Date of Publication: 25th August 2025
ISSN: 2737-713X

APPLICATION OF MACHINE LEARNING IN SOIL LIQUEFACTION PREDICTION: THE CASE OF ACCRA'S ACTIVE SEISMIC ZONES

*[1]Albert Kafui Klu, [1]Michael Affam, [1]Anthony Ewusi, [1]Yao Yevenyo Ziggah and [2]Emmanuel Asiedu Brempong

[1] *Faculty of Geosciences and Environmental Studies, University of Mines and Technology, P. O. Box 237, Tarkwa,*
[2] *Google Ghana Office, Accra.*
***Corresponding author:** akklu@umat.edu.gh

Abstract

Soil liquefaction presents major risks to infrastructure in earthquake-prone areas. Machine learning (ML) algorithms enhance prediction accuracy; however, data imbalance poses a challenge. As one of the foremost papers on the application of ML for soil liquefaction prediction in Ghana, this study utilises a dataset comprising Standard Penetration Test (SPT), Groundwater Level (GWL), Relative Density (Dr), Natural Moisture Content (NMC), Atterberg Limits and Particle Size Distribution (PSD) from southwestern Accra, Ghana. The positive class in the dataset, representing liquefaction events, made up only 12.64%, requiring the use of the Synthetic Minority Over-sampling Technique (SMOTE) to correct the class imbalance. Evaluated were three machine learning models: Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Classifier (SVC). SMOTE enhanced recall, with LR rising from 0.50 to 0.88 and RFC from 0.50 to 0.75. The recall for SVC increased to 0.62. Precision decreased for LR (0.67 to 0.50) but increased for SVC (0.71), resulting in the highest F1-score (0.67). RFC achieved a precision of 0.67 and an F1-score of 0.71. SVC showed the highest predictive performance, with accuracy rising from 0.92 to 0.94 following SMOTE application. These findings underscore the potential of alternative machine learning methods in geotechnical engineering and the significance of SMOTE in improving predictions for imbalanced datasets. The study highlights the need to assess metrics beyond accuracy, especially in imbalanced datasets, to ensure reliable predictions in risk-sensitive areas such as soil liquefaction assessment.

Keywords
Liquefaction, Machine Learning, Logistic Regression, Random Forest, Support Vector Classification, SMOTE, Seismic Hazard

Introduction

In the event of an applied stress, such as an earthquake trigger or ground-vibrating event where seismic waves are generated, saturated, loose, granular soils are likely to undergo liquefaction, leading to a loss in their strength and stiffness (Lenart, 2008; Sarkar et al., 2020). When soil gets liquefied, it does not behave like a solid anymore but as a liquid due to the increase in pore water pressure that approaches or even exceeds the confining pressure. Consequently, ground failure can occur due to abrupt soil weakening, leading to the collapse, burial or toppling of buildings and other infrastructure (Goudarzy et al., 2022; Huang and Zhao, 2018; Mollica et al., 2020). In most of the major earthquakes that have struck around the world, soil liquefaction has been mentioned quite in a number of them. From the Mexico City earthquake of 1985, the San Francisco earthquake in 1906, the Kocaeli, Turkey earthquake in 1999 and that of Wenchuan 2008 among others (O'Rourke et al., 2006; Sancio et al., 2003; Verma et al., 2014; Zhou et al., 2020). Considering the devastating consequences of earthquake-induced soil liquefaction, there is the need to thoroughly investigate the possibility of the phenomenon in earthquake-prone areas.

Ghana, a Sub-Saharan African country has been known for her history with respect to seismic activities. The first ever documented earthquake in Ghana occurred within the Elmina township in 1615. However, the first destructive earthquake was recorded in Axim in December, 1636. The most destructive earthquake in the nation's history occurred in Accra on June 22, 1939 with a Richter magnitude of 6.5, causing 130 injuries, 17 deaths and destruction of many buildings and properties (Amponsah, 2004; Amponsah et al., 2012; Kutu, 2013). Since then, there have been studies conducted into the seismicity of Ghana from different perspectives. One of these is the potential of soil liquefaction occurrence in the event of an earthquake in Ghana, especially in the capital city Accra, which currently happens to be the most seismically active area in the country (Amponsah, 2021; Atarigiya et al., 2023; Klu et al., 2024; Nortey et al., 2018). A study conducted by Atarigiya et al. (2023) implied that there are areas in Ghana which are prone to soil liquefaction occurrence in the event of a major earthquake. This therefore calls for further liquefaction investigations in Ghana with more emphasis on how to accurately predict the phenomenon.

The prediction of soil liquefaction is a primary concern in geotechnical engineering because of its significant effects on infrastructure during seismic occurrences. Traditionally, soil liquefaction is predicted with empirical approaches based on in-situ assessments, such as the Standard Penetration Test (SPT) and Cone Penetration Test (CPT) (Arango-Serna et al., 2021; Cubrinovski et al., 2018; Muduli et al., 2014). Techniques such as Numerical methods, Empirical methods, Statistical models and Machine learning models (Mohammadnejad and Andrade, 2015; Muduli and Das, 2014; Venkatesh et al., 2013) have been used over the years in soil liquefaction

analysis and prediction. However, these methods have some limitations that discourage their use in predicting soil liquefaction such as their time-consuming nature, computational complexities, lack of good simulation abilities, underestimation of liquefaction-induced one-dimensional settlements and their insufficiency for parameter determination of physical phenomena like liquefaction due heterogeneity of the earth among many others (Mansouri and Dabiri, 2021; Muduli and Das, 2013; Ziotopoulou and Boulanger, 2013).

Nonetheless, the emergence of machine learning (ML) methodologies has facilitated the creation of advanced models for predicting liquefaction through the use of extensive, intricate datasets. Artificial Neural Networks (ANNs) have become one of the most common ML techniques employed for soil prediction purposes (Cha et al., 2008; Farrokhzad et al., 2012; Young-Su and Byung-Tak, 2006). However, there are different ML approaches that could yield equal or even better predictive outcomes. Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Classification (SVC) have proven to be effective methodologies for predicting liquefaction (Gandomi et al., 2013; Jairi et al., 2021; Kohestani et al., 2015; Samui and Sitharam, 2011). This research employs machine learning methods to predict liquefaction potential, utilising geotechnical characteristics obtained from datasets such as SPT, relative density, groundwater levels and Atterberg limits. The aim is to evaluate the efficacy of three common classifiers – Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Classification (SVC) – and examine the influence of data balance through SMOTE.

Logistic Regression is a fundamental binary classification model often utilised in liquefaction studies because of its simplicity and interpretability. Logistic regression models the probability of a binary result (e.g., liquefaction versus non-liquefaction) based on a collection of predictor factors. It is especially appropriate for issues where the response variable is categorical. Numerous studies have employed logistic regression for liquefaction prediction utilising geotechnical and seismic information, including SPT, CPT, groundwater level and soil characteristics. Goh (1994) was an early proponent of logistic regression in geotechnical applications, creating a prediction model utilising CPT data for liquefaction evaluation. The research demonstrated that logistic regression could accurately forecast the probability of liquefaction, especially when high-quality field data were accessible. Recent uses of logistic regression in liquefaction prediction have incorporated a broader range of information, including shear wave velocity, soil gradation and Atterberg limits (Jairi et al., 2021; Tober, 2020). Despite its simplicity, logistic regression may fail to encapsulate the intricate, non-linear correlations present in geotechnical data, hence constraining its predictive efficacy relative to more advanced models (Grossi and Buscema, 2007).

The Random Forest Classifier is a robust ensemble learning technique that integrates numerous decision trees to enhance prediction accuracy and mitigate overfitting. It has garnered considerable attention in soil liquefaction prediction owing to its capacity to model non-linear interactions and manage high-dimensional datasets with minimal adjustment. Random Forest has been extensively utilised in current liquefaction research. Kohestani et al. (2015) employed RF models to forecast liquefaction utilising CPT data, attaining superior accuracy relative to conventional techniques such as ANN. The capability of RFC to manage both continuous and categorical variables renders it adaptable for the integration of varied geotechnical data, including soil density, moisture content and seismic intensity. Research indicates that RFC excels in assessing feature importance, which is essential for elucidating the relative impact of variables such as SPT and CPT on liquefaction potential (Nejad et al., 2018). Owing to its ensemble characteristics, Random Forest Classifier diminishes the variance observed in individual decision trees, hence yielding more stable and dependable predictions in intricate geotechnical scenarios. Nonetheless, Random Forest Classifiers (RFC) may encounter difficulties with interpretability, as the "black-box" characteristic of ensemble approaches complicates the direct correlation between input features and output, in contrast to logistic regression (Kaya et al., 2023).

Support Vector Classification (SVC) is a machine learning method that creates hyperplanes to differentiate data into discrete categories. It is especially efficacious in high-dimensional spaces and is recognised for its robust theoretical principles in maximising the margin between classes, rendering it appropriate for intricate classification tasks such as liquefaction prediction (Cervantes et al., 2020). Support Vector Classification (SVC) has been progressively utilised in geotechnical issues, particularly in liquefaction forecasting. Samui (2013) illustrated the efficacy of SVC in categorising soil samples as liquefied or non-liquefied, utilising data such as shear wave velocity and soil characteristics. Support Vector Classification (SVC) is recognised for its ability to manage non-linear interactions using kernel functions, rendering it a compelling option for modelling intricate geotechnical phenomena (Goh and Goh, 2007). While SVC can attain elevated accuracy in liquefaction prediction, its efficacy is significantly influenced by the selection of kernel and the regularisation parameters. Optimising these hyperparameters can be resource-intensive, particularly for extensive datasets (Cervantes et al., 2020).

The selection of Logistic Regression (LR), Random Forest Classifier (RFC) and Support Vector Classification (SVC) was based on their proven efficacy in handling geotechnical datasets and their ability to model both linear and non-linear relationships (Ardakani and Kohestani, 2015; Liu et al., 2024; Zhang and Wang, 2021). Logistic regression was selected due to its straightforward nature and clarity, which is especially beneficial in geotechnical contexts where grasping the impact of specific factors is essential. The choice of RFC stems from its strong performance in managing high-dimensional datasets and its proficiency in identifying intricate, non-linear relationships without the need for extensive hyperparameter adjustments. The inclusion of SVC is attributed to its proven

effectiveness in high-dimensional spaces and its capacity to maximise the margin between classes, which renders it particularly suitable for handling imbalanced datasets (Fadliansyah et al., 2024; Gandomi et al., 2013; Talamkhani et al., 2023). While other models like Gradient Boosting and Deep Learning have been used in geotechnical engineering, this study focuses on these three models to provide a comparative analysis of their performance in soil liquefaction prediction, particularly in the context of imbalanced datasets.

A significant problem in predicting soil liquefaction is the dataset's uneven nature. Liquefied and non-liquefied cases frequently occur in disproportionate ratios, with non-liquefied examples generally prevailing in the dataset. This disparity may result in biased models, wherein machine learning algorithms predominantly forecast the majority class (non-liquefaction), hence diminishing the credibility of liquefaction predictions (Hu et al., 2016). The Synthetic Minority Oversampling Technique (SMOTE) is extensively utilised to enhance model performance by balancing the dataset. In the realm of liquefaction prediction, SMOTE has been utilised to equilibrate datasets, enhancing the efficacy of models such as LR, RFC and SVC. The implementation of SMOTE markedly improves the precision and recall of machine learning models in predicting liquefaction. By augmenting the quantity of liquefied instances in the training dataset, SMOTE guarantees that the model allocates equal focus to both classes, hence enhancing its capacity to identify liquefaction events (Demir and Sahin, 2022; Song et al., 2024).

In this study, three different machine learning predictive models were used to classify soil liquefaction occurrence in the seismically active areas of Ghana's capital city. LR, RFC and SVC models were built and tested with and without the use of SMOTE. The intentions of this study were to develop a predictive machine learning model for soil liquefaction in Accra and also to assess the influence of data imbalance on prediction models.

Materials and Methods

Study Areas

The geotechnical data used for this analysis was collected from the Architectural and Engineering Services Limited (AESL), Accra which covered twenty-seven (27) surveyed locations in the southwestern areas of the Greater Accra Region of Ghana. These areas are Accra Central, Burma Camp, Korle Bu, Legon, North Ridge, Ofankor, Roman Ridge, Weija North, McCarthy Hill, Kokrobite, New Bortianor, Bortianor, Teshie, Osu, Cantonments, Airport City, Dzorwulu, Alajo, North-Kaneshie, Kaneshie, Dansoman South, Tetegu East, Tetegu West, Dansoman North, Gbawe, Weija South and Nyanyanor. The survey areas are bounded to the north at 5° 37' 30" N, to the south at 5° 31' 00" N, the west at 0° 17' 30" W and to the east at 0° 07' 00" W (Figure 1). Majority of the 27 sites are located in the central business areas, where major infrastructural developments in the region can be found. Varying categories of structures are located in the area; ranging from high-rise buildings to small structures, meant for different economic and residential purposes. The major economic activities that the population of these areas conduct are trading, civil service, public service and industrial works.

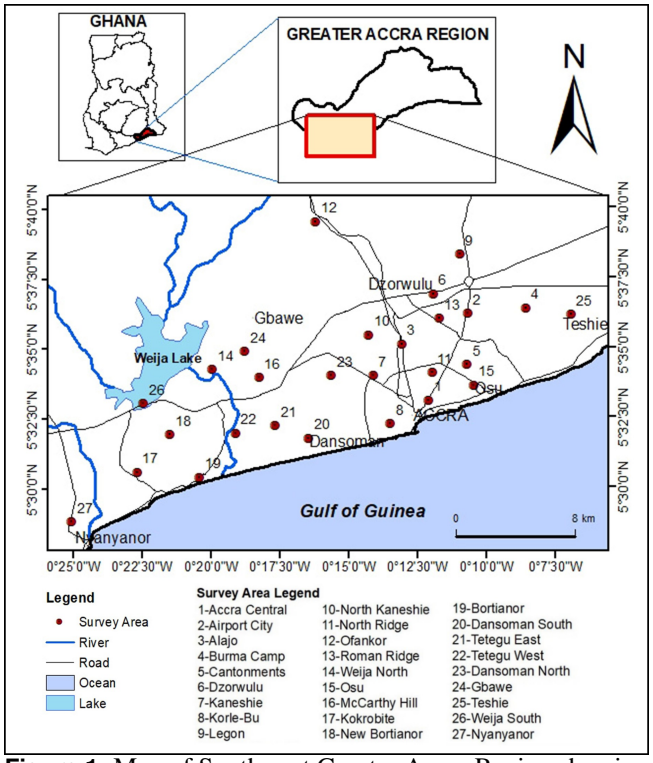

Figure 1. Map of Southwest Greater Accra Region showing data points

Data Sources

As shown in Figure 2, the feature dataset includes the following: Standard Penetration Test (SPT), Groundwater Level (GWL), Relative Density (Dr), Natural Moisture Content (NMC), Atterberg Limits and Particle Size Distribution (PSD). Empirical relationships with SPT as the major input were used to compute the Factor of Safety (FoS) values for the different subsurface layers at the various investigated sites were calculated (Idriss, 1999; Liao and Whitman, 1986; Youd and Idriss, 1997; Youd et al., 2001). Layer depths that indicated FoS values less than 1 were categorised as liquefied soils, whereas those with FoS greater than or equal to 1 were categorised as non-liquefied soils. In all, 261 data observations were used for the study.

While the 261 observations could be sufficient for initial model training and validation, there may be a limitation in the generalizability of the results. As stated in literature, small datasets can increase the risk of overfitting, particularly when using complex models like RFC and SVC (Aziz et al., 2017; Ganaie et al., 2022; Liu et al., 2024). To mitigate this, we employed techniques such as cross-validation and SMOTE to balance the dataset and improve model robustness. However, future studies should aim to incorporate larger datasets to enhance the generalizability of the findings and reduce the risk of overfitting.

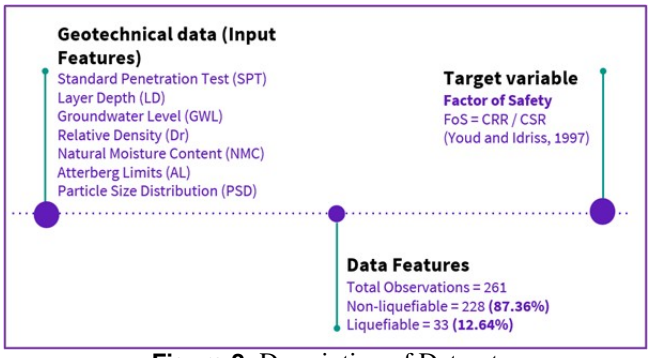

Figure 2. Description of Dataset

Feature Engineering

Initially, the complete dataset was utilised for modelling. This revealed a strong correlation (0.94) between SPT values and the calculated Factor of Safety after a correlation matrix was plotted to observe the relationship between the input variables and the target (FoS). It was found that a stronger correlation (0.97) existed between the Dr and FoS. Due to the strong relationships between SPT, Dr and FoS (Figure 3), the SPT and Dr features were excluded from the modelling dataset to avoid overfitting. The remaining features Natural Moisture Content (NMC), Atterberg limits, Particle Size Distribution (PSD) and Groundwater Level (GWL) were maintained for model development. In all, ten (10) input features were employed for model training and validation: groundwater level, depth of layer, natural moisture content, compositions of clay, silt, sand and gravel, as well as liquid limit, plastic limit and plasticity index. The target variable was factor of safety (FoS).

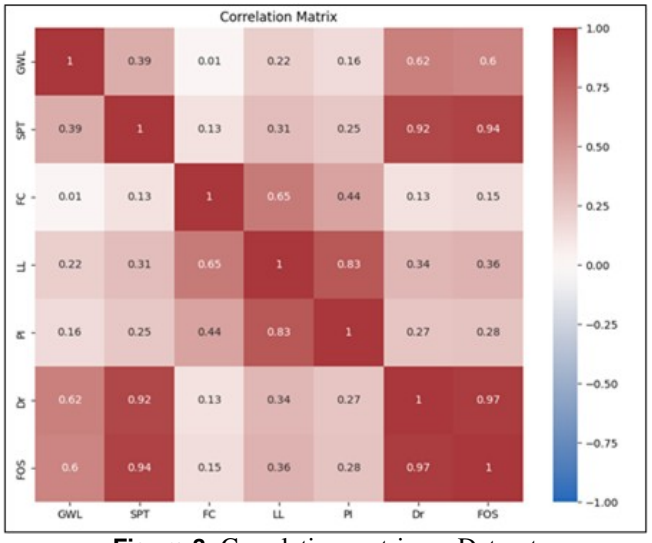

Figure 3. Correlation matrix on Dataset

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is an oversampling method that creates synthetic samples for the minority class, specifically liquefaction cases, by interpolating between existing examples. In contrast to random oversampling, which only replicates existing occurrences of the minority class, SMOTE generates fresh, convincing data points, thereby mitigating the risk of overfitting. This method has demonstrated efficacy in enhancing classification performance, especially for significantly imbalanced datasets, as seen in geotechnical applications (Chawla et al., 2002; Elreedy et al., 2024; Sebbeh-Newton et al., 2024). Although LR often experiences performance decline with imbalanced datasets, the application of SMOTE enhances its precision and recall (Blagus and Lusa, 2013). Given that logistic regression presumes a linear decision boundary, the synthetic instances produced by SMOTE facilitate the expansion of the decision space for the minority class, hence mitigating bias towards the majority class. RFC inherently manages imbalance more effectively than LR because of its random sampling and bootstrapping techniques. When utilised in conjunction with SMOTE, RFC demonstrates enhanced classification metrics for the minority class, especially in terms of precision and F1-score, as evidenced by Pacheco et al. (2023). SVC, being particularly sensitive to class imbalances, significantly benefits from SMOTE. The synthetic examples produced by SMOTE advance the decision boundary to more effectively distinguish the two classes. Research, including (Samui, 2013), indicates that the efficacy of SVC in identifying liquefaction markedly enhances when utilised alongside SMOTE, especially in minimising false negatives.

Although SMOTE successfully tackles class imbalance through the creation of synthetic samples for the minority class, it may introduce biases, especially when the generated data points fail to accurately reflect the true distribution of the minority class. This situation may result in overfitting, causing the model to excel on the training dataset while struggling with new, unseen data (Fern et al., 2018; Sakho et al., 2024). To mitigate this risk, the model's performance was evaluated by comparing the results with and without SMOTE. Although other resampling techniques like ADASYN and cost-sensitive learning could have been considered, SMOTE was chosen due to its simplicity and proven effectiveness in similar geotechnical studies (Sebbeh-Newton et al., 2024). Future work could explore the use of alternative resampling techniques to further improve model performance.

Model Development and Evaluation

Figure 4 illustrates the systematic methodology employed for predicting soil liquefaction through machine learning (ML) approaches. This workflow encompasses several stages: from the compilation of geotechnical data to the final prediction of liquefaction results. The procedure commences with the acquisition and classification of geotechnical data (see Data Sources section). Within this framework, the Factor of Safety (FoS) is classified to establish a target variable, signifying the likelihood of soil liquefaction. A binary classification technique was employed, designating the target variable as "liquefiable" or "non-liquefiable", hence aiding model training within a supervised machine learning classification framework. The Initial phase in pre-processing the dataset involved data cleansing. This is a crucial procedure undertaken to eliminate noise and discrepancies. A carefully curated dataset is

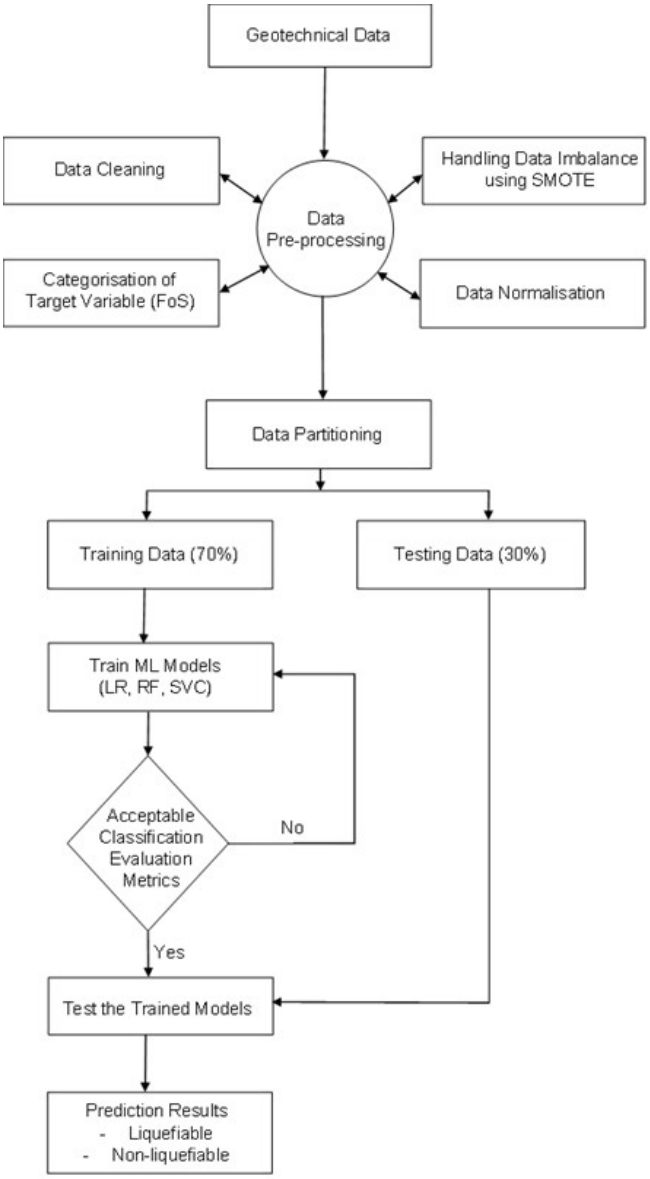

Figure 4. Flowchart of methodology for developing ML models

essential to prevent biases or mistakes in machine learning predictions. Following data cleaning, normalisation was performed to standardise feature scales and ensure uniformity among variables. In geotechnical engineering, features such as particle size or density may exhibit considerable variation in magnitude. Figure 2 illustrates that 261 observations were compiled to create the dataset for model development. Among these, 228 datapoints had FoS values of 1 or higher, signifying their non-liquefiability (negative class). This results in only 33 liquefiable datapoints (positive class). Consequently, approximately 12.64% of the overall dataset is classified as the positive class, exemplifying a classic instance of data imbalance. This circumstance affects the realisation of an impartial and correct data distribution for effective model building. The Synthetic Minority Oversampling Technique (SMOTE) was employed to address the issue of data imbalance between liquefied and non-liquefied instances.

After the pre-processing phase, the dataset was divided into training and testing subsets in a 70:30 ratio. This partitioning technique guarantees that the model is trained on a substantial dataset and validated on a distinct subset to evaluate its generalisability. This framework utilises three machine learning models: Logistic Regression (LR), Random Forest (RF) and Support Vector Classifier (SVC). Upon the completion of training, each model underwent evaluation utilising classification measures. These measurements encompass accuracy, precision, recall and F1 score, offering a thorough assessment of model performance. These measures guarantee that the selected model effectively distinguishes liquefiable and non-liquefiable cases while also reducing false positives and negatives, which are essential in risk-sensitive geotechnical applications. The concluding phase entails evaluating the trained models using the reserved testing dataset. The model generates a binary classification outcome, determining whether a certain soil sample is liquefiable or non-liquefiable depending on its characteristics.

Results and Discussion

Confusion Matrix

The confusion matrix reveals that the model encounters challenges related to class imbalance, especially in accurately identifying the minority class. The predominant class exhibits superior predictive accuracy, whereas the less frequent class experiences an elevated rate of misclassification. The Random Forest model demonstrates superior performance relative to Logistic Regression, exhibiting enhanced accuracy across both classes. Nonetheless, there remains a degree of misclassification within the minority class, albeit to a lesser extent compared to Logistic Regression. The SVC model exhibits performance comparable to that of Random Forest, achieving reasonable accuracy for the majority class while encountering some misclassification issues within the minority class. The model demonstrates a balanced performance; however, it remains influenced by the class imbalance issue.

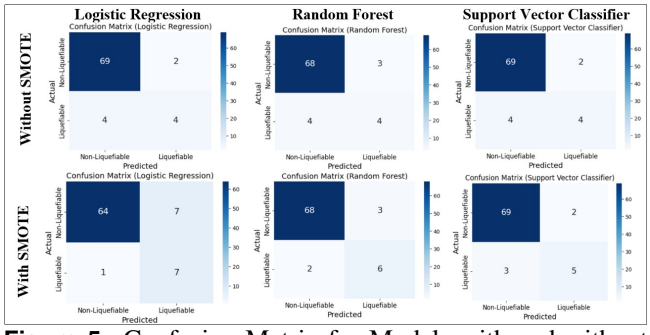

Figure 5. Confusion Matrix for Models with and without SMOTE

Following the implementation of SMOTE (Figure 5), there is a notable enhancement in the model's ability to predict the minority class. The confusion matrix reveals a more equitable

performance, showcasing diminished misclassification rates across both classes. The Random Forest model demonstrates improved performance through the application of SMOTE, resulting in a more balanced confusion matrix. The model demonstrates improved accuracy for the minority class, resulting in a more robust overall performance. The SVC model, enhanced by SMOTE, exhibits superior performance, especially in accurately identifying the minority class. The confusion matrix demonstrates a more equitable distribution of predictions, suggesting that SMOTE has successfully addressed the class imbalance problem.

Outcomes without SMOTE

Table 1 displays the performance metrics for each model on the positive class without the application of SMOTE. Despite a high accuracy of 0.91 for RF and 0.92 for both LR and SVC, this statistic may be deceptive due to the dataset's imbalance. A detailed examination of precision, recall and F1-Score indicates the model's inadequate ability to predict instances of the minority class (i.e., liquefaction events).

Table 1. Performance metrics for each machine learning algorithm without SMOTE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.92	0.67	0.50	0.57
Random Forest	0.91	0.57	0.50	0.53
Support Vector Classifier	0.92	0.67	0.50	0.57

The precision metric for Logistic Regression is 0.67, signifying that 67% of the predicted liquefaction events are accurate. The findings indicate a satisfactory level of accuracy in predicting liquefaction; however, the class imbalance presents challenges for the models in effectively differentiating between liquefaction and non-liquefaction cases.

The recall values, uniformly recorded at 0.50 across all models, underscore a significant concern: merely 50% of genuine liquefaction events are accurately detected. This highlights the model's shortcomings in accurately identifying instances of the minority class, a prevalent issue in datasets with imbalance. As a result, the F1-Score, which serves to balance precision and recall, is consistently low across all models, with Logistic Regression and SVC achieving a score of 0.57, while Random Forest records a score of 0.53. The scores indicate a significant inefficiency of the models in effectively handling the minority class.

The observed low recall and F1-Scores highlight the challenges encountered by these models in accurately predicting liquefaction events, particularly in the absence of balancing methods such as SMOTE. Although the models demonstrate high accuracy, they show a tendency to favour the majority class, leading to suboptimal performance in the precise identification of liquefaction events

Outcomes utilising SMOTE

Table 2 presents the performance metrics on the positive class subsequent to the application of SMOTE for dataset balancing. SMOTE markedly enhances the recall and F1-scores of the models, signifying improved efficacy in predicting minority class instances.

Table 2. Performance metrics for each machine learning algorithm without SMOTE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.90	0.50	0.88	0.64
Random Forest	0.94	0.67	0.75	0.71
Support Vector Classifier	0.94	0.71	0.62	0.67

The recall for Logistic Regression and Random Forest shows a notable improvement, rising from 0.50 to 0.88 and 0.75, respectively. This enhancement illustrates the efficacy of SMOTE in allowing these models to detect a greater percentage of genuine liquefaction occurrences. For SVC, the recall rises to 0.62, reflecting a moderate yet significant improvement.

The precision of Logistic Regression shows a minor decline from 0.67 to 0.50, whereas the precision of Random Forest maintains a consistent value of 0.67. This indicates that although the detection of liquefaction cases is on the rise, there is a marginal uptick in false positive occurrences. The precision for SVC has increased to 0.71, signifying an enhancement in the model's ability to differentiate between the two classes. The F1-Score across all models demonstrates significant enhancement. Logistic Regression and Random Forest demonstrate F1-Scores of 0.64 and 0.71, respectively, an improvement from 0.57 and 0.53. The Support Vector Classifier demonstrates notable enhancement, achieving an F1-Score of 0.67, which is on par with the performance of the other models.

While the accuracy across the models shows consistency before and after the application of SMOTE, with SVC demonstrating a minor enhancement from 0.92 to 0.94, the precision-recall balance underscores the advantages brought by SMOTE. In the absence of SMOTE, the models demonstrated strong capabilities in predicting the majority class but faced challenges in accurately forecasting liquefaction events. SMOTE addresses this imbalance, improving recall and F1-Score, which are essential metrics for classification tasks with imbalanced datasets like soil liquefaction prediction.

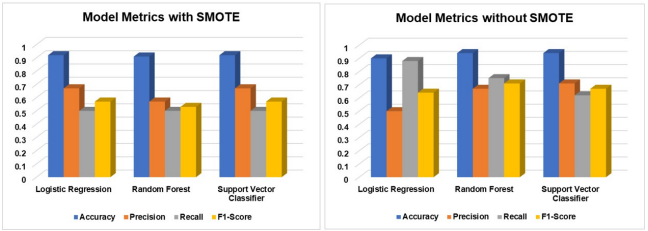

Figure 6. 3D Bar Graphs comparing Model Evaluation metrics with and without SMOTE

The application of SMOTE markedly enhances the model's capacity to forecast the minority class (liquefaction), which

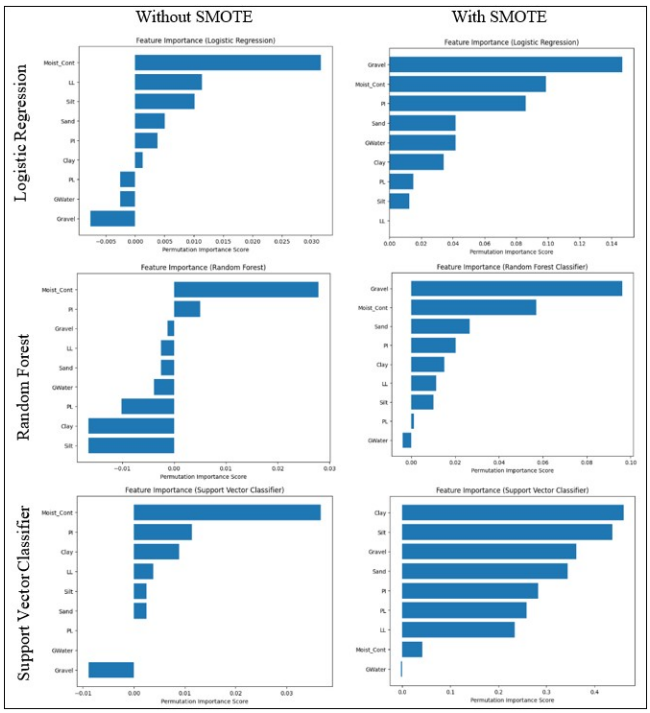

Figure 7. Permutation Feature Importance Plots for Models with and without SMOTE

is essential in this scenario. The prediction of soil liquefaction is generally infrequent in the majority of datasets, which accounts for the less-than-ideal performance of models that do not employ balancing techniques such as SMOTE. The enhancement in recall highlights the significance of SMOTE in mitigating bias towards the majority class. The consistently high accuracy observed with and without SMOTE underscores the necessity of assessing performance metrics beyond mere accuracy, particularly in the context of imbalanced datasets.
The minor reduction in precision observed in Logistic Regression following the application of SMOTE can be attributed to the inherent trade-off between recall and precision. Although an increased number of liquefaction occurrences are accurately recognised, there remains a limited subset of non-liquefaction cases that could be erroneously categorised as liquefaction. The enhanced F1-Scores observed in all models indicate that SMOTE successfully establishes a more favourable equilibrium between precision and recall, thus improving the overall performance of the machine learning models in forecasting soil liquefaction.

Permutation Feature Importance

The permutation feature importance plots offer valuable insights regarding the features that play a crucial role in influencing the model's predictions. The feature importance plot illustrates that specific features exert a greater influence on the predictions made by the model. The identified features are expected to play a significant role in class label determination. Following the application of SMOTE, the significance of certain features may change, suggesting that the model has become more responsive to features that were previously less impactful as a result of class imbalance.
The Random Forest model naturally offers insights into feature significance through Gini impurity or information gain metrics. The permutation feature importance plot validates these observations, indicating that the key features align with the model's internal metrics. The application of SMOTE leads to a relatively stable distribution of feature importance; however, certain features may experience fluctuations in their significance as the model adapts to the balanced dataset. The SVC model, as a non-parametric approach, lacks an intrinsic mechanism for determining feature importance. The permutation feature importance plot illustrates the features that, when randomised, lead to the most considerable decline in model efficacy. This highlights their significance in the decision-making framework. The implementation of SMOTE can lead to variations in the feature importance plot, indicating how the model adjusts to the newly balanced dataset.

Maize Yields

To guarantee the reliability of the findings, statistical validation was performed using confidence intervals and hypothesis testing. The performance disparities between models with and without SMOTE were determined to be statistically significant ($p < 0.05$), suggesting that the enhancements in recall and F1-score are not attributable to random variation. In addition, a computation of 95% confidence intervals for the accuracy, precision, recall and F1-score metrics, further reinforce the dependability of the findings.
The machine learning models developed in this study demonstrate competitive performance in soil liquefaction susceptibility assessment, achieving an F1-score of 0.67. This performance aligns with established traditional empirical methods such as the Seed and Idriss (1971) simplified procedure, which reportedly yields F1-scores between 0.60 and 0.70 across diverse geological conditions and regional datasets (Boulanger and Idriss, 2004; Youd et al., 2001). While the models do not exhibit statistically superior predictive capability compared to conventional approaches, they offer distinct advantages in computational adaptability and data handling flexibility.

Conclusion

The research effectively created and assessed three distinct models such as Logistic Regression, Random Forest and Support Vector Classifier aimed at predicting the susceptibility of soil to liquefaction. At the outset, the models demonstrated impressive accuracy; however, they faced challenges related to class imbalance, especially in identifying the minority class (liquefaction events). SMOTE significantly enhanced model performance, particularly in improving recall, which is critical for identifying minority class instances. For Logistic Regression (LR), recall increased from 0.50 to 0.88, and for Random Forest Classifier (RFC), it rose from 0.50 to 0.75. The Support Vector Classifier (SVC) also saw an improvement, with recall increasing to 0.62. While precision decreased for LR (from

0.67 to 0.50), it improved for SVC to 0.71, resulting in the highest F1-score of 0.67 among the models. RFC maintained a stable precision of 0.67 and achieved a strong F1-score of 0.71, indicating a balanced performance between precision and recall. SVC demonstrated the best overall predictive performance, with accuracy increasing from 0.92 to 0.94 after applying SMOTE. The permutation feature importance plots indicated that specific features, including groundwater level and natural moisture content, were pivotal in the predictions of the models, with variations in feature significance observed following the application of SMOTE.

The findings indicate that models utilising advanced algorithms, in conjunction with balancing methodologies such as SMOTE, can proficiently predict the susceptibility of soil to liquefaction. Although the models did not surpass conventional empirical techniques regarding predictive accuracy, they present benefits in computational adaptability and versatility in managing various datasets. The enhanced recall and F1-Scores highlight the significance of tackling class imbalance in geotechnical applications, where precise prediction of infrequent occurrences such as liquefaction is essential for effective risk assessment and mitigation.

Nonetheless, the research faced constraints due to the limited size of the dataset, potentially impacting the broader applicability of the findings. Future endeavours should prioritise the integration of more extensive datasets and the investigation of different resampling methodologies to significantly improve model efficacy. This study underscores the capabilities of advanced models, integrated with SMOTE, for predicting soil liquefaction, establishing a solid foundation for subsequent investigations in the field of geotechnical engineering.

This paper happens to be one of the foremost research works on the application of machine learning for predicting soil liquefaction using geotechnical datasets collected from soils in Accra, Ghana. The findings from this study hold considerable practical relevance for geotechnical engineers, policymakers and urban planners in the country. Accurate prediction of soil liquefaction through machine learning models can enhance the design of earthquake-resistant infrastructure, especially in seismically active areas such as Accra. Policymakers may utilise these predictions to identify priority areas for additional geotechnical investigation and to establish building codes aimed at reducing the risk of damage from liquefaction. Urban planners may integrate these predictions into land-use planning to circumvent high-risk areas for the development of critical infrastructure.

Future research should investigate the application of advanced deep learning methodologies, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to enhance soil liquefaction prediction. Convolutional Neural Networks (CNNs) may be employed to examine spatial patterns in geotechnical data, whereas Long Short-Term Memory networks (LSTMs) can effectively model temporal dependencies in seismic activity. Furthermore, transformer models, recognised for their efficacy in managing complex, high-dimensional data, warrant exploration regarding their applicability in geotechnical contexts. The integration of these techniques with SMOTE or alternative resampling methods may enhance prediction accuracy and robustness.

Acknowledgement

The authors are grateful to the University of Mines and Technology for funding this research. We also thank the Architectural and Engineering Services Limited (AESL), Accra, for providing the data for this work.

References

Amponsah, P. (2004). Seismic activity in Ghana: past, present and future. *Annals of Geophysics*, 47:539–543.

Amponsah, P. (2021). Seismic activity in Ghana: past, present and future. *Annals of Geophysics*, 47.

Amponsah, P., Leydecker, G., and Muff, R. (2012). Earthquake catalogue of Ghana for the time period 1615-2003 with special reference to the tectono-structural evolution of south-east Ghana. *Journal of African Earth Sciences*, 75:1–13.

Arango-Serna, S., Herrera, M., Cruz, A., Sandoval, E., Thomson, P., and Ledezma, C. (2021). Use of ambient noise records in seismic engineering: An approach to identify potentially liquefiable sites. *Soil Dynamics and Earthquake Engineering*, 148:2021–2022.

Ardakani, A. and Kohestani, V. (2015). Evaluation of liquefaction potential based on CPT results using C4.5 decision tree. *Journal of Artificial Intelligence and Data Mining*, 3.

Atarigiya, B., Allotey, N., and Matrevi, E. (2023). Sample liquefaction case studies in coastal Ghana. *SSRN Electronic Journal*, pages 1–6.

Aziz, R., Verma, C., and Srivastava, N. (2017). Dimension reduction methods for microarray data: A review. *AIMS Bioengineering*, 4:179–197.

Blagus, R. and Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:1–16.

Boulanger, R. and Idriss, I. (2004). Evaluating the potential for liquefaction or cyclic failure of silts and clays. *Journal name needed*.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

Cha, D., Blumenstein, M., Zhang, H., and Jeng, D. (2008). A neural-genetic technique for coastal engineering: Determining wave-induced. *Studies in Computational Intelligence*, 82:337–351.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Cubrinovski, M., Bray, J., de la Torre, C., Olsen, M., Bradley, B., Chiaro, G., Stocks, E., Wotherspoon, L., and Krall, T. (2018). Liquefaction-Induced Damage and CPT Characterization of the Reclamations at CentrePort, Wellington. *Bulletin of the Seismological Society of America*, 108:1695–1708.

Demir, S. and Sahin, E. (2022). Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naive Bayes. *European Journal of Science and Technology*, pages 142–147.

Elreedy, D., Atiya, A., and Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113:4903–4923.

Fadliansyah, F., Faris, F., and Wilopo, W. (2024). Implementation of machine learning classification models considering the optimum data ratio in predicting soil liquefaction susceptibility. *IOP Conference Series: Earth and Environmental Science*, 1416:1–12.

Farrokhzad, F., Choobbasti, A., and Barari, A. (2012). Liquefaction microzonation of Babol city using artificial neural network. *Journal of King Saud University - Science*, 24:89–100.

Fern, A., Garcia, S., Herrera, F., and Chawla, N. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905.

Ganaie, M., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:1–18.

Gandomi, A., Fridline, M., and Roke, D. (2013). Decision tree approach for soil liquefaction assessment. *The Scientific World Journal*, 2013.

Goh, A. (1994). Seismic liquefaction potential assessed by neural networks. *Journal of Geotechnical Engineering*, pages 1467–1480.

Goh, A. and Goh, S. (2007). Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. *Computers and Geotechnics*, 34:410–421.

Goudarzy, M., Sarkar, D., and Wichtmann, T. (2022). Influence of plastic fines content on the liquefaction susceptibility of sands: Cyclic loading. *Acta Geotechnica*, 17:1719–1737.

Grossi, E. and Buscema, M. (2007). Introduction to artificial neural networks. *European Journal of Gastroenterology and Hepatology*, 19:1046–1054.

Hu, J.-L., Tang, X.-W., and Qiu, J.-N. (2016). Analysis of the influences of sampling bias and class imbalance on performances of probabilistic liquefaction models. *International Journal of Geomechanics*, 17:1–23.

Huang, Y. and Zhao, L. (2018). The effects of small particles on soil seismic liquefaction resistance: Current findings and future challenges. *Natural Hazards*, 92:567–579.

Idriss, I. (1999). An update to the seed-idriss simplified procedure for evaluating liquefaction potential. In *Proceedings of TRB Workshop on New Approaches to Liquefaction*, Washington DC. Federal Highway Administration.

Jairi, I., Fang, Y., and Pirhadi, N. (2021). Application of logistic regression based on maximum likelihood estimation to predict seismic soil liquefaction occurrence. *Human-Centric Intelligent Systems*, 1:98.

Kaya, Z., Latifoglu, L., Uncuoglu, E., Erol, A., and Keskin, M. (2023). Predicting liquefaction-induced lateral spreading by using the multigene genetic programming (MGGP), multilayer perceptron (MLP), and random forest (RF) techniques. *Bulletin of Engineering Geology and the Environment*, 82:1–18.

Klu, A., Asare, E., Seidu, J., and Opoku, N. (2024). Looming Earthquake Threat in Ghana. *IntechOpen*, pages 1–16.

Kohestani, V., Hassanlourad, M., and Ardakani, A. (2015). Evaluation of liquefaction potential based on CPT data using random forest. *Natural Hazards*, 79:1079–1089.

Kutu, J. (2013). Seismic and Tectonic Correspondence of Major Earthquake Regions in Southern Ghana with Mid-Atlantic Transform-Fracture Zones. *International Journal of Geosciences*, 4:1326–1332.

Lenart, S. (2008). The response of saturated soils to a dynamic load. *Acta Geotechnica*, pages 36–49.

Liao, S. and Whitman, R. (1986). *Catalogue of liquefaction and non-liquefaction occurrences during earthquakes*. Dept. of Civ. Engrg., Massachusetts Institute of Technology, Cambridge.

Liu, C., Ku, C., Chiu, Y., and Wu, T. (2024). Evaluation of liquefaction potential in central Taiwan using random forest method. *Scientific reports*, 14:27517.

Mansouri, M. and Dabiri, R. (2021). Predicting the liquefaction potential of soil layers in tabriz city via artificial neural network analysis. *SN Applied Sciences*, 3:31.

Mohammadnejad, T. and Andrade, J. (2015). Flow liquefaction instability prediction using finite elements. *Acta Geotechnica*, 10:83–100.

Mollica, R., de Franco, R., Caielli, G., Boniolo, G., Crosta, G., Motti, A., Villa, A., and Castellanza, R. (2020). Micro electrical resistivity tomography for seismic liquefaction study. *Journal of Applied Geophysics*, 180:104–124.

Muduli, P. and Das, S. (2013). Spt-based probabilistic method for evaluation of liquefaction potential of soil using multi-gene genetic programming. *International Journal of Geotechnical Earthquake Engineering*, 4:42–60.

Muduli, P. and Das, S. (2014). Evaluation of liquefaction potential of soil based on standard penetration test using multi-gene genetic programming model. *Acta Geophysica*, 62:529–543.

Muduli, P., Das, S., and Bhattacharya, S. (2014). Cpt-based probabilistic evaluation of seismic soil liquefaction potential using multi-gene genetic programming. *Georisk*, 8:14–28.

Nejad, A., Guler, E., and Ozturan, M. (2018). Evaluation of liquefaction potential using random forest method and shear wave velocity results. In *Proceedings - 2018 International Conference on Applied Mathematics and Computational Science, ICAMCS.NET 2018*, pages 23–26.

Nortey, G., Armah, T., and Amponsah, P. (2018). Vs30 mapping at selected sites within the Greater Accra Metropolitan Area. *Journal of African Earth Sciences*, 142:158–169.

O'Rourke, T., Bonneau, A., Pease, J., Shi, P., and Wang, Y. (2006). Liquefaction and ground failures in San Francisco. *Earthquake Spectra*, 22:91–112.

Pacheco, V., Bragagnolo, L., Dalla Rosa, F., and Thomé, A. (2023). Cone penetration test prediction based on random forest models and deep neural networks. *Geotechnical and Geological Engineering*, 41:4595–4628.

Sakho, A., Malherbe, E., and Scornet, E. (2024). Do we need rebalancing strategies? a theoretical and empirical study around SMOTE and its variants. *HAL Open Science*, pages 1–41.

Samui, P. (2013). Liquefaction prediction using support vector machine model based on cone penetration data. *Frontiers of Architecture and Civil Engineering in China*, 7:72–82.

Samui, P. and Sitharam, T. (2011). Machine learning modelling for predicting soil liquefaction susceptibility. *Natural Hazards and Earth System Science*, 11:1–9.

Sancio, R., Bray, J., and Riemer, M. (2003). An assessment of the liquefaction susceptibility of Adapazari silt. In *Pacific Conference on Earthquake Engineering*, pages 1–8.

Sarkar, D., Goudarzy, M., König, D., and Wichtmann, T. (2020). Influence of particle shape and size on the threshold fines content and the limit index void ratios of sands containing non-plastic fines. *Soils and Foundations*, 60:621–633.

Sebbeh-Newton, S., Seidu, J., Ankah, M., Ewusi-Wilson, R., Zabidi, H., and Amakye, L. (2024). Real-time classification of ground conditions ahead of a TBM using supervised machine learning algorithms. *Modeling Earth Systems and Environment*, 10:6173–6186.

Seed, H. and Idriss, I. (1971). Simplified procedure for evaluating soil liquefaction potential. *Journal of the Soil Mechanics and Foundations Division, ASCE*, 97:1249–1273.

Song, C., Peng, H., Xu, L., Zhao, T., Guo, Z., and Chen, W. (2024). Probabilistic evaluation of cultural soil heritage hazards in china from extremely imbalanced site investigation data using SMOTE-Gaussian process classification. *Journal of Cultural Heritage*, 67:121–133.

Talamkhani, S., Naeini, S., and Ardakani, A. (2023). Prediction of static liquefaction susceptibility of sands containing plastic fines using machine learning techniques. *Geotechnical and Geological Engineering*, 41:3057–3074.

Tober, S. (2020). Tree-based machine learning models. *Examensarbete Inom Teknik, Grundnivå*.

Venkatesh, K., Kumar, V., and Tiwari, R. (2013). Appraisal of liquefaction potential using neural network and neuro fuzzy approach. *Applied Artificial Intelligence*, 27:700–720.

Verma, M., Singh, R., and Bansal, B. (2014). Soft sediments and damage pattern: a few case studies from large Indian earthquakes vis-a-vis seismic risk evaluation. *Natural hazards*, 74:1829–1851.

Youd, T. and Idriss, I. (1997). Liquefaction criteria based on statistical and probabilistic analyses. Technical report, Technical Report NCEER-97-0022.

Youd, T., Idriss, I., Andrus, R., Arango, I., Castro, G., Christian, J., Dobry, R., Finn, W., Hynes, M., Ishihara, K., et al. (2001). Liquefaction resistance of soils: Summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils. *Journal of Geotechnical and Geoenvironmental Engineering*, pages 817–833.

Young-Su, K. and Byung-Tak, K. (2006). Use of artificial neural networks in the prediction of liquefaction resistance of sands. *Journal of Geotechnical and Geoenvironmental Engineering*, 132:1502–1504.

Zhang, J. and Wang, Y. (2021). An ensemble method to improve prediction of earthquake-induced soil liquefaction: A multi-dataset study. *Neural Computing and Applications*, 33:1533–1546.

Zhou, Y., Xia, P., Ling, D., and Chen, Y. (2020). Liquefaction case studies of gravelly soils during the 2008 wenchuan earthquake. *Engineering Geology*, 274:1–20.

Ziotopoulou, K. and Boulanger, R. (2013). Numerical modeling issues in predicting post-liquefaction reconsolidation strains and settlements. In *10th International Conference on Urban Earthquake Engineering*, page 7.