Machine-learning approach to site classification

L.A. Brits
PeraGage, Cape Town, South Africa

C.J. MacRobert
Stellenbosch University, Stellenbosch, South Africa

ABSTRACT: This paper presents a Machine Learning approach for site classification according to the Geotechnical Site Investigations for Housing Developments (GSFH-2). Based on input from practising engineers and geologists a flowchart was developed to classify 430 individual soil layers according to GSFH-2. These soil layers were classified according to the expected soil movement by considering the moisture condition, colour, consistency, structure, texture and origin of the soil (MCCSTO). Three machine learning models, namely: Support Vector Machine (SVM), Decision Tree, and Random Forest models, were then developed using the database. Term Frequency – Inverse Document Frequency (TF-IDF) was the primary Natural Language Processing (NLP) method used to process and analyse the text input in combination with N-grams and other pre-processing techniques. Evaluation used Feature Importance, Confusion Matrices, and statistical metrics. The results indicated that the Random Forest-model in combination with lower-casing achieved the best performance with an accuracy-score of 71%. The accuracy of the proposed Random Forest-model could be increased by user verification of predictions made on unlabelled data.

1 INTRODUCTION

To date, machine learning in geotechnical engineering has primarily focused on quantitative or numerical data. However, geotechnical engineers often rely on descriptions, qualitative or language-based information to guide decisions. This paper presents a machine learning approach to classify soil layers that takes test pit profile descriptions to classify soil layers according to the South African Geotechnical Site Investigations for Housing Developments (GSFH-2) standard.

2 SITE CLASSIFICATION

It is required by the Geotechnical Site Investigations for Housing Developments (GFSH-2 2002) that a provisional site classification should be derived by interpreting the soil profile and founding recommendations made according to the site class. Site classes are derived from an estimation of the range of expected soil movement experienced by single-storey and double-storey type 1 masonry buildings, where the foundation width is limited to 0.6 m for single-storey buildings and 0.8 m for double-storey buildings and the load on the foundation does not cause the soil bearing pressure to exceed 50 kPa. Table 1 lists the different site classes with their respective expected soil movement.

Table 1. Site class designations of single- and double-storey type 1 masonry buildings according to SANS 10400-H, 2012.

Nature of founding material	Expected soil movement (mm)	Site class
Stable	Negligible	R
	<7.5	H
	7.5 to 15	H1
Expansive	15 to 30	H2
	>30	H3
Compressible and collapsible	<5	C
	5 to 10	C1
	>10	C2
Compressible	<10	S
	10 to 20	S1
	>20	S2
Fill	Variable	P

Consecutive soil layers in a profile should be described using terms provided by Guidelines for Soil and Rock Logging in South Africa published by the South African Institute for Engineers and Engineering Geologists. Each soil layer is described according to its moisture condition, colour, consistency, structure, soil texture, and origin (MCCSTO).

3 MCCSTO SOIL PROFILING

Moisture condition is a relative indication of water content, ranging between dry and wet and therefore depends on the soil type (Brink & Bruin 2002, Jennings et al. 1973). Soil colour aids in the identification of expansive clays which are often described as dark grey, black, maroon, and mottled yellow grey (DPW 2007). Soils with collapse potential are well-drained soils and are often red-brown, brown or light brown (Day 2016). Consistency is the measure of a soil's hardness or toughness. Since drainage and permeability influence shear strength, a distinction between the consistency descriptions of cohesive and non-cohesive material should be made (Jennings et al. 1973).

Soil structure refers to the presence and nature of joints in the soil (Jennings et al. 1973). Soil type is described according to the proportional composition of the soil based on grain size. Particle size classes should be written in increasing order of abundance with the predominant size class written in uppercase (Dippenaar et al. 2024). Knowledge of the local geology and recognition of preserved primary rock structures are useful for the identification of the origin of residual soils while a close relationship exists between the landform and the origin of a transported soil (Jennings et al. 1973).

4 CLASSIFICATION FLOWCHART

A structured approach was used to perform the classification of soil layers in the soil profiles from test pit logs to ensure consistency in the labelled data that was used to train the machine learning models. A flowchart was developed in collaboration with experts in the field. The flowchart follows a step-by-step path with logical tests to identify the appropriate site class and starts by selecting the dominant soil texture.

4.1 *Clay*

The flowchart for clays is displayed in Figure 1. For a clay-dominant soil, the first test was to decide whether a clay was potentially expansive based on the origin of the soil. If the clay origin was not potentially expansive, settlement was considered rather than heave and the flowchart for Class S used. The structure of the clay was then considered. If slickensiding and shattering (Netterberg 2019) were indicated, the layer was assigned to Class H3. For fissured and intact layers, the soil colour was considered. If a fissured clay was black, dark-grey, maroon, or mottled, it was assigned Class H2; if not, it was assigned Class H1. A black, dark-grey, maroon, or mottled intact clay was also assigned to Class H1. If an intact clay was not black, dark-grey, maroon, or mottled, the moisture condition was considered; very moist or wet clay of low expansive potential was assumed to undergo greater settlement (i.e. Class S) than heave (Class H).

4.2 *Sand*

The classification flowchart, shown in Figure 2, is based on the typical field characteristics of collapsible soils. Collapsible soils are clayey or silty sands with low moisture content, dense consistency, and pinhole-voided structure. The origin of the sand was also included as criteria for classification. The origin of collapsible soils includes various transported soils and certain residual soils such as the residual granites of the Basement Complex. If these conditions were not met, the flowchart for Class S was used. Subdivision within Class C was made based on consistency and the description of the voids. Soils classified as Class C were loose and weakly voided, while those in Class C1 were medium-dense with weak voiding. Class C2 soils were dense with a pinhole-voided structure.

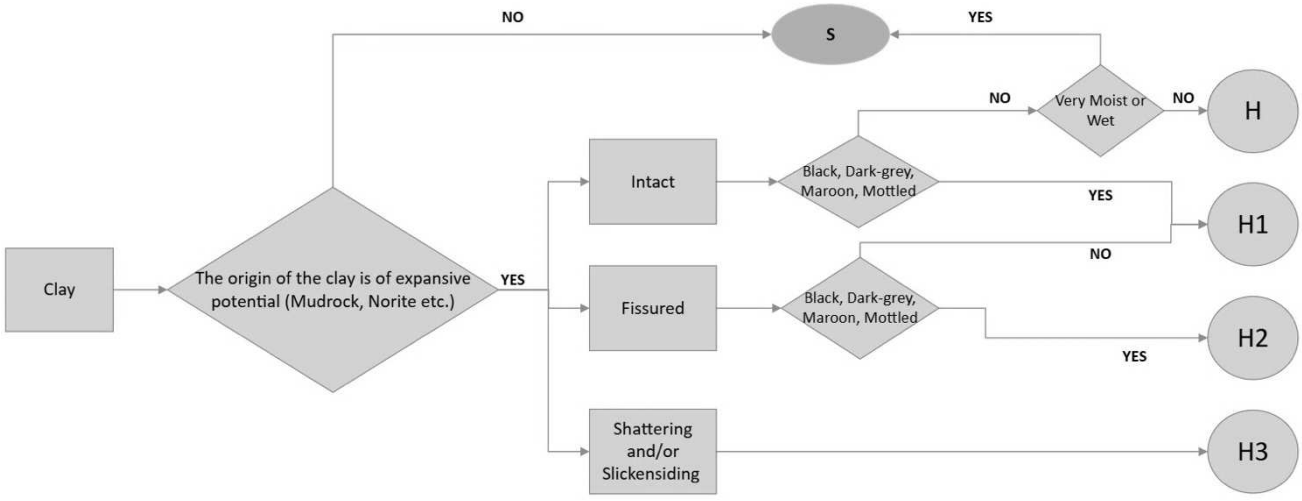


Figure 1. Classification flow-chart for class H, H1, H2 and H3

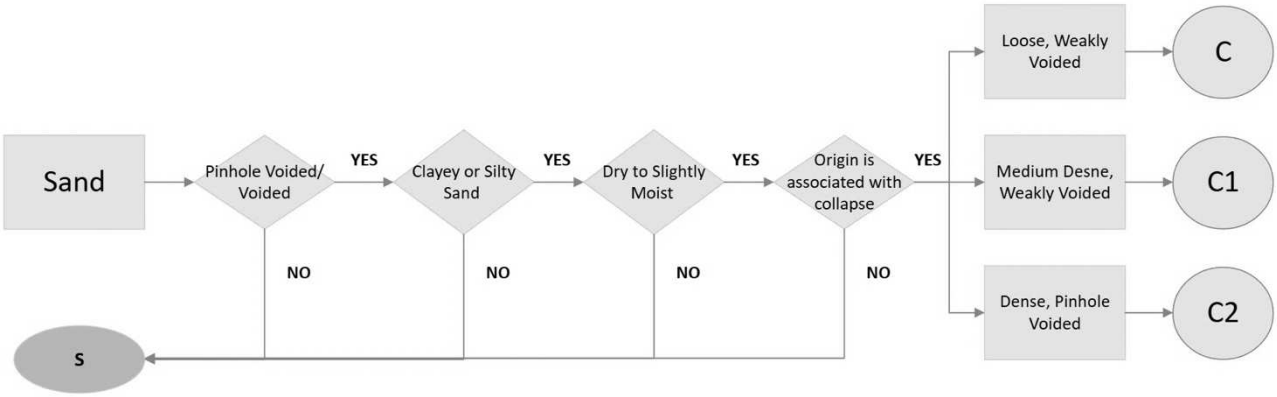


Figure 2. Classification flow-chart for Class C, C1 and C2

4.3 *Silts, sands not in Class C and clays not in Class C*

For silts and sands and clays that do not form part of classes C and H respectively, the criteria for classification were based solely on the consistency of the soil (Figure 3). If the consistency was described as a range (e.g., medium dense to dense), the lower bound description (medium dense) was used for classification.

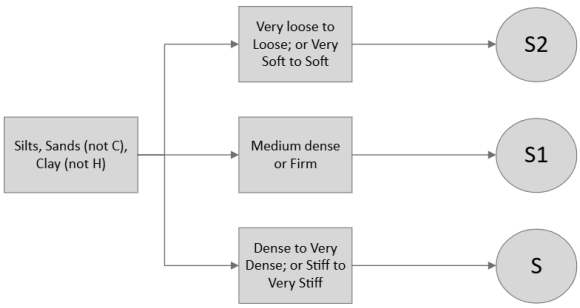


Figure 3. Classification flow-chart for Class S, S1, S2

4.4 *Fill*

When classifying fill, a distinction was made between controlled and uncontrolled fill. Controlled fill such as compacted engineered fill was labelled as Class PC. Uncontrolled fill such as construction rubble and end tipped soil was labelled as Class PU.

5 MACHINE LEARNING METHODOLOGY

5.1 *Data labelling*

Using soil profiles obtained from various South African consultants, 430 soil layers were classified to obtain a uniform split between Classes. A subset of 384 layers was used for training and testing of the machine-learning models. The validation dataset consisted of 46 soil layers.

An excerpt of input data is illustrated in Figure 4. A semicolon was used to separate attributes within the MCCSTO descriptions, while a hashtag was employed to delimit the layer description from the class label.

```
Slightly moist; dark reddish orange; soft; intact; silty CLAY;
Residual Norite#H
Slightly moist; reddish dark brown speckled light brown; firm;
shattered and slickensided; sandy CLAY; Residual Norite#H3
```

Figure 4. Model input data

5.2 *Pre-processing*

Lowercasing was applied to all three machine learning models to reduce noise within the data. Bigrams and lemmatization were additionally used for the Decision Tree-model. Bigrams capture pairs of consecutive words to preserve word order (e.g. slightly moist is a bigram). Lemmatization transforms a word to its root form (lemma) which preserves the semantics as well as the definition of a word (e.g., fissuring becomes fissure).

5.3 *Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer*

For a machine to read the descriptive data, words need to be transformed to numerical data that represents words in the shape of low-dimensional vectors. The term frequency measures how often a word appears in a document:

$$TF = \frac{Number\ of\ times\ word\ appears\ in\ document}{Total\ number\ of\ words\ in\ document} \quad (1)$$

The inverse document frequency measures the number of documents in a corpus that contains a certain word *w* and is therefore a measure of importance of a word:

$$IDF = log\left(\frac{Total\ number\ of\ documents\ in\ corpus}{Number\ of\ documents\ containing\ word\ w}\right) \quad (2)$$

The TF-IDF score is expressed as the product of the TF and IDF:

$$TF - IDF = TF \times IDF \quad (3)$$

5.4 *Machine learning models*

Three machine learning models were tested: Support vector machine learning, Decisions Trees and Random Forests. The Support Vector Machine (SVM) learning is suitable for linear and nonlinear classification of complex small- to medium-sized datasets. SVM finds a decision boundary (hyperplane) that best separates training instances from different classes. A hyperplane is a $n - 1$ dimensional plane that separates the n dimensional feature space of the training instances into two distinct regions (Nguyen et al. 2022). The training instances located on two parallel hyperplanes are called support vectors (Géron 2019).

A Decision Tree is a supervised model used for classification tasks by recursively splitting the instance space into subspaces. This process simplifies a complex decision-making task into a series of binary decisions, leading to a final prediction in the form of a class label or termination node (Rokach & Maimon 2005). Splitting of internal nodes is done according to a single attribute and the model will consequently search for the best attribute upon which to split the node (Rokach & Maimon 2005).

Random Forests are Decision Tree-based ensemble machine learning models with each tree being trained on different random samples of the data (Cutler et al. 2011). Instead of finding the best attribute at a node for splitting, Random Forests search for the best attribute among a random subset of features (feature bagging). Some of the instances may be sampled multiple times for any given predictor, while other instances are not sampled at all. The unsampled training instances are referred to as out-of-bag instances and can be used to evaluate the model without the need of a separate validation set since the model have not seen these samples (Géron 2019).

5.5 *Performance measures*

Precision, recall, F1-score and accuracy were metrics used to evaluate the classifications performance of the three machine learning models. Precision is defined as the ratio of true positive predictions to the total predicted positives. Recall measures the ratio of true positive predictions to the total actual positives.

The F1-score is the harmonic mean of precision and recall. Therefore, a high F1-score is an indication of a good precision and recall. Accuracy is defined as the proportion of true predictions (positive and negative) decided by all predictions.

Confusion matrices plot model-predicted labels against the true-labels and were used during the final model evaluation to obtain a visual representation of the classifier performance. Learning curves were used to predict the size a training set must be for the model accuracy to stabilize.

6 MODEL SELECTION

The subset of data used for training and testing was split into a training dataset (70%) and testing dataset (30%). The machine learning models were trained and tested using different combinations of NLP-techniques. Both accuracy and F1-score were considered for the selection of the final model that will be tested against the validation dataset. The best performing SVM-model, with lower-casing as the only NLP-technique, achieved an accuracy of 70%. The use of bigrams and lemmatization resulted in the decision tree model with the highest accuracy with a score of 70%. The random forest model with only lower-casing achieved the best overall performance with accuracy of 71% and a F1-score standard deviation of 0.18. The results of the three models are summarized in Table 2.

Table 2. Comparison of F1-score performance between models

Model	SVM	Decision Tree	Random Forest
Accuracy	70%	70%	71%
Average F1	0.7	0.7	0.7
F1 standard deviation	0.24	0.22	0.18

The trained Random Forest-model with only lower-casing was selected as the best model based on the higher accuracy score compared to the SVM and Decision Tree-model, as well as a lower standard deviation of F1-score. The lower standard deviation indicates a more consistent performance across all classes compared to the SVM and Decision Tree. The performance metrics for the chosen model is displayed in Table 3.

Table 3. Accuracy report of Random Forest-model on test dataset

Class	Precision	Recall	F1-score
C	1.00	0.60	0.75
C1	0.55	0.75	0.63
C2	0.71	0.83	0.77
H	0.40	0.40	0.40
H1	0.60	0.43	0.50
H2	1.00	0.80	0.89
H3	0.90	1.00	0.95
S	0.67	0.86	0.75
S1	0.50	0.29	0.36
S2	0.78	0.88	0.82
PC	0.75	0.75	0.75
PU	0.83	0.83	0.83
Accuracy	71%		

7 FINAL MODEL EVALUATION

The validation dataset was used to evaluate the performance of the random forest model. The accuracy of the model, evaluated against the validation dataset reduced from 71% to 59% which suggest that the model was overfitted to the training data and does not generalize well when seeing new instances. The performance measures of the random forest model for all classes are summarized in Table 4 and a confusion matrix is plotted in Figure 5 to visually represent classifications.

Table 4. Accuracy report of Random Forest-model on validation dataset

Class	Precision	Recall	F1-score
C	0.75	0.75	0.75
C1	0.33	0.25	0.29
C2	0.60	1.00	0.75
H	0.50	0.50	0.50
H1	0.25	0.25	0.25
H2	1.00	0.75	0.86
H3	0.80	1.00	0.89
S	0.38	1.00	0.55
S1	0.50	0.25	0.33
S2	0.67	0.5	0.57
PC	1.00	0.67	0.80
PU	1.00	0.33	0.50
Accuracy	59%		

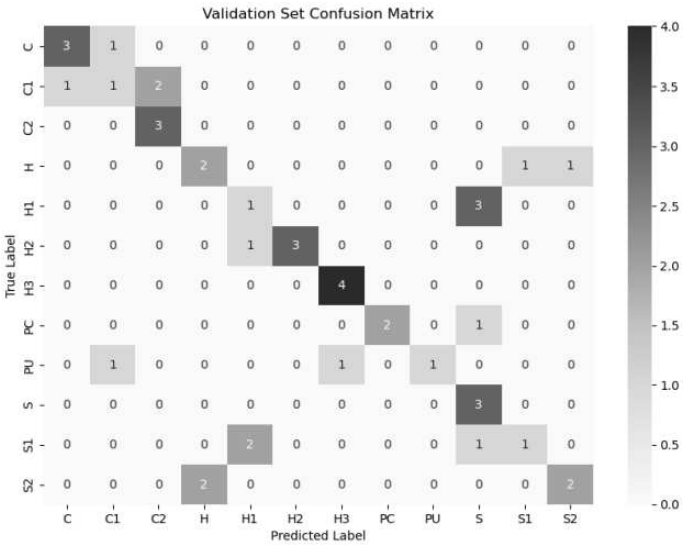


Figure 5. Validation dataset confusion matrix

The model generalizes well for new instances of Class C and C2 as no significant decrease in F1-score is observed. The F1-score for classes H2 and H3 remained high which confirms that the model recognized the terms fissured and slickensided and can classify most instances of class H2 and all instances of class H3 correctly.

However, some scatter between sub-C-classes are present, the model can generally distinguish C-classes from other classes. Confusion occurs between class H1 and subclasses of Class S. The criteria for the classification of class H in Figure 1, which is based on the expected soil movement from heaving, does not consider the soil consistency. Therefore, if the origin of the expansive clay is not captured, the model can easily classify a class H1 instance as class S when the
model considers consistency as a more important feature and the clay is described as stiff to very stiff. This can possibly be resolved by increasing the size of the training dataset so that the model has enough instances to capture the origins of expansive soils as important features to better distinguish between classes H and S.

To predict the number of instances required to achieve a model accuracy of 80%, a logarithmic- and power-law-curve is fitted over the validation curve (see Fig. 6). For the logarithmic- and power-law curve it can be estimated that 1470 and 514 training instances are required respectively to achieve a model accuracy of 80%.

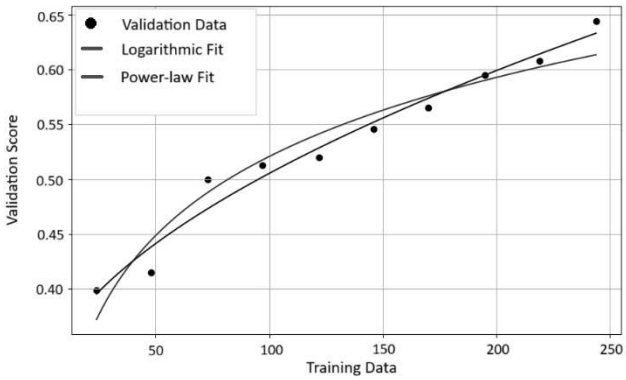


Figure 6. Logarithmic- and log-power- learning curves

8 CONCLUSIONS

A flowchart was developed to classify soil layers from test pit-log descriptions according to Geotechnical Site Investigations for Housing Developments (GFSH-2 2002). The classified data was used to train a Support Vector Machine-, Decision Tree- and Random Forest-model. The Random Forest model, with only lowercasing applied as a language pre-processing technique, achieved the best overall performance on the test dataset. The reduction in accuracy when the model was evaluated against a validation dataset indicates that the model is overfitted to the training data. The model classifies instances of class H3 with success which indicates that the model considers the terms shattering and slickensiding as important features. The Random Forest model had difficulty to distinguish between subclasses of Class S and Class H1. This can be resolved by increasing the number of training instances. By fitting a logarithmic and power-law learning curve to the validation curve, it is predicted that between 1470 and 514 validation

instances would be required to achieve model accuracy of 80%.

ACKNOWLEDGEMENTS

Professor Peter Day (Stellenbosch University, Jones and Wagner) and Mr. Tony A'Bear (Bear Geo Consultants) are acknowledged for their guidance in developing the classification flowcharts. Mr Eldon Burger (Stellenbosch University) provided guidance on machine learning.

REFERENCES

Brink, A.B.A. & Bruin, R.M.H. 2002. *Guidelines for Soil and Rock Logging in South Africa*. Pretoria: South African Institution of Civil Engineering (SAICE).

Day, P. 2016. Soil behaviour: A practical perspective 2: Collapsible soils.

DPW (Department of Public Works). 2007. *Identification of problematic soils in Southern Africa: Technical notes for civil and structural engineers*. Pretoria: Department of Public Works.

Dippenaar, M., Van Rooyen, J. & Davis, G. 2004. *Engineering geological soil and rock description*. South African Institute for Engineering and Environmental Geologists (SAIEG).

Géron, A. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media Incorporated.

GFSH-2 2002. Geotechnical Site Investigations for Housing Developments. National Department of Housing.

Jennings, J., Brink, A. & Williams, A. 1973. Revised guide to soil profiling for civil engineering purposes in southern Africa. *The Civil Engineer in South Africa* (15).

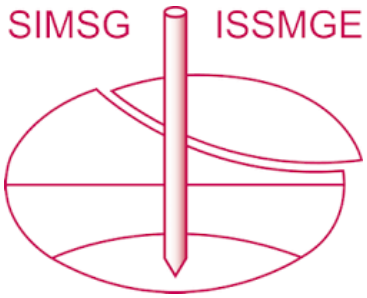
Netterberg, F. 2019. Identification of potentially expansive clay soils from soil structure. *Proceedings of the 17th African Regional Conference on Soil Mechanics and Geotechnical Engineering. 7-9 October 2019.* Cape Town.

Nguyen, M., Costache, R., Sy, A., Ahmadzadeh, H. Le, V. Prakash, I. Binh, A. and Pham, T. 2022. Novel approach for soil classification using machine learning methods. *Bulletin of Engineering Geology and the Environment* 81: 1-17.

Rokach, L. & Maimon, O. 2005. *Data Mining and Knowledge Discovery Handbook*. Boston: Springer US.

SABS Standards Division. 2012. SANS 10400-H: The application of the National Building Regulations Part H: Foundations.

INTERNATIONAL SOCIETY FOR
SOIL MECHANICS AND
GEOTECHNICAL ENGINEERING



This paper was downloaded from the Online Library of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE). The library is available here:

https://www.issmge.org/publications/online-library

This is an open-access database that archives thousands of papers published under the Auspices of the ISSMGE and maintained by the Innovation and Development Committee of ISSMGE.

The paper was published in the proceedings of the 2nd Southern African Geotechnical Conference (SAGC2025) and was edited by SW Jacobsz. The conference was held from May 28th to May 30th 2025 in Durban, South Africa.