotechnical interpretation where suitable data are available. In this thesis, ML is treated as an engineering decision-support method that can assist with data exploration, parameter prediction, pattern recognition, and preliminary screening. It is not treated as a standalone replacement for laboratory testing, geological interpretation, conventional geotechnical risk assessment, or professional judgement. This distinction is important because a statistical model can only learn from the data, assumptions, and validation procedures used to develop it.

In recent times, interest in using ML methods within engineering disciplines has increased substantially. This is evident by the rapid annual growth of the published peer-reviewed journal articles and conference papers in an extensive range of Science and Engineering fields. Civil Engineering and Geotechnical Engineering have shown a significant rise in ML related research as is evident in Figures 1.1 and 1.2, summarising the number of search results extracted from SCOPUS, the scientific abstract and citation database by Elsevier.

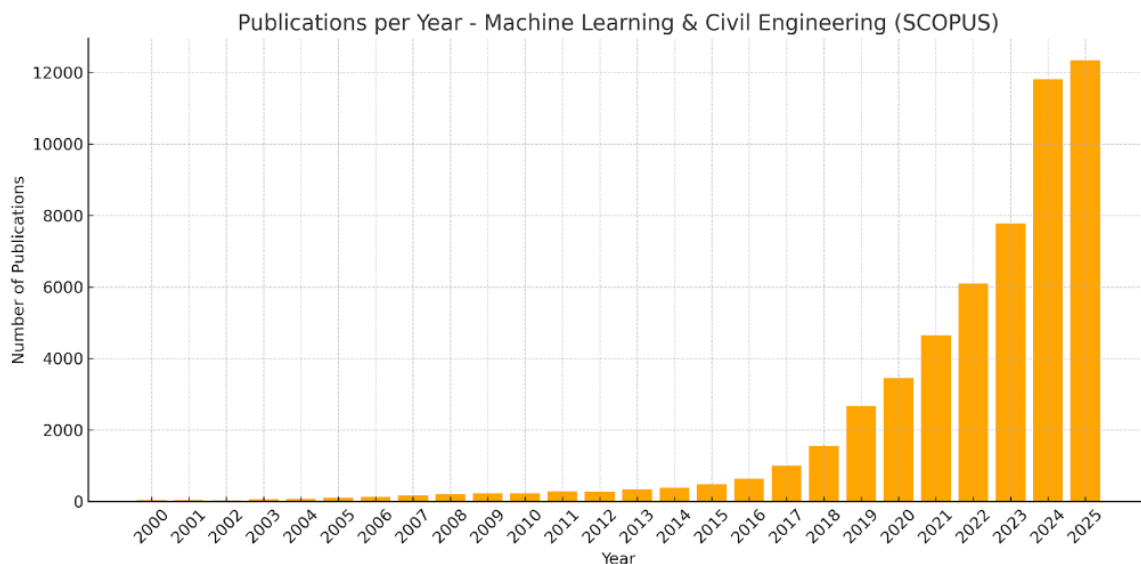


Figure 1.1 Bar graph of search results for the number of SCOPUS publications per year that include the key terms "Machine Learning" & "Civil Engineering". (Source: SCOPUS, Literature search in December 2025)

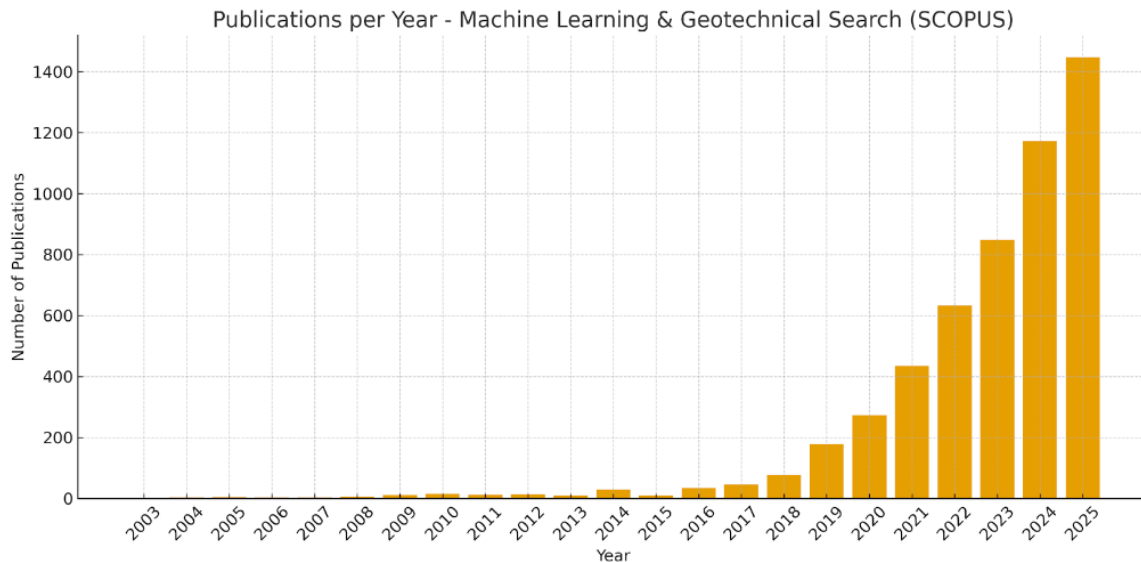


Figure 1.2 Bar graph of search results for the number of SCOPUS publications per year that include the terms "Machine Learning" & "Geotechnical Engineering". (Source: SCOPUS, Literature search in December 2025.)

1.2 Problem Statement

This Thesis attempts to develop a reproducible ML-supported workflow using the available laboratory dataset, from recent engineering projects in Doha, Qatar. In this thesis, the term geotechnical risk is used in a screening-level sense. The available dataset does not support a full quantitative geotechnical risk assessment involving formal probability, consequence, exposure, or risk-ranking calculations. Instead, the study investigates whether ML can support the preliminary assessment of strength-related geotechnical risk indicators, particularly through the prediction and interpretation of UCS using lithology, depth-related variables, and laboratory-derived parameters. The work should therefore be interpreted as a data-driven screening and interpretation workflow, not as a complete geotechnical risk assessment framework.

The immediate problem is therefore more limited and more practical than direct risk prediction. The available data need to be compiled, cleaned, checked, and converted into modelling-ready forms before they can support defensible analysis. Strength-related variables such as UCS, density, moisture content, porosity, lithology, and depth can then be explored to identify whether ML models can predict missing or untested strength-related values, identify

multivariate relationships, rank influential parameters, and flag anomalous samples or borehole-unit groups for review.

A transparent rule-based Python tool also has value in this workflow. Such a tool can compare laboratory results against published Doha rock parameter ranges and assign simple screening categories. These categories should be interpreted as preliminary hazard indicators or weak-ground susceptibility flags, not as final risk ratings. Their main benefit is traceability: the rule, threshold, source range, input value, and assigned category can be reviewed by an engineer. Machine learning can add complementary value when the relationships between variables are multivariate, non-linear, or difficult to assess consistently by manual review alone. In this thesis, ML is used to support preliminary geotechnical risk screening by identifying data patterns, predicting strength-related parameters, interpreting feature importance, and highlighting samples or groups that may require closer engineering review. The outputs are therefore intended to support engineering judgement rather than replace it. This complementary position is consistent with Duncan's (2000) argument that additional analytical tools can add value when used alongside, rather than in place of, established geotechnical measures.

1.3 Aim of the Thesis

This thesis aims to investigate how machine learning can support preliminary geotechnical risk screening by analysing laboratory test data, predicting strength-related parameters, and identifying weak-ground indicators within a Doha rock engineering context. The objectives of the thesis are to:

- compile and clean geotechnical laboratory data extracted from Doha factual reports;
- undertake exploratory analysis of available laboratory parameters and data completeness;
- select suitable features for strength-related prediction, with UCS as the first modelling target;
- develop and compare initial ML models for UCS prediction;
- interpret feature importance, model performance, residuals, and engineering plausibility;

- compare the workflow with published Doha rock parameter ranges through transparent rule-based screening; and
- identify limitations and future data requirements for more complete geotechnical risk screening.

The scope of the thesis is intentionally cautious. The current dataset supports preliminary screening of strength-related indicators rather than direct prediction of full geotechnical risk. Model performance must therefore be interpreted in both statistical and engineering terms, and all outputs must be checked against geological plausibility, source-data quality, and the intended decision context. The workflow is intended to support engineering judgement rather than replace formal risk assessment.

1.4 Thesis Structure

The thesis is structured to first introduce the background and motivation for the research, followed by a review of relevant literature on machine learning applications in geotechnical engineering. The methodology then describes the data preparation, exploratory analysis, model development, and evaluation process. The machine learning workflow presents the completed UCS feature selection, Ridge Regression baseline, and Random Forest Regression comparison. The results and discussion interpret the model outputs in terms of both statistical performance and geotechnical engineering relevance. Finally, conclusions and recommendations are provided for future work, with emphasis on data requirements for more complete preliminary risk screening.

2. Literature Review

2.1 Geotechnical Engineering and Risk Assessment

Geotechnical engineering is inherently characterised by uncertainty arising from spatial variability of ground conditions, limited site investigation data, and simplified modelling assumptions. Baecher and Christian (2003) pointed out that, as opposed to other engineering practitioners like structural and mechanical engineers, geotechnical engineers and engineering geologists deal with the geometries and materials that nature provides, and with uncertainties that are mostly deductive in their nature, using limited observations, geological modelling and statistical methods to infer and best approximate the ground behaviour. The authors assigned the uncertainties found in geotechnical engineering into three classes,

- (i) the “natural” or “aleatory uncertainty” that is associated with either temporal and/or spatial variations and processes,
- (ii) the “knowledge” or “epistemic uncertainty” that is attributed to the lack of data, or lack of information on events and processes, or the limited understanding of physical laws that all combined can limit the ability to model, and
- (iii) the “operational and decision uncertainties” that are associated with construction and human factors.

Phoon and Kulhawy (1999) highlighted that geotechnical variability is much more complicated and arises from several sources of uncertainty including the inherent variability of the geological materials due to the interactions of various geological processes that act upon from their formation to their present in-situ state, the measurement errors due to either equipment, or procedure, but also operator and arbitrary effects, and the transformation uncertainty which the authors described as the uncertainty introduced when measurements are changed to design parameters based on either empirical or other correlation models. The *Observational Method* described in the classic paper by Peck, R. B. (1969) was the commonly applied method in dealing with the geological uncertainties.

2.2 Machine Learning in Geotechnical Engineering

Machine learning (ML) has become increasingly visible in the geotechnical engineering practice because many geotechnical problems involve nonlinear material behaviour, spatial variability, incomplete site information and substantial uncertainty. Conventional geotechnical interpretation remains grounded in site investigation, laboratory testing, field observations, empirical correlations, numerical analysis, geological understanding and engineering judgement. ML does not remove the need for those foundations; rather, it offers additional tools for extracting patterns from available data, comparing alternative predictors and supporting faster preliminary assessment where adequate data exist (Liu et al., 2024; Marcher et al., 2020; Yaghoubi et al., 2024).

The literature reviewed in this chapter shows that ML has been applied across a wide range of geotechnical tasks, including:

- soil and rock classifications,
- site characterisation,
- liquefaction assessment,
- slope stability,
- landslide and subsidence susceptibility,
- settlement prediction,
- tunnelling performance,
- foundation capacity, and
- prediction of strength and deformation parameters.

These applications are attractive because many geotechnical relationships are difficult to capture using simple linear correlations. However, the same studies also indicate that model reliability is strongly influenced by the size, quality, and representativeness of the training dataset (Ayawah et al., 2022; Goetz et al., 2015; Rauter & Tschuchnigg, 2021; Song et al., 2024).

It is important to note that geotechnical datasets differ from many datasets in other engineering fields because they are often sparse, expensive to obtain, spatially clustered, locally biased and affected by different testing procedures, geological histories and construction methods. A dataset may contain thousands of records from one project but still represent only a narrow

geological domain. On the other hand, a regional or published case-history database may contain broader experience but inconsistent input definitions. These characteristics mean that validation, interpretability and judgement about the training domain are especially important in geotechnical ML (Goetz et al., 2015; Marcher et al., 2020; Zhao & Wang, 2020).

This chapter reviews recent citations on the application of ML in geotechnical engineering with emphasis on the types of data used, common methods, application areas, model evaluation and interpretability. The review is organised thematically rather than as a bibliography listing. A central premise is that statistical accuracy alone is not sufficient, model outputs must also be physically plausible, consistent with the geological setting and used within an appropriate domain of applicability (Abdollahi et al., 2024; Goetz et al., 2015; Nguyen et al., 2025; Sun et al., 2024).

2.3 Review of ML Applications in Geotechnical Engineering

ML is increasingly used in geotechnical engineering because it can represent complex input-output relationships without requiring a closed-form equation. This is useful where behaviour is influenced by interacting variables such as soil state, stratigraphy, groundwater, construction sequence, geometry and loading history.

Reviews of artificial intelligence in geotechnical computational mechanics indicate that:

- “Artificial Neural Networks” (ANNs),
- “Random Forests” (RFs) and
- “Support Vector Machines” (SVMs)

are among the most common methods, while other methods like “Deep Learning” (DL), or “Convolutional Neural Networks” (CNN), or “Long Short-term Memory” (LSTM) networks and hybrid models are increasingly used for specific problems (Liu et al., 2024; Yaghoubi et al., 2024).

Liu et al. (2024) reviewed the use of AI-based computational mechanics in geotechnical engineering and classified the applications according to physical and mechanical properties, constitutive modelling, and other geotechnical material characteristics. Their review showed that ANN, RF, and SVM were among the algorithms most commonly reported in the computational mechanics literature, with mechanical-property prediction being the main area

of application. This finding is relevant to this thesis because it indicates that the field has progressed beyond isolated proof-of-concept studies. Machine learning is now being applied to estimate parameters that are directly relevant to design and analysis, including strength, stiffness, and stress–strain response. At the same time, Liu et al. (2024) noted that reliability depends on the type of dataset and the selected model inputs, and identified physically guided and adaptive learning as future research directions.

Yaghoubi et al. (2024) provided a systematic review with a meta-analysis of ANN, ML, DL and “Ensemble Learning” in geotechnical engineering. They reported that ANN remains widely used, partly because many geotechnical studies rely on real-world laboratory and field datasets that suit supervised input-output learning. However, their review also noted that ensemble learning often performs strongly for geotechnical prediction tasks. This broader review supports a trend seen in the individual studies discussed below: earlier applications often used ANN or SVM, while more recent work increasingly uses RF, gradient boosting, XGBoost, LightGBM, automated ML (AutoML) and explainable AI tools.

ML models have been used to estimate undrained shear strength from index and stress-history variables, to predict rock strength from indirect test results, and to forecast tunnel settlement from monitoring time series (Cao et al., 2021; Nguyen et al., 2025; Sun et al., 2024; W. Zhang et al., 2021). In these cases, ML can reduce reliance on expensive or time-consuming tests, provided that the model is trained on data relevant to the intended application. It can also support rapid screening or design comparison, although final design decisions still require geotechnical review.

The tunnelling literature illustrates why ML is used in modern geotechnical practice. Tunnelling projects can generate large quantities of operational, monitoring and geological data, including TBM parameters, face observations, settlement records, laser scanning outputs and instrumentation readings. Marcher et al. (2020) argued that ML can improve efficiency and consistency in tunnelling workflows, but also emphasised the need for verified data sources, preprocessing and ethical use. Their discussion is particularly relevant because it frames ML as part of a broader digitalisation process in tunnelling, rather than as an isolated prediction tool. Similarly, Ayawah et al. (2022) found that models for predicting ground conditions ahead of TBMs were promising, but that a single universal model is unlikely to perform reliably across all rock masses, TBM types and classification systems.

ML in geotechnical engineering should therefore be considered like a decision-support tool instead of a replacement for engineering interpretation. The literature consistently indicates that good model performance on a particular dataset does not guarantee physical validity or reliable extrapolation to a new geological environment (Ayawah et al., 2022; Goetz et al., 2015; Marcher et al., 2020; Rauter & Tschuchnigg, 2021). This distinction is especially important in risk assessment, where the consequences of misclassification or overconfident prediction can be significant.

2.4 Types of geotechnical data used in ML

The data used in geotechnical ML studies vary substantially according to the engineering problem. Common data sources include:

- in situ testing,
- laboratory testing,
- geological logging,
- monitoring instrumentation,
- construction records,
- numerical simulations and
- spatial datasets.

The reviewed studies show that successful ML applications depend not only on the algorithm but also on careful input parameter selection. Inputs must be measurable, relevant to the mechanism being modelled, consistent across the dataset and available at the time the prediction is required.

In site investigation and soil classification studies, common inputs include cone penetration test (CPT) measurements such as tip resistance, sleeve friction and derived cone indices, as well as borehole observations, depth and soil class labels. Rauter and Tschuchnigg (2021) used 1339 CPTs to classify soils using SVM, ANN and RF models. Their study considered different combinations of direct CPT data and derived indices, with RF classifiers giving the best mix of prediction accuracy and time for training. They reported accuracies of about 75% for grain-size-based soil classes and about 97-99% for Robertson-type soil behaviour classes, a CPT-based (Cone Penetration Test) classification system. A key caution was that performance

differed between Austrian and Dutch CPT data, which the authors attributed to the locally limited nature of the training data.

Zhao and Wang (2020) addressed a related site characterisation question, with interpolation and stratification of different multilayer soil property profiles from few measurements. Their proposed procedure used Bayesian supervised learning to interpolate non-stationary measurements from different soil layers, followed by the use of modified k-means clustering to define the stratification. The study used numerical examples and real-life data, including soil property measurements such as effective friction angle. It showed that an interpolation-first procedure can outperform a conventional stratification-first approach where each layer contains too few measurements for reliable layer-by-layer interpolation. The important point for geotechnical practice is that ML was used not just to predict a target value, but to support interpretation under sparse-data conditions.

In strength and parameter prediction studies, input data commonly include laboratory index properties, stress-history variables and indirect rock test results. W. Zhang et al. (2021) predicted undrained shear strength of soft clays using vertical effective and preconsolidation stresses, the liquid and plastic limit and the natural water content. Elsaywy et al. (2022) used soft clay properties including void ratio, moisture content, dry unit weight, specific gravity, the liquid limit, plasticity and consistency indexes and data from pocket penetration shear. Nguyen et al. (2025) used stress-history and index variables for undrained shear strength prediction and then applied SHAP to explain the feature contributions. These studies show that conventional geotechnical descriptors remain central to ML, the algorithm does not remove the need to select physically meaningful inputs.

For rock properties, Matin et al. (2018) used point load index, P-wave rock mass velocity and also used the Schmidt hammer rebound number to evaluate Uniaxial Compressive Strength (UCS) and modulus of elasticity. Sun et al. (2024) similarly focused on rock UCS prediction, using indirect rock property indicators and explainable ML to interpret the prediction mechanism. Niu et al. (2024), in a review of rock classification systems, discussed how ML can complement established rock mass classification approaches such as RMR, GSI and Q, especially when geological surveys, sensors and imaging tools provide larger datasets.

For foundation and excavation problems, inputs often combine soil parameters with geometry and construction descriptors. Amjad et al. (2022) predicted pile bearing capacity using 200 driven pile static load test case histories, with input variables related to pile dimensions and standard penetration test (SPT) blow counts. Their sensitivity analysis found that SPT blow count along the extent of the pile shaft had the greatest influence on predicted pile capacity. R. Zhang et al. (2021) used finite element analyses of braced excavations in anisotropic clays and included inputs such as anisotropic shear strength and unloading shear modulus ratios, support system stiffness, excavation width to depth and wall penetration depth to depth ratios. D. Zhang et al. (2022) used seven deep excavation variables covering soil properties and spatial characteristics, including excavation depth and width, to predict excavation-induced tunnel displacement.

Tunnelling and monitoring applications use a broader mix of data. H. Xu et al. (2019) predicted TBM penetration rate from rock strength, Brazilian tensile strength, the rock quality designation (RQD), weathering zone data, and TBM thrust force and revolutions per minute. D. Xu et al. (2023) used slurry pressure balanced shield machine parameters, slurry system variables and cutter information to predict geological condition and formation percentages. Hu et al. (2019) and Cao et al. (2021) used tunnel settlement monitoring time series, while Song et al. (2024) reviewed settlement prediction studies involving dataset establishment, input feature selection and hyperparameter optimisation. These studies show the growing importance of monitoring and construction data as ML inputs.

Hazard and susceptibility studies often use spatially distributed environmental and geomorphological predictors. Goetz et al. (2015) used spatial predictors for landslide susceptibility mapping and emphasised spatial cross-validation and map appearance, while Bui et al. (2018) modelled land subsidence sensitivity using conditioning factors in a geographic information system. Liquefaction studies commonly use SPT-based variables, cyclic stress ratio, peak ground acceleration and soil index properties (Ozsagir et al., 2022; Samui & Sitharam, 2011). Across these applications, the data are rarely neutral, they reflect site investigation density, measurement uncertainty, local geology and prior decisions about what to record.

2.5 Common machine learning methods

The reviewed studies show repeated use of tree-based methods, support vector methods, neural networks, ensemble learning and hybrid optimisation approaches. The choice of method is often influenced by:

- the dataset size,
- the type of prediction target,
- the desired interpretability and
- the need to represent nonlinear relationships.

Simpler models can be easier to interpret and implement, while more complex models may improve statistical performance but require careful validation and explanation.

ANNs are common because they can approximate nonlinear relationships and have been used for tunnel settlement forecasting, rectangular tunnel stability, soft soil strength prediction and liquefaction susceptibility (Cao et al., 2021; Keawsawasvong et al., 2022; Pham et al., 2018; Samui & Sitharam, 2011). Keawsawasvong et al. (2022), for example, trained ANN models using finite element limit analysis results to predict the factor of stability at rectangular tunnels in Hoek-Brown type of rock masses. The inputs were six dimensionless parameters describing cover-depth ratio, width ratio, normalised UCS, geological strength index, material constant and disturbance degree, while the output was tunnel stability factor. The study reported high predictive accuracy but also noted that predictions should be kept within the parameter ranges used to train the model.

Support vector machines and support vector regression are widely used where datasets are moderate in size and nonlinear boundaries or regression functions are needed. They have been applied to liquefaction susceptibility, CPT interpretation, tunnel settlement forecasting and soft soil shear strength prediction (Hu et al., 2019; Ozsagir et al., 2022; Pham et al., 2018; Rauter & Tschuchnigg, 2021; Samui & Sitharam, 2011). These methods can perform strongly, but model interpretation can be difficult, and performance depends on kernel and hyperparameter selection. In the liquefaction work of Samui and Sitharam (2011), SVM outperformed ANN for predicting liquefaction susceptibility from SPT-related inputs. In contrast, in other

applications such as pile capacity and some ensemble comparisons, SVM was not the best model (Amjad et al., 2022; W. Zhang et al., 2021).

Tree-based and ensemble methods are prominent in recent geotechnical ML. Decision trees are interpretable and have been effective in classification tasks, including fine-grained soil liquefaction prediction, where Ozsagir et al. (2022) reported 90% accuracy for the best decision tree model. RF, gradient boosting, XGBoost, LightGBM and related ensemble methods often improve predictive performance by combining many weak learners and reducing variance or bias. Examples include pile capacity prediction with XGBoost, slope stability classification with gradient boosting, rock-head prediction with hybrid N-XGBoost and undrained shear strength prediction using XGBoost and RF models (Amjad et al., 2022; J. Zhou et al., 2019; W. Zhang et al., 2021; Zhu et al., 2021).

Hybrid and optimised models are also common. Several studies combined ML algorithms with particle swarm optimisation, genetic algorithms, Bayesian optimisation, tree-structured Parzen estimators or other metaheuristic strategies to improve hyperparameter tuning or feature selection (Hu et al., 2019; Pham et al., 2018; Sun et al., 2024; W. Zhang et al., 2021; M. Zhou et al., 2021). These approaches can improve measured performance, but they can also increase model complexity and make it harder to distinguish genuine generalisation from dataset-specific optimisation. This is one reason why the best-performing model in a paper should not automatically be assumed to be the best model for another site or thesis dataset.

Deep learning appears mainly in applications with time series, images, monitoring records or large datasets. Cao et al. (2021) proposed a CEEMDAN-LSTM model for tunnel settlement forecasting from limited univariate settlement data, while reviews identify growing use of LSTM, CNN and other DL models in geotechnical mechanics and tunnelling (Liu et al., 2024; Song et al., 2024). However, DL is not automatically the preferable method for geotechnical datasets, many of which remain small, heterogeneous and site-specific. The literature therefore suggests a pragmatic approach, simpler and more transparent models should be considered alongside complex models, especially where the intended use is design support or risk assessment.

2.6 Applications in classification and prediction

Machine learning applications in geotechnical engineering can be grouped into several subjects:

- tunnelling and settlement prediction;
- slope, landslide and subsidence assessment;
- foundation engineering;
- liquefaction;
- soil and rock property prediction; and
- site investigation interpretation.

2.6.1 Tunnelling and settlement prediction

Tunnel-related applications are among the most developed areas in the reviewed literature because tunnelling projects generate operational, geological and monitoring data over time. Hu et al. (2019) compared particle swarm optimised support vector regression (PSO-SVR), back-propagation neural networks and extreme learning machines for univariate tunnel settlement forecasting using real-world settlement data from two Chinese cities. The input was historical settlement time series data, and the predicted output was future tunnel surface settlement. PSO-SVR gave the best forecasting accuracy among the tested models. The study is useful because it compared several ML approaches on the same forecasting task, but it also shows a limitation of univariate settlement forecasting: models based only on historical settlement may not explicitly include construction changes, geological variation or groundwater effects.

Cao et al. (2021) examined a similar tunnel-settlement forecasting problem using a hybrid CEEMDAN-LSTM model based on sensor data. The study noted that deep-learning models generally need extensive datasets for training, while tunnel monitoring records can be limited and based on a single variable. Their method first separated one settlement time series into several components using CEEMDAN and then predicted each component with LSTM networks. The predicted output was short-period tunnel surface settlement. The authors reported improved prediction accuracy with reasonable computational efficiency compared with other machine-learning techniques, particularly under small-sample conditions. This

study shows how deep learning can be adapted to geotechnical monitoring data, although its reliability still depends on the continuity and representativeness of the monitoring record.

Song et al. (2024) reviewed AI based prediction of tunnelling-induced surface settlement. Rather than proposing one model, the review focused on dataset establishment, input feature selection and hyperparameter optimisation. The review identified ANN and support vector methods as widely used in settlement prediction, while also discussing newer AI variants and optimisation methods. Its main contribution for this chapter is a synthesis of the practical challenges in settlement prediction: settlement is affected by ground conditions, tunnel geometry, construction parameters and monitoring configuration, so the usefulness of any ML model depends on whether those features are adequately represented in the dataset.

Zhang et al. (2021) investigated surface settlement associated with braced excavations in anisotropic clay ground. Their database was produced from 1,479 finite element analyses, with input variables describing anisotropic soil strength and stiffness, support-system stiffness, excavation geometry, and wall penetration depth. The predicted output was maximum ground surface settlement. XGBoost and RF regression models were benchmarked against conventional regression approaches and achieved better performance. The study therefore highlights the usefulness of ensemble methods for modelling nonlinear interactions in excavation problems. However, because the dataset was generated from finite element analyses, the ML predictions also reflect the assumptions of the numerical model, including the selected constitutive model and parameter ranges.

Zhang et al. (2022) proposed an AutoML framework to predict tunnel displacement induced by excavation. The model used seven input variables covering soil properties and the spatial characteristics of deep excavation, including excavation depth, excavation width, and ratios related to shear strength and stiffness. The target output was maximum horizontal tunnel displacement. The authors applied 10-fold cross-validation, compared the AutoML model with six genetic-algorithm-optimised ML models, and further checked the model against practical engineering monitoring data. They reported that AutoML provided a good balance between efficiency and robustness. The stiffness–strength ratio, excavation depth, and excavation width were identified as key controlling variables. This paper is particularly relevant to geotechnical risk assessment because it links prediction with validation using monitoring data.

Xu et al. (2019) focused on TBM penetration rate prediction. The study used field observations and laboratory tests from a Malaysian tunnelling project, with inputs including UCS, Brazilian tensile strength, RQD, weathering zone, thrust force and revolutions per minute. The output was TBM penetration rate. KNN, CHAID, SVM, CART and neural network models were compared using R^2 , RMSE, VAF and MAE. KNN was ranked as the best model, with testing values of $R^2=0.962$, RMSE = 0.116, VAF = 96.226 and MAE = 0.081. The study also identified UCS as the most important input and RPM as the least important input for that dataset. The result shows that relatively simple instance-based methods can perform well where similar historical cases exist.

Xu et al. (2023) predicted geological conditions for a large-diameter slurry pressure balance shield tunnel using field data from the Heyan Road river-crossing tunnel. The study compared RF, AdaBoost, and SVM for rock-mass classification, and tested several single-target, regressor-chain, and multi-objective models for predicting formation percentages. Input variables included shield operational parameters, slurry-system parameters, and cutter information, while the outputs were rock-mass class and formation percentages. RF achieved the best performance for rock-mass classification, whereas multi-objective random forest performed best for formation-percentage prediction. The authors also found that slurry-system variables and cutter type improved forecasting performance. This provides a useful example of input-parameter selection, showing that machine-specific and construction-process variables can have a meaningful effect on geological prediction.

Zhou et al. (2021) assessed rock-mass quality at NATM tunnel faces using hybrid ensemble-learning models. The study developed a multi-source dataset containing 133 tunnel faces and used thirteen quantitative and qualitative input variables to describe tunnel-face conditions. The target output was RMR, calculated using an empirical formula. CART, multilayer perceptron, gradient-boosted regression trees, and RF models were combined with tree-structured Parzen estimator hyperparameter tuning. The hybrid ensemble approaches, particularly TPE-GBRT and TPE-RF, outperformed the meta-ML models based on metrics such as correlation coefficient, MAE, RMSE, and MAPE. Sensitivity analysis identified influential variables including groundwater, rock-structure category, and trace parameters. The study demonstrates the value of multi-source tunnel-face data, although the predicted target remains linked to an empirical RMR calculation.

Keawsawasvong et al. (2022) studied the stability of rectangular tunnels in Hoek-Brown rock masses. Using Finite Element limit analysis they generated upper and lower bound solutions, and ANN models were trained using six dimensionless input parameters: cover-depth ratio, width ratio, normalised UCS, GSI, Hoek-Brown material constant and disturbance factor. The output was tunnel stability factor. The authors reported high model accuracy using R^2 , MAE and RMSE, and noted that only one hidden layer was sufficient for the optimal ANN model. This study is useful as a rapid design-support tool, but its applicability is limited to the parameter ranges and failure criterion used in the numerical dataset.

Taken together, the tunnelling studies show a progression from time-series settlement forecasting to multi-source construction intelligence. Monitoring data, TBM operational data, geometry and ground classification variables are all used, but the strongest studies also recognise limitations associated with site specificity, numerical assumptions and input availability.

2.6.2 Slope stability, landslide assessment and subsidence

Slope stability and geohazard studies often formulate the problem as classification, factor of safety prediction or susceptibility mapping. J. Zhou et al. (2019) used gradient boosting for circular slope stability prediction based on 221 case histories. Inputs used included parameters like unit weight, cohesion, friction and slope angles, slope height and also the pore water pressure coefficient; the output was a stable or failed class. The model used five-fold cross-validation and was evaluated using AUC, classification accuracy and Cohen's kappa. Reported values of $AUC = 0.900$, $accuracy = 0.8654$ and $kappa = 0.7324$ indicated strong performance. The authors also used variable importance and partial dependence plots, which is valuable because slope stability predictions must be interpreted in relation to physical mechanisms.

Nanehkaran et al. (2023) compared multilayer perceptron, SVM, decision tree and RF models for slope stability analysis. The study used 100 records of earth slopes from Iran, with factor of safety values generated or checked using limit equilibrium analysis and GeoStudio software. The output was factor of safety or stability classification, and performance was assessed using confusion matrices, loss functions and ROC curves. The multilayer perceptron performed best, with $accuracy = 0.901$ and $precision = 0.900$, followed by SVM. This study reinforces that neural networks can perform strongly for slope datasets, but it also illustrates a common limitation: the ML target is closely linked to conventional limit equilibrium calculations.

Abdollahi et al. (2024) extended slope-related ML to road embankment factor of safety prediction with an explicit explainability focus. The study used accepted geotechnical inputs for embankment core and foundation soils, embankment height, California bearing ratio and newly codified geometric features such as slope type ratios and number of berms. A pool of 19 ML techniques was trained using an AutoML pipeline, and SHAP was used for post-hoc interpretability. For the Bishop-method dataset, LightGBM outperformed other approaches with $MAE = 0.1472$, $RMSE = 0.2276$ and $R^2=0.7738$. For other limit equilibrium methods, XGBoost showed very strong performance in the reported comparisons. SHAP analysis identified height, California bearing ratio, slope type and cohesion as influential variables. The study is important because it demonstrates how ML can be combined with explainable AI to build trust in safety-related predictions.

Goetz et al. (2015) compared statistical and machine-learning methods for modelling landslide susceptibility across three areas in Lower Austria. The models included logistic regression, generalised additive models, weights of evidence, SVM, RF, and bootstrap-aggregated classification trees with penalised discriminant analysis. The input data consisted of spatial predictor variables, while the predicted output was landslide susceptibility. The authors assessed model performance using spatial k-fold cross-validation, variable-importance analysis, and visual inspection of the resulting susceptibility maps. RF and bootstrap-aggregated classification gave the strongest overall results, although the differences between methods were generally small. Slope angle, surface roughness, and plan curvature were consistently important predictors. An important caution from the study was that models with high predictive accuracy can still produce spatial artefacts, whereas smoother models may be easier to interpret or more suitable for mapping.

Bui et al. (2018) used machine learning for land-subsidence susceptibility mapping in South Korea. The study tested four models: Bayesian logistic regression, SVM, logistic model tree, and alternating decision tree. Eight conditioning factors were used within a GIS-based framework, and the output was a land-subsidence susceptibility map. Model performance was assessed using sensitivity, specificity, accuracy, kappa, RMSE, ROC-based measures, and success and prediction-rate curves. Bayesian logistic regression produced the most acceptable and reliable map in that study. This result is useful because it shows that, for some geohazard

mapping problems, a probabilistic and interpretable model may perform better than more complex alternatives.

2.6.3 Foundations, bearing capacity and pile capacity

Foundation-related problems are suitable for supervised learning when sufficient load-test records or case histories are available. Amjad et al. (2022) used 200 static load-test case histories from driven piles to predict pile bearing capacity. The input variables included pile geometry and SPT blow-count information, while the output was ultimate pile bearing capacity. XGBoost, AdaBoost, RF, decision tree, and SVM models were assessed using R^2 , MAE, RMSE, mean absolute relative error, Nash–Sutcliffe efficiency, and relative strength ratio. XGBoost gave the best testing performance, with $R^2 = 0.955$, MAE = 59.929, RMSE = 80.653, MARE = 6.6, NSE = 0.950, and RSR = 0.225. Sensitivity analysis showed that SPT blow count along the pile shaft was the most important input.

This study shows that ensemble-learning methods can support empirical design, particularly when pile capacity can be related to measurable site-investigation and geometry variables. It also demonstrates the practical value of SPT data in ML applications, because SPT results are commonly available on geotechnical projects. However, pile-capacity prediction remains strongly affected by pile type, installation method, ground conditions, and the load-test database used. A model trained on one set of driven-pile case histories should therefore be validated before being applied to other pile types or geological settings.

2.6.4 Liquefaction and seismic-related applications

Liquefaction studies demonstrate how ML can be used for classification problems where historical earthquake and SPT datasets are available. Samui and Sitharam (2011) developed ANN and SVM models to assess liquefaction susceptibility using SPT data from the 1999 Chi-Chi earthquake. The models used corrected SPT value $(N1)_{60}$ and cyclic stress ratio as inputs. The authors also tested a simplified input combination based on $(N1)_{60}$ and peak ground acceleration. The predicted output was liquefaction susceptibility. The study found that SVM performed better than ANN, including for global case histories, and suggested that the simplified two-input model could be used as a rapid assessment tool. However, simplified input sets may exclude important liquefaction controls, such as fines content, soil fabric, groundwater conditions, and depositional environment.

Ozsagir et al. (2022) compared seven ML algorithms for predicting liquefaction potential in fine-grained soils using field data from the Adapazari region after the 1999 earthquake. The models included decision tree, logistic regression, SVM, k-nearest neighbours, stochastic gradient descent, RF, and ANN. Input variables were obtained from field and laboratory data, including SPT-related and soil physical properties, while the output was liquefaction potential. Model performance was assessed using accuracy, recall, precision, F1 score, and receiver operating characteristic measures. The decision-tree model achieved the best result, with an overall accuracy of 90%, and mean grain size D50 was identified as the most influential feature. The study is notable because the best-performing model was also relatively interpretable, which is useful for decision-makers evaluating liquefaction potential.

The liquefaction studies show both the strength and weakness of ML in seismic geotechnical applications. ML can capture nonlinear empirical patterns in historical earthquake datasets, but predictions remain conditional on the earthquake records, soil types, triggering conditions and class definitions included in the training data. Extrapolation to new seismic settings should therefore be treated cautiously.

2.6.5 Soil and rock property prediction

Considerable work has focused on estimating strength, stiffness and classification targets from more readily available data. Pham et al. (2018) compared PANFIS, GANFIS, SVR and ANN for prognosticating the shear strength of soft soils. The dataset contained 188 plastic clay samples from two bridge projects in Vietnam. Inputs included soil index and composition variables such as plasticity index, liquid limit, moisture content and clay content, while the target was soft soil shear strength. PANFIS achieved the best performance, with RMSE = 0.038 and $R=0.601$, followed by GANFIS, SVR and ANN. The study suggests that hybrid neuro-fuzzy and optimisation methods can be useful for soft soil strength prediction, although the moderate correlation also indicates that the problem remains difficult and dataset-specific.

Zhang et al. (2021) used Bayesian-optimised XGBoost and RF models to estimate the undrained shear strength of soft clay. The input variables included preconsolidation stress, vertical effective stress, liquid limit, plastic limit, and natural water content. The target output was undrained shear strength. The authors compared the ensemble methods with SVM, multilayer perceptron, MARS, and two transformation models using fivefold cross-validation. XGBoost and RF performed better than the comparison methods, and XGBoost also provided

feature-importance rankings. This paper is important because it directly compares modern ensemble approaches with both conventional ML models and empirical transformation methods.

Elsawy et al. (2022) modelled the undrained shear strength of sensitive alluvial soft clay. The dataset included eight input variables: moisture content, specific gravity, void ratio, dry unit weight, LL, PI, consistency index, and pocket-penetrometer shear data. Several ML formulations were trained, including linear regression, regression trees, Gaussian process regression, ensemble trees, and SVR. The target output was vane-test undrained shear strength. The fine Gaussian SVR model achieved the best agreement, with a reported testing $R^2 = 0.96$. The study highlighted that realistic, information-rich datasets, together with suitable algorithm selection, can significantly improve prediction accuracy.

Nguyen et al. (2025) also examined undrained shear strength, with a stronger focus on model explainability. ANN, XGBoost, and RF models were trained and assessed using MSE, MAE, and R^2 . SHAP was then applied to explain the contribution of each input feature. RF produced the strongest result, with $R^2 = 0.82$, compared with ANN ($R^2 = 0.75$) and XGBoost ($R^2 = 0.78$). SHAP identified preconsolidation stress as an important predictor, while dependence plots indicated that RF represented nonlinear interactions more effectively than ANN. This study is useful because it moves beyond predictive accuracy and also considers model transparency.

Matin et al. (2018) applied RF for variable selection and for estimating UCS and modulus of elasticity in rock mechanics. The input variables included point load index, P-wave velocity, and Schmidt hammer rebound number. The predicted outputs were UCS and modulus of elasticity. RF was used both to rank the importance of the variables and to predict the target properties. The authors reported R^2 values of 0.93 for UCS and 0.91 for modulus of elasticity, with P-wave velocity identified as an influential predictor. The study shows that RF can be useful for both prediction and feature screening when several indirect rock-test measurements are available.

Sun et al. (2024) proposed an interpretable AOA-XGBoost model for rock UCS prediction. The study used an Archimedes optimisation algorithm to tune XGBoost and applied SHAP to explain the prediction mechanism. The target output was UCS, and the model achieved RMSE = 14.69 and testing $R^2 = 0.9117$. SHAP analysis was used to identify how input factors

contributed to predictions. The study is relevant because rock strength prediction is often affected by expensive laboratory testing and black-box concerns; explainable boosting models can provide both accuracy and interpretive support.

Niu et al. (2024) reviewed rock classification systems with emphasis on artificial intelligence techniques. The paper discussed traditional systems such as RSR, RMR, RMI, GSI and Q, and then reviewed AI applications in rock classification. The review concluded that AI and ML can improve consistency and make use of larger datasets from geological surveys, sensors and imaging tools, but should complement rather than replace established empirical systems. This conclusion aligns with the broader theme of this chapter: ML may improve prediction and classification, but it should remain linked to geological meaning and engineering purpose.

2.6.6 Classification and site characterisation

Classification and site characterisation studies are especially important for geotechnical risk assessment because uncertainty in soil type, stratigraphy and rockhead level can strongly influence design decisions. Rauter and Tschuchnigg (2021) showed that CPT data can be used to classify soils with ML, with RF outperforming SVM and ANN in their study. The relatively high accuracy for Robertson-type soil behaviour classes compared with grain-size-based classes suggests that ML performance depends strongly on how the target class is defined. The authors' observation that local training data affected transferability is directly relevant to any thesis workflow based on site investigation records.

Zhao and Wang (2020) addressed sparse multilayer soil property profiles by reversing the conventional sequence of interpretation. Instead of stratifying first and then interpolating within each layer, their method interpolated using measurements from all layers and then performed clustering-based stratification. The output was a stratified soil property profile, and the method was most beneficial when layer-specific data were sparse. This is a useful example of ML supporting the geotechnical interpretation process, rather than only predicting a single design parameter.

Zhu et al. (2021) predicted rockhead position using a hybrid N-XGBoost framework based on limited borehole and spatial geographic information. The study compared XGBoost with multivariate linear regression, ANN, SVM, LightGBM and gradient boosting regression tree models. The predicted output was rockhead elevation or depth. XGBoost achieved $R^2=0.89$

and RMSE = 5.8 m, and the probabilistic N-XGBoost framework also provided uncertainty estimates. The authors noted that the method was useful for limited borehole data and that SMOGN helped address imbalanced target distributions, especially rare extreme values. This is particularly relevant to geotechnical risk because rare deep or shallow rockhead conditions can be critical even if they are underrepresented in the dataset.

Ayawah et al. (2022) reviewed and tested ML models for ground condition prediction ahead of TBMs. The study considered models such as RF, extremely randomised trees, CART, AdaBoost-CART, KNN, ANN and support vector classifiers, and evaluated whether a single model could predict ground conditions across different TBM types, rock mass conditions and classification systems. Using Pahang-Selangor Raw Water Transfer Tunnel data, the authors found that models could perform well but generally performed less strongly than reported in their original settings. They concluded that a single universal model is unlikely and recommended site-specific modelling and careful selection of relevant inputs. This conclusion is a useful caution for all classification tasks in geotechnical ML.

2.7 Summary of review findings

Table 2.1 below summarises the main applications, inputs, targets and cautions from the above reviewed literature. The table shows three broad patterns. First, geotechnical ML is strongly problem-specific: the same algorithm family may be used for pile capacity, settlement, liquefaction or landslide mapping, but the inputs and validation strategy differ substantially. Second, ensemble methods are increasingly common because they often perform well on nonlinear tabular data. Third, the most useful studies do more than report accuracy; they also examine feature importance, uncertainty, data limitations or interpretability.

The table also highlights that the strongest geotechnical ML studies are those that link model choice to engineering context. For tabular design and case-history datasets, tree-based ensembles such as RF, gradient boosting and XGBoost often perform well and provide feature importance. For settlement and monitoring data, time-series and hybrid models are common because temporal structure matters. For spatial susceptibility mapping, validation must consider spatial dependence and the quality of the final map, not just a single accuracy statistic. Across the table, SPT, CPT, laboratory index data, geometry, geological condition, monitoring records and construction parameters all play different roles depending on the target variable.

This reinforces the importance of input selection: a high-performing algorithm cannot compensate for missing or poorly defined geotechnical predictors.

Table 2.1 Summary of machine learning applications in geotechnical engineering

Reference	Geotechnical application	ML methods	Input parameters/data	Predicted output	Main outcome/findings	Notes/limitations
Liu et al. (2024); Yaghoubi et al. (2024)	General Geotechnical AI review	ANN, RF, SVM, DL, ensemble learning	Published geotechnical datasets and studies	Trends in applications and methods	ANN, RF and SVM are common; ensemble and deep learning methods are increasing	Reliability depends on data type, model inputs and task definition
Marcher et al. (2020); Ayawah et al. (2022); Song et al. (2024)	Tunnelling data, ground conditions, and settlement	Supervised and Unsupervised ML; RF, ANN, SVM, KNN, ensemble models	TBM data, tunnel monitoring, geological and classification data	Ground class, settlement and construction support information	ML can improve consistency and forecasting in tunnelling workflows	Site-specific modelling, preprocessing and engineering review remain essential
Hu et al. (2019); Cao et al. (2021); R. Zhang et al. (2021); D. Zhang et al. (2022)	Tunnel settlement and excavation-induced displacement	PSO-SVR, BPNN, ELM, CEEMDAN- LSTM, XGBoost, RF, AutoML	Settlement records, excavation geometry, soil properties, monitoring data	Settlement or tunnel displacement	Hybrid and ensemble models improved prediction in the studied datasets	Small samples and numerical-model assumptions can limit generalisation
H. Xu et al. (2019); D. Xu et al. (2023)	TBM performance and shield geology prediction	KNN, SVM, NN, CART, CHAID, RF, AdaBoost, MORF	Rock strength, RQD, weathering, TBM thrust, RPM, slurry and cutter data	Penetration rate, rock class, formation percentage	KNN, RF and MORF performed best in the reported comparisons	Model transfer depends on machine type and ground conditions
M. Zhou et al. (2021); Keawsawasvong et al. (2022)	Tunnel face rock quality and tunnel stability	CART, MLP, GBRT, RF, ANN	Multi-source face data; Hoek-Brown and geometric parameters	RMR or stability factor	Ensemble and ANN models gave rapid predictions	Targets partly depend on empirical or numerical source models
Zhao and Wang (2020); Rauter and Tschuchnigg (2021); Zhu et al. (2021)	Site characterization, CPT interpretation and rockhead prediction	Bayesian supervised learning, clustering, SVM, ANN, RF, N- XGBoost	Sparse soil properties, CPT data, borehole/geospatial data	Soil layers, soil class, rockhead level	ML supported interpretation where direct data were limited	Local geology and class imbalance affect transferability

Reference	Geotechnical application	ML methods	Input parameters/data	Predicted output	Main outcome/findings	Notes/limitations
Matin et al. (2018); Sun et al. (2024); Niu et al. (2024)	Rock strength and rock classification	RF, AOA-XGBoost, SHAP, review of AI methods	Point load index, P-wave velocity, Schmidt rebound, rock classification indicators	UCS, modulus, rock class	RF and XGBoost achieved strong predictive performance and feature ranking	AI should complement established classification systems
Pham et al. (2018); W. Zhang et al. (2021); Elsayy et al. (2022); Nguyen et al. (2025)	Soft clay and undrained shear strength prediction	PANFIS, GANFIS, SVR, ANN, XGBoost, RF, SHAP	Index properties, stress history, water content, limits, pocket shear	Shear strength or undrained shear strength	Ensemble and SVR-based methods performed well; SHAP improved interpretability	Datasets are often regional and sensitive to sampling quality
Amjad et al. (2022)	Pile bearing capacity	XGBoost, AdaBoost, RF, DT, SVM	Static load test cases, pile geometry, SPT data	Pile bearing capacity	XGBoost gave the best testing performance; shaft SPT was most influential	Applicability depends on pile type and case-history coverage
J. Zhou et al. (2019); Nanekaran et al. (2023); Abdollahi et al. (2024)	Slope stability and road embankment factor of safety	GBM, MLP, SVM, DT, RF, AutoML, SHAP	Unit weight, cohesion, friction angle, slope geometry, pore pressure, CBR	Stability class or factor of safety	GBM, MLP and AutoML models produced strong case-study performance	Physical plausibility and explainability are critical for safety use
Goetz et al. (2015); Bui et al. (2018)	Landslide and land subsidence susceptibility	GLM, GAM, WOE, SVM, RF, BPLDA, BLR, LMT, ADTree	Spatial, geomorphological and conditioning factors	Susceptibility class/map	RF/BPLDA and BLR performed well in the respective studies	Spatial validation, map artefacts and interpretability affect usefulness
Samui and Sitharam (2011); Ozsagir et al. (2022)	Liquefaction susceptibility	ANN, SVM, DT, LR, KNN, SGD, RF	SPT data, CSR, peak ground acceleration, grain-size and soil properties	Liquefaction class/potential	SVM and DT models were effective in the reported datasets	Regional earthquake and soil conditions limit direct transfer

2.8 Model evaluation and interpretability

Model evaluation in geotechnical ML usually relies on regression or classification performance measures. Regression studies commonly report R^2 , RMSE, MAE, MSE, MAPE, and other efficiency-based indicators. Classification studies typically use metrics such as accuracy, precision, recall, F1 score, AUC, and confusion matrices.

Examples include pile-capacity prediction assessed using R^2 , MAE, RMSE, MARE, NSE, and RSR (Amjad et al., 2022), slope-stability classification assessed using AUC, accuracy, and Cohen's kappa (J. Zhou et al., 2019), and liquefaction classification assessed using accuracy, recall, precision, F1 score, and ROC-based indicators (Ozsagir et al., 2022).

Cross-validation (splitting the dataset several times in different ways) is widely used to reduce dependence on a single train-test split, particularly where datasets are small. J. Zhou et al. (2019) used five-fold cross-validation for gradient boosting slope stability prediction, while D. Zhang et al. (2022) used 10-fold cross-validation to improve robustness for excavation-induced tunnel displacement prediction. Rauter and Tschuchnigg (2021) and Keawsawasvong et al. (2022) also discussed cross-validation in the context of limited or classification-based geotechnical datasets. However, cross-validation only tests performance under the sampling scheme used. If data from the same site, geological unit or project appear in both training and testing folds, performance may still overstate generalisation to a genuinely new site.

Spatial and temporal validation are especially important in geotechnical applications. Goetz et al. (2015) emphasised spatial k-fold cross-validation for landslide susceptibility mapping because random splitting can overestimate performance when nearby samples are spatially autocorrelated. For monitoring and settlement forecasting, temporal ordering matters because a model that uses future-influenced patterns during training may appear more accurate than it would be in real-time application (Cao et al., 2021; Hu et al., 2019; Song et al., 2024). These issues are directly relevant to risk assessment, where the model should be judged by its likely future performance under field deployment conditions.

Interpretability, that is understanding why a model made a certain prediction or which input variables influenced the prediction most, has become a major theme in recent literature. Tree-based models can provide feature importance measures, and several studies used this to identify influential variables such as slope geometry, cohesion, excavation geometry, SPT blow count,

preconsolidation stress and rock test indicators (Abdollahi et al., 2024; Amjad et al., 2022; D. Zhang et al., 2022; Nguyen et al., 2025; Sun et al., 2024; J. Zhou et al., 2019). SHAP has been used to provide more detailed post-hoc explanations for factor of safety, undrained shear strength and rock strength predictions (Abdollahi et al., 2024; Nguyen et al., 2025; Sun et al., 2024). These tools help connect model behaviour to engineering reasoning, but they do not prove causality. A feature may be important in a model because it is a proxy for another unmeasured variable or because of correlations within the training data.

Uncertainty is another important evaluation issue. Zhu et al. (2021) combined XGBoost with NGBost to provide probabilistic rock-head predictions, which is useful because rock-head uncertainty can materially affect tunnelling and foundation decisions. Zhao and Wang (2020) also incorporated interpolation uncertainty in multilayer soil profile interpretation. These approaches are valuable because geotechnical decisions often require not only a best estimate but also an understanding of confidence, variability and potential adverse conditions.

The reviewed studies also show that performance metrics must be interpreted in relation to engineering consequences. A high R^2 for settlement prediction may still be inadequate if the model underestimates rare large settlements. A high classification accuracy for liquefaction, landslide or slope stability may be misleading if the failed or liquefied class is rare. A susceptibility map with good AUC may still be unsuitable if it contains spatial artefacts or if the underlying predictors are not causally meaningful (Bui et al., 2018; Goetz et al., 2015; Ozsagir et al., 2022). For this reason, model evaluation should combine statistical metrics, visual diagnostics, sensitivity analysis, uncertainty checks and geotechnical judgement.

2.9 Limitations and research gap

The main limitation across the reviewed literature is data dependence. ML models learn from the data provided to them, and geotechnical datasets are frequently limited, unevenly distributed, site-specific and affected by inconsistent testing, logging and construction practices. A model trained on one geological environment may not extrapolate reliably to another, even if the same nominal input variables are available (Ayawah et al., 2022; Goetz et al., 2015; Rauter & Tschuchnigg, 2021). This is particularly important for rare or high-consequence events such as slope failure, liquefaction, excessive tunnel settlement or adverse ground conditions.

A second limitation is that high statistical accuracy does not automatically imply physical validity. Some models are trained on empirical targets, numerical simulation outputs or historical classifications. In such cases, the ML model may reproduce the assumptions and biases of the source method rather than independently discovering ground behaviour (D. Zhang et al., 2022; Keawsawasvong et al., 2022; R. Zhang et al., 2021; M. Zhou et al., 2021). For example, a model trained on finite element results is only as credible as the constitutive assumptions, boundary conditions and parameter ranges used to generate those results.

A third limitation is interpretability. Black-box models can produce accurate predictions but may be difficult to justify in design, construction or risk-management settings. Recent studies using SHAP, feature importance and partial dependence plots are a useful response to this problem, but interpretation tools should be used carefully and checked against geotechnical understanding (Abdollahi et al., 2024; Nguyen et al., 2025; Sun et al., 2024; J. Zhou et al., 2019). Model explanations can identify associations in the training data, but they do not replace geological reasoning, mechanism-based analysis or independent validation.

Input-parameter selection remains an important gap in both geotechnical ML research and practice. Several studies showed that model performance improved when relevant construction-related or geological variables were included. Examples include slurry-system and cutter parameters for shield tunnelling, pile-shaft SPT blow counts for pile-capacity prediction, and preconsolidation stress for undrained shear strength prediction (Amjad et al., 2022; D. Xu et al., 2023; Nguyen et al., 2025). However, many datasets still lack variables that would be desirable from a geotechnical-mechanism perspective. Missing groundwater information, inconsistent soil descriptions, limited geological-unit data, and sparse monitoring records can all reduce the practical usefulness of ML models.

Further research is required in several areas. First, dataset descriptions, preprocessing steps, missing-data treatment, and validation strategies should be reported more clearly so that results can be reproduced. Second, benchmark datasets for common geotechnical tasks would allow fairer comparison between methods. Third, probabilistic and uncertainty-aware models need further development, especially for stratigraphy, rock-head prediction, settlement, and geohazard applications. Fourth, physics-based and hybrid modelling approaches may help link the predictive flexibility of ML with established geotechnical principles, although these

methods still require careful validation (Liu et al., 2024; Niu et al., 2024; Song et al., 2024; Zhu et al., 2021).

For this thesis, the reviewed literature supports a balanced position. ML can support geotechnical engineering by improving prediction, classification, and pattern recognition when suitable data are available. It can also contribute to geotechnical risk assessment by identifying influential variables, highlighting uncertain zones, and comparing alternative interpretations. However, the reliability of a model depends on several factors:

- the quality and representativeness of the training data,
- the appropriateness of the selected method,
- the validation strategy and
- the extent to which results are interpreted through engineering judgement and geological understanding.

3. Theoretical Background of Machine Learning

This chapter provides the theoretical background needed to understand how machine-learning methods can be applied in geotechnical engineering practice. It introduces the main types of geotechnical data, key geotechnical parameters, basic machine-learning concepts, model-evaluation principles, and limitations relevant to geotechnical risk assessment. The chapter supports the methodology and results chapters by defining the concepts used for data preparation, model training, performance assessment, and interpretation of model outputs.

3.1 Basic Concepts of Machine Learning

Artificial intelligence (AI) is a broad discipline that includes computational methods designed to perform tasks commonly linked to human reasoning, learning, perception, and decision-making. Machine learning (ML) is a branch of AI in which algorithms identify patterns in data and use those patterns to support prediction, classification, grouping, or data-driven interpretation. Deep learning (DL) is a further branch of ML that uses neural-network architectures with multiple layers. ML also draws on statistics, mathematics, computer science, and data science, and is generally regarded as one of the main applied branches of AI. A widely used formal definition of ML was provided by Mitchell (1997). This definition identifies three main components of an ML system: the Task, the Experience, and the Performance measure.

- The task, T, refers to the type of problem that the ML algorithm is expected to solve. Common ML tasks include classification, regression, clustering, dimensionality reduction, and anomaly detection. Classification assigns observations to separate categories, whereas regression predicts continuous numerical values.
- The experience, E, refers to the data from which the algorithm learns. In most practical applications, data are organised into datasets, commonly in tabular form. Rows usually represent individual observations or cases, while columns represent variables or features. In geotechnical engineering, for example, each row may represent a soil or rock sample, a field test location, a borehole interval, or a monitoring record. The columns may include properties such as depth, unit weight, water content, plasticity

index, standard penetration test value, cone penetration resistance, geological descriptors, or instrumentation readings.

- The performance measure, P , is the quantitative criterion used to assess how well the algorithm completes the task. The selected metric depends on the problem type. Classification models can be evaluated using accuracy, precision, recall, or F1-score, while regression models can be evaluated using metrics such as mean absolute error, root mean squared error, or coefficient of determination.

In supervised ML, one variable is normally selected as the “target variable”, which represents the value the model aims to estimate. The other variables are treated as “input features” or “predictor variables”.

A machine-learning model can be represented as a function that converts input features into a predicted output:

$$\hat{y} = f(X)$$

where X represents the input-feature set, f is the function learned from the data, and $\hat{y}$ is the predicted output. In geotechnical applications, X may include field-test results, laboratory measurements, depth, geological descriptors, or monitoring values. The predicted output $\hat{y}$ may be a continuous parameter, a categorical class, or an engineering indicator.

Traditional computer algorithms usually follow predefined instructions to process input data and produce an output. By contrast, ML algorithms learn from data by detecting patterns and building a model that can generate predictions or support decisions for new input cases. The aim is not only to achieve good performance on the training dataset, but also to generalise to data that were not used during training. Generalisation is a key challenge in ML because a model that matches the training data too closely may perform poorly on new cases. This problem is commonly known as overfitting.

The development of an ML model involves three important concepts:

- The first is the “representation space”, also known as the “hypothesis space”, which refers to the set of possible models that the algorithm can select from.

- The second is the “evaluation function”, often expressed as a loss or cost function, which measures how well a model performs.
- The third is the “optimisation process”, which searches through the hypothesis space to identify a model that minimises error or maximises predictive performance.

ML methods are often classified into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

- “Supervised learning” applies when input features and corresponding target values are both available. It is used for regression tasks, where the target is continuous, and classification tasks, where the target is categorical. In geotechnical engineering, regression examples include predicting settlement, undrained shear strength, permeability, deformation modulus, or factor of safety. Classification examples include assigning soil type, rock-mass quality class, liquefaction susceptibility category, or geotechnical risk class (Geem et al., 2026a; Sarker, 2021; Zhang et al., 2022).
- “Unsupervised learning” applies when there is no predefined target variable. Instead, the model identifies patterns, clusters, or structures in the input data. In geotechnical engineering, unsupervised methods may be used to group similar soil samples, support geological-unit interpretation, detect anomalous monitoring responses, or reduce high-dimensional datasets for visualisation.
- “Semi-supervised learning” uses a limited set of labelled data together with a larger set of unlabelled data. This approach can be useful when reliable labels are difficult, expensive, or time-consuming to obtain. This is relevant to geotechnical engineering because labelled datasets often depend on laboratory testing, expert interpretation, or detailed site-investigation records. However, semi-supervised learning requires careful consideration of label quality, uncertainty, and data representativeness.
- “Reinforcement learning” is based on learning through interaction with an environment. The algorithm takes actions and receives rewards or penalties depending on the outcome of those actions. Over time, it learns a strategy, or policy, that maximises the expected reward. Reinforcement learning is less commonly used in routine geotechnical practice than supervised and unsupervised learning, but it may

have potential applications in optimisation, construction planning, and decision-support problems.

Common ML algorithms used in geotechnical studies include linear regression, logistic regression, decision trees, random forests, support vector machines, artificial neural networks, gradient boosting methods, clustering techniques, principal component analysis, and anomaly detection methods (Geem et al., 2026a; Sarker, 2021; Shao et al., 2023). These methods differ in their assumptions, data requirements, interpretability, and ability to model non-linear relationships.

- “Linear regression” is a supervised regression method that predicts a continuous numerical output by fitting a linear relationship between input variables and the target variable.
- “Logistic regression” is a supervised classification method used to estimate the probability that an observation belongs to a particular class.
- “Decision trees” use rule-based splits to classify or predict outcomes, while “random forests” combine a large number of decision trees for improving the predictive accuracy and the robustness.
- “Support vector machines” and “support vector regression” are typically used the latter for classification and the former for regression, respectively, and can model non-linear relationships through kernel functions.
- “Artificial neural networks” are used to model complex non-linear relationships through interconnected layers of computational nodes.
- “Gradient boosting methods”, including Extreme Gradient Boosting, build models in succession, with each of the new models attempting to correct errors made by previous models.
- “k-means clustering” is used to group similar observations,
- “Principal component analysis” is used for dimensionality reduction, and
- “Isolation forest” is commonly used for anomaly detection.

Optimisation is closely related to ML but has a distinct role. In ML modelling, optimisation is used to estimate model parameters, tune hyperparameters, select features, or minimise prediction error. In engineering applications, optimisation may also support design search, calibration, inverse analysis, and risk-informed decision-making under constraints. However, optimisation results should still be checked against geotechnical mechanisms and practical design requirements, because a mathematically optimal solution may not be physically reasonable, constructible, or sufficiently robust for engineering use (Onyelowe et al., 2023; Zhang et al., 2022).

In geotechnical engineering, ML methods should be viewed as data-driven complements to engineering judgement rather than substitutes for geological and geotechnical interpretation. Their reliability depends on data quality, appropriate feature selection, model validation, and interpretation in the context of ground behaviour and engineering mechanisms.

3.2 Geotechnical engineering data

Geotechnical engineering relies on data collected to describe ground conditions and to support design, construction, and risk management decisions. Typical sources include borehole logs, trial pits, cone penetration tests (CPT), standard penetration tests (SPT), laboratory test results, geological mapping records, geophysical surveys, instrumentation and monitoring data, and construction observations (Gao, 2024; Phoon & Zhang, 2023). These data sources provide different levels of detail and reliability, and they are commonly combined during engineering interpretation.

Geotechnical data are strongly influenced by natural ground variability. Soil and rock conditions may change over short distances because of depositional history, weathering, faulting, groundwater variation, and anthropogenic disturbance. In addition, investigation data may be affected by sampling disturbance, test method limitations, inconsistent logging terminology, missing values, transcription errors, and differences between field and laboratory conditions (Gao, 2024; Phoon & Zhang, 2023). These issues are important for machine learning because the quality of a model depends on the quality, consistency, and representativeness of the data used to train it.

For machine learning applications, geotechnical data are usually converted into structured datasets. Input features may include depth, unit weight, water content, Atterberg limits, fines

content, SPT blow count, CPT cone resistance, sleeve friction, pore pressure, laboratory strength parameters, deformation parameters, rock quality designation, discontinuity descriptions, groundwater level, and monitoring measurements. Categorical information, such as soil type, weathering grade, geological unit, or construction stage, may also be included if it is encoded consistently. Typical target outputs include soil or rock classification, parameter prediction, settlement prediction, slope stability indicators, tunnel deformation trends, susceptibility classes, and risk categories (Gao, 2024; Geem et al., 2026a).

The spatial and site-specific nature of geotechnical data distinguishes it from many general machine learning datasets. A model trained on one geological setting may not generalise to another if the underlying ground conditions, investigation methods, or engineering controls differ substantially. For this reason, data provenance, site context, and the variety of conditions accounted for in the training dataset should be documented before model predictions are used for engineering interpretation (Phoon & Zhang, 2023).

3.3 Geotechnical parameters

Geotechnical parameters describe the physical, hydraulic, strength, and deformation behaviour of soil and rock materials. Common soil parameters include bulk density, dry density, water content, particle size distribution, plasticity index, undrained shear strength, effective cohesion, effective friction angle, compression index, permeability, and stiffness modulus. Common rock and rock mass descriptors include intact rock strength, rock quality designation, fracture spacing, weathering grade, discontinuity condition, groundwater condition, and rock mass classification values (Gao, 2024; Geem et al., 2026a).

These parameters are obtained from a combination of field investigation, laboratory testing, empirical correlations, back-analysis, and engineering judgement. Some parameters are measured directly, while others are inferred from index tests or correlations. For example, SPT and CPT results may be used to estimate strength, stiffness, density, or liquefaction-related indicators, depending on the soil type and the selected interpretation framework (Gao, 2024; Geem et al., 2026a). Such correlations are useful but uncertain, and their applicability depends on the geological setting, test quality, calibration data, and assumptions used in the original method.

Parameter variability is central to geotechnical risk assessment. Unlike manufactured materials, soil and rock are naturally variable and only partially observed through investigation points. A single design parameter may therefore represent an interpreted value derived from scattered measurements, geological judgement, statistical summaries, and design conservatism. Machine learning models can assist by identifying patterns across multiple variables, but the predicted outputs should still be checked against plausible engineering ranges and known site behaviour (Phoon & Zhang, 2023).

In a machine learning dataset, geotechnical parameters may act either as input features or as target variables. For example, CPT measurements, depth, and soil behaviour type may be used as inputs to predict undrained shear strength, while measured settlements or deformation rates may be used as target values for performance prediction. The distinction between input and target variables must be made carefully, because leakage can occur if a feature contains information that would not be available at the time of prediction.

3.4 Model training and evaluation

A typical machine-learning workflow involves data preparation, feature selection or feature engineering, division of the dataset into training and testing subsets, model fitting, validation, prediction, and performance assessment (Geem et al., 2026a; Sarker, 2021). Data preparation may involve unit correction, terminology standardisation, treatment of missing values, duplicate removal, categorical-variable encoding, and screening for unrealistic values. These steps are particularly important in geotechnical datasets because raw information may originate from different investigation campaigns, laboratories, logging conventions, or monitoring systems.

The training dataset is used to fit the model, while an independent testing dataset is used to evaluate performance on data not seen during training. Validation data or cross-validation may also be applied during model selection and hyperparameter tuning. Separating the data into training, validation, and testing subsets helps assess whether the model has learned relationships that can generalise, rather than simply memorising the training records (Geem et al., 2026a; Sarker, 2021).

Overfitting occurs when a model learns noise, local irregularities, or site-specific artefacts instead of the main relationship of interest. An overfitted model may perform well on the

training dataset but poorly on new or unseen data. Underfitting occurs when the model is too simple to capture the important relationships within the dataset. In geotechnical engineering, overfitting is a significant concern because datasets are often small, spatially correlated, imbalanced, or limited to a narrow range of geological conditions (Onyelowe et al., 2023; Phoon & Zhang, 2023).

For regression problems, common evaluation metrics include the “Mean Absolute Error” (MAE), or the “Root Mean Squared Error” ($RMSE$), and the “Coefficient of determination” (R^2) (Geem et al., 2026a). The mean absolute error is defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where MAE is the mean absolute error, y_i is the observed value, $\hat{y}_i$ is the predicted value, and with n representing the number of observations. The root mean squared error is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where $RMSE$ is the root mean squared error, and y_i , $\hat{y}_i$, and n have the same meanings as above. The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where R^2 is the coefficient of determination, and $\bar{y}$ is the mean of the observed values. A higher R^2 indicates that the model explains a greater proportion of the variance in the observed target variable, although it should not be interpreted without considering error magnitude, bias, and engineering relevance.

For classification tasks, commonly used metrics include accuracy, precision, recall, F1-score, and the confusion matrix (Geem et al., 2026a). Accuracy represents the proportion of predictions that are correct overall, while precision and recall describe different aspects of performance for each class. Precision is important when false positive predictions need to be limited, whereas recall is important when false negative predictions are more critical. The F1-score combines precision and recall into one metric:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where F_1 is the F_1 -score, *Precision* is the percentage of predicted positive cases that are correct, and *Recall* is the percentage of actual positive cases that are correctly identified. A confusion matrix summarises correct and incorrect predictions for each class and is particularly useful when classes are imbalanced.

Model evaluation in geotechnical engineering should include both statistical and engineering assessment. A model with favourable numerical metrics may still be unsuitable if it produces physically unrealistic predictions, relies on biased training data, performs poorly for critical minority classes, or is applied outside the range of the original dataset (Gao, 2024; Phoon & Zhang, 2023).

3.5 Relevance to geotechnical engineering

Machine learning is relevant to geotechnical engineering because many geotechnical decisions depend on interpreting complex, uncertain, and incomplete datasets. Machine learning methods can support exploratory data analysis, identification of patterns, prediction of parameters, classification of ground conditions, interpretation of monitoring trends, and preliminary geotechnical risk assessment. These applications are especially useful where multiple variables interact in nonlinear ways or where large volumes of investigation or monitoring data are available (Baghbani et al., 2022; Gao, 2024; Geem et al., 2026a).

However, geotechnical engineering is not only a statistical prediction problem. Ground behaviour is controlled by geological history, stress state, groundwater conditions, material fabric, drainage conditions, construction sequence, and scale effects. A model that performs well within a dataset may fail if it is transferred to a different geological environment or used outside the range of training data. Therefore, model outputs should be reviewed in relation to geological plausibility, known mechanisms, design assumptions, and consequences of error (Gao, 2024; Phoon & Zhang, 2023).

For geotechnical risk assessment, machine learning may assist with identifying zones of higher uncertainty, ranking susceptibility indicators, detecting unusual monitoring behaviour, or supporting classification of risk categories. The results should be used as decision-support

information rather than deterministic truth. Appropriate use requires clear documentation of assumptions, model limitations, data coverage, uncertainty, and the role of engineering judgement (Gao, 2024; Phoon & Zhang, 2023).

Reliability analysis provides a probabilistic framework for considering geotechnical uncertainty, including variability in soil or rock properties, measurement error, and uncertainty in model form or transformation relationships. Traditional reliability methods often estimate failure probability or reliability index using repeated deterministic analyses, such as Monte Carlo simulation, which can be computationally expensive for complex slope, excavation, or tunnel problems. ML surrogate models, including ANN, SVM, MARS, ELM, XGBoost, and CNN-based approaches, can improve computational efficiency by approximating the relationship between uncertain inputs and performance measures, but they require careful training, validation, and interpretation before being used in risk assessment (Zhang et al., 2023).

Interpretability is particularly important in engineering contexts because design and risk decisions must be explainable. Simple models, such as linear regression or shallow decision trees, may be easier to interpret but may not capture complex nonlinear behaviour. More complex models, like random forests, or gradient boosting methods, and neural networks, may improve predictive performance but can be harder to explain. Feature importance, partial dependence, and SHAP values are examples of tools that may support interpretation, although their assumptions and limitations must be understood before they are used in engineering decision-making (Geem et al., 2026b; Phoon & Zhang, 2023).

Therefore, machine learning should be considered as a supporting tool rather than a replacement for engineering judgement. Its value depends on data quality, appropriate model selection, transparent interpretation, validation against independent data where possible, and careful consideration of uncertainty and limitations (Geem et al., 2026b; Phoon & Zhang, 2023). Future work in geotechnical ML is likely to depend less on algorithm choice alone and more on the "data first, practice central" agenda: working with "ugly" real project data and small datasets, keeping geotechnical context central, using human-in-the-loop decision-making, improving explainability, and linking models to digital twins or monitoring systems where this adds practical value (Phoon & Zhang, 2023).

3.6 Classification of machine learning methods

Machine-learning methods can be grouped according to the type of learning task and the form of the available data. In geotechnical engineering, the main relevant categories are supervised learning, unsupervised learning, semi-supervised learning, and, in some applications, deep-learning methods. This distinction is important because each method type requires different data conditions and is suited to different engineering questions (Geem et al., 2026a; Sarker, 2021).

3.6.1 Supervised learning

Supervised learning applies when the dataset contains input variables together with known target values. The model uses these examples to learn how the input features relate to the target output. This is one of the most widely applied approaches in geotechnical machine learning, because many engineering problems involve estimating a known parameter or assigning a known class from available investigation, laboratory, or monitoring data (Geem et al., 2026a; Sarker, 2021).

Supervised learning is commonly separated into regression and classification. Regression is applied when the target variable is numerical and continuous. In geotechnical engineering, regression models may be used to predict undrained shear strength, friction angle, compression index, settlement, permeability, factor of safety, or deformation modulus. Classification is applied when the target variable represents a category. Examples include soil-type classification, rock-mass quality class, liquefaction susceptibility class, deformation warning level, or geotechnical risk category (Geem et al., 2026a).

The reliability of supervised learning depends strongly on the quality of target labels. In geotechnical datasets, target values may themselves contain uncertainty because they may be based on laboratory disturbance, empirical correlations, interpreted geological units, or judgement-based classes. This means that supervised model performance should be interpreted in relation to both model error and uncertainty in the underlying labels (Phoon & Zhang, 2023).

3.6.2 Unsupervised learning

Unsupervised learning is used when the dataset does not include a predefined target variable. Instead, the model attempts to identify patterns, clusters, or lower-dimensional structures within the data. These methods are often useful during exploratory data analysis, especially

where geological units, behavioural groups, or anomalous observations are not yet clearly defined (Geem et al., 2026a; Sarker, 2021).

In geotechnical engineering, unsupervised learning can support grouping of similar soil or rock samples, identification of data trends with depth, screening of monitoring data, and comparison between interpreted geological units and data-driven clusters. Clustering results should not be treated as direct geological truth; they require interpretation using borehole logs, geological mapping, engineering geology knowledge, and site context.

Common unsupervised learning approaches include clustering and dimensionality reduction. Clustering techniques, such as k-means clustering, organise data points into groups according to their similarity. Dimensionality-reduction techniques, such as principal component analysis, reduce the number of variables in a dataset while preserving the most important information (Geem et al., 2026a; Sarker, 2021).

In k-means clustering, the aim is to reduce the distance between each data point and its assigned cluster centre.

$$J = \sum_{j=1}^k \sum_{i=1}^{n_j} |x_i^{(j)} - c_j|^2$$

where J is the objective function, k defined as the number of clusters, n_j defined as the number of points assigned to cluster j , $x_i^{(j)}$ is the i th data point assigned to cluster j , c_j is the centre of cluster j , and $\|\cdot\|$ denotes the distance measure.

3.6.3 Deep learning

Deep learning is a branch of machine learning that uses artificial neural networks with several layers. These methods are able to represent complex nonlinear relationships and may be suitable for large datasets, image data, time-series records, and text-based information. In geotechnical engineering, deep learning can be applied to image-based soil or rock classification, tunnel-face interpretation, landslide susceptibility mapping, settlement forecasting, and monitoring-data analysis (Baghbani et al., 2022; Geem et al., 2026a; Montesinos López et al., 2022; Zhang et al., 2021).

A simple artificial neural network can be considered as a function that transforms input variables through weighted connections and activation functions:

$$\hat{y}=f(WX+b)$$

where X is the input data, W represents model weights, b is a bias term, f is an activation function, and $\hat{y}$ is the predicted output. In practice, deep learning models may contain many layers and parameters, which can increase predictive flexibility but also increase data requirements and reduce interpretability (Montesinos López et al., 2022).

Common deep-learning architectures used in geotechnical research include feedforward neural networks (FNNs), long short-term memory networks (LSTMs), recurrent neural networks (RNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs). FNNs are commonly used for structured parameter prediction, RNNs and LSTMs are appropriate for sequential or time-series monitoring data, CNNs are useful for image, spatial, and gridded datasets, and GANs may be used for synthetic data generation or image-based modelling where appropriate (Zhang et al., 2021; Zhang et al., 2022).

Although deep learning can provide strong predictive performance and feature-learning capability in suitable settings, its use in geotechnical engineering must be assessed carefully. Many geotechnical datasets are small, site-specific, irregularly sampled, and affected by uncertainty. Deep learning methods may therefore be inappropriate unless there is sufficient data, a clear prediction objective, robust validation, and a practical need for the added model complexity (Phoon & Zhang, 2023; Zhang et al., 2021).

3.7 Commonly used machine learning methods in geotechnical engineering

Several machine learning methods are commonly used in geotechnical engineering. The selection of method depends on the problem type, dataset size, feature quality, nonlinear behaviour, interpretability requirements, and intended engineering application. This section summarises the main methods relevant to the thesis (Baghbani et al., 2022; Geem et al., 2026a; Shao et al., 2023).

3.7.1 Linear regression

Linear regression is one of the simplest supervised learning methods. It assumes that the target variable can be represented as a linear combination of input variables. Although many geotechnical problems are nonlinear, linear regression is useful as a baseline model because it

is transparent, computationally simple, and relatively easy to interpret (Geem et al., 2026a; Sarker, 2021).

A multiple linear regression model can be written as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

where $\hat{y}$ is the predicted value, $x_1, x_2, \dots, x_p$ are input variables, p defined as the number of input variables, β_0 the intercept, and $\beta_1, \beta_2, \dots, \beta_p$ are regression coefficients. The coefficients provide a direct indication of the direction and magnitude of the relationship between the input variables and the target variable, subject to the assumptions of the model.

In geotechnical engineering, linear regression has commonly been used to develop empirical correlations between measured soil or rock properties and engineering parameters. Examples include the prediction of undrained shear strength from CPT or SPT data, estimation of compression index from index properties, settlement prediction, pile capacity estimation, permeability estimation, and correlations between rock strength and indirect testing results. Although these relationships may not fully capture complex nonlinear ground behaviour, linear models are useful because they provide transparent baseline predictions and allow engineers to understand the influence of each input variable.

3.7.2 Decision trees

Decision trees make predictions by splitting the data into branches based on input variables. They can represent nonlinear relationships and interactions between variables, and their rule-based structure can be easier to communicate than more complex models. In geotechnical engineering, decision trees may be used for soil classification, risk category prediction, or interpretation of parameter ranges (Geem et al., 2026a; Sarker, 2021).

The main advantage of decision trees is interpretability. However, single decision trees can overfit the training data if they are allowed to grow too deep or split on small numbers of samples. Controls such as maximum tree depth, minimum samples per split, and pruning can reduce overfitting (Geem et al., 2026a; Sarker, 2021).

3.7.3 Random forests

Random forests are ensemble models that combine the outputs of multiple decision trees to produce a final prediction. Each tree is trained using a different subset of the data and input

features. For regression tasks, the final output is usually calculated by averaging the predictions from the individual trees, while for classification tasks it is commonly based on majority voting. This structure helps reduce the instability of single decision trees and can improve overall predictive performance (Geem et al., 2026a; Sarker, 2021).

For regression, the random forest prediction can be expressed as:

$$y^{\wedge} = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

where B is defined as the number of trees, $T_b(X)$ is the prediction from tree b for input features X , and $y^{\wedge}$ is the final averaged prediction.

Random forests are useful for structured geotechnical datasets because they can model nonlinear relationships, handle mixed input variables after appropriate preprocessing, and provide feature importance measures. They may be applied to parameter prediction, susceptibility mapping, material classification, monitoring interpretation, and risk assessment. Feature importance results should be interpreted carefully because they can be affected by correlated variables, data imbalance, and the variety of values in the training dataset (Geem et al., 2026a; Shao et al., 2023).

3.7.4 Support vector machines

Support vector machines are supervised learning methods that can be applied to both classification and regression tasks. In classification, they aim to define a decision boundary that separates classes with the maximum possible margin. In regression, support vector regression predicts continuous values while allowing errors within a defined tolerance. Kernel functions can also be used to represent nonlinear relationships in transformed feature spaces (Baghbani et al., 2022; Geem et al., 2026a; Sarker, 2021).

Support vector machines can be effective for geotechnical datasets with complex class boundaries or limited sample sizes. They have potential applications in liquefaction assessment, slope stability classification, soil type classification, and geotechnical parameter prediction. However, model performance depends strongly on kernel selection, feature scaling, and hyperparameter tuning (Baghbani et al., 2022; Geem et al., 2026a; Shao et al., 2023).

3.7.5 Artificial neural networks

Artificial neural networks are algorithmic models made up of connected layers of computational nodes. Input data are processed through weighted connections and activation functions to generate an output. Neural networks are able to represent complex nonlinear relationships, which makes them useful for geotechnical prediction problems where input–output relationships cannot be easily described using simple equations (Baghbani et al., 2022; Montesinos López et al., 2022; Onyelowe et al., 2023).

In geotechnical engineering, artificial neural networks can be applied to settlement prediction, bearing-capacity estimation, slope-stability assessment, deformation forecasting, and the prediction of soil and rock parameters. Their main advantage is predictive flexibility, particularly where relationships are nonlinear and difficult to express using simple empirical equations. Their main limitations include reduced interpretability, sensitivity to model architecture and training settings, and the need for sufficient data to support reliable generalisation (Baghbani et al., 2022; Onyelowe et al., 2023; Yaghoubi et al., 2024).

3.7.6 Gradient boosting methods

Gradient boosting methods build predictive models by combining many weak learners, that is means simple prediction models that are not very accurate on their own, usually decision trees, in a sequential process. Each new separate tree attempts to reduce errors made by the previous trees. Examples include gradient boosting machines, XGBoost, LightGBM, and CatBoost (Geem et al., 2026a; Sarker, 2021).

These methods are often effective for structured tabular datasets and can capture nonlinear relationships and interactions between variables. In geotechnical engineering, they may be used for parameter prediction, classification, susceptibility assessment, settlement prediction, and performance forecasting. Although gradient boosting and other ensemble methods can provide high predictive accuracy, they require careful tuning and should be interpreted with consideration of data quality, overfitting risk, and engineering reasonableness (Geem et al., 2026a; Shao et al., 2023; Yaghoubi et al., 2024).

3.7.7 Clustering methods

Clustering methods are unsupervised learning techniques that group similar data points together. In geotechnical engineering, clustering may be used to identify groups of similar soil

or rock samples, support geological unit interpretation, screen monitoring behaviour, or explore patterns before supervised modelling (Geem et al., 2026a; Sarker, 2021).

Common clustering approaches include k-means clustering, hierarchical clustering, and density-based clustering. These methods can support exploratory data analysis, especially when the internal structure of the dataset is not fully understood. However, the resulting clusters should be reviewed using engineering judgement, because mathematical similarity does not always represent a meaningful geological or geotechnical unit (Geem et al., 2026a; Sarker, 2021).

3.8 Summary

The machine-learning methods most commonly applied in geotechnical engineering include:

- linear regression,
- decision trees,
- random forests,
- support vector machines,
- artificial neural networks,
- gradient boosting methods, and
- clustering techniques.

Simpler models are useful for transparency and baseline comparison, while more complex models may improve prediction performance where sufficient data are available (Baghbani et al., 2022; Geem et al., 2026a; Shao et al., 2023). Reviews comparing ANN, ML, DL, and ensemble learning indicate that no single method is universally best: suitable model choice depends on data quantity, data quality, nonlinear behaviour, interpretability needs, computational cost, and the geotechnical task being addressed (Yaghoubi et al., 2024).

For geotechnical applications, model selection should not be based only on statistical accuracy. The selected method should also be appropriate for the available data, interpretable where possible, consistent with geotechnical understanding, and suitable for the intended level of engineering risk (Geem et al., 2026b; Phoon & Zhang, 2023).

4. Data and Methodology

This chapter sets out the methodological framework used to prepare geotechnical data for machine learning. It explains why each stage of the workflow is required and how the approach is supported by geotechnical machine learning literature. The chapter is intentionally methodological: the detailed implementation, scripts, notebooks, inventory tables, model outputs, figures, and first UCS modelling results are presented in Chapter 5.

The workflow recognises that geotechnical datasets are usually heterogeneous, sparse, locally biased, and affected by uncertainty in sampling, testing, geological interpretation, and project context (Gao, 2024; Phoon & Zhang, 2023; Zhang et al., 2022). Machine learning can support geotechnical prediction and risk assessment only when the data are traceable, cleaned conservatively, evaluated with appropriate validation procedures, and interpreted using engineering judgement (Baghbani et al., 2022; Phoon & Zhang, 2023; Sarker, 2021; Zhang et al., 2023).

4.1 Exploratory data inventory

The first stage of the methodology is an exploratory data inventory. This stage is an initial audit of the available dataset before cleaning, feature selection, or modelling. In geotechnical machine learning studies, such an inventory is necessary because useful data may be distributed across factual reports, laboratory schedules, field investigation logs, monitoring systems, spatial layers, geological descriptions, and derived engineering summaries rather than stored in a single analysis-ready database (Baghbani et al., 2022; Gao, 2024; Zhang et al., 2022).

The inventory should identify the available data sources, the number of records, the main table types, the data formats, and the observation unit represented by each table. A row may represent a laboratory specimen, a borehole interval, a CPT interval, a monitoring timestamp, a pile test, a tunnel face, or a spatial grid cell. This distinction is important because combining incompatible observation units can create misleading relationships or data leakage. The inventory should also record metadata such as source file, report identifier, original table name, original column names, units, depth or coordinate reference, and whether each variable is directly measured, interpreted, or derived.

Completeness and consistency are also assessed during the inventory stage. This includes checking missing data, duplicate fields, inconsistent tables, repeated headers, unit rows, ambiguous column names, and possible extraction artefacts. Real-world datasets may contain ambiguous values, missing values, outliers, and irrelevant records, all of which can affect model training and interpretation (Sarker, 2021). In geotechnical applications, these issues are compounded by differences in field practice, laboratory procedures, geological terminology, and reporting conventions (Phoon & Zhang, 2023).

The purpose of the exploratory inventory is therefore to determine whether the available data are traceable and suitable for later machine learning preparation. It does not, by itself, produce modelling results.

4.2 Category-specific cleaning of data

Geotechnical data require category-specific cleaning because different data groups have different formats, meanings, and engineering constraints. Laboratory test results, field investigation data, monitoring records, spatial layers, and geological descriptions cannot be cleaned reliably using a single generic procedure. For example, a laboratory strength value, a borehole lithology description, a groundwater chemistry result, and a monitoring timestamp each require different checks before they can be used in a machine learning workflow.

Category-specific cleaning begins by preserving raw or source information. Source files should not be overwritten, and derived datasets should retain enough traceability to allow a cleaned value to be checked against its original record. This is particularly important for report-derived geotechnical data, where tables may contain merged cells, continuation headers, unit rows, footnotes, or text qualifiers. Conservative preservation of source fields is consistent with the broader need for transparent data preparation in geotechnical machine learning, where small cleaning decisions can materially affect model behaviour (Phoon & Zhang, 2023; Sarker, 2021; Zhang et al., 2022).

Geological and material descriptions require special care because they often contain inconsistent spelling, abbreviations, local terminology, mixed capitalisation, and variable descriptive detail. Standardising lithology, weathering grade, material type, and descriptive classes can improve model usability, but the cleaning should not remove meaningful geological distinctions. Similar descriptions may be grouped where the engineering meaning is equivalent,

while ambiguous terms should be retained for review rather than forced into an uncertain class. This is relevant to rock and soil classification tasks, CPT interpretation, tunnel-face assessment, and geotechnical risk screening, where categorical descriptions may act as important predictors or targets (Niu et al., 2024; Rauter & Tschuchnigg, 2021; Zhou et al., 2021).

Numerical cleaning includes converting values to numeric format where they are clear scalar measurements, removing text symbols where this does not change the meaning, checking units, identifying impossible values, and preserving qualified values where the meaning is uncertain. Examples include values reported with inequality symbols, ranges, footnotes, or non-standard unit labels. Unit checks are essential because geotechnical variables may be recorded in MPa, kPa, metres, millimetres, percentages, density units, or project-specific depth references. Suspected impossible values should be flagged and checked against the source record rather than automatically removed, because unusual values may represent either data errors or genuine ground variability.

Cleaning decisions should be documented by field category. Laboratory parameters may require range checks and unit conversions; field investigation logs may require depth-interval checks and borehole-level consistency checks; monitoring records may require timestamp ordering and sensor-quality checks; spatial data may require coordinate reference and resolution checks; and geological descriptions may require terminology standardisation. Bui et al. (2018) and Goetz et al. (2015) illustrate the need for careful handling of spatial and geological conditioning factors in geohazard prediction, while Rauter and Tschuchnigg (2021) demonstrate the importance of preprocessing and outlier review in CPT-based classification.

4.3 Parameter summary and data suitability assessment

After category-specific cleaning, each variable should be summarised before it is considered for modelling. Parameter summary provides a structured view of data availability, missingness, distribution, and plausibility. It helps distinguish variables that are suitable for use as model inputs from variables that are too sparse, inconsistent, or uncertain for reliable modelling.

For numerical variables, useful summaries include valid count, missing count, missing percentage, minimum, maximum, mean, median, standard deviation, quartiles, and possible outlier indicators. For categorical variables, useful summaries include valid count, missing percentage, number of unique categories, most frequent category, and category frequencies.

These summaries are descriptive rather than conclusive. They provide evidence for later feature-selection decisions, but they do not replace engineering judgement.

Missingness must be interpreted in relation to the geotechnical context. A value may be missing because the test was not scheduled, the result was not applicable to the material, the sample was unsuitable, the result was rejected, or the information was lost during report extraction. Rows with missing target values cannot be used for supervised training of that target, although they may remain useful for exploratory analysis or other modelling tasks. Sparse variables may still be engineering-important, but their use must be justified carefully because high missingness can reduce model reliability and produce biased training datasets (Ozsagir et al., 2022; Phoon & Zhang, 2023).

Parameter summaries also support plausibility checks. Statistical outliers should be assessed alongside engineering range checks, material descriptions, units, and source records. Previous geotechnical machine learning studies have used outlier and missing-value treatment in different ways depending on the dataset and modelling purpose. W. Zhang et al. (2021) used interquartile-range screening in an undrained shear strength database, Rauter and Tschuchnigg (2021) removed measurement errors and outliers from a CPT classification dataset, and Ozsagir et al. (2022) reviewed boxplot outliers but did not remove them where they were not extreme. These examples support a transparent, case-specific approach rather than automatic deletion.

4.4 Feature selection and target definition

Feature selection involves identifying the input parameters that are useful for a machine-learning model. In geotechnical engineering, this selection should consider engineering relevance, data availability, measurement reliability, correlation structure, redundancy, leakage risk, and whether the variables would be known at the time a prediction is made. Baghbani et al. (2022) emphasised that selecting appropriate input parameters is a critical step in geotechnical artificial-intelligence models, while Sarker (2021) described feature selection as a method for reducing unnecessary complexity, excluding weak variables, and limiting overfitting.

Engineering relevance means that a feature should have a plausible relationship to the target or to the process being modelled. For example, depth, index properties, density, stress state, lithology, weathering grade, groundwater condition, or rock mass quality may be relevant

depending on the prediction task. However, relevance alone is not sufficient. A variable with strong engineering meaning may be unsuitable for modelling if it is too sparse, inconsistently measured, poorly standardised, or unavailable in the same records as the target.

Redundancy and leakage must also be reviewed. Highly correlated variables may provide overlapping information and may increase model complexity without improving generalisation. Leakage occurs when an input contains information that would not be available at prediction time or directly encodes the target. For example, a derived parameter calculated from the target, a same-test companion value, or a post-event observation may produce optimistic model performance that would not transfer to real decision-making. This is especially important in geotechnical workflows, where interpreted parameters and derived classifications may embed engineering assumptions.

The first target in this thesis is unconfined compressive strength (UCS), because it is an engineering-relevant intact rock strength parameter and has suitable data availability for an initial supervised regression workflow. The detailed UCS feature-selection results are presented in Chapter 5. The same workflow can later be extended to other geotechnical targets where suitable data are available, including RQD, TCR, weathering grade, permeability, groundwater-related indicators, and borehole-unit classifications. These extensions would support a broader transition from sample-level strength prediction toward rock-mass and geotechnical risk screening.

4.5 Machine learning model development and training

Model development begins after the target and candidate features have been defined. The first task is to construct a modelling table in which each row represents a consistent observation unit, each input feature is available or treated in a documented way, and the target variable is clearly defined. Identifier and traceability fields, such as source file, report ID, borehole ID, or sample ID, should normally be retained outside the model input matrix unless there is a justified engineering reason for including them.

Categorical variables must be encoded before most machine learning algorithms can use them. One-hot encoding is suitable for unordered categories such as lithology or material group where no ordinal relationship is assumed. Ordinal encoding may be appropriate for ordered classes such as weathering grade or risk category, provided the ordering is meaningful and

documented. Numerical variables may require scaling for algorithms that are sensitive to feature magnitude, such as support vector machines, neural networks, regularised linear models, and distance-based methods. Tree-based methods are generally less sensitive to scaling, but preprocessing should still be documented (Rauter & Tschuchnigg, 2021; Sarker, 2021).

Data splitting is needed to assess how well a model can generalise to new data. A common approach is to separate the dataset into training and testing subsets. The training subset is used to fit the model, while the testing subset is kept aside for performance assessment. Some geotechnical studies use 70/30 or 80/20 train-test splits, while others apply k-fold cross-validation when datasets are small (Amjad et al., 2022; Bui et al., 2018; Pham et al., 2018; W. Zhang et al., 2021; Zhu et al., 2021). The most suitable approach depends on dataset size, class balance, spatial grouping, temporal order, and the intended engineering application.

Where observations are spatially grouped by borehole, site, geological unit, or project, random splitting may overestimate model generalisation because nearby observations are not independent. Spatial, grouped, or time-aware validation may be required for datasets with strong spatial or temporal structure. Goetz et al. (2015) demonstrated the importance of spatial cross-validation in landslide-susceptibility modelling, and Hu et al. (2019) used time-aware training and testing for tunnel-settlement forecasting. In this thesis, any splitting strategy should be documented with the random seed, grouping assumptions, selected features, preprocessing steps, and model settings.

Training involves fitting the selected algorithms to the training dataset, producing predictions, and saving the trained model files and prediction outputs. Preprocessing should be fitted only on the training data and then applied to the testing data through a reproducible pipeline. This reduces the risk of information leakage from the testing set into scaling, imputation, or encoding parameters. The specific algorithms and saved model outputs used in the thesis implementation are reported in Chapter 5.

4.6 Model performance evaluation and comparison

Trained models should be assessed using unseen testing data or an appropriate validation procedure. This stage evaluates whether the model can generalise beyond the dataset used during fitting. For regression problems, including UCS prediction, commonly used

performance measures include root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination, R^2 . RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where n is the number of observations, y_i is the measured value, and $\hat{y}_i$ is the predicted value. RMSE is expressed in the same units as the target and gives greater weight to larger errors. MAE is defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

MAE is also expressed in the target units and is easier to interpret as an average absolute prediction error. The coefficient of determination, R^2 , indicates the proportion of target variance explained by the model predictions, although it should not be interpreted alone because a high R^2 does not guarantee acceptable engineering error.

Different models should be compared using the same data split, target definition, feature set, and performance metrics. Otherwise, differences in performance may reflect changes in the dataset rather than differences in model capability. Train-test performance gaps should also be examined because a model that performs very well on training data but poorly on testing data may be overfitted, whereas a model that performs poorly on both may be underfitted. The detailed model comparison table and UCS results for this thesis are presented in Chapter 5.

4.7 Error analysis and diagnostic interpretation

Performance metrics summarise model accuracy, but they do not explain where predictions fail. Error analysis is therefore required to examine model reliability, limitations, and possible engineering causes of poor prediction. This stage is especially important in geotechnical engineering because rare or poorly represented ground conditions may be more important for risk assessment than the average prediction error.

Predicted-versus-actual plots are used to check whether predictions follow the measured values across the target range. Residual plots show whether errors are biased, heteroscedastic, or

associated with particular predicted values, measured values, or input features. Large-error cases should be inspected separately because they may indicate data-quality issues, unusual geological conditions, unit errors, sparse feature coverage, or limitations in the selected model. Outliers should be interpreted with both statistical diagnostics and geotechnical judgement.

Error trends can also be examined against key features such as depth, density, moisture content, lithology, weathering grade, groundwater condition, or rock mass indicators. Feature influence methods, including tree-based feature importance, permutation importance, sensitivity checks, or SHAP-based interpretation, may help explain which inputs are influencing predictions, but their results should be interpreted as model- and dataset-specific rather than universal measures of geotechnical importance (Abdollahi et al., 2024; Matin et al., 2018; W. Zhang et al., 2021; Zhu et al., 2021).

The purpose of diagnostic interpretation is not only to rank models but also to understand when a model should not be trusted. If errors concentrate in high-strength samples, rare lithological groups, deeper intervals, weathered zones, or groundwater-affected conditions, the model may require additional features, more representative data, grouped validation, or a narrower domain of application. The detailed residual plots, large-error tables, and diagnostic observations from the initial UCS workflow are reported in Chapter 5.

4.8 Engineering interpretation and workflow limitations

Machine learning outputs must be interpreted within the broader geotechnical engineering context. A model may perform statistically well while still being limited by sparse data, site-specificity, uncertain units, non-independent observations, incomplete geological representation, or poor performance on rare but important cases. These limitations are central in geotechnical applications because engineering decisions depend on geological understanding, ground variability, uncertainty, and consequence, not only on statistical fit (Phoon & Kulhawy, 1999; Phoon & Zhang, 2023; Zhang et al., 2023).

The engineering interpretation should ask whether the model behaviour is plausible. Predictions should be checked against known geological conditions, expected material behaviour, laboratory and field test limitations, and the range of the training data. A model should not be applied outside the domain represented by its training data without additional validation. If the target is an interpreted class, empirical rating, or risk category, the model

should be described as learning that interpretation framework rather than independently verifying ground truth.

For geotechnical risk assessment, particular care is required for rare but high-consequence conditions. A model that performs well on common cases may still be unsuitable if it fails to identify weak zones, high groundwater inflow potential, problematic lithologies, karstic features, or other critical ground hazards. Risk-related use also requires clear communication of uncertainty, assumptions, data gaps, and the difference between screening support and design verification. Guidance on geotechnical risk assessment similarly emphasises structured hazard identification, consequence assessment, and professional judgement rather than reliance on a single numerical output (Transport for NSW, 2014).

The initial thesis implementation focuses on UCS prediction because it provides a practical first supervised regression target. However, UCS is an intact sample-scale property and does not represent complete geotechnical risk. The workflow remains general and can be extended to RQD, TCR, weathering grade, permeability, groundwater-related indicators, and borehole-unit classifications where suitable data are available. Such extensions would allow later work to connect sample-level strength behaviour with rock-mass condition and risk-related screening more directly.

4.9 Reproducibility and data storage

Reproducibility is maintained by separating raw or source data, cleaned data, processed modelling datasets, scripts, notebooks, figures, tables, and model results. Source files should be preserved unchanged. Cleaned and processed datasets should be generated through documented scripts or notebooks so that each output can be recreated from the same inputs and assumptions.

A reproducible geotechnical machine learning workflow should record source-file paths, cleaning assumptions, unit conversions, excluded records, missing-value treatment, feature definitions, encoding rules, scaling methods, train-test splits, random seeds, model settings, package versions, and output locations. Version-controlled scripts and documented notebooks support traceability and reduce the risk that later model results become disconnected from the data-preparation decisions that produced them.

Outputs should be stored in a structured way, with figures, tables, and model artefacts separated from raw and cleaned data. Model evaluation outputs should state the dataset version, target variable, selected features, preprocessing steps, algorithm, hyperparameters, data split, metrics, and any important limitations. This documentation is particularly important in geotechnical engineering because datasets are often sparse, heterogeneous, and project-specific, and because model outputs must remain interpretable as decision-support evidence rather than uncontrolled numerical artefacts (Phoon & Zhang, 2023; Sarker, 2021; Zhang et al., 2022).

5. Machine Learning Workflow

This chapter describes the machine learning workflow that was used in the thesis and provides the methodological link between the factual Doha geotechnical dataset, feature screening, model development, performance evaluation, diagnostic review, and engineering interpretation. The first implementations are centred on Unconfined Compressive Strength (UCS) and Point Load Strength (Is_{50}) predictions, with later workflow stages also examining selected strength or stiffness related parameters where sufficient paired data are available. The results are therefore interpreted as screening-level outputs from a laboratory dataset, not as direct predictions of complete geotechnical risk.

The dataset used in this thesis was compiled from factual geotechnical investigation reports from engineering projects in Doha, Qatar, in which the author participated professionally. These reports provided real project-based borehole, lithological, depth-related, and laboratory testing data, rather than a synthetic or benchmark dataset. As a result, the dataset reflects the practical characteristics of factual project records, including uneven parameter availability, variable reporting formats, missing paired data, and differences in sampling density between lithological units. The dataset is therefore interpreted as a project-derived laboratory dataset suitable for screening-level machine learning assessment. It is set within the geological and geotechnical context of Doha, Qatar, where the shallow engineering geology is dominated by gently dipping to sub-horizontal Palaeogene to Quaternary deposits. The main bedrock sequence relevant to this study comprises carbonate and evaporitic formations, including the ‘Simsima Limestone’ of the ‘Upper Damman’ Formation, the ‘Midra’ Shale of the ‘Lower Damman’ Formation, and the underlying ‘Rus’ Formation, represented in engineering terms by ‘Rus Limestone’ and ‘Rus Gypsum’ units. These units are important because the shallow rock mass is affected by lithological succession, discontinuities, weathering, groundwater flow, and karstic or dissolution features, with permeability governed more by discontinuities, karst, and weathering than by primary porosity alone (Karagkounis et al., 2016).

The Simsima Limestone requires caution because it may not behave as a uniform engineering material. Complex weathering patterns, variable dolomitisation, dissolution cavities, and secondary infill by weakly cemented carbonate or argillaceous material can influence rock-mass strength and stiffness (Fourniadis, 2010). For this thesis, these conditions justify treating

lithology, depth, weathering-related descriptors, density, moisture, and related laboratory or logging fields as contextual variables when interpreting UCS and other geotechnical parameters. They also explain why lithology-based grouping and error review are useful in the machine-learning workflow: the models are tested against data that reflect a variable geological system rather than a single homogeneous rock type.

Table 5.1 Summarises the main lithological units considered in the Chapter 5 dataset interpretation.

Lithological unit	General description
Simsima Limestone (SL)	Upper Dammam carbonate unit and main shallow founding stratum in Doha; variably dolomitic, weathered, fractured, and locally affected by dissolution features.
Midra Shale (MSH)	Lower Dammam shale-bearing unit that commonly acts as an aquitard between the Simsima Limestone and Rus Formation.
Dukhan Limestone (DL)	Thin Lower Dammam limestone unit, described in the Doha stratigraphy as white, compact, fossiliferous chalky limestone with reported thickness of approximately 1 m.
Rus Limestone (RUS)	Carbonate facies of the Rus Formation, including weak carbonate marls, siltstones, thin limestone, and chalky materials.
Rus Gypsum (RG)	Evaporitic sulfate facies of the Rus Formation, dominated by gypsum and anhydrite with marl and thin limestone.

The computational workflow was implemented in **Python**, with **Jupyter Notebook** used as the main analysis platform and **Visual Studio Code** used for code editing and project management. **GitHub** was used as the project repository and version-control platform for tracking scripts, notebooks, data-processing outputs, figures, and table-generation work. **OpenAI Codex** was used as a coding agent to support Python scripts, notebook development, data-processing routines, figure and table generation, and review and refinement of outputs. The main Python libraries were **pandas** and **numpy** for data handling and numerical operations, **matplotlib** for plotting and graphical outputs, and **scikit-learn** for train-test

splitting, preprocessing, model training, prediction, and performance evaluation. These tools supported the staged Chapter 5 workflow from exploratory data inventory, category-specific cleaning, parameter summary, feature selection, model development and training, and performance evaluation through to error analysis and engineering interpretation.

5.1 Workflow overview

The purpose of the workflow is to keep the analysis reproducible, transparent, and suitable for geotechnical engineering interpretation. The workflow is organised into the following stages, each of which is discussed in a separate section below:

1. Exploratory data inventory, which is the first audit of the dataset to understand what data exists, how complete it is, and whether it is useful for machine learning. Focus on number of records, missing data, duplicate fields etc.
2. Category-specific cleaning of geotechnical field and laboratory data, cleaning each type of data in the way that best suits its format and meaning. For example, group similar descriptions, such as lithology or weathering terms, standardise labels, spelling, abbreviations, convert to numbers, remove text symbols, check units, identify impossible values.
3. Parameter summary, which shows what each variable looks like in the dataset and whether it has enough reliable data for analysis. Helps to explain which parameters are suitable as model inputs and which parameters may be too incomplete, sparse, or inconsistent to use confidently. Focus on statistical parameters like count, missing percentage, minimum, maximum, mean, median, and standard deviation.
4. Feature selection, which means choosing the most useful input parameters for the machine learning model. For example, deciding which geotechnical variables should be used to predict the target parameter, such as UCS.
5. Machine learning model development and training, which is the stage where selected geotechnical parameters are used to build and train a model that can predict a target parameter such as UCS. It includes steps like, data preparation and splitting into training and testing sets, model training using the training set of data, generating predictions, saving outputs.

6. Model performance evaluation and comparison, which is the stage where the trained models are tested on the unseen testing set of data, their prediction errors are measured, and the best-performing model is identified. Comparison of different models based on metrics like RMSE, MAE, and R^2 .

7. Error analysis and diagnostic interpretation, which is the stage where the model errors are examined to understand the model's reliability, limitations, and possible causes of poor predictions. For example, in this stage it is examined predicted vs actual values, large errors or outliers, error trends, feature influence that checks how the input parameters affect the predictions.

8. Engineering interpretation and workflow limitations, finally the outcomes need to be checked if they make engineering sense. Discussing what the model results mean for geotechnical engineering assessments and clearly stating what the model cannot yet do reliably.

The initial implementation focuses on UCS prediction because UCS had suitable data availability and engineering relevance for the first modelling attempt. However, the workflow structure remains general so that future work can extend the feature-selection and modelling stages to other geotechnical parameters, such as RQD, TCR, weathering grade, permeability, groundwater-related indicators, or borehole-unit classifications where suitable data are available.

5.2 Stage 1: Exploratory data inventory

The purpose of this stage is to document the structure, completeness, duplication status, and target availability of the dataset before any further cleaning, feature selection, imputation, or model training. This stage should therefore be interpreted as an initial dataset audit. It supports transparent workflow decisions but does not perform additional cleaning, feature selection, imputation, or model training.

The first workflow application used the following prepared dataset `data_cleaned/rock_lab_cleaned.csv`. This file was prepared before Stage 1 through extraction, consolidation, and initial cleaning of the rock laboratory test tables. Stage 1 did not create a new cleaned modelling dataset. Instead, it created an inventory of the prepared dataset,

including the data inventory, missingness summary, duplicate summary, target availability summary, and Markdown summary. These outputs were saved in `outputs/tables`.

The dataset contained 5,638 records and 67 columns, comprising 39 numerical columns and 28 categorical or text columns. The target availability summary showed that UCS was available for 1,102 records, making it the most suitable initial supervised modelling target within this dataset. Point load index Is_{50} was available for 2,672 records and missing from 2,966 records (52.61% missing). It was not selected as the target for the first workflow application, but its comparatively higher availability than several other strength and index-property parameters made it suitable for later review as an additional modelling target.

The following additional workflow application, WD2, was therefore added to test whether the same staged workflow could also support prediction of Is_{50} . WD2 used the same extracted Doha rock laboratory data source, but the later feature-screening and modelling stages used a workflow-specific copied dataset:

`data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv`

The WD2 target was defined as the cleaned numerical field Is_{50} . The reviewed WD2 input retained 5,638 records, of which 2,619 records had valid Is_{50} values, and 3,019 records were missing the target (53.55%). The reduction from the Stage 1 inventory count of 2,672 valid Is_{50} records reflects the Stage 3 review, in which 53 lithology-grouped extreme Is_{50} values were set to missing before WD2 feature screening. The main WD2 candidate fields reviewed for target-paired availability were `depth`, `lithology`, `bulk_density`, and `moisture_content`, with `depth` and `lithology` retained for the later WD2 workflow. In the Is_{50} valid subset, `depth` was available for 2,618 records and missing from 1 record, while `lithology` was available for 1,689 records and missing from 930 records. The `lithology` field contained 7 unique values in the Is_{50} valid subset, with `RUS` as the dominant recorded category (845 records, 50.03% of valid lithology records). These values show that WD2 had useful depth coverage but incomplete and uneven lithology coverage.

The WD2 inventory also identified important sparsity limitations. The WD1-equivalent four-predictor set of `depth`, `lithology`, `bulk_density`, and `moisture_content` produced only 14 complete target-valid records before imputation, so it was not suitable as the main WD2 feature set. UCS was also excluded from the main WD2 predictor set so that the workflow

remained a standalone Is_{50} prediction exercise rather than a direct UCS- Is_{50} conversion. Detailed WD2 feature-selection, modelling, and diagnostic outcomes are reported in the later workflow stages rather than in this exploratory inventory section.

Table 5.2 Summarises the dataset, source data, prepared dataset, and target parameter used in each workflow application.

Workflow ID	Dataset description	Source/input data	Prepared dataset used	Target parameter
WD1	Rock laboratory data from Doha	Extracted rock laboratory test tables from Doha factual reports	data_cleaned/rock_lab_cleaned.csv	UCS
WD2	Rock laboratory data from Doha	Extracted rock laboratory test tables from Doha factual reports	data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv	Is_{50}

5.3 Stage 2: Category-specific cleaning

The purpose of this stage is to standardise dataset values in a traceable way before later parameter summary, feature selection, and modelling. Cleaning is applied according to field type and engineering meaning, rather than by applying one generic rule to all columns. This stage preserves original engineering context where required for audit and interpretation.

For the first workflow application, Stage 2 was applied to the cleaned rock laboratory dataset used in Stage 1 data_cleaned/rock_lab_cleaned.csv. The stage produced the dataset data_cleaned/rock_lab_stage_2_category_cleaned.csv.

Obvious missing-value markers and blank text values were standardised. Clear scalar numerical values in depth, strength, and laboratory parameter fields were converted to numerical format where appropriate. Ambiguous values, ranges, inequalities, and non-scalar entries were retained through raw or note fields and flagged where required. Lithology descriptors were cleaned as geological descriptor fields. The original lithology text was retained in the added lithology_raw column, while clear inconsistencies in the cleaned lithology field, such as case, spacing, abbreviation, and spelling differences, were standardised where the meaning was unambiguous. No rows were deleted simply because values were

missing, and suspicious or outlying values were flagged for later engineering review rather than automatically removed.

Table 5.3 Summarises the input dataset, and output dataset for Stage 2 in each workflow application.

Workflow ID	Dataset description	Input dataset	Output dataset	Target parameter
WD1/WD2	Rock laboratory data from Doha	data_cleaned/rock_lab_cleaned.csv	data_cleaned/rock_lab_stage_2_category_cleaned.csv	UCS/ I_{s50}

5.4 Stage 3: Parameter summary

The purpose of this stage is to summarise parameter availability, missingness, and basic distribution before feature selection and modelling. Numerical and categorical fields are reviewed according to their data type and engineering role. This stage also supports the identification of clearly inconsistent or extreme values before the dataset is used in later workflow stages.

For the first workflow application, Stage 3 was applied to the category-cleaned dataset produced by Stage 2 `data_cleaned/rock_lab_stage_2_category_cleaned.csv`. The stage produced parameter summary tables, descriptive statistics, histograms, and parameter-versus-depth plots using the Stage 3 Python script and the exploratory Jupyter notebook. Numerical parameters were summarised using count, missing count, missing percentage, mean, median, standard deviation, minimum, maximum, quartiles, and interquartile range. Categorical parameters were summarised using valid count, missing count, missing percentage, unique-value count, most common value, and frequency.

The descriptive statistics and graphical checks identified isolated extreme values in parameters such as point load index, moisture content, bulk density, Young's modulus, Poisson's ratio, and porosity. These values were reviewed as potential outliers, unit inconsistencies, or extraction errors. Where values were clearly inconsistent with the expected engineering range of the parameter, they were removed at value level from the reviewed dataset, while the original Stage 2 dataset was retained for traceability. The reviewed output dataset from this stage was saved as:

data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv

The Stage 3 summary tables were saved under:

outputs/tables/stage_3_parameter_summary.

The outlier review tables were saved under:

outputs/tables/stage_3_outlier_review.

The graphical outputs were saved under:

outputs/figures/stage_3_parameter_review and

outputs/figures/stage_3_lithology_depth_plots

This stage did not perform feature selection, train/test splitting, or model training. Its output dataset was used as the reviewed input dataset for the following stage of the workflow.

Table 5.4 Summarises the input dataset, output dataset, summary outputs, and target parameter used for Stage 3 in each workflow application.

Workflow ID	Dataset description	Input dataset	Output dataset/ summaries	Target parameter
WD1/WD2	Rock laboratory data from Doha	data_cleaned/rock_lab_stage_2_category_cleaned.csv	data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv; summary tables outputs/tables/stage_3_parameter_summary outlier review tables in outputs/tables/stage_3_outlier_review figures in outputs/figures	UCS/ I_{s50}

Selected parameter-versus-depth plots grouped by lithology are presented below to support the Stage 3 graphical review. The comparisons with published geological descriptions are descriptive only and do not provide formal stratigraphic validation.

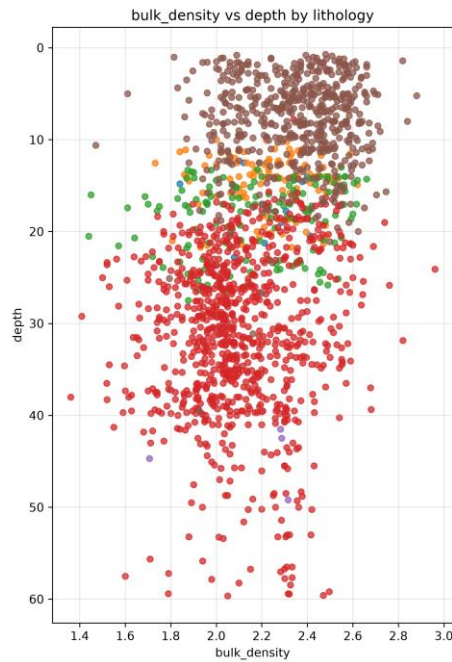


Figure 5.1 Bulk density versus depth grouped by lithology.

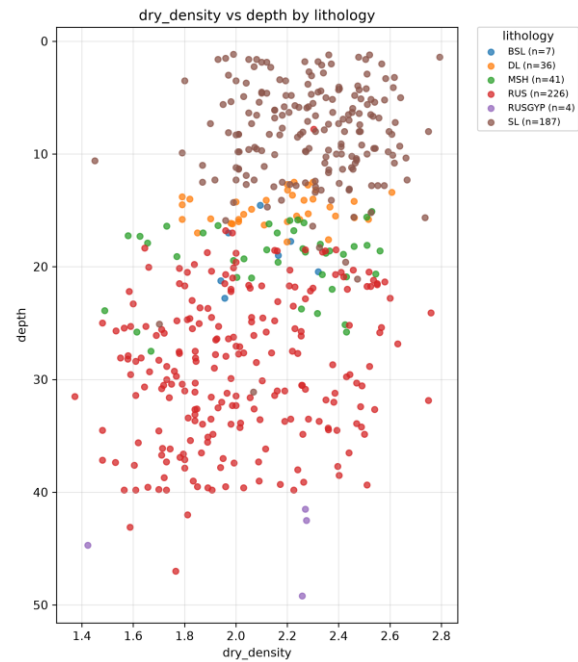


Figure 5.2 Dry density versus depth grouped by lithology.

The Figure 5.1 plot shows bulk density against depth for the WD1 lithology groups. Simsima Limestone (SL) is concentrated mainly in the shallower part of the profile, while Rus Formation (RUS) records dominate the deeper interval, with Midra Shale (MSH) and Dukhan Limestone (DL) mostly occupying intermediate depths. This broad lithology-depth separation is compatible with the Doha succession described by Karagkounis et al. (2016), while the scatter within SL is broadly consistent with the reported variability of Simsima Limestone due to dolomitisation, weathering, and karstic dissolution (Fourniadis, 2010; Karagkounis et al., 2016). The plot should be interpreted cautiously because density values overlap between lithologies and the sample counts are uneven.

The figure 5.2 plot shows dry density against depth. The same broad vertical ordering visible in the bulk density plot is present, with SL mostly shallow, RUS mostly deeper, and DL and MSH concentrated between these intervals. This is broadly consistent with the Doha formation sequence described by Karagkounis et al. (2016), while the within-unit spread is compatible with the variable engineering behaviour described for Simsima Limestone and weathered carbonate materials (Fourniadis, 2010; Karagkounis et al., 2016). The smaller number of dry

density records and the overlap between lithologies mean that this figure should not be used to infer boundaries.

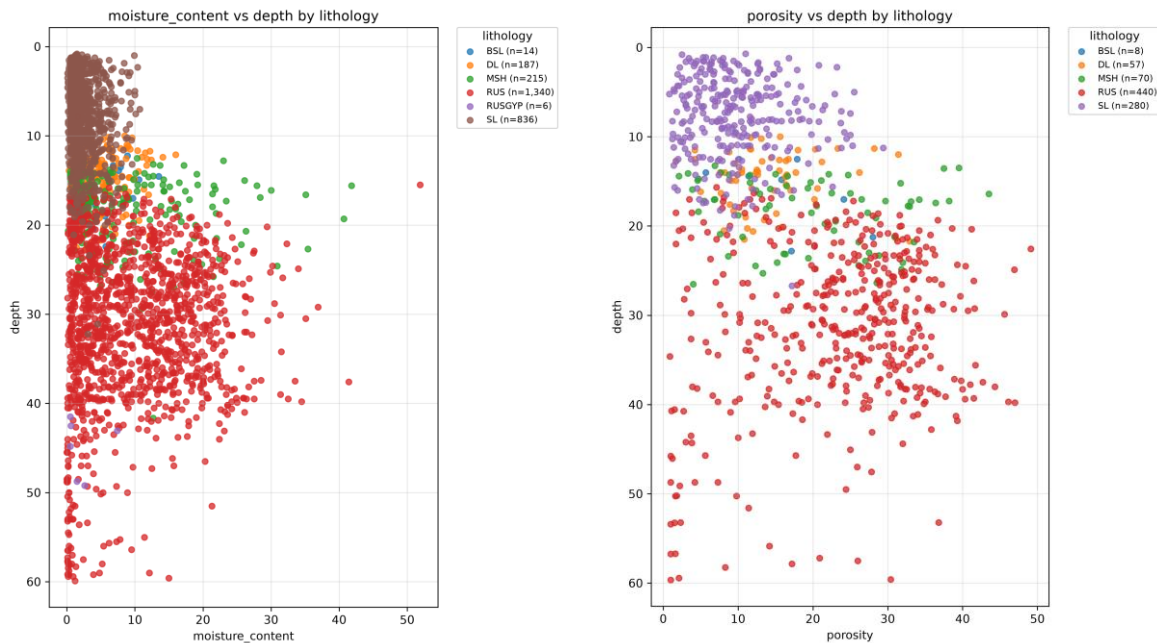


Figure 5.3 Moisture content versus depth grouped by lithology. **Figure 5.4 Porosity versus depth grouped by lithology.**

The Figure 5.3 plot shows moisture content against depth. SL has many low-moisture values in the shallow interval, while MSH and RUS include wider moisture-content scatter, including some higher values at intermediate to greater depths. This is compatible with the description of Midra Shale as a shale-bearing aquitard and with the reported influence of discontinuities, weathering, karst features, and groundwater flow on Doha rock masses (Karagkounis et al., 2016). The pattern is not a simple depth trend and should be interpreted cautiously because moisture content may reflect sampling condition, local groundwater behaviour, and uneven lithology representation.

The Figure 5.4 plot shows porosity against depth. SL and RUS occupy distinct depth ranges overall, while MSH and DL overlap the transition between the shallow and deeper groups; RUS also shows a broad porosity spread. This is broadly consistent with the described formation sequence and with the geological variability attributed to weathering, karstification, dolomitisation, and interlayered materials in Doha rock masses (Fourniadis, 2010; Karagkounis et al., 2016). However, Karagkounis et al. (2016) note that rock-mass permeability is governed

more by discontinuities, karst, and weathering than by primary porosity, so this plot should not be over-interpreted as a hydraulic indicator.

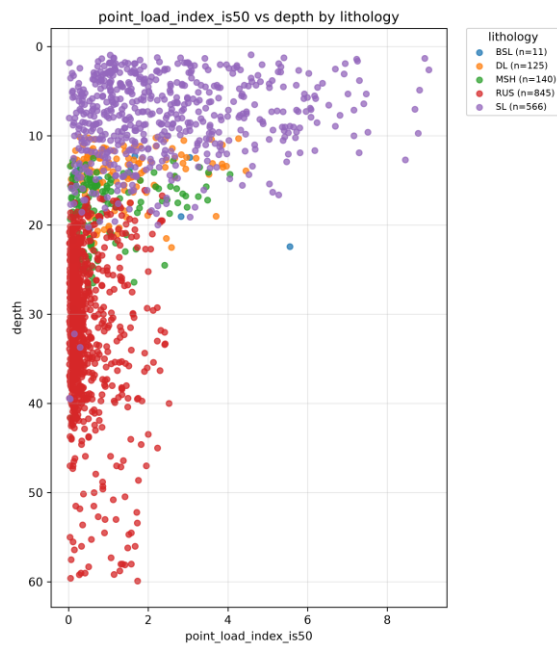


Figure 5.5 Point load index I_{s50} versus depth grouped by lithology.

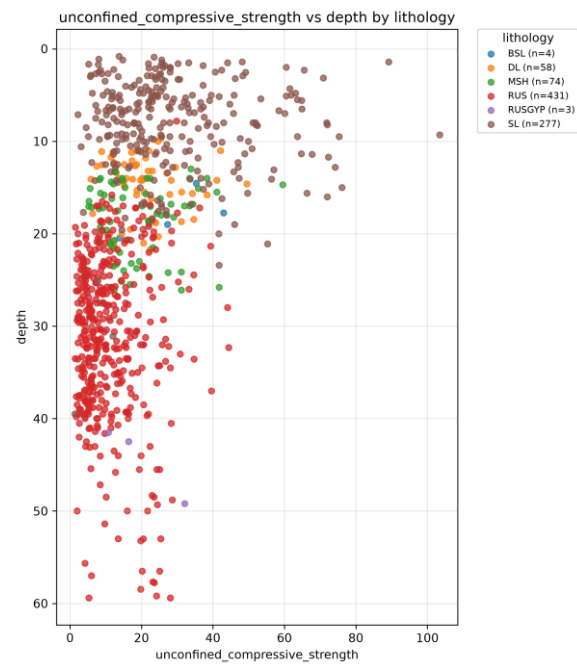


Figure 5.6 Unconfined compressive strength versus depth grouped by lithology.

The Figure 5.5 plot shows point load index I_{s50} against depth as an index-strength measure. SL spans a wider range and includes higher I_{s50} values, whereas RUS is mostly deeper and concentrated at lower values, with MSH and DL generally between these groups. This supports the geological distinction between lithological units and is broadly consistent with the source descriptions that emphasise variable Simsima Limestone rock quality and differing strength/stiffness behaviour across Doha formations (Fourniadis, 2010; Karagkounis et al., 2016). The figure remains exploratory because point-load testing is unevenly distributed and does not by itself validate formation layering.

The Figure 5.6 plot shows unconfined compressive strength (UCS) against depth. SL includes a wide UCS range and several higher-strength values in the shallower interval, whereas RUS is mostly deeper and commonly lower in UCS, with MSH and DL occupying intermediate depth and strength ranges. This is broadly consistent with the published descriptions that warn against treating Simsima Limestone as a single uniform strength unit and distinguish units with different geotechnical behaviour (Fourniadis, 2010; Karagkounis et al., 2016). The plot remains

descriptive because UCS availability is incomplete and lithology categories are unevenly represented.

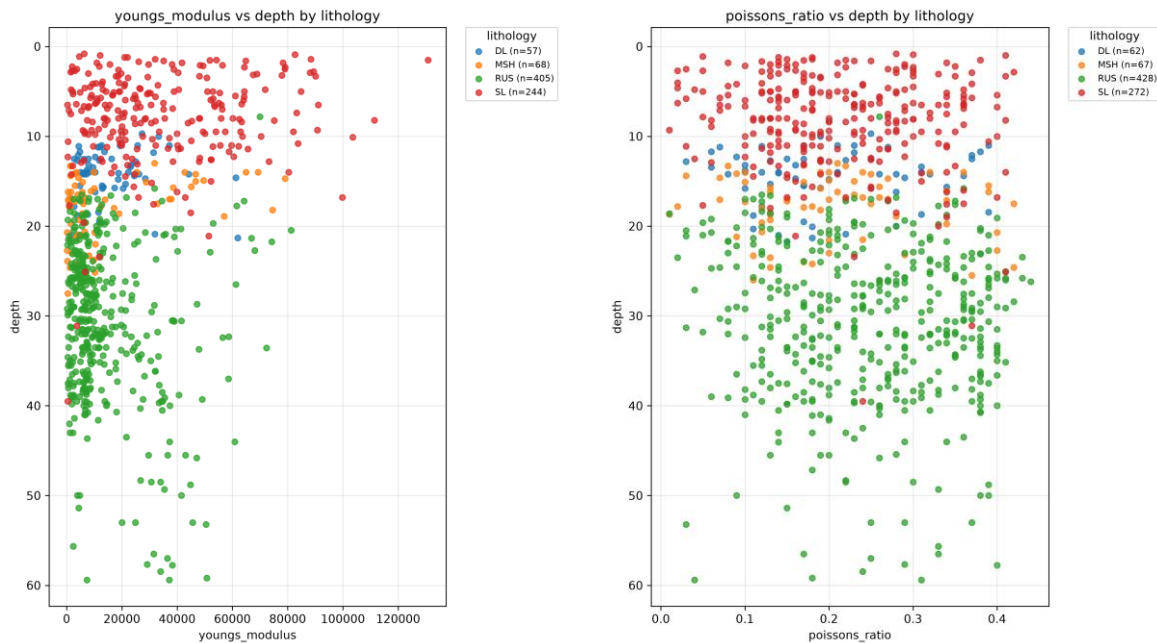


Figure 5.7 Young's modulus versus depth grouped by lithology. **Figure 5.8 Poisson's ratio versus depth grouped by lithology.**

The Figure 5.7 plot shows Young's modulus against depth. SL contains the widest stiffness range and the highest modulus values in the shallow interval, while RUS is mostly deeper and more concentrated at lower modulus values, with DL and MSH forming smaller intermediate groups. This pattern is compatible with the source descriptions that relate Simsima Limestone rock quality to strength and stiffness, and that distinguish Doha units by characteristic geotechnical behaviour (Fourniadis, 2010; Karagkounis et al., 2016). The figure should be interpreted as an exploratory comparison only because modulus data are incomplete and laboratory stiffness values do not fully represent rock-mass behaviour.

The Figure 5.8 plot shows Poisson's ratio against depth. The clearest pattern is lithology-depth grouping, with SL mainly shallow and RUS mainly deeper, rather than a strong separation by Poisson's ratio value alone. This is compatible with the largely subhorizontal Simsima-Midra-Rus succession described for Doha, but it does not provide strong evidence of lithology-specific elastic behaviour in this dataset (Karagkounis et al., 2016). The interpretation should

be treated cautiously because several lithology groups overlap and some categories have fewer valid elastic-property records.

5.5 Stage 4: Feature selection

The purpose of Stage 4 is to be used as a screening and feature-selection stage before model development. It reviews parameter availability, target-feature pairing, and practical modelling suitability so that later modelling could use a compact and reproducible set of candidate predictors.

For the first workflow application, Stage 4 was rerun after Stage 3 produced the reviewed dataset with outliers treated:

`data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv`.

The first modelling target remained UCS. The Stage 4 input contained 5,638 rows and 20 columns, of which 1,095 records had valid UCS values. Feature screening was therefore focused on the UCS-valid subset, because candidate predictors must be available in the same records as the target to support later supervised modelling. The selected compact initial feature set for WD1 was `bulk_density`, `depth`, `lithology`, and `moisture_content`. These selected feature sets had 844 complete records from the 1,095 UCS-valid records before any imputation.

The output of Stage 4 was therefore a documented feature-selection decision, not a completed modelling result. The main Stage 4 outputs included:

- `feature_selection_summary_WD1.csv`,
- `selected_features_WD1.csv`,
- `excluded_features_WD1.csv`,
- `feature_sets_WD1.csv`,
- `target_feature_availability_WD1.csv`, and
- `stage_4_summary_WD1.md`.

For the second workflow application, WD2, a workflow-specific copy of the same dataset reviewed Stage 3 dataset was added renamed:

`data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv`

The target parameter for WD2 is Is_{50} . The WD1-equivalent predictor set of depth, lithology, bulk_density, and moisture_content was initially tested for consistency with WD1. However, this four-predictor combination produced only 14 complete target-valid records before any imputation. The final WD2 feature set was therefore reduced to depth and lithology. bulk_density and moisture_content were excluded from the main WD2 feature set due to insufficient paired data availability with Is_{50} . UCS was also not used as a predictor, so that WD2 remains a practical standalone point-load-index prediction workflow rather than a direct UCS- Is_{50} conversion. The WD2 Stage 4 input contained 5,638 rows, of which 2,619 records had valid Is_{50} values and 3,019 records were missing the target value. The difference from the Stage 3 numerical summary count of 2,672 reflects the reviewed Stage 3 outlier treatment, where 53 lithology-grouped extreme Is_{50} values were set to missing before the reviewed WD2 dataset was used in Stage 4. This output documents feature availability and complete-case suitability only; it is not a modelling result.

Table 5.5 Summarises the input dataset, output summaries, target parameter, and selected features used in Stage 4.

Workflow ID	Dataset description	Input dataset	Output summaries	Target parameter	Selected features
WD1	Rock laboratory data from Doha	<i>data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv</i>	Stage 4 feature-selection tables in <i>outputs/tables/stage_4_feature_selection</i>	UCS	bulk_density, depth, lithology, moisture_content
WD2	Rock laboratory data from Doha	<i>data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv</i>	Stage 4 WD2 feature-selection tables in <i>outputs/tables/stage_4_feature_selection</i>	Is_{50}	depth, lithology

5.6 Stage 5: Machine learning model development and training

Stage 5 used the reviewed dataset and selected features from Stage 4 to develop and train the initial supervised regression models for each workflow target parameter. Its purpose was to

convert the feature-selection decision into reproducible model inputs, apply appropriate preprocessing, fit initial model types, and save model-development outputs for later evaluation.

For the first workflow application, WD1, Stage 5 used the Stage 3 reviewed dataset:

`data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv`

The modelling dataset was restricted to records with valid UCS values and complete values for the selected predictors. This produced 844 complete records from the 1,095 UCS-valid records before model development. The selected predictors were `bulk_density`, `depth`, `lithology`, and `moisture_content`. `bulk_density`, `depth`, and `moisture_content` were treated as continuous numerical input variables, while `lithology` was treated as a categorical predictor requiring encoding before model training.

Two initial UCS regression model types were developed: **Model 1** used **Ridge Regression** and **Model 2** used **Random Forest Regression**. Stage 5 should therefore be interpreted as model development and training only. Detailed performance comparison, residual review, and engineering interpretation are reserved for the following workflow stages.

For the second workflow application, WD2, Stage 5 used the reviewed point-load-index dataset `data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv`.

The WD2 target parameter was Is_{50} . The final Stage 4 selected predictors were `depth` and `lithology`, corresponding to the reduced WD2-A feature set. The Stage 5 model input dataset was limited to complete records for `point_load_index_is50`, `depth`, and `lithology`, producing 1,689 complete records. `depth` was treated as a continuous numerical predictor, while `lithology` was treated as a categorical predictor and encoded within the modelling pipeline. The main WD2 model did not use `bulk_density`, `moisture_content`, `unconfined_compressive_strength`, or `youngs_modulus` as predictors.

The same two model types used for WD1 were then applied to WD2: **Ridge Regression** as the simple baseline model, and **Random Forest Regression** as the nonlinear tree-based model. No additional model types were introduced, so that WD1 and WD2 remain comparable workflow applications. As for WD1, the WD2 Stage 5 outputs represent model setup and training only; detailed performance comparison, error analysis, and engineering interpretation are reserved for later stages.

Table 5.6 Summarises the input dataset and selected features, model input records, target parameter, and Stage 5 output for WD1 and WD2.

Workflow ID	Dataset description	Input dataset	Target parameter	Selected predictors	Model 1	Model 2
WD1	Rock laboratory data from Doha	data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv	UCS	bulk_density, depth, lithology, moisture_content	Ridge Regression	Random Forest Regression
WD2	Rock laboratory data from Doha	data_cleaned/rock_lab_stage_3_parameter_reviewed_WD2.csv	Is ₅₀	depth, lithology	Ridge Regression	Random Forest Regression

5.7 Stage 6: Model performance evaluation and comparison

Stage 6 evaluates the trained Stage 5 models using saved prediction outputs and held-out test data. The purpose is to calculate consistent regression performance metrics, compare candidate models, and provide a reproducible basis for later residual analysis and engineering interpretation. The models are not retrained in this stage.

For the first workflow application, WD1, Stage 6 was rerun using the updated Stage 5 WD1 prediction outputs and model artefacts. The target parameter was UCS, and the evaluation used the same 80/20 train-test split defined in Stage 5, with 169 test records. The two models evaluated were Ridge Regression and Random Forest Regression.

The rerun showed that Random Forest Regression had the lower test error, with a test RMSE of approximately 9.566, test MAE of approximately 6.606, and test R^2 of approximately 0.569. Ridge Regression had a test RMSE of approximately 10.149, test MAE of approximately 7.148, and test R^2 of approximately 0.515. The Random Forest model therefore ranked first by test RMSE and MAE, although its larger train-test performance gap indicates possible overfitting and requires further diagnostic review in Stage 7.

Stage 6 was also applied to WD2 using the same procedure. The target parameter was Is_{50} , and the selected predictors were depth and lithology. Evaluation used the held-out test set from the Stage 5 split, which contained 338 test records. Model 1 used Ridge Regression and Model 2 used Random Forest Regression. Ridge Regression produced a test RMSE of approximately 1.108, test MAE of approximately 0.794, and test R^2 of approximately 0.363. Random Forest Regression produced a test RMSE of approximately 1.280, test MAE of approximately 0.861, and test R^2 of approximately 0.151. Ridge Regression therefore performed better for WD2 on all three test-set metrics. The difference was modest for RMSE and MAE but more pronounced for R^2 , and both WD2 models showed weak predictive performance based on the test-set R^2 values. These results are reported as model-performance evaluation only; residual diagnosis and engineering interpretation are addressed in later workflow stages.

Table 5.7 Summarises in table sections (a) and (b) the Stage 6 performance evaluation metrics, outputs, workflow script, and target parameter used for each workflow.

(a)					
Workflow ID	Target parameter	Predictors	Test-set records	Stage 6 output location	Workflow Python script
WD1	UCS	bulk_density, depth, lithology, moisture_content	169	Tables in outputs/tables/stage_6_model_performance_evaluation; figures in outputs/figures/stage_6_model_performance_evaluation/WD1	scripts/stage_6_model_performance_evaluation.py
WD2	Is ₅₀	depth, lithology	338	Tables in outputs/tables/stage_6_model_performance_evaluation; figures in outputs/figures/stage_6_model_performance_evaluation/WD2	scripts/stage_6_model_performance_evaluation.py
(b)					
Workflow ID	Model 1	Model 1 test RMSE	Model 1 test MAE	Model 1 test R²	
WD1	Ridge Regression	10.149	7.148	0.515	
WD2		1.108	0.794	0.363	
Workflow ID	Model 2	Model 2 test RMSE	Model 2 test MAE	Model 2 test R²	
WD1	Random Forest Regression	9.566	6.606	0.569	
WD2		1.280	0.861	0.151	

5.8 Stage 7: Error analysis and diagnostic interpretation

Stage 7 investigates model errors after the Stage 6 performance comparison. Its purpose is diagnostic interpretation rather than model training or model selection. It examines residuals, absolute errors, large-error cases, UCS-range behaviour, and available feature-related error patterns so that model reliability and limitations can be described before engineering interpretation.

For the first workflow application, WD1, Stage 7 was rerun using the updated Stage 6 WD1 prediction and performance outputs. The target parameter remained UCS, and the diagnostics reviewed the same 169 test records for each model. The models reviewed were Ridge Regression and Random Forest Regression. No models were retrained, and the selected predictors, train-test split, and Stage 6 performance outputs were not altered.

The diagnostic outputs showed that Random Forest Regression retained the lower test error, with a test MAE of approximately 6.606 and test RMSE of approximately 9.566, compared with approximately 7.148 and 10.149 for Ridge Regression. However, Random Forest Regression also had the larger MAE generalisation gap, supporting the Stage 6 warning of possible overfitting. Ridge Regression showed stronger prediction range compression. The largest-error cases were mainly high-UCS underpredictions, indicating that the upper part of the UCS range remained difficult for both models to reproduce. Lithology diagnostics were generated, but the lithology groups were unevenly supported, so these results should be treated as diagnostic rather than conclusive.

Table 5.8 Summarises the input diagnostic files, models, diagnostic outputs, and target parameter used in Stage 7.

Workflow ID	Target parameter	Predictors	Test-set records	Main diagnostic model	Comparison model	Diagnostic outputs	Output table location	Output figure location
WD1	UCS	bulk_density, depth, lithology, moisture_content	169	Random Forest Regression	Ridge Regression	WD1 residual plots, measured-versus-predicted plot, residual distribution, largest-error plot, and per-lithology error table	outputs/tables/stage_7_error_analysis_diagnostics	outputs/figures/stage_7_error_analysis_diagnostics/WD1
WD2	Is ₅₀	depth, lithology	338	Ridge Regression	Random Forest Regression	WD2 residual plots, measured-versus-predicted plot, residuals-versus-depth plot, residual distribution plot, and per-lithology error table	outputs/tables/stage_7_error_analysis_diagnostics	outputs/figures/stage_7_error_analysis_diagnostics/WD2

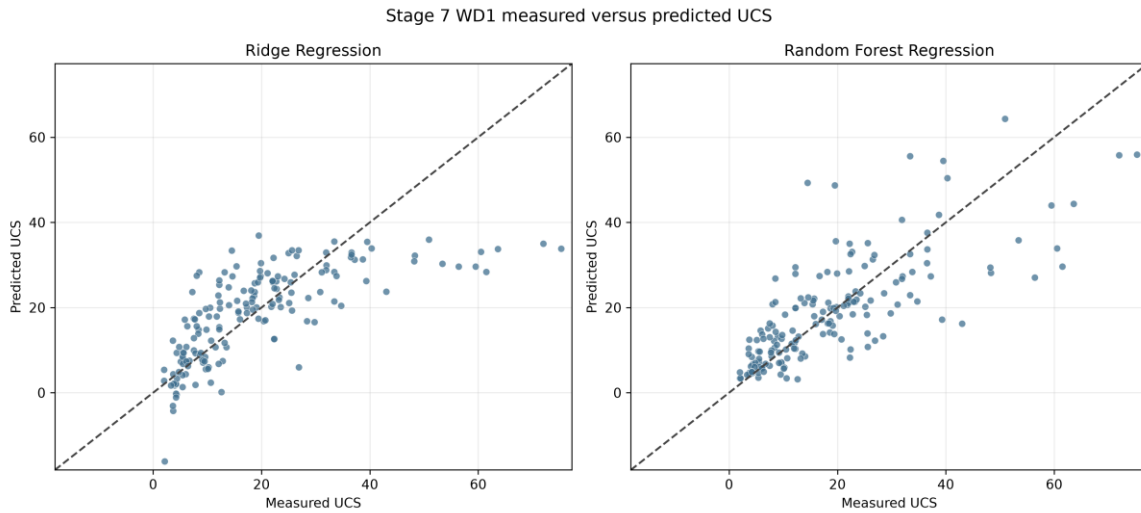


Figure 5.9 Measured versus predicted UCS for WD1.

Figure 5.9 compares measured UCS with predicted UCS for the WD1 test set. The dashed line represents perfect prediction. Ridge Regression shows stronger prediction-range compression, while Random Forest Regression follows the measured values more closely but still underpredicts some high-UCS records.

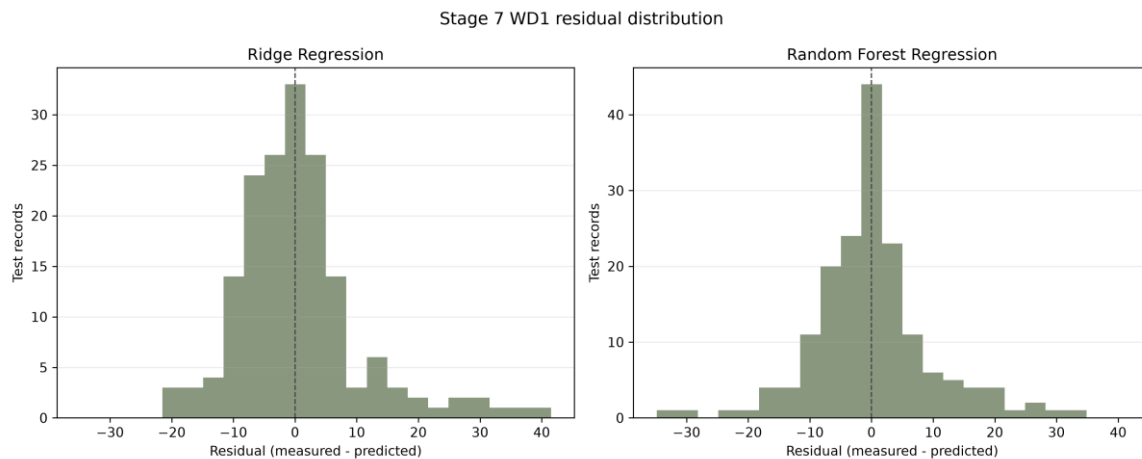


Figure 5.10 Residual distribution for WD1.

Figure 5.10 shows the distribution of residuals, where residual is defined as measured UCS minus predicted UCS. The residuals are generally centred around zero, but the positive residual tail indicates underprediction of some higher UCS samples. Random Forest Regression shows slightly more compact residual behaviour than Ridge Regression.

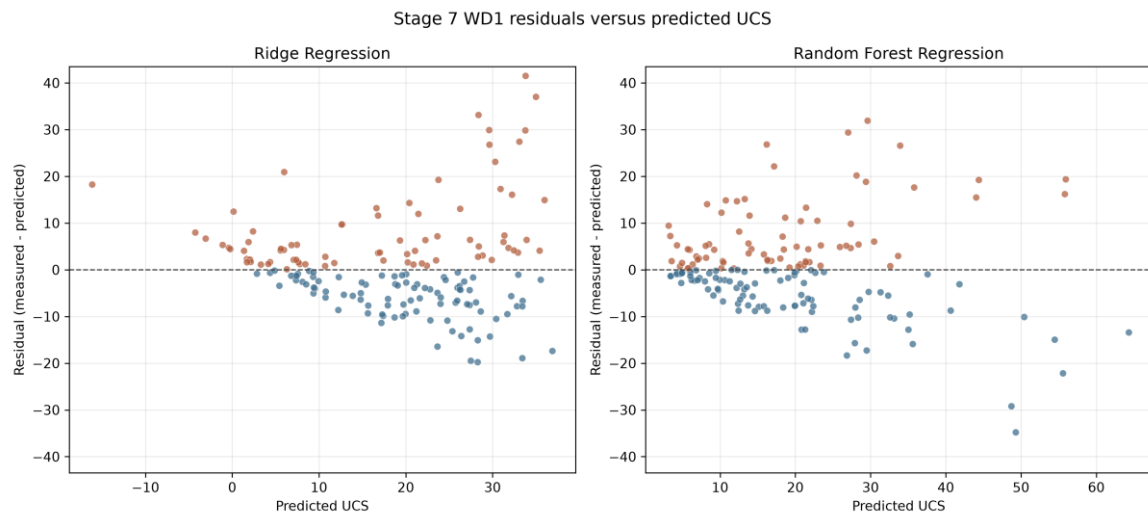


Figure 5.11 Residuals versus predicted UCS for WD1.

Figure 5.11 provides a diagnostic check for residual patterns against predicted UCS. The residuals are not fully random, with Ridge Regression showing stronger structure and range-compression behaviour. Random Forest Regression improves the pattern but still has large residuals for some high-strength samples.

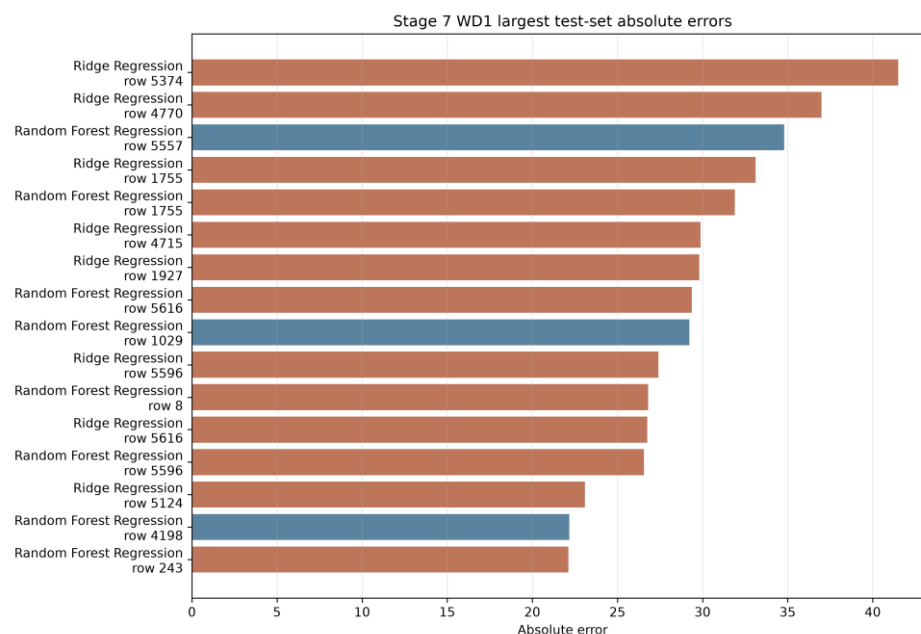


Figure 5.12 Largest absolute errors for WD1.

Figure 5.12 summarises the largest individual absolute errors in the WD1 test set. These cases are mainly high-UCS underpredictions, showing that both reviewed models had difficulty reproducing the upper UCS range.

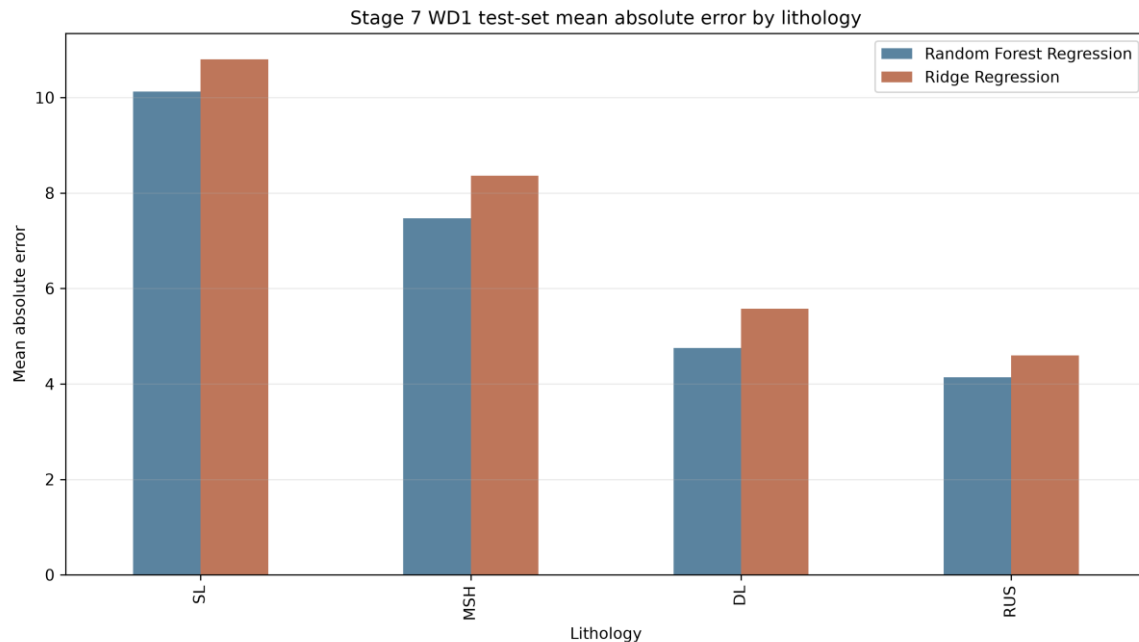


Figure 5.13 Mean absolute error by lithology for WD1.

Figure 5.13 compares mean absolute error by lithology. Error magnitude differs between lithological groups, and Random Forest Regression generally gives lower lithology-level MAE than Ridge Regression. This interpretation should remain cautious because the lithology groups are unevenly represented.

To further assess lithology-related error behaviour, the saved WD1 test-set predictions were grouped by lithology, and supplementary metrics were calculated for each model. These metrics were used diagnostically rather than for model selection, because the main Stage 6 comparison was based on the full combined WD1 test set and the lithology groups were unevenly represented. Unlike the overall Stage 6 R^2 values, the lithology-specific metrics in this section are supplementary diagnostic summaries calculated after grouping the test-set predictions by lithology.

Table 5.9 Supplementary WD1 test-set error metrics by lithology.

Lithology	Model	Test records	MAE	RMSE	R²
BSL	Ridge Regression	1	19.234	19.234	NA
BSL	Random Forest Regression	1	26.803	26.803	NA
DL	Ridge Regression	9	5.575	7.005	-1.141
DL	Random Forest Regression	9	4.751	6.505	-0.846
MSH	Ridge Regression	16	8.362	11.671	0.254
MSH	Random Forest Regression	16	7.470	9.226	0.534
RUS	Ridge Regression	87	4.596	5.886	0.417
RUS	Random Forest Regression	87	4.139	5.696	0.454
SL	Ridge Regression	56	10.803	14.269	0.262
SL	Random Forest Regression	56	10.127	13.483	0.341

The supplementary metrics indicate that prediction error was not uniform across lithologies. Based on MAE and RMSE, RUS showed the strongest lithology-level error control, supported by the largest number of test records. DL also showed relatively low MAE and RMSE, although its negative R² and smaller test count indicate that this result should be interpreted cautiously. MSH produced intermediate errors, while SL showed higher MAE and RMSE than RUS, DL, and MSH. BSL should be read only as an individual-case diagnostic because it contained one test record. Random Forest Regression produced lower lithology-level MAE and RMSE than Ridge Regression for DL, MSH, RUS, and SL, although the strength of this comparison varies by group and should not be treated as a separate model-selection result. Lithology-specific R²

is supplementary and may be unstable for small or narrow-range groups, as illustrated by the DL results and the NA values for BSL.

For the second workflow application, WD2, Stage 7 was also applied using the saved Stage 6 held-out test prediction outputs. The target parameter was I_{s50} , and the selected predictors were depth and lithology. The diagnostic review focused mainly on Ridge Regression because it performed better than Random Forest Regression in Stage 6, with a test RMSE of approximately 1.108, test MAE of approximately 0.794, and test R^2 of approximately 0.363. Random Forest Regression was retained as the comparison model, with a test RMSE of approximately 1.280, test MAE of approximately 0.861, and test R^2 of approximately 0.151.

The WD2 residual review used observed-versus-predicted plots, residuals against predicted values, residuals against depth, residual distributions, largest-error summaries, and lithology-grouped error tables. The model-level diagnostics showed a near-zero mean residual for both models, but the largest listed errors were concentrated mainly in the high I_{s50} range and were mostly underpredictions. Ridge Regression showed a compressed prediction range ratio of approximately 0.362, while Random Forest Regression showed a larger prediction range ratio of approximately 1.198 and a larger train-test error gap. The per-lithology diagnostic table was created to check whether errors were concentrated in particular lithological groups. For Ridge Regression, BSL had only one test record and should be treated as an individual-case diagnostic; DL showed an over-prediction tendency, while MSH, RUS, and SL showed under-prediction tendencies. These observations are diagnostic only and are not used here as final engineering interpretation.

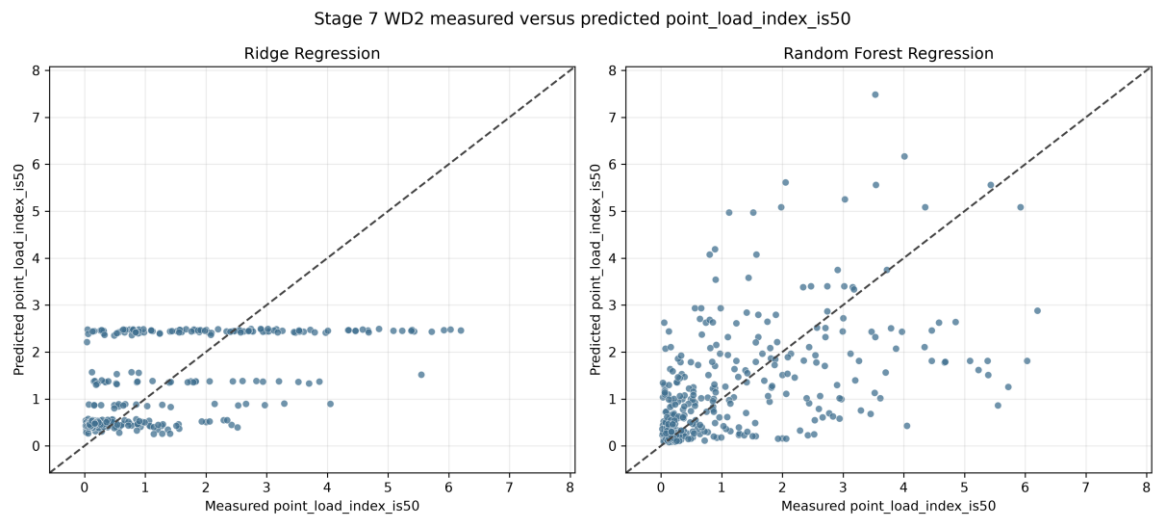


Figure 5.14 Measured versus predicted Is50 for WD2.

Figure 5.14 compares measured point load index Is_{50} with predicted Is_{50} for the WD2 test set. The dashed line represents perfect prediction. Ridge Regression gives the stronger overall test-set performance, but the spread around the 1:1 line shows that the selected predictors do not fully reproduce the observed Is_{50} variability.

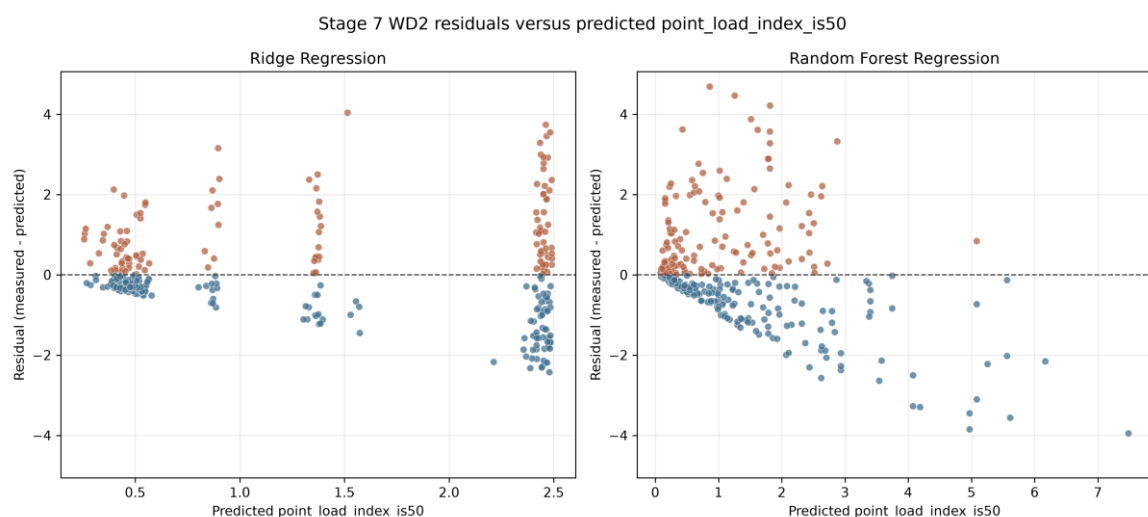


Figure 5.15 Residuals versus predicted Is50 for WD2.

Figure 5.15 provides a diagnostic check for residual structure against predicted Is_{50} . The residual pattern indicates prediction-range compression for Ridge Regression, while Random Forest Regression shows a wider prediction range but weaker overall test-set performance.

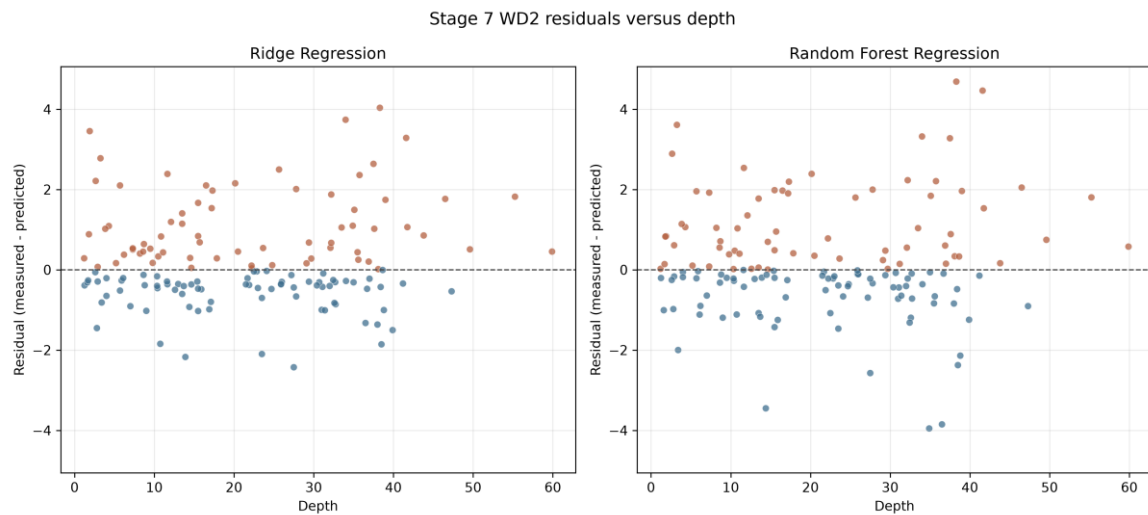


Figure 5.16 Residuals versus depth for WD2.

Figure 5.16 compares WD2 residuals against depth. The plot was used to check whether prediction errors were concentrated at particular depths. Any observed depth-related pattern is diagnostic only and should not be interpreted as a final engineering trend.

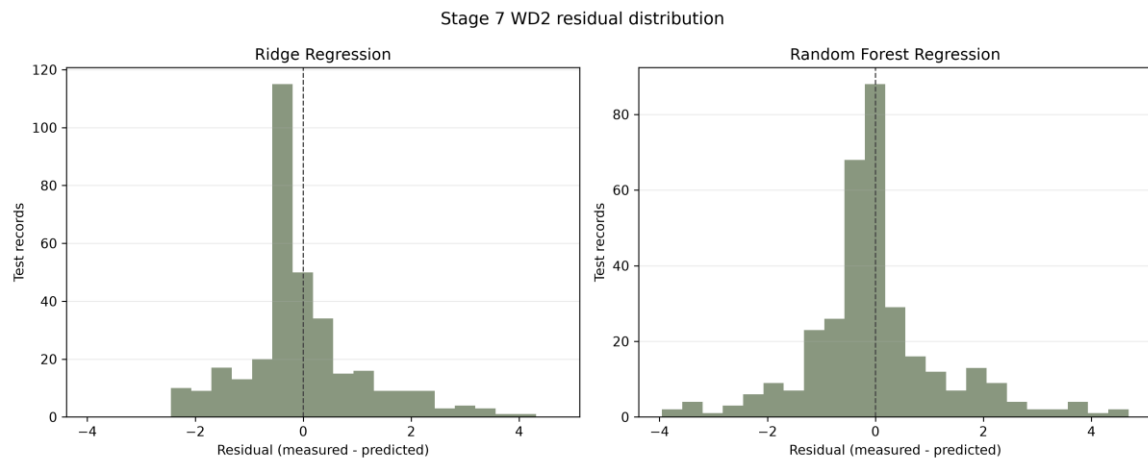


Figure 5.17 Residual distribution for WD2.

Figure 5.17 shows the distribution of WD2 residuals, where residual is defined as measured Is_{50} minus predicted Is_{50} . The residuals are generally centred close to zero, but the positive residual tail indicates under-prediction of some higher Is_{50} samples.

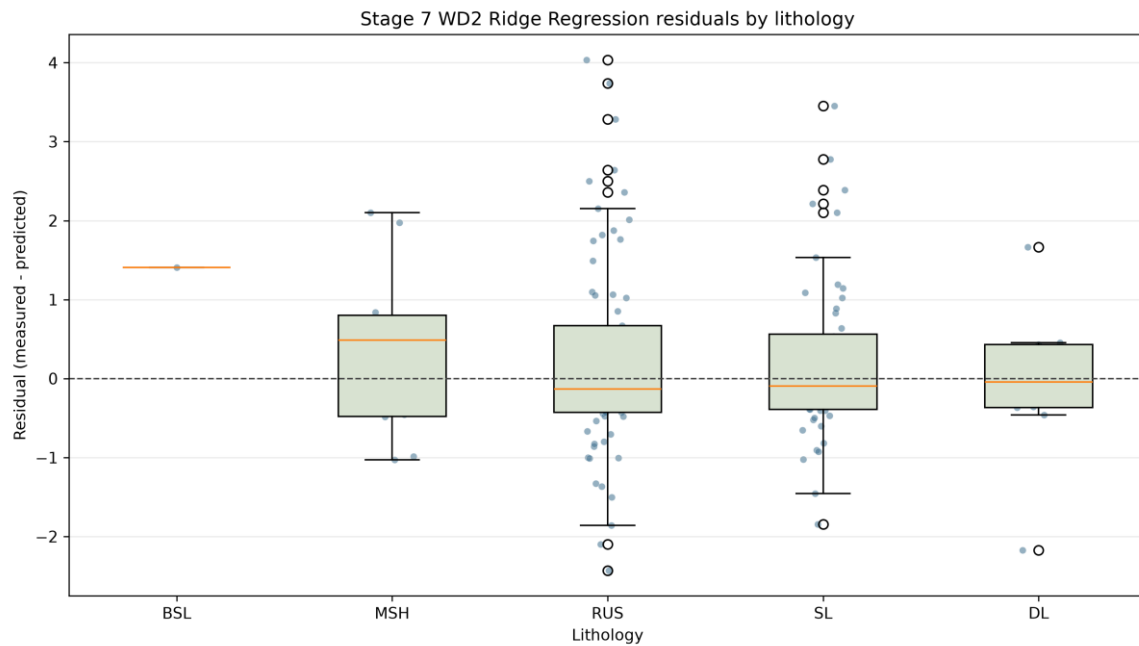


Figure 5.18 Ridge Regression residuals by lithology for WD2.

Figure 5.18 compares WD2 residual behaviour by lithology for Ridge Regression. The lithology-grouped diagnostics indicate that BSL is represented by only one test record, DL trends toward over-prediction, and MSH, RUS, and SL trend toward under-prediction. These results should be interpreted cautiously because lithology groups are unevenly represented.

Table 5.10 Supplementary WD2 test-set error metrics by lithology.

Lithology	Model	Test records	MAE	RMSE	R ²
BSL	Ridge Regression	1	1.405	1.405	NA
BSL	Random Forest Regression	1	1.773	1.773	NA
DL	Ridge Regression	9	0.666	0.964	-0.503
DL	Random Forest Regression	9	0.579	0.819	-0.084
MSH	Ridge Regression	10	0.952	1.118	-0.190
MSH	Random	10	0.945	1.183	-0.333

Lithology	Model	Test records	MAE	RMSE	R ²
	Forest Regression				
RUS	Ridge Regression	73	0.911	1.280	0.378
RUS	Random Forest Regression	73	1.035	1.523	0.119
SL	Ridge Regression	52	0.760	1.050	0.414
SL	Random Forest Regression	52	0.832	1.201	0.235

5.9 Stage 8: Engineering interpretation and workflow limitations

Stage 8 interprets the modelling results from an engineering and workflow perspective. It does not introduce new modelling, feature selection or diagnostics. Instead, it draws together the reviewed dataset, selected predictors, model performance outputs and diagnostic evidence to explain what the workflow can support, where it remains limited, and how the results should be used in geotechnical assessment.

For the first workflow application, WD1, the reviewed rock laboratory dataset `data_cleaned/rock_lab_stage_3_parameter_reviewed_WD1.csv` was used to assess initial prediction of UCS (unconfined_compressive_strength). The selected predictors were `bulk_density`, `depth`, `lithology`, and `moisture_content`. The modelling dataset contained 844 complete records, split into 675 training records and 169 test records, and was used to train Ridge Regression and Random Forest Regression models. Stage 6 showed that Random Forest Regression had the lower test error, with RMSE approximately 9.566, MAE approximately 6.606, and R^2 approximately 0.569, compared with Ridge Regression RMSE approximately 10.149, MAE approximately 7.148, and R^2 approximately 0.515.

The WD1 results indicate that the reviewed rock laboratory dataset can support an initial UCS prediction workflow. Random Forest Regression gave the stronger test performance overall, but its larger generalisation gap means possible overfitting should be considered before accepting the model uncritically. Ridge Regression provided a useful linear baseline, although it showed stronger prediction-range compression. The largest errors remained mainly high-UCS underpredictions. This is important for geotechnical interpretation because underprediction of high-strength samples can reduce the ability of the model to distinguish stronger rock intervals from moderate-strength intervals. Random Forest Regression followed the measured values more closely than Ridge Regression overall, but both models had difficulty reproducing the upper UCS range.

The supplementary per-lithology diagnostics reported in the Stage 7 table indicate that prediction error was not uniform across lithological groups. The evidence is strongest for RUS because it had the largest test-set support and low MAE and RMSE relative to the other groups. DL appeared low-error in MAE and RMSE terms, but this should be interpreted cautiously

because it was based on only nine test records and had negative lithology-level R^2 . SL, and to a lesser extent MSH, showed higher errors, suggesting that these groups may contain greater UCS variability, more complex lithological behaviour, or less favourable representation by the available predictors. BSL cannot be interpreted as a lithology-level performance group because it contained only one test record. Although Random Forest Regression generally had lower lithology-level MAE and RMSE than Ridge Regression for DL, MSH, RUS and SL, this is a supplementary diagnostic observation rather than a separate model-selection conclusion.

For the second workflow application, WD2, the same staged workflow was reused to test prediction of Is_{50} as a second geotechnical target parameter. The final WD2 model used only `depth` and `lithology`, because the WD1-equivalent four-predictor set of `depth`, `lithology`, `bulk_density`, and `moisture_content` had insufficient complete paired data with Is_{50} . The Stage 5 model input contained 1,689 complete records, and the held-out Stage 6 test set contained 338 records. This makes WD2 useful as a test of a sparse but practical workflow, but it also limits predictive strength because the retained predictors represent only broad depth and lithology context.

Stage 6 showed that Ridge Regression performed better than Random Forest Regression for WD2 on all three test-set metrics, with RMSE approximately 1.108, MAE approximately 0.794, and R^2 approximately 0.363, compared with Random Forest Regression RMSE approximately 1.280, MAE approximately 0.861, and R^2 approximately 0.151. This result suggests that, for this WD2 test case, the available predictor set supported only broad depth/lithology trends rather than more complex nonlinear modelling. The Ridge Regression test R^2 indicates weak to modest predictive capability, not reliable prediction of design Is_{50} values.

The WD2 diagnostics showed that the largest listed errors were mostly high Is_{50} underpredictions, and Ridge Regression showed prediction-range compression. This indicates difficulty reproducing stronger rock samples or the upper Is_{50} range. Point load index Is_{50} is influenced by lithology, sample condition, intact rock fabric, testing orientation, weathering, and local geological variability, which are not fully captured by `depth` and `lithology`. `bulk_density` and `moisture_content` were not retained because of insufficient paired data availability with Is_{50} , so potentially useful explanatory variables were excluded. UCS was

intentionally not used as a predictor so that WD2 remained a standalone prediction workflow rather than a UCS- I_{s50} conversion.

The WD2 lithology-grouped diagnostics suggest non-uniform error behaviour between lithologies. For Ridge Regression, BSL had only one test record and should be treated as an individual-case diagnostic only; DL showed an over-prediction tendency, while MSH, RUS, and SL showed under-prediction tendencies. These findings are supplementary because the lithology groups are unevenly represented. WD2 should therefore be used only for screening-level interpretation of broad data patterns, predictor-availability limitations, and lithology/depth-related variability. It should not be used as a replacement for laboratory point-load testing, laboratory strength assessment, or design strength selection.

Table 5.11 - Summarises the Stage 8 engineering interpretation and workflow limitations for WD1 and WD2.

Workflow ID	Target parameter	Predictors	Test-set records	Best-performing model	Main diagnostic finding	Main limitation	Engineering use
WD1	UCS	bulk_density, depth, lithology, moisture_content	169	Random Forest Regression	Moderate predictive capability; high-UCS underprediction; lithology-dependent error tendencies	Uneven lithology support, limited predictor set, and incomplete rock-mass controls	Screening-level interpretation only, not standalone design values
WD2	Is ₅₀	depth, lithology	338	Ridge Regression	Weak to modest predictive capability; high-Is ₅₀ underprediction; lithology-dependent error tendencies	Sparse predictor set and uneven lithology representation	Screening-level interpretation only, not replacement for laboratory testing or design values

The engineering interpretation is broadly consistent with the geological and geotechnical characteristics described for Doha rock formations. The retention of lithology as a predictor, together with the lithology-dependent error behaviour observed in Stage 7, supports the importance of lithological variability between the Simsim Limestone, Midra Shale, Rus Formation and associated units described by Karagkounis et al. (2016). For WD1, the inclusion of `depth`, `bulk_density`, and `moisture_content` is also compatible with the role of lithological succession, material condition, groundwater or moisture effects, and variation in density, strength and stiffness. For the Simsim Limestone specifically, Fourniadis (2010) describes complex weathering, karstification, dolomitisation, variable cementation and differences in rock quality that affect rock-mass strength and stiffness. These characteristics are compatible with the WD1 finding that high-UCS values were difficult to predict, because local cementation, dolomitisation, discontinuities, karst features and sample-scale variability may not be represented adequately by the available row-level laboratory predictors.

For WD2, the difficulty in predicting high Is_{50} values and the lithology-dependent residual behaviour are also consistent with variable engineering behaviour across Doha rock units. This supports the need to treat lithology and local geological conditions as important controls when interpreting point load index Is_{50} . However, the weak to modest WD2 performance shows that depth and lithology alone are not sufficient to fully characterise intact rock strength variability. The comparison with Fourniadis (2010) and Karagkounis et al. (2016) should therefore remain qualitative and cautious, rather than being treated as a direct validation of the WD2 model.

This comparison should not be interpreted as formal geological validation of the models or as evidence that the workflow has identified geological boundaries. Both workflow applications use extracted row-level laboratory data, incomplete paired testing, uneven lithology representation and limited predictor sets, and neither includes direct rock-mass controls such as RQD, TCR, SCR, weathering grade, discontinuity condition, cementation class, karst observations or groundwater descriptors. WD1 is therefore useful as an exploratory and screening tool for identifying broad UCS trends, and WD2 is useful for screening-level review of broad Is_{50} patterns, predictor-availability limitations and lithology/depth-related variability. Neither workflow should be used as a standalone design tool. Predictions should be checked against geological logging, laboratory test results, engineering judgement and the relevant

Doha rock formation literature before being used in geotechnical interpretation or design assessment.

5.10 Chapter summary

Chapter 5 applied the staged machine-learning workflow to a cleaned Doha rock laboratory dataset. Stages 1 to 3 prepared, reviewed and summarised the data; Stage 4 selected target-specific feature sets; Stage 5 trained Ridge Regression and Random Forest Regression models; Stage 6 compared model performance; Stage 7 diagnosed residual and lithology-related error patterns; and Stage 8 interpreted the results from an engineering and workflow-limitation perspective.

Two workflow applications were completed. WD1 used UCS as the target and showed moderate predictive capability for UCS. Random Forest Regression performed better overall than Ridge Regression, although the diagnostics indicated possible overfitting and under-prediction of high-UCS samples. WD2 used Is_{50} as the target and showed weaker predictive capability. Ridge Regression performed better than Random Forest Regression for WD2, suggesting that depth and lithology alone captured only broad Is_{50} trends rather than the full variability of the point load index.

The contrast between WD1 and WD2 shows that model performance depends strongly on the target parameter, paired data availability, and predictor completeness. The selected predictors and diagnostic patterns are broadly compatible with the importance of lithological variability and material condition described for Doha rock formations and Simsima Limestone (Fourniadis, 2010; Karagkounis et al., 2016), but this should be treated as qualitative context rather than formal geological validation. Neither workflow should be interpreted as a standalone design tool. The outputs are useful for screening-level interpretation, identifying data gaps, assessing predictor availability, and understanding broad lithology- and depth-related patterns, but predictions should be checked against laboratory testing, geological logging, engineering judgement, and the Doha geological context.

6. Results

This chapter presents the results from the machine-learning workflow described in Chapter 5. The main modelling target was unconfined compressive strength (UCS), using cleaned and screened rock laboratory data extracted from Doha factual reports. A second workflow application was also completed for point load index (Is_{50}), but the UCS workflow remains the primary result because UCS was the first target parameter with suitable availability and direct engineering relevance.

The results should be interpreted as screening-level modelling outcomes, not as design values or formal geological validation. The reported model-performance values refer to the full held-out test datasets used in Chapter 5 unless stated otherwise. Lithology-level results are supplementary diagnostics only, because the lithology groups are unevenly represented and some contain very few test records.

6.1 Exploratory data analysis results

The cleaned Doha rock laboratory dataset contained 5,638 records. UCS was the most suitable initial supervised modelling target because it had the strongest combination of data availability and engineering relevance among the reviewed strength parameters. After Stage 3 review and outlier treatment, 1,095 records had valid UCS values for feature screening.

Exploratory review showed that the dataset was useful but incomplete. Several geotechnically relevant variables, including point load index (Is_{50}), dry density, porosity, elastic properties, abrasivity, and shear-strength parameters, had limited paired availability with UCS or possible leakage concerns. These restrictions reduced the range of variables that could be used reliably in the first UCS model.

The exploratory results also confirmed that lithology and depth were important for interpretation. The records showed broad lithology-depth grouping, with Simsima Limestone (SL), Midra Shale (MSH), Dukhan Limestone (DL), Rus Formation (RUS), and related units not represented evenly. This pattern is consistent with the expected Doha geological succession and the variable engineering behaviour described for the Simsima Limestone and associated Doha rock formations (Fourniadis, 2010; Karagkounis et al., 2016), but it should be treated as qualitative context rather than formal stratigraphic validation.

6.2 Feature selection results

For the UCS workflow, Stage 4 selected a compact practical predictor set made up of `bulk_density`, `depth`, `lithology`, and `moisture_content`. These variables were retained because they provided a workable balance between geotechnical meaning and paired data availability. The selected UCS modelling dataset contained 844 complete records from the 1,095 UCS-valid records.

The selected variables reflect the main controls available in the cleaned laboratory dataset. Depth provides broad geological context, lithology represents the cleaned geological descriptor, bulk density provides a material-property indicator, and moisture content may reflect sample condition, weathering, or groundwater-related influences. These variables do not fully represent rock-mass behaviour, because direct rock-mass controls such as RQD, TCR, SCR, weathering grade, discontinuity condition, karst observations, and groundwater descriptors were not included in the completed UCS model.

For the supplementary WD2 workflow, the target was Is_{50} . The WD1-equivalent four-predictor set had insufficient complete paired data with (Is_{50}), so the final WD2 model used only `depth` and `lithology`. UCS was not used as a predictor in WD2 so that the workflow remained a standalone prediction exercise rather than a UCS- Is_{50} conversion.

6.3 Model 1 Ridge Regression results

Ridge Regression was used as the linear baseline model. For the UCS workflow, it was trained on 675 records and evaluated on 169 held-out test records using the selected predictors `bulk_density`, `depth`, `lithology`, and `moisture_content`.

On the combined UCS test set, Ridge Regression produced a test RMSE of approximately 10.149, a test MAE of approximately 7.148, and a test (R^2) of approximately 0.515. This indicates moderate predictive capability for the screened UCS dataset, but the model did not fully capture the observed UCS variability.

The diagnostic results in Chapter 5 show that Ridge Regression had stronger prediction-range compression than Random Forest Regression. This means that the linear baseline tended to under-represent the spread of measured UCS values, particularly in the higher-strength range.

6.4 Model 2 Random Forest Regression results

Random Forest Regression was used as the nonlinear comparison model for the UCS workflow. It used the same selected predictors and the same 675 training and 169 held-out test records as the Ridge Regression model.

On the combined UCS test set, Random Forest Regression produced a test RMSE of approximately 9.566, a test MAE of approximately 6.606, and a test (R^2) of approximately 0.569. These values were better than the Ridge Regression results on the same test split, so Random Forest Regression was the stronger of the two UCS models in the workflow.

The result should not be overstated. Random Forest Regression showed a larger train-test performance gap than Ridge Regression, indicating possible overfitting. It followed the measured UCS values more closely overall, but both models had difficulty reproducing the upper UCS range.

6.5 Comparison of model performance and diagnostic interpretation

For the main UCS workflow, Random Forest Regression performed better than Ridge Regression on the combined held-out test set. The improvement was modest but consistent across RMSE, MAE, and (R^2). The Random Forest test (R^2) was approximately 0.569, compared with approximately 0.515 for Ridge Regression.

The comparison supports the use of Random Forest Regression as the better-performing model for this dataset, while Ridge Regression remains a useful transparent baseline. The main limitation is that the Random Forest result is accompanied by a larger train-test gap, so the model should be treated as a screening tool rather than a reliable standalone predictor of UCS.

The error analysis showed that the largest UCS errors were mainly high-UCS underpredictions. This limits the ability of the models to distinguish stronger rock intervals from moderate-strength intervals. The issue is consistent with the restricted predictor set, sparse paired data for some useful variables, and the absence of direct rock-mass descriptors.

The supplementary lithology-level diagnostics showed that model behaviour was not uniform across rock units. RUS had the strongest support in the test set and showed lower errors than several other groups. SL and, to a lesser extent, MSH showed higher errors, while DL had few test records and negative lithology-level (R^2), so it should be interpreted cautiously. BSL is

retained here as a dataset lithology code and had only one test record, so it cannot be treated as a lithology-level performance group. These subgroup results should not be used as separate model-selection results.

The engineering interpretation is broadly consistent with the expected behaviour of the Doha formations. Lithology, depth, density, and moisture content are reasonable controls for screening UCS because the Doha rock sequence includes materials with different strength, stiffness, weathering, dolomitisation, karstification, and groundwater-related behaviour. The Simsima Limestone is known to be variable and should not be treated as a single uniform strength unit (Fourniadis, 2010; Karagkounis et al., 2016). The model results are compatible with this context, but they do not prove geological boundaries or replace engineering judgement.

The supplementary WD2 workflow gave weaker results than the UCS workflow. Ridge Regression performed better than Random Forest Regression for Is_{50} , with a combined test-set RMSE of approximately 1.108, MAE of approximately 0.794, and (R^2) of approximately 0.363. This result indicates only weak to modest predictive capability, reflecting the sparse predictor set of depth and lithology and the limited coverage of other relevant parameters.

6.6 Summary

The workflow shows that the cleaned Doha rock laboratory dataset can support an initial, reproducible UCS prediction workflow. The main UCS model used `bulk_density`, `depth`, `lithology`, and `moisture_content` as selected predictors. Random Forest Regression performed better than Ridge Regression on the combined UCS test set, but the result remains moderate and affected by possible overfitting and high-UCS underprediction.

The results also show that model reliability varies between lithologies and should not be treated as uniform across all rock units. Lithology-level diagnostics are useful for interpretation, but they are supplementary because of sparse data, missing values, uneven lithology representation, and restricted parameter coverage.

Overall, the Chapter 6 results support screening-level interpretation of broad UCS trends in the cleaned Doha laboratory dataset. They do not support standalone design strength selection, formal geological validation, or uniform prediction reliability across the Simsima Limestone, Midra Shale, Rus Formation, Dukhan Limestone, and related units. Predictions should be

checked against laboratory testing, geological logging, engineering judgement, and the Doha geological context discussed in the existing thesis references.

7. Discussion

This chapter discusses the completed machine-learning results presented in Chapters 5 and 6. The discussion focuses on the meaning of the results rather than repeating the detailed metrics. The main modelling target was unconfined compressive strength (UCS), using the cleaned Doha rock laboratory dataset and the final selected predictors `bulk_density`, `depth`, `lithology`, and `moisture_content`. The supplementary WD2 workflow for Is_{50} is discussed only as a secondary comparison.

7.1 Interpretation of results

The completed workflow shows that the cleaned Doha rock laboratory dataset can support a reproducible screening-level machine-learning assessment of UCS. The staged process of inventory, cleaning, parameter review, feature selection, modelling, evaluation, diagnostics, and engineering interpretation provided a transparent route from extracted laboratory records to interpretable model outputs. UCS was the most appropriate main target because it had sufficient availability in the reviewed dataset and is directly relevant to intact rock strength interpretation.

The final UCS predictor set was deliberately compact. `bulk_density`, `depth`, `lithology`, and `moisture_content` were retained because they provided a practical balance between paired data availability and geotechnical meaning. Depth and lithology provide broad geological context, while bulk density and moisture content provide material and sample-condition information. These predictors are useful for screening, but they cannot fully represent rock-mass behaviour because direct descriptors such as RQD, TCR, SCR, weathering grade, discontinuity condition, karst observations, and groundwater information were not included in the completed UCS model.

Random Forest Regression performed better than Ridge Regression on the combined held-out UCS test set. The reported UCS performance values in Chapters 5 and 6 refer to this combined test set unless stated otherwise. The improvement was consistent across the main performance metrics, but it should be treated cautiously because the Random Forest model also showed a larger train-test performance gap. This indicates that the model captured more structure in the

available data than the linear baseline, but it should not be interpreted as a design-level prediction tool.

The UCS model therefore demonstrated moderate predictive capability rather than high reliability. The most important diagnostic limitation was the underprediction of high-UCS samples and the compression of the predicted strength range. This reduces the model's ability to separate stronger rock intervals from moderate-strength intervals. From an engineering perspective, this limitation is important because the upper strength range may be controlled by local cementation, fabric, dolomitisation, weathering state, discontinuities, or sample-scale variability that are not fully described by the selected row-level predictors.

7.2 Comparison with traditional geotechnical interpretation

The modelling results are broadly consistent with traditional geotechnical interpretation of the Doha rock formations. The importance of lithology, depth, density, and moisture-related information is compatible with the expected variability of Simsim Limestone (SL), Midra Shale (MSH), Dukhan Limestone (DL), Rus Formation (RUS), and associated units. The observed variability is also consistent with the existing thesis references, which describe the influence of lithological succession, weathering, karstification, dolomitisation, groundwater conditions, and variable rock quality in the Doha area (Fourniadis, 2010; Karagkounis et al., 2016).

This consistency should be interpreted qualitatively. The models do not validate geological boundaries, define formation contacts, or replace geological logging. Instead, they show that the cleaned laboratory dataset contains broad patterns that are reasonable in the known geological context. The results are therefore useful for exploratory interpretation, but the model outputs still require checking against borehole logs, laboratory records, geological descriptions, and engineering judgement.

The lithology-level diagnostic results add useful context, but they should not be treated as separate model-selection results. Model reliability varied between lithologies because the dataset had sparse records in some groups, missing values, uneven lithology representation, and restricted parameter coverage. RUS had stronger test-set support than several other groups, while DL had few records and BSL, retained as a dataset lithology code, had only one record for lithology-level interpretation. SL and MSH showed larger errors in parts of the diagnostic

review, which is compatible with variable material behaviour but should not be over-interpreted as a lithology-specific performance conclusion.

7.3 Practical relevance

The practical value of the workflow is that it converts a complex extracted laboratory dataset into a documented and reproducible modelling process. Each stage records what data were available, which variables were usable, how predictors were selected, how models were evaluated, and where the main errors occurred. This makes the workflow useful for screening, exploratory interpretation, and identifying broad trends in UCS behaviour.

The workflow is also useful because it highlights where the dataset is not strong enough. The modelling results show that predictor availability, paired data completeness, and lithology representation strongly affect model reliability. The diagnostic outputs identify large residuals, high-UCS underprediction, and lithology-dependent error patterns that can guide further review of source records and future data collection.

The supplementary WD2 workflow for Is_{50} had weaker predictive performance than the UCS workflow. Ridge Regression performed better than Random Forest Regression for WD2, but the result indicated only weak to modest predictive capability. This is consistent with the limited WD2 predictor set of depth and lithology, which captured broad context but not enough material or rock-mass information to reproduce the full variability of point load index values. WD2 should therefore be treated as a secondary workflow showing the effect of sparse paired data, not as an equivalent result to the UCS model.

7.4 Limitations

The main limitations arise from data sparsity, missing values, uneven lithology representation, and restricted parameter coverage. Although the cleaned rock laboratory dataset was large enough to support a reproducible UCS workflow, many potentially useful parameters were not sufficiently paired with UCS or were unsuitable because of leakage risk. As a result, the final model used a small set of predictors and excluded several variables that may be important for rock strength interpretation.

The completed UCS model also lacks direct rock-mass and field-condition descriptors. It does not include RQD, TCR, SCR, weathering grade, discontinuity condition, fracture spacing, karst

observations, groundwater descriptors, or other additional field and rock-mass descriptors. This means that the model is mainly an intact-rock laboratory screening tool. It cannot represent the full behaviour of a rock mass or the engineering risk associated with excavation, groundwater inflow, weak zones, or construction performance.

Prediction-range compression remains a major limitation. Both modelling workflows had difficulty with the upper strength or index-strength range, although the issue was most important for UCS interpretation. High-UCS underprediction means that the model may smooth stronger samples toward the centre of the dataset. This behaviour is acceptable as a diagnostic warning in a screening workflow, but it is not acceptable as a basis for design strength selection.

The validation scope is also limited. The held-out test sets provide useful internal checks, but they do not prove generalisation to new projects, boreholes, reports, testing programmes, or geological settings. The lithology-level metrics are supplementary only and may be unstable for small groups or groups with narrow observed ranges. The results should therefore be treated as evidence of moderate screening capability within the cleaned dataset, not as proof of broad predictive reliability.

7.5 Engineering judgement

Engineering judgement remains necessary at every stage of the workflow. It is needed to assess whether extracted values are credible, whether missing values can be handled responsibly, whether predictors are appropriate, whether apparent patterns make geological sense, and whether model errors are acceptable for the intended use. The workflow can support this judgement, but it cannot replace it.

The model outputs are best used as decision-support information. They can help identify broad UCS trends, compare a transparent linear baseline with a more flexible nonlinear model, flag large residuals, and show where model reliability varies between lithologies. They should not be used to assign design UCS values without reference to laboratory testing, geological logging, source report context, and project-specific geotechnical assessment.

For geotechnical risk assessment, the current workflow remains indirect. It supports laboratory data review and screening-level strength interpretation, but it does not model complete geotechnical risk, rock-mass quality, groundwater behaviour, or construction consequences.

Future application to risk screening would require richer borehole- and rock-unit-level variables, including rock-mass and groundwater descriptors, in addition to the laboratory parameters used here.

7.6 Summary

The completed workflow demonstrates that the cleaned Doha rock laboratory dataset can support a reproducible screening-level UCS modelling process. The final UCS model used `bulk_density`, `depth`, `lithology`, and `moisture_content` as selected predictors. Random Forest Regression performed better than Ridge Regression on the combined held-out UCS test set, but the improvement should be interpreted cautiously because the model showed only moderate predictive capability and possible overfitting.

The main limitations are high-UCS underprediction, prediction-range compression, sparse paired data, missing values, uneven lithology representation, and the absence of direct rock-mass controls. Lithology-level diagnostics are useful for interpretation, but they are supplementary and should not be used as separate model-selection results.

Overall, the workflow is useful for screening, exploratory interpretation, and identifying broad trends in the Doha rock laboratory dataset. It is not a substitute for laboratory testing, geological logging, or engineering judgement. The findings are broadly consistent with the expected variability of Simsim Limestone and associated Doha rock units discussed by Fourniadis (2010) and Karagkounis et al. (2016), but this consistency should be treated as geological context rather than formal model validation.

8. Conclusions and Recommendations

This chapter provides brief conclusions and recommendations based on the completed machine-learning workflow reported in Chapters 5 to 7. The conclusions are limited to the cleaned Doha rock laboratory dataset, the UCS workflow, the supplementary I_{s50} workflow, and the Ridge Regression and Random Forest Regression models evaluated in this thesis. The results should be read as screening-level evidence, not as design values, formal geological validation, or a complete geotechnical risk assessment method.

8.1 Conclusions

This thesis developed and applied a reproducible machine-learning workflow to cleaned Doha rock laboratory data. The workflow moved from data inventory and category-specific cleaning through parameter review, feature selection, model training, model evaluation, diagnostic interpretation, and engineering review. This staged process made the modelling decisions traceable and allowed the results to be interpreted in terms of both statistical performance and geotechnical meaning.

UCS was the main modelling target because it had suitable data availability and direct engineering relevance as an intact rock strength parameter. The final UCS workflow used `bulk_density`, `depth`, `lithology`, and `moisture_content` as selected predictors. These variables provided a practical balance between paired data availability and engineering meaning, but they do not fully represent rock-mass behaviour or project-scale geotechnical risk.

Random Forest Regression performed better than Ridge Regression on the combined held-out UCS test set. This indicates that the nonlinear model captured more structure in the available data than the linear baseline. However, the improvement should be interpreted cautiously because the Random Forest model also showed a larger train-test performance gap, and both models had difficulty reproducing the upper UCS range.

The completed UCS workflow therefore demonstrated moderate screening-level predictive capability, not design-level reliability. The workflow can support data review, exploratory interpretation, identification of broad UCS trends, and screening of possible large-error cases.

It does not replace laboratory testing, geological logging, source report review, or engineering judgement.

The lithology-level diagnostics were useful for interpreting where errors were concentrated, but they should remain supplementary. They do not provide separate model-selection results because the lithology groups were unevenly represented and some groups contained few test records. The findings were broadly consistent with the expected variability of Simsim Limestone and associated Doha rock units described in the existing thesis references, but this consistency should be treated as geological context rather than formal model validation (Fourniadis, 2010; Karagkounis et al., 2016).

The supplementary WD2 workflow for Is_{50} was weaker than the UCS workflow. It used only depth and lithology because richer paired predictors were not sufficiently available. The WD2 results are therefore secondary and mainly show the effect of sparse paired data and restricted predictor coverage on model reliability.

8.2 Main limitations

The main limitations of the completed workflow are sparse paired data, missing values, uneven lithology representation, restricted parameter coverage, and limited validation scope. Although the cleaned rock laboratory dataset was large enough to support a reproducible UCS workflow, many potentially useful variables were not sufficiently paired with UCS or were unsuitable for modelling because of data coverage or leakage concerns.

The selected UCS predictors provided only a compact row-level laboratory description. The workflow did not include additional borehole, geological, groundwater, and rock-mass descriptors that would be needed to represent field conditions more fully. This limits the ability of the models to represent rock-mass behaviour, groundwater effects, weak zones, or construction-related geotechnical risk.

The study demonstrated that ML could assist in the preliminary assessment of strength-related geotechnical risk indicators from project-derived laboratory data. However, the results do not represent a full quantitative geotechnical risk assessment. The dataset did not include sufficient information on probability of occurrence, consequence, exposure, construction response, or formal risk ranking. The workflow is therefore best understood as a screening-level tool for identifying patterns, supporting UCS interpretation, and highlighting lithology-related

variability that may inform subsequent engineering judgement and more detailed risk assessment.

The most important model limitation was high-UCS underprediction and compression of the predicted strength range. This reduces confidence in the model's ability to distinguish stronger rock intervals from moderate-strength intervals. The held-out test results provide a useful internal check, but they do not prove generalisation to new projects, boreholes, testing programmes, geological units, or source reports.

8.3 Recommendations for future work

Future work should continue the staged workflow while improving the geotechnical content of the input data. Priority should be given to adding richer borehole, geological, groundwater, and rock-mass descriptors where they can be extracted reliably. These may include improved lithology, weathering, discontinuity, groundwater, and rock-quality information where those data are available and consistently recorded.

The workflow should also be tested on additional datasets, projects, or reporting packages to check whether the observed patterns are stable beyond the cleaned dataset used in this thesis. Validation should be strengthened across lithologies and geological units, especially where current test-set support is limited or uneven.

Additional model types may be explored if supported by the thesis scope and by sufficient data, but model expansion should not be treated as a substitute for better geological and geotechnical descriptors. Any future comparison should use consistent target definitions, data splits, evaluation metrics, diagnostic checks, and engineering review.

8.4 Final statement

The thesis demonstrates that machine learning can support a reproducible screening-level workflow for cleaned Doha rock laboratory data. The main UCS workflow identified broad strength trends and highlighted important data and model limitations, while the supplementary WD2 workflow showed the reduced reliability expected when predictor coverage is sparse. The workflow is useful for screening, exploratory interpretation, and data review, but it remains a decision-support tool only. It must be used alongside laboratory testing, geological logging, source report review, and professional engineering judgement.

References

- Abdollahi, A., Li, D., Deng, J., & Amini, A. (2024). An explainable artificial-intelligence-aided safety factor prediction of road embankments. *Engineering Applications of Artificial Intelligence*, 136, Article 108854. <https://doi.org/10.1016/j.engappai.2024.108854>
- Amjad, M., Ahmad, I., Ahmad, M., Wroblewski, P., Kaminski, P., & Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation. *Applied Sciences*, 12(4), Article 2126. <https://doi.org/10.3390/app12042126>
- Ayawah, P. E. A., Sebbeh-Newton, S., Azure, J. W. A., Kaba, A. G. A., Anani, A., Bansah, S., & Zabidi, H. (2022). A review and case study of artificial intelligence and machine learning methods used for ground condition prediction ahead of tunnel boring machines. *Tunnelling and Underground Space Technology*, 125, Article 104497. <https://doi.org/10.1016/j.tust.2022.104497>
- Baghbani, A., Choudhury, T., Costa, S., & Reiner, J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews*, 228, Article 103991. <https://doi.org/10.1016/j.earscirev.2022.103991>
- Bui, D. T., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., Khosravi, K., Panahi, M., Ahmad, B. B., & Saro, L. (2018). Land subsidence susceptibility mapping in South Korea using machine learning algorithms. *Sensors*, 18(8), Article 2464. <https://doi.org/10.3390/s18082464>
- Cao, Y., Zhou, X., & Yan, K. (2021). Deep learning neural network model for tunnel ground surface settlement prediction based on sensor data. *Mathematical Problems in Engineering*, 2021, Article 9488892. <https://doi.org/10.1155/2021/9488892>
- Duncan, J. M. (2000). Factors of safety and reliability in geotechnical engineering. *Journal of Geotechnical and Geoenvironmental Engineering*, 126(4), 307–316. [https://doi.org/10.1061/\(ASCE\)1090-0241\(2000\)126:4\(307\)](https://doi.org/10.1061/(ASCE)1090-0241(2000)126:4(307))

- Elsawy, M. B. D., Alsharekh, M. F., & Shaban, M. (2022). Modeling undrained shear strength of sensitive alluvial soft clay using machine learning approach. *Applied Sciences*, 12(19), Article 10177. <https://doi.org/10.3390/app121910177>
- Fourniadis, I. (2010). Geotechnical characterization of the Simsima Limestone (Doha, Qatar). In Q. He & S.-L. Shen (Eds.), *Geoenvironmental engineering and geotechnics: Progress in modeling and applications* (pp. 273-278). American Society of Civil Engineers. [https://doi.org/10.1061/41105\(378\)38](https://doi.org/10.1061/41105(378)38)
- Gao, W. (2024). The application of machine learning in geotechnical engineering. *Applied Sciences*, 14(11), Article 4712. <https://doi.org/10.3390/app14114712>
- Geem, Z. W., Bekdaş, G., Nigdeli, S. M., Aydın, Y., Işıkdag, Ü., & Kim, T.-H. (2026a). A review of machine learning (ML) methods for geotechnical engineering. In *Introduction and applications of machine learning in geotechnics*. Elsevier. <https://doi.org/10.1016/B978-0-443-41481-7.00013-4>
- Geem, Z. W., Bekdaş, G., Nigdeli, S. M., Aydın, Y., Işıkdag, Ü., & Kim, T.-H. (2026b). Explainable artificial intelligence (XAI) in geotechnical engineering. In *Introduction and applications of machine learning in geotechnics* (pp. 31–52). Elsevier. <https://doi.org/10.1016/B978-0-443-41481-7.00005-5>
- Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences*, 81, 1–11. <https://doi.org/10.1016/j.cageo.2015.04.007>
- Hu, M., Li, W., Yan, K., Ji, Z., & Hu, H. (2019). Modern machine learning techniques for univariate tunnel settlement forecasting: A comparative study. *Mathematical Problems in Engineering*, 2019, Article 7057612. <https://doi.org/10.1155/2019/7057612>
- Karagkounis, N., Latapie, B., Sayers, K., & Mulinti, S. R. (2016). Geology and geotechnical evaluation of Doha rock formations. *Geotechnical Research*, 3(3), 119-136. <https://doi.org/10.1680/jgere.16.00010>

- Keawsawasvong, S., Seehavong, S., & Ngamkhanong, C. (2022). Application of artificial neural networks for predicting the stability of rectangular tunnels in Hoek-Brown rock masses. *Frontiers in Built Environment*, 8, Article 837745. <https://doi.org/10.3389/fbuil.2022.837745>
- Karagkounis, N., Latapie, B., Sayers, K., & Mulinti, S. R. (2016). Geology and geotechnical evaluation of Doha rock formations. *Geotechnical Research*, 3(3), 119–136. <https://doi.org/10.1680/jgere.16.00010>
- Liu, H., Su, H., Sun, L., & Dias-da-Costa, D. (2024). State-of-the-art review on the use of AI-enhanced computational mechanics in geotechnical engineering. *Artificial Intelligence Review*, 57, Article 196. <https://doi.org/10.1007/s10462-024-10836-w>
- Marcher, T., Erharter, G. H., & Winkler, M. (2020). Machine learning in tunnelling: Capabilities and challenges. *Geomechanics and Tunnelling*, 13(2), 191–198. <https://doi.org/10.1002/geot.202000001>
- Matin, S. S., Farahzadi, L., Makaremi, S., Chehreh Chelgani, S., & Sattari, G. (2018). Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. *Applied Soft Computing*, 70, 980–987. <https://doi.org/10.1016/j.asoc.2017.06.030>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 379–425). Springer. https://doi.org/10.1007/978-3-030-89010-0_10
- Nanehkaran, Y. A., Licai, Z., Chengyong, J., Chen, J., Anwar, S., Azarafza, M., & Derakhshani, R. (2023). Comparative analysis for slope stability by using machine learning methods. *Applied Sciences*, 13(3), Article 1555. <https://doi.org/10.3390/app13031555>
- Nguyen, H.-H.-D., Nguyen, T.-N., Phan, T.-A.-T., Huynh, N.-T., Huynh, Q.-D., & Trieu, T.-T. (2025). Explainable artificial intelligence model for the prediction of undrained shear strength. *Theoretical and Applied Mechanics Letters*, 15, Article 100578. <https://doi.org/10.1016/j.taml.2025.100578>

- Niu, G., He, X., Xu, H., & Dai, S. (2024). Development of rock classification systems: A comprehensive review with emphasis on artificial intelligence techniques. *Eng*, 5(1), 217–245. <https://doi.org/10.3390/eng5010012>
- Onyelowe, K. C., Mojtahedi, F. F., Ebid, A. M., Rezaei, A., Osinubi, K. J., Eberemu, A. O., Salahudeen, B., Gadzama, E. W., Rezazadeh, D., Jahangir, H., Yohanna, P., Onyia, M. E., Jalal, F. E., Iqbal, M., Ikpa, C., Obianyo, I. I., & Rehman, Z. U. (2023). Selected AI optimization techniques and applications in geotechnical engineering. *Cogent Engineering*, 10(1), Article 2153419. <https://doi.org/10.1080/23311916.2022.2153419>
- Ozsagir, M., Erden, C., Bol, E., Sert, S., & Ozocak, A. (2022). Machine learning approaches for prediction of fine-grained soils liquefaction. *Computers and Geotechnics*, 152, Article 105014. <https://doi.org/10.1016/j.compgeo.2022.105014>
- Pham, B. T., Son, L. H., Hoang, T.-A., Nguyen, D.-M., & Bui, D. T. (2018). Prediction of shear strength of soft soil using machine learning methods. *Catena*, 166, 181–191. <https://doi.org/10.1016/j.catena.2018.04.004>
- Phoon, K.-K., & Kulhawy, F. H. (1999). Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36, 612–624.
- Phoon, K.-K., & Zhang, W. (2023). Future of machine learning in geotechnics. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 17(1), 7–22. <https://doi.org/10.1080/17499518.2022.2087884>
- Rauter, S., & Tschuchnigg, F. (2021). CPT data interpretation employing different machine learning techniques. *Geosciences*, 11(7), Article 265. <https://doi.org/10.3390/geosciences11070265>
- Samui, P., & Sitharam, T. G. (2011). Machine learning modelling for predicting soil liquefaction susceptibility. *Natural Hazards and Earth System Sciences*, 11, 1–9. <https://doi.org/10.5194/nhess-11-1-2011>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2, Article 160. <https://doi.org/10.1007/s42979-021-00592-x>

Shao, W., Yue, W., Zhang, Y., Zhou, T., Zhang, Y., Dang, Y., Wang, H., Feng, X., & Chao, Z. (2023). The application of machine learning techniques in geotechnical engineering: A review and comparison. *Mathematics*, 11(18), Article 3976. <https://doi.org/10.3390/math11183976>

Song, M., Yang, M., Yao, G., Chen, W., & Lyu, Z. (2024). Artificial intelligence driven tunneling-induced surface settlement prediction. *Automation in Construction*, 168, Article 105819. <https://doi.org/10.1016/j.autcon.2024.105819>

Sun, J., Zhang, R., Zhang, A., Wang, X., Wang, J., Ren, L., Zhang, Z., & Zhang, Z. (2024). Rock strength prediction based on machine learning: A study from prediction model to mechanism explanation. *Measurement*, 238, Article 115373. <https://doi.org/10.1016/j.measurement.2024.115373>

Transport for NSW. (2014). *Geotechnical risk assessment and hazard management* (T HR CI 12100 ST, Version 1.0). State of NSW through Transport for NSW.

Xu, D., Wang, Y., Huang, J., Liu, S., Xu, S., & Zhou, K. (2023). Prediction of geology condition for slurry pressure balanced shield tunnel with super-large diameter by machine learning algorithms. *Tunnelling and Underground Space Technology*, 131, Article 104852. <https://doi.org/10.1016/j.tust.2022.104852>

Xu, H., Zhou, J., Asteris, P. G., Armaghani, D. J., & Tahir, M. M. (2019). Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate. *Applied Sciences*, 9(18), Article 3715. <https://doi.org/10.3390/app9183715>

Yaghoubi, E., Yaghoubi, E., Khamees, A., & Vakili, A. H. (2024). A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering. *Neural Computing and Applications*, 36, 12655–12699. <https://doi.org/10.1007/s00521-024-09893-7>

Zhang, D., Shen, Y., Huang, Z., & Xie, X. (2022). Auto machine learning-based modelling and prediction of excavation-induced tunnel displacement. *Journal of Rock Mechanics and Geotechnical Engineering*, 14(4), 1100–1114. <https://doi.org/10.1016/j.jrmge.2022.03.005>

- Zhang, R., Li, Y., Goh, A. T. C., Zhang, W., & Chen, Z. (2021). Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models. *Journal of Rock Mechanics and Geotechnical Engineering*, 13(6), 1478–1484. <https://doi.org/10.1016/j.jrmge.2021.08.001>
- Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021). Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review*, 54, 5633–5673. <https://doi.org/10.1007/s10462-021-09967-1>
- Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, 12(1), 469–477. <https://doi.org/10.1016/j.gsf.2020.03.007>
- Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., & Zhang, Y. (2022). Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. *Gondwana Research*, 109, 1–17. <https://doi.org/10.1016/j.gr.2022.03.015>
- Zhang, W., Gu, X., Hong, L., Han, L., & Wang, L. (2023). Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges. *Applied Soft Computing*, 136, Article 110066. <https://doi.org/10.1016/j.asoc.2023.110066>
- Zhao, T., & Wang, Y. (2020). Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods. *Engineering Geology*, 265, Article 105430. <https://doi.org/10.1016/j.enggeo.2019.105430>
- Zhou, J., Li, E., Yang, S., Wang, M., Shi, X., Yao, S., & Mitri, H. S. (2019). Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. *Safety Science*, 118, 505–518. <https://doi.org/10.1016/j.ssci.2019.05.046>
- Zhou, M., Chen, J., Huang, H., Zhang, D., Zhao, S., & Shadabfar, M. (2021). Multi-source data driven method for assessing the rock mass quality of a NATM tunnel face via hybrid ensemble learning models. *International Journal of Rock Mechanics and Mining Sciences*, 147, Article 104914. <https://doi.org/10.1016/j.ijrmms.2021.104914>

Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W., & Chiam, K. (2021). Prediction of rockhead using a hybrid N-XGBoost machine learning framework. *Journal of Rock Mechanics and Geotechnical Engineering*, 13(5), 1231–1245. <https://doi.org/10.1016/j.jrmge.2021.06.012>

Author's Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.