

Journal of Rehabilitation in Civil Engineering

Journal homepage: <https://civiljournal.semnan.ac.ir/>

M5 Soft Computing Techniques for Assessment of Soil Liquefaction

Abu Sayed¹; Md. Mahabub Rahman^{2,*} 

1. Lecturer, Department of Civil Engineering, Pundra University of Science and Technology, Bangladesh

2. Lecturer, Department of Civil Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh

* Corresponding author: mmr.civil@hstu.ac.bd

ARTICLE INFO

Article history:

Received: 06 July 2024

Revised: 12 November 2024

Accepted: 14 January 2025

Keywords:

Soil liquefaction;

Machine learning techniques;

Regression model;

Classification model;

Performance metrics.

ABSTRACT

It is essential to precisely estimate the liquefaction potential because soil liquefaction is a factor that raises the quantity and intensity of losses in an earthquake. In the past, the prediction of soil liquefaction was based on multiple analytical inferences. The purpose of this research is to develop an M5 model for both classification and regression in order to investigate the suitability of the M5 decision tree for liquefaction assessment. Additionally, the divisional approaches of fuzzy clustering means (FCM), kfold clustering, and grid search cross-validation (Gridsearch CV) are investigated in order to create effective regression and classification models. In this work, specific models are developed using a data set of 200 boreholes from standard penetration tests on soils in the Dinajpur region. The efficacy of the constructed models is assessed using several performance measures, such as root mean squared error (RMSE), mean absolute error (MAE), and coefficient of correlation (R) for regression models, and accuracy, precision, and AUC value for classification models. Based on the results, it was found that the M5 decision tree regression model shows $R = 0.95$, $IoA = 0.86$, and $IoS = 0.96$ for testing and $R = 0.93$, $IoA = 0.88$, and $IoS = 0.96$ for training data. On the other hand, the classification model shows accuracy = 95%, recall = 1, and F1 score = 0.97 for testing and 98.75%, 1, and 0.99 for training, respectively. Both of these results were found for the Kfold technique, which predicts a more accurate value than other divisional approaches.

E-ISSN: 2345-4423

© 2025 The Authors. Journal of Rehabilitation in Civil Engineering published by Semnan University Press.

This is an open access article under the CC-BY 4.0 license. (<https://creativecommons.org/licenses/by/4.0/>)

How to cite this article: Sayed, A. and Rahman, M. Mahabub (2025). M5 Soft Computing Techniques for Assessment of Soil Liquefaction. Journal of Rehabilitation in Civil Engineering, 13(3), 199-214. <https://doi.org/10.22075/jrce.2025.34669.2134>

1. Introduction

Earthquake-induced liquefaction is a disastrous type of geohazard that may result in slope failure, foundation damage, and various types of structural failure [1]. This catastrophic phenomenon occurs when cyclic loading goes through the unpacked granular soil with water table cases, causing shear strength loss and transforming the soil into a liquid state [2]. A well-known example of liquefaction is the 1964 Niigata (Japan) and 1964 Great Alaskan Earthquake, in which large-scale soil liquefaction occurred, causing wide-spread damage to building structures and underground facilities [3].

In soil mechanics and geotechnical engineering, soil liquefaction is still a challenging but popular research issue, mostly because of the soil structure's complexity, unpredictability, and uncertainty [4]. To determine soil liquefaction, various researchers have proposed different approaches, such as field approaches [5,6], empirical approaches [7,8], laboratory approaches [9], numerical methods [10], artificial intelligence, and machine learning models [11–14], and the like. Innovative computing methods like machine learning and artificial intelligence have become increasingly popular in geotechnical engineering in the past few years [15]. Soft computing techniques are also becoming popular in other subdivisions of civil engineering [16,17]. A neural network-based model was suggested by Goh [18], to predict and assess the chance of liquefaction in saturated, sandy soil. Later, other geotechnical researchers developed several machine learning approaches, including neural network training, support vector machines (SVM), genetic code programming (GP), least squares support vector machines (LSSVM), and stochastic gradient boosting (SGB), to do the liquefaction study [19–21]. Artificial neural networks (ANNs) were employed by Ramakrishnan et al. [22] to forecast the vulnerability of unconsolidated sediments to liquefaction. There were twenty-three datasets used to train the backpropagation neural network, including the liquefaction severity index, liquefaction sensitivity index, cyclic resistance ratio, and cyclic stress ratio. The field data and anticipated results were similar, suggesting that the ANN is feasible for mapping liquefaction susceptibility. Venkatesh et al. [23] analyzed liquefaction phenomena using a multilayer perceptron network and a feed-forward backpropagation technique. 159 geotechnical data points were gathered, and neural network models with the best-hidden layers and transfer functions were trained. Better prediction skills were revealed when the investigation compared the projected liquefaction potential values of neural networks and neuro-fuzzy models. To assess a soil's potential for seismic liquefaction based on shear wave velocity, Lui et al. [13] presented the random forest (RF) approach. Five training parameters and the Andrus database were used in the model's development. The Chinese code and Andrus techniques were contrasted with the model, which was verified using the Kayen database. The findings demonstrated a satisfactory overall success rate of more than eighty percent and an excellent rate of success for liquefied locations. The extreme learning machine (ELM), as developed by Zhang et al. [24], uses CPT data to evaluate the liquefaction potential of soil deposits. In the past, liquefaction has been effectively predicted using the ANN model, the most widely used machine learning technique in this sector [25][22]. A number of the key disadvantages of ANN models are weak generalization capabilities, a sluggish convergence rate, and model overfitting, all of which can influence result prediction. Because existing ML-based liquefaction assessment models stress accuracy above explainability, they are intrinsically opaque. Because the currently available liquefaction datasets are small in size and contain proportionately more liquefaction events than non-liquefaction events, these models behave differently from databases from other areas of the world. Later, further researchers created several

innovative machine learning methods, including decision trees, M5, support vector regression (SVR), lazily K-star (LKS), random forest (RF), gradient boosting machine (GBM), and the like [26][27][28]. When handling enormous quantities of data and enhancing prediction accuracy, the current ML-based models are better alternatives. Each machine learning approach has its own set of constraints owing to the parameters and model uncertainty [29]. The majority of Bangladesh's underlying lithology is made up of loose, sandy, and clayey sediments from floodplains [30]. After the Great Indian Earthquake in 1897, the Bengal Earthquake in 1885, and the Srimangal Earthquake in 1918, there were records of extensive liquefaction in Bangladesh's alluvial deposits [31–34]. According to paleo-seismic investigations, a sequence of earthquakes along the Dauki fault is thought to have caused the liquefaction evidence that was seen in the country's northeast and north [35,36] Furthermore, it is situated near the tectonically active Arakan megathrust and Himalayan orogenic belt, which contain at least five significant active fault zones and have been linked to evidence of massive earthquakes. A locked megathrust, according to Steckler et al. [37], is present along the Indo-Burman Himalaya boundaries, supporting the idea that these areas will be resistant to large earthquakes in the future. Thus, the nation must do more research on the examination of the liquefaction resistance of the soils in the major cities.

The primary objective of this research is to apply the M5 machine learning algorithm for liquefaction prediction using standard penetration test data. To prepare the data set, liquefaction vulnerability is evaluated utilizing empirical formulas using Borelog data. The formulas determine LPI based on the dependent variables such as depth of soil, correct SPT-N value, fineness content, overburden pressure, effective overburden pressure, and PGA. Finally, 200 borelog-dependent variables are utilized in the M5 model's development to evaluate the LPI. This study develops two sorts of models, the regression model being the one used to determine the target variable's value. In contrast, the target variable's categorization is confirmed using a classification model. For the classification and regression models of M5, many divisional techniques are used, including FCM, kfold clustering, and Gridsearch Cv. In order to determine which divisional strategy is most effective for M5 in predicting liquefaction, comparisons between each divisional technique are completed using various performance scores.

2. Description of dataset

2.1. Descriptive statistics for input data

In this study, soil liquefaction is predicted for the Dinajpur area in Bangladesh's northern district. To conduct this, standard bore-log charts of various boreholes reliant on penetration tests are gathered from different portion of Dinajpur district to compile the datasets. To determine liquefaction potential, datasets are used with the Idriss and Boulanger approach [38]. Finally, full data is sorted into 200 nos. according to the important parameters to predict soil liquefaction such as depth of soil $Z(m)$, correct SPT-N value $N1(60_{CS})$, fineness content ($F < 0.0075$), overburden pressure (σ_{av}), effective overburden pressure (σ'_{av}), Peak ground acceleration (PGA), and liquefaction potential index (LPI). Full datasets were utilized in the M5 model's development. Table 1 represents the statistical information of input parameters i.e. Total number of borehole data (200 Nos.), Average, Standard deviation, Minimum, Maximum etc. used for regression model. At a time, Fig. 1 represents the Pearson correlation coefficient for the input parameters. Pearson correlation coefficient shows that the parameters have relationship with each other and with the target variable. The relation may either positive or negative (+1=strong positive and -1=strong negative).

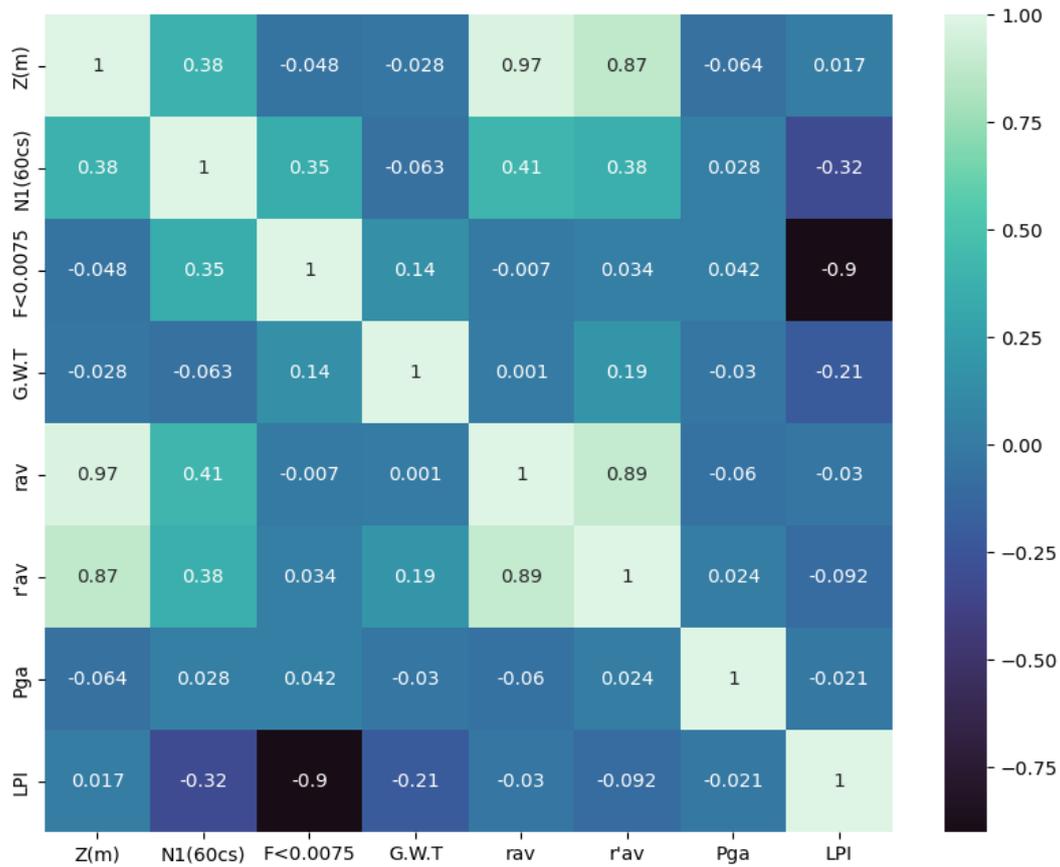


Fig. 1. Pearson correlation matrix for dataset.

For Classification model, the target variable i.e. Liquefaction potential index (LPI) is classified into two classes (Liquefiable and no liquefiable). The sample which has LPI value of zero are taken as no liquefiable sample and which have the LPI value greater than zero are taken as liquefiable. That means,

When,

$LPI=0$, No liquefiable, (59 Nos. Sample)

and

$LPI.>0$, Liquefiable (141 Nos. of Sample)

Based on this theory, among the 200 sample, 59 samples are found as no liquefiable and 141 samples are classified as liquefiable.

Table 1. Statistical information of input parameters.

Input Parameters	Count	Average	Statistical Parameters		
			Standard Deviation	min	max
Z(m)	200	3.69	2.51	0.76	10.67
Nl(60cs)	200	11.16	6.61	2.29	70.11
F<0.0075	200	72.77	23.43	11	99
GWT	200	2.60	0.749	1.5	4.75
σ_{av}	200	68.51	46.38	18.24	199.15
σ'_{av}	200	45.08	22.93	1.10	115.35
PGA	200	0.18	0.062	0.03	0.36
LPI	200	2.44	2.45	0	9.67

2.2. Data visualization and feature scaling

The pairwise scatter plot of all input features is represented in Fig. 2. In this figure, the upper triangle represents a scatter plot including the Pearson correlation coefficient, i.e., each cell in the upper triangle has a scatter plot of the feature pair linked to its orientation. Furthermore, each scatter plot will include a caption with the Pearson correlation coefficient between the two features. Each diagonal cell contains a histogram for the corresponding feature, showing the distribution of the values for each input feature. The lower triangle represents the kernel density estimate (KDE) plot showing the density of each data point, i.e., every cell in the lower triangle has a kernel density estimate (KDE) plot of the feature pair that corresponds to its position. This plot depicts the estimated distribution of data points.

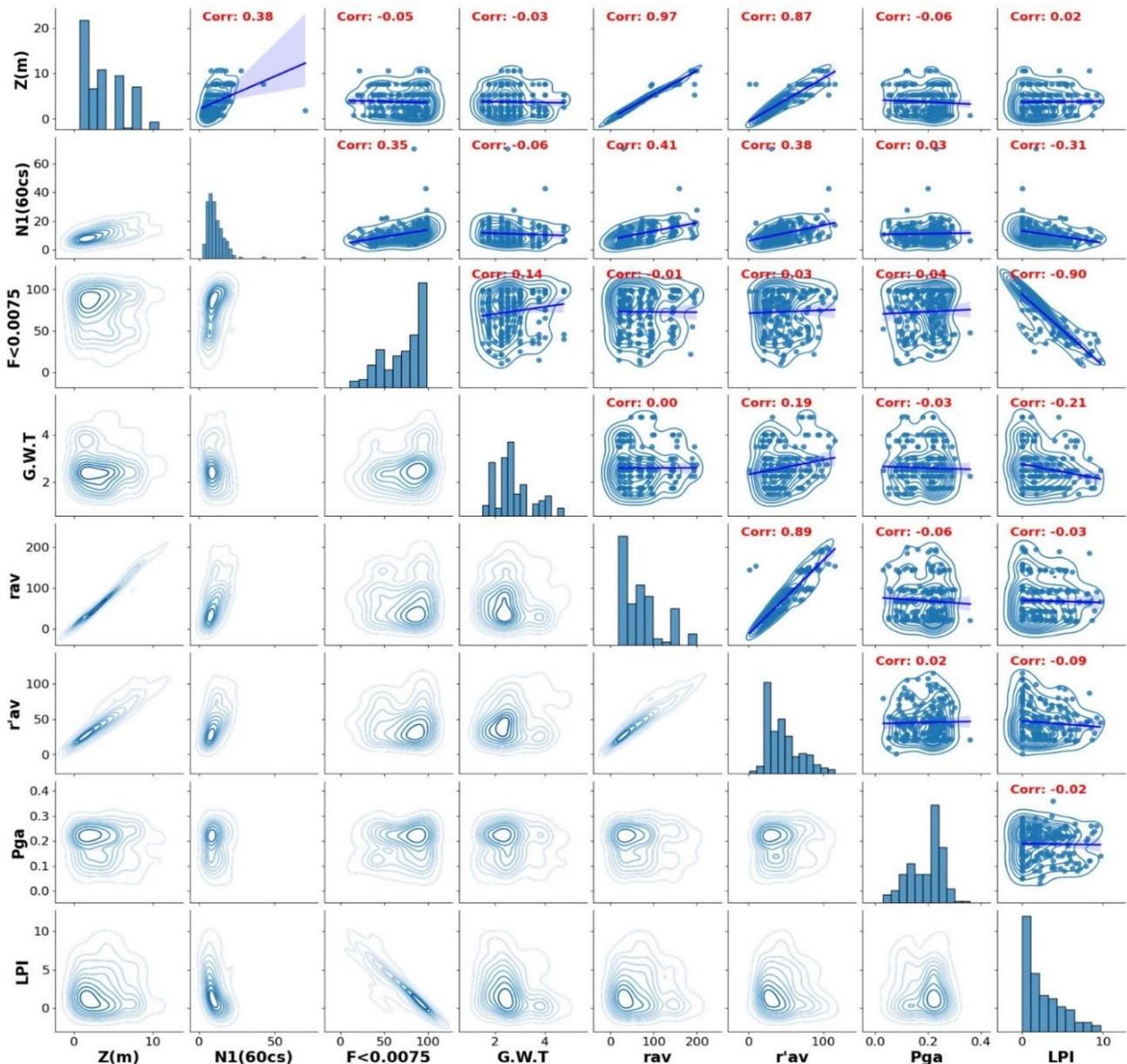


Fig. 2. Pairwise scatter plot with correlation coefficient: histograms and kernel density estimate (KDE) plot of all input features.

The range of magnitudes of those input features varies widely, as seen in Table 1. As a result, using Euclidian distances in M5 machine learning algorithms may not be acceptable. As a preprocessing

step before building the model, the feature scaling approach is used to equalize the scale of the input features [39]. To rescale inputs to a defined range, typically [0, 1], the min-max scaling method (`sklearn.preprocessing.MinMaxScaler`) is chosen in this application. Smaller standard deviations are a benefit of this restricted range, which also reduces the impact of outliers and enhances machine learning predictive power on small datasets.

3. Methodology

3.1. Overall methodology

To achieve the goal, the first step was to compile a dataset of liquefaction vulnerability for the selected study region. After evaluating the liquefaction vulnerability, the final data set was created and evaluated on Google Colaboratory employing Python. Programming was used to validate null values, do statistical analysis (Table 1), and generate a heat map based on the Pearson correlation coefficient (Figure 1). Since the classification model classifies the target variable and predicts its value, the data are evaluated in such a way that the target variable has only two alternatives. This work created a binary classification model (Model I). In this case, zero denotes that the dirt in the borehole is not liquefiable, whereas one denotes that it is. However, since the regression model (Model II) predicts the value of the output data, there is no need to categorize the target variable. Then input and target variables are selected, and the scaled dataset is divided into training (80%) and testing (20%) by using the pre-selected divisional approaches. Using the best hyperparameters (obtained by the trial-and-error method) fitted for predicting soil liquefaction, a classification and regression model are developed. The hyperparameters are selected based on the accuracy score in the classification model and the higher value of the Pearson correlation coefficient (R) in the regression model (Trial and error method: for which parameters the accuracy and R value are closer to the ideal one). Table 2 represents the hyperparameters used to develop each model. Based on the performance indicators calculated using the trial-and-error method, the best-predicted model was selected. Fig. 3 depicts the flowchart of the research methodology.

Table 2. Hyper parameters used to develop the models.

Divisional Approach	Classification Model	Regression Model
GridSearch CV	max_depth=10, min_samples_leaf=4, min_samples_split=10 scoring='accuracy', cv=5, n_jobs=-1, verbose=2	max_depth=10, min_samples_leaf=4, min_samples_split=10
kFold	max_depth=5, min_samples_leaf=2), max_samples=0.9, n_estimators=30, random_state=42	max_depth=10, min_samples_leaf=4, max_samples=0.7, random_state=42
FCM	max_depth=5, min_samples_leaf=2), max_samples=0.9, n_estimators=30, random_state=42	max_depth=10, min_samples_leaf=4, max_samples=0.7, random_state=42

3.2. M5 machine learning algorithm

The most popular classifier of the decision tree family, the M5 algorithm, creates regression trees with leaves made up of multivariate linear models. The M5 model tree is a numerical prediction algorithm wherein the nodes of the tree are selected based on an attribute that optimizes the expected reduction in error as a function of the output parameter's standard deviation [40]. Quinlan [41] was the first to introduce M5 model trees, and Wang and Witten [42] refined their concept to

create the M50 method. Regression trees and model trees both work well with big data sets. But compared to regression trees, model trees are typically much smaller and more accurate [40].

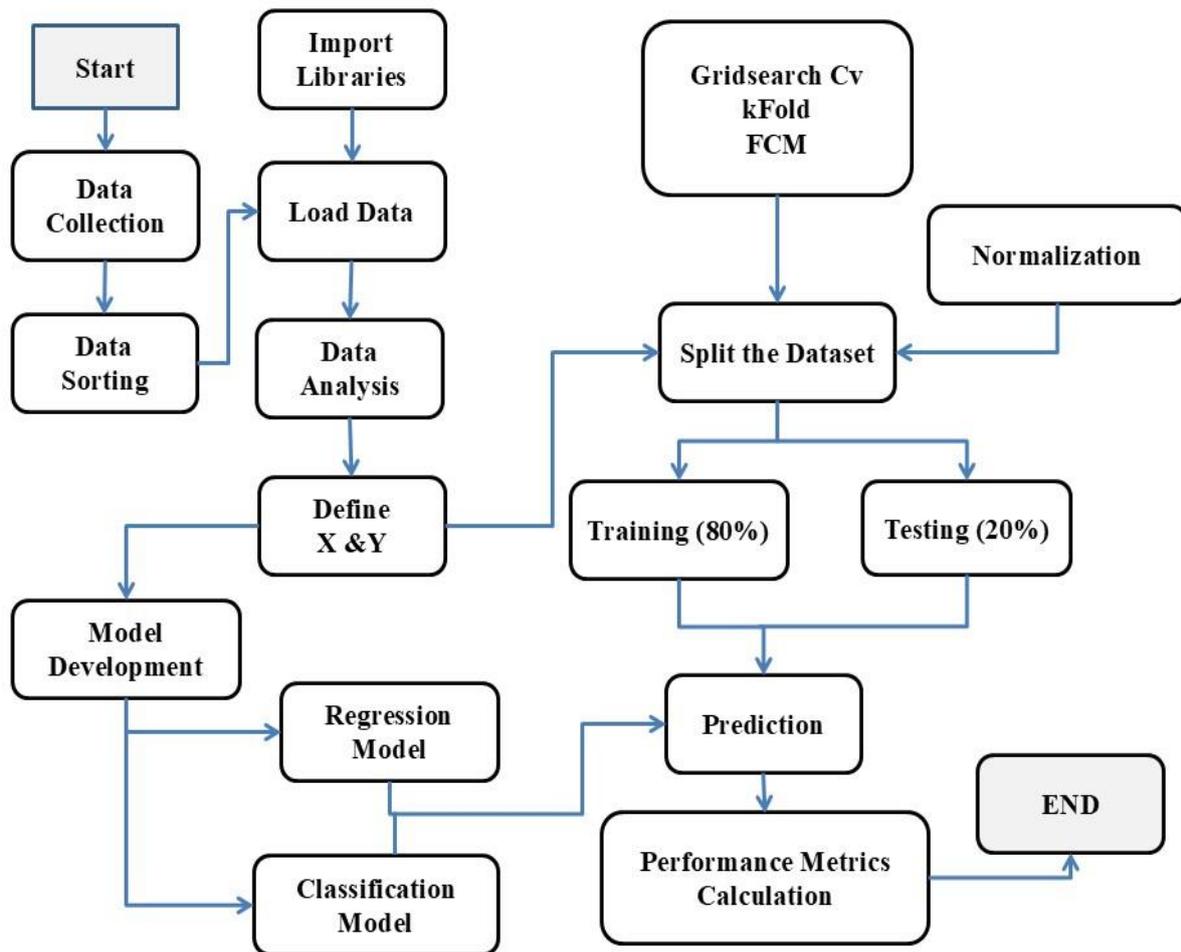


Fig. 3. Flowchart of the research methodology.

The M5 model tree algorithm first splits the instance space recursively in order to create a regression tree. In order to reduce intra-subset variability in the values as they flow from the root through the branch to the node, the splitting condition is applied. After testing each attribute at that node, the expected reduction in error is calculated, and the variability is expressed as the standard deviation of the values that reach that node from the root through the branch. The characteristic that maximizes the anticipated reduction in error is selected. If all of the instances that reach the node have output values that differ by a small amount or if there are only a few instances left, the splitting process is over. Following the growth of the tree, each inner node has a linear multiple regression model constructed for it using its associated data as well as all the attributes that are tested in the sub-tree rooted at that node. If removing an attribute lowers the expected error for future data, the linear regression models are then simplified. Following this simplification, pruning is taken into consideration for each subtree. When the estimated error for the linear model at a subtree's root is less than or equal to the sub-tree's expected error, pruning takes place. Following the final stage of pruning, discontinuities between adjacent linear nodes in the tree's leaves are compensated for through a regularization process. After the tree has been pruned, this procedure is initiated, particularly for models that are based on training sets with few instances (data points). Usually, the predictions are improved by this smoothing process [41].

3.3. Divisional approaches

GridSearchCV, KFold, and FCM were selected because of their complimentary advantages in clustering, validation, and model selection. When the dataset is tiny or there are just a few hyperparameters, GridSearchCV is quite useful for hyperparameter tuning since it can thoroughly test every possible combination of parameters. Although this exhaustive search has a good chance of discovering the ideal configuration, it runs the danger of overfitting if there is insufficient cross-validation. KFold cross-validation, which divides the dataset into k-folds, mitigates this problem. In this way, the model may be trained and validated on several subsets, reducing overfitting and producing a trustworthy performance score. Even though KFold's balanced approach takes a lot of time, it helps estimate model generalization, especially when "k" is chosen well. Last but not least, FCM provides flexible clustering by giving data points in several clusters partial memberships, making it appropriate for datasets with overlapping data. This "soft" clustering can identify intricate patterns that K-Means and other rigid clustering methods can miss. FCM increases the flexibility of data segmentation, despite its computational complexity and sensitivity to initial circumstances.

4. Results and discussions

In order to classify liquefaction and anticipate probable values—two crucial tasks for geotechnical and seismic engineering—this study intends to evaluate the possible applicability of the M5 decision tree soft computing model. So far, machine learning techniques have been used in the literature for liquefaction analysis. In this study, nine performance parameters—precision, recall, accuracy, F1_Score, log loss, Kappa coefficient, Mathew's correlation coefficient, receiver operating curves, or ROC curves, area under coverage, or AUC value, and specificity—are employed for classification. During training and testing, models (based on several divisional techniques) have shown remarkable performance in terms of the selected fitness criteria. The classification model's fitness parameters for training and testing data are shown in Table 3. The kFold divisional approach shows the best-fitting parameters among the techniques when comparing the actual values of performance metrics with the ideal values of each performance measure.

Table 3. Performance indices for classification model.

Performance indices	GridSearch Cv		kFold		FCM		Ideal Value
	Training	Testing	Training	Testing	Training	Testing	
Precision	0.9818	0.9629	0.9823	0.9375	0.9826	1	1
Recall	0.9729	0.8667	1	1	1	0.9	1
Accuracy	0.9688	0.875	0.9875	0.95	0.9875	0.923	1
F1_Score	0.9773	0.9122	0.991	0.9677	0.9912	0.9231	1
Log loss	1.1263	4.5054	0.4505	1.8022	0.4505	3.6044	lower
k	0.9269	0.6969	0.9702	0.8571	0.9695	0.7826	1
MCC	0.9269	0.7087	0.9706	0.866	0.9699	0.8017	1
AUC	0.9661	0.8833	0.9795	0.9	0.9787	0.929	1
Specificity	0.9592	0.9	0.9592	0.8	0.9574	1	1

Though precision (1.0), AUC value (0.929) and Specificity (1.0) for testing data are maximum at fuzzy clustering, which means a divisional approach, But other performance metrics such as Recall (1.0), accuracy (0.95), F1_score (0.9677), log loss (1.8022), kappa coefficient (0.8571), MCC (0.866) etc. are closer to the ideal value than the other divisional approach in the case of testing data, as are precision (0.9375), AUC (0.9) and Specificity (0.8) are closer to the accepted value for

testing data. For training data, the kFold divisional approach depicts satisfying performance indices values for training data as well (Precision=0.9823, Recall =1.0, Accuracy= 0.9875, F1_Score= 0.991, Log loss= 0.4505, k=0.9702, MCC= 0.9706, AUC=0.9795, Specificity =0.9592).

Figs. 4-6 represent the receiver operating curves for testing data for three divisional approaches in the classification model. The performance of a binary classification model can be assessed graphically using the ROC curve. Under various threshold settings, it displays the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). In this graphical representation, AUC, or area under the curve, is a single scalar value that represents the whole two-dimensional area beneath the whole ROC curve, which represents the performance of a binary classification model. One is the optimal classifier's AUC. All positive and negative instances are correctly distinguished by the model. No capacity to discriminate (AUC = 0.5). The model functions similarly to arbitrary guesswork. When $0.5 < \text{AUC} < 1$, the model has some discriminating ability. The model performs better at differentiating between positive and negative classes, the closer its AUC value is to 1. So, the ROC curve represents that the developed models have the differentiating power to predict soil liquefaction.

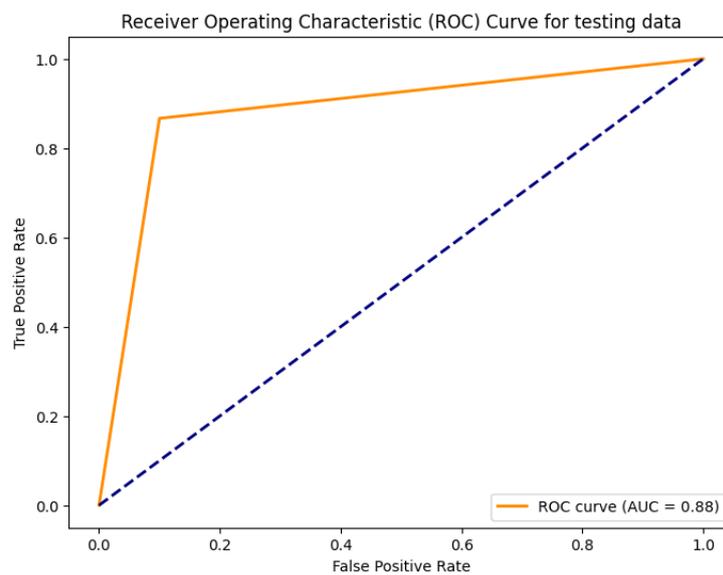


Fig. 4. ROC curve of testing data set for Gridsearch Cv divisional approach.

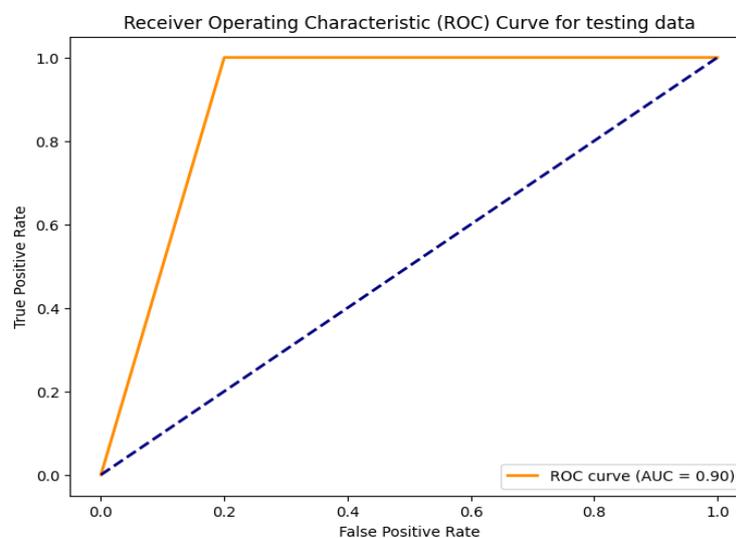


Fig. 5. ROC curve of testing data set for kFold divisional approach.

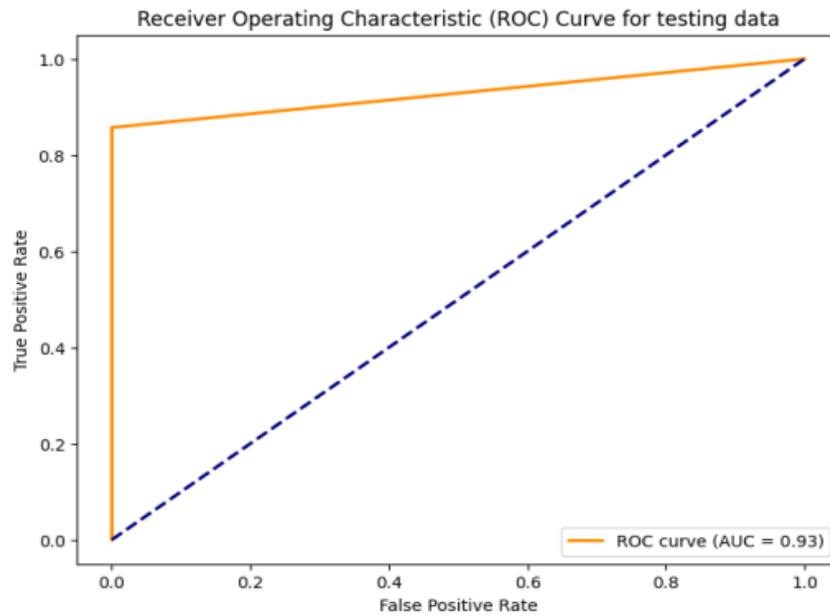


Fig. 6. ROC curves of testing data for FCM divisional approach.

On the other hand, eight performance indices, i.e., mean absolute area (MAE), root mean squared error (RMSE), coefficient of determination (R^2), Pearson correlation coefficient (R), adjusted coefficient of determination (adj R^2), variance adjusted for (VAF%), index of agreement (IoA), index of similarity (IoS), etc., are used to find the best regression approach to predict soil liquefaction in the selected zone. The values of performance metrics depict the success of the model in predicting soil liquefaction in cases of training and testing. Table 4 represents the performance indices of the regression model for training and testing data for each divisional approach.

Table 4. Performance indices for regression model.

Performance indices	GridSearch Cv		kFold		FCM		Ideal Value
	Training	Testing	Training	Testing	Training	Testing	
MAE	0.2718	0.6891	0.4685	0.5085	0.4362	0.6959	0
RMSE	0.5221	1.5097	0.9085	0.8581	0.7936	1.4186	0
R^2	0.9572	0.4372	0.8624	0.8765	0.9011	0.5031	1
R	0.9785	0.7175	0.9297	0.9471	0.95	0.7515	1
adj R^2	0.9569	0.4224	0.8615	0.8732	0.9005	0.49	1
VAF (%)	95.72	43.73	86.25	88.25	90.17	50.34	100
IoA	0.9353	0.7787	0.8806	0.8623	0.8947	0.7775	1
IoS	0.9889	0.8435	0.9605	0.9622	0.9727	0.8647	1

Among the divisional approaches, when the comparison of values of fitness parameters in the case of testing data is done, it is found that the best model is the kFold regression model, showing the values of indices closer to the ideal value than other divisional approaches (MAE = 0.5085, RMSE = 0.8581, R^2 = 0.8765, R = 0.9471, adj R^2 = 0.8732, VAF (%) = 88.25, IoA = 0.8623, and IoS = 0.9622). For training data, the kFold divisional approach also depicts accepted performance indices values (MAE = 0.4685, RMSE = 0.9085, R^2 = 0.8624, R = 0.9297, adj R^2 = 0.8615, VAF (%) =

86.25, IoA = 0.8806 and IoS = 0.9605). Table 5 represents the statistical information in the case of actual and predicted data found by each regression model, which indicates the ability to predict the liquefaction potential index by the deceived regression model.

Table 5. Statistical analysis actual and predicted LPI.

Statistical Parameters	Actual LPI	Predicted LPI		
		Gridsearch cv	kfold	FCM
Count	200	200	200	200
Average	2.442068	2.7579	2.3822	2.4007
Standard Deviation	2.453866	2.1481	2.1444	2.2756
Minimum	0	0	0	0
Maximum	9.6723	8.1222	7.1627	7.4368
Median	1.6576	2.3374	1.802	1.7652

Fig. 7-9 represent the graphical representations of actual and predicted data obtained for the testing dataset for different divisional approaches. Visual inspection shows that the graph is more similar in the case of the k-fold divisional approach. Thus, it can be inferred from the performance metrics that the M5 decision tree algorithm is unquestionably beneficial for forecasting the likelihood of liquefaction.

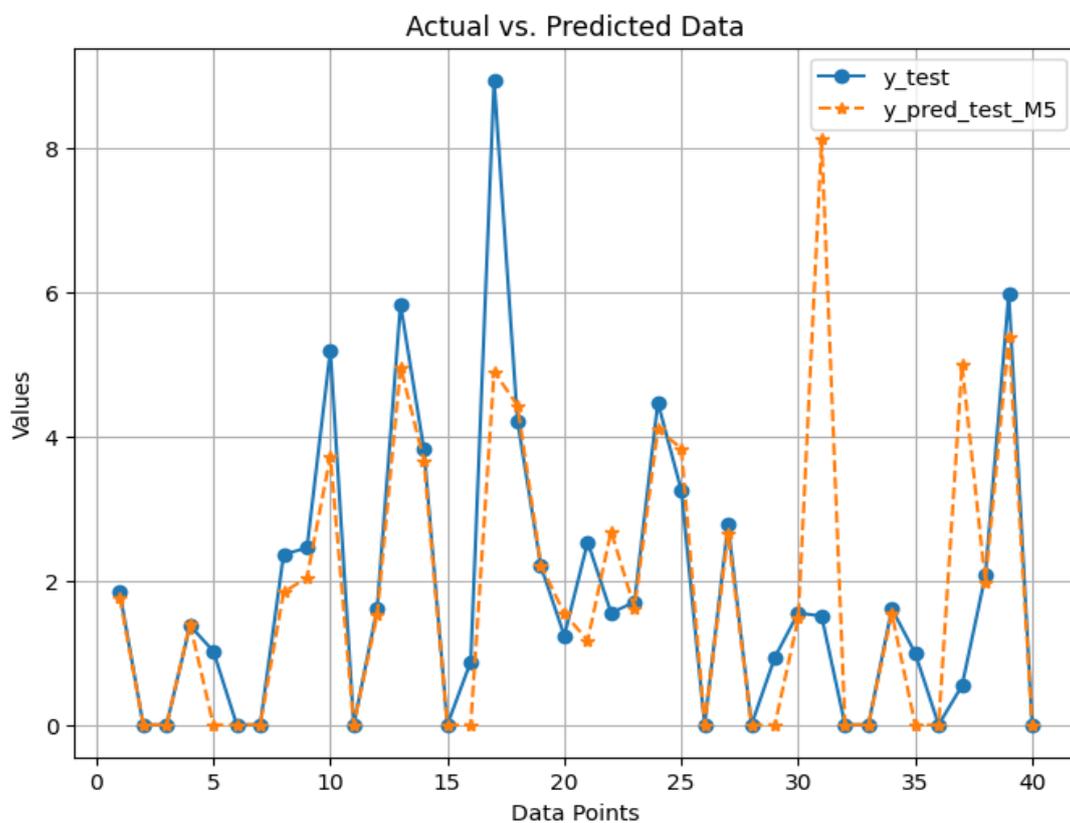


Fig. 7. Graphical illustration of actual and predicted data for Gridsearch Cv divisional approach.

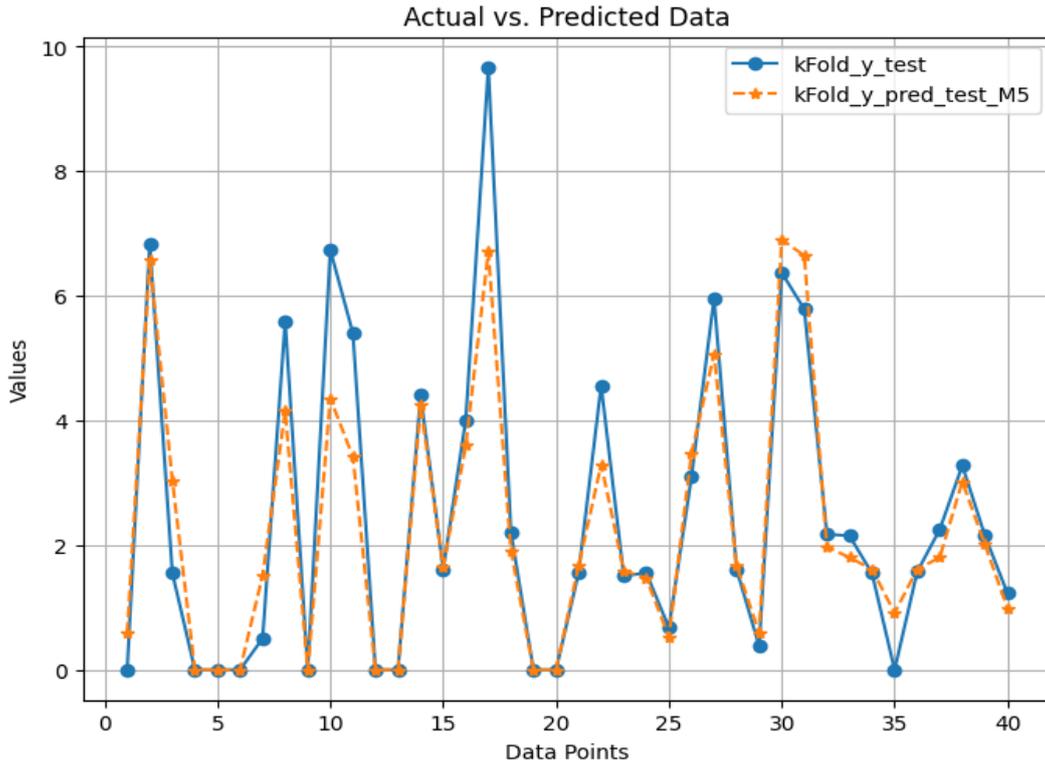


Fig. 8. Graphical illustration of actual and predicted data for kFold divisional approach.

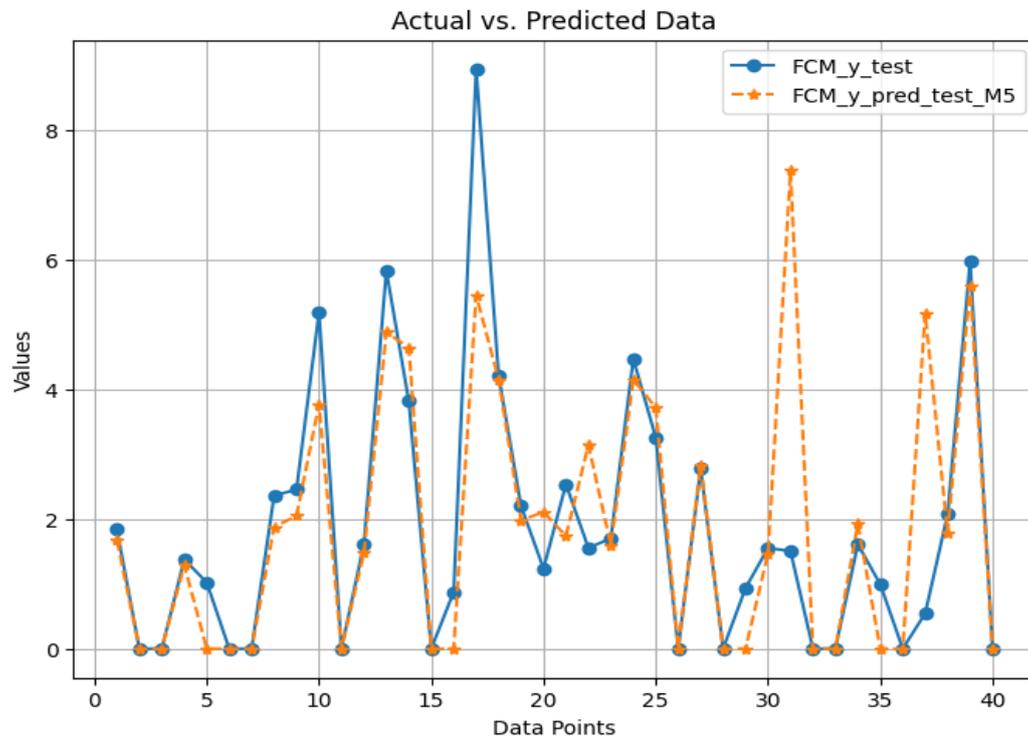


Fig. 9. Graphical illustration of actual and predicted data for FCM divisional approach.

5. Conclusions

In this study, M5 regression and classification models based on three divisional approaches were advanced and validated to predict the LPI of soil. To assess LPI of soil the dependent variables were collected for 200 locations. Min-max scaler was used to scaling all features. The complete dataset

was first split into training and testing sets. The model is then built using the training dataset, and the developed models are validated using the testing sets. The performance of both models in predicting soil LPI is summarized below.

- ❖ For the M5 regression model, the kfold clustering divisional approach shows better performance indices than other divisional approaches.
- ❖ The performance indices of testing dataset for kfold clustering are $R^2 = 0.88$, $R = 0.95$, $\text{adj } R^2 = 0.87$, $\text{VAF} (\%) = 88.25$, $\text{IoA} = 0.86$, and $\text{IoS} = 0.96$.
- ❖ An additional indication of the strong positive linear relationship between the predicted and actual values is the high Pearson correlation coefficient (R).
- ❖ For the M5 regression model, the Gridsearch CV shows lower performance indices for testing data with $R^2 = 0.44$, $R = 0.72$, $\text{adj } R^2 = 0.42$, $\text{VAF} (\%) = 43.73$, $\text{IoA} = 0.78$, and $\text{IoS} = 0.84$.
- ❖ For the M5 classification model, the kFold clustering achieved the highest predictive efficiency during the testing phase. The performance indices are precision = 0.9375, recall = 1.0, accuracy = 0.95, F1_score = 0.9677, log loss = 1.8022, kappa coefficient = 0.8571, MCC = 0.866, AUC = 0.9, and specificity = 0.8.
- ❖ Regarding the classification model, all performance metrics indicate a robust model with excellent predictive power for both positive and negative cases. i.e., the high accuracy and F1 score demonstrate that the model is accurate and successfully strikes a balance between precision and recall; the high precision and perfect recall show that the model effectively identifies positive cases with very few false positives and no false negatives.
- ❖ Additionally, the classification model's performance indices are worse on the Gridsearch CV. In contrast, FCM is shown to have the greatest AUC for the classification model.
- ❖ The training phase revealed no significant fluctuations or undesirable values for both the regression and classification models, indicating the generalization capabilities and resilience of the model.

The M5 decision tree method with the kFold divisional technique is a helpful tool for measuring the LPI of soil, according to the investigation's findings. To develop dependable and efficient prediction models, regression and classification soft computing techniques based on M5 decision tree algorithms were developed. The created models are distinguished by their strong generalization potential, cheap computing costs, and minimal over-fitting problems. The created models may be used as a useful tool to help geotechnical experts predict the soil liquefaction potential based on their overall performance.

Funding

For the research, authoring, and/or publishing of this work, the authors received no financial assistance.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Abu Sayed: Conceptualization, Methodology, Data collection, Data analysis, original draft writing.

Md. Rahman: Conceptualization, Supervision, Data analysis, original draft writing, Review & editing.

References

- [1] Huang Y, Yu M. Review of soil liquefaction characteristics during major earthquakes of the twenty-first century. *Nat Hazards* 2013;65:2375–84. <https://doi.org/10.1007/s11069-012-0433-9>.
- [2] Castro G, Poulos SJ. Factors Affecting Liquefaction and Cyclic Mobility. *J Geotech Eng Div* 1977;103:501–16. <https://doi.org/10.1061/AJGEB6.0000433>.
- [3] Youd T. Major cause of earthquake damage is ground failure. *Civ Eng* 1978;48:47–51.
- [4] Bao X, Jin Z, Cui H, Chen X, Xie X. Soil liquefaction mitigation in geotechnical engineering: An overview of recently developed methods. *Soil Dyn Earthq Eng* 2019;120:273–91. <https://doi.org/10.1016/j.soildyn.2019.01.020>.
- [5] Youd TL, Idriss IM. Liquefaction Resistance of Soils: Summary Report from the 1996 NCEER and 1998 NCEER/NSF Workshops on Evaluation of Liquefaction Resistance of Soils. *J Geotech Geoenvironmental Eng* 2001;127:297–313. [https://doi.org/10.1061/\(ASCE\)1090-0241\(2001\)127:4\(297\)](https://doi.org/10.1061/(ASCE)1090-0241(2001)127:4(297)).
- [6] Ayasrah M, Qiu H, Zhang X, Daddow M. Prediction of Ground Settlement Induced by Slurry Shield Tunnelling in Granular Soils. *Civ Eng J* 2020;6:2273–89. <https://doi.org/10.28991/cej-2020-03091617>.
- [7] Youd JM, Newman JMB, Clark MG, Appleby GJ, Rattigan S, Tong ACY, et al. Increased metabolism of infused 1-methylxanthine by working muscle. *Acta Physiol Scand* 1999;166:301–8. <https://doi.org/10.1046/j.1365-201x.1999.00572.x>.
- [8] Rahman MM, Hossain MB, Roknuzzaman M. Effect of peak ground acceleration (PGA) on liquefaction behavior of subsoil: A case study of Dinajpur Sadar Upazila, Bangladesh. *AIP Conf. Proc.*, 2023, p. 030002. <https://doi.org/10.1063/5.0129770>.
- [9] Monkul MM, Gültekin C, Gülver M, Akin Ö, Eseller-Bayat E. Estimation of liquefaction potential from dry and saturated sandy soils under drained constant volume cyclic simple shear loading. *Soil Dyn Earthq Eng* 2015;75:27–36. <https://doi.org/10.1016/j.soildyn.2015.03.019>.
- [10] V. Galavi, A. Petalas RBJB. No Title n.d.
- [11] Goh AT. Probabilistic neural network for evaluating seismic liquefaction potential. *Can Geotech J* 2002;39:219–32. <https://doi.org/10.1139/t01-073>.
- [12] Chen Z, Li H, Goh ATC, Wu C, Zhang W. Soil Liquefaction Assessment Using Soft Computing Approaches Based on Capacity Energy Concept. *Geosciences* 2020;10:330. <https://doi.org/10.3390/geosciences10090330>.
- [13] Liu L, Zhang S, Yao X, Gao H, Wang Z, Shen Z. Liquefaction Evaluation Based on Shear Wave Velocity Using Random Forest. *Adv Civ Eng* 2021;2021:1–9. <https://doi.org/10.1155/2021/3230343>.
- [14] Talamkhani S, Naeini SA, Ardakani A. Prediction of Static Liquefaction Susceptibility of Sands Containing Plastic Fines Using Machine Learning Techniques. *Geotech Geol Eng* 2023;41:3057–74. <https://doi.org/10.1007/s10706-023-02444-2>.
- [15] MY F, LA A-H, MM A, AA J, SA A-H. Coupled Finite Element and Artificial Neural Network Analysis of Interfering Strip Footings in Saturated Cohesive Soils. *Transp Infrastruct Geotechnol* 2024;11:2168–85. <https://doi.org/10.1007/s40515-023-00369-0>.
- [16] Barkhordari MS, Fattahi H, Armaghani DJ, Khan NM, Afrazi M, Asteris PG. Failure mode identification in reinforced concrete flat slabs using advanced ensemble neural networks. *Multiscale Multidiscip Model Exp Des* 2024. <https://doi.org/10.1007/s41939-024-00554-9>.
- [17] Rahman M, Hossain M, Sayed A, Thakur S. Assesment of Liquefaction Potential Based on the Logistic Regression Machine Learning Algorithm. *7th Int. Conf. Civ. Eng. Sustain. Dev. (ICCESD 2024)*, 2024, p. 1–11. <https://doi.org/10.13140/RG.2.2.24047.65440>.
- [18] Goh ATC. Seismic Liquefaction Potential Assessed by Neural Networks. *J Geotech Eng* 1994;120:1467–80. [https://doi.org/10.1061/\(ASCE\)0733-9410\(1994\)120:9\(1467\)](https://doi.org/10.1061/(ASCE)0733-9410(1994)120:9(1467)).

- [19] Pal M. Support vector machines-based modelling of seismic liquefaction potential. *Int J Numer Anal Methods Geomech* 2006;30:983–96. <https://doi.org/10.1002/nag.509>.
- [20] Samui P, Hariharan R. A unified classification model for modeling of seismic liquefaction potential of soil based on CPT. *J Adv Res* 2015;6:587–92. <https://doi.org/10.1016/j.jare.2014.02.002>.
- [21] Xue X, Liu E. Seismic liquefaction potential assessed by neural networks. *Environ Earth Sci* 2017;76:192. <https://doi.org/10.1007/s12665-017-6523-y>.
- [22] Ramakrishnan D, Singh TN, Purwar N, Barde KS, Gulati A, Gupta S. Artificial neural network and liquefaction susceptibility assessment: a case study using the 2001 Bhuj earthquake data, Gujarat, India. *Comput Geosci* 2008;12:491–501. <https://doi.org/10.1007/s10596-008-9088-8>.
- [23] Venkatesh K, Kumar V, Tiwari RP. APPRAISAL OF LIQUEFACTION POTENTIAL USING NEURAL NETWORK AND NEURO FUZZY APPROACH. *Appl Artif Intell* 2013;27:700–20. <https://doi.org/10.1080/08839514.2013.823326>.
- [24] Zhang Y, Qiu J, Zhang Y, Wei Y. The adoption of ELM to the prediction of soil liquefaction based on CPT. *Nat Hazards* 2021;107:539–49. <https://doi.org/10.1007/s11069-021-04594-z>.
- [25] Samui P, Sitharam TG. Machine learning modelling for predicting soil liquefaction susceptibility. *Nat Hazards Earth Syst Sci* 2011;11:1–9. <https://doi.org/10.5194/nhess-11-1-2011>.
- [26] Kumar DR, Samui P, Burman A. Prediction of Probability of Liquefaction Using Soft Computing Techniques. *J Inst Eng Ser A* 2022;103:1195–208. <https://doi.org/10.1007/s40030-022-00683-9>.
- [27] Kohestani VR, Hassanlourad M, Ardakani A. Evaluation of liquefaction potential based on CPT data using random forest. *Nat Hazards* 2015;79:1079–89. <https://doi.org/10.1007/s11069-015-1893-5>.
- [28] Zou M, Jiang W-G, Qin Q-H, Liu Y-C, Li M-L. Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting. *Materials (Basel)* 2022;15:5298. <https://doi.org/10.3390/ma15155298>.
- [29] Momeni E, Jahed Armaghani D, Hajihassani M, Mohd Amin MF. Prediction of uniaxial compressive strength of rock samples using hybrid particle swarm optimization-based artificial neural networks. *Measurement* 2015;60:50–63. <https://doi.org/10.1016/j.measurement.2014.09.075>.
- [30] Mostazid M, Rahman M, Rahman M. Seismic Vulnerability Assessment of Existing RCC Buildings in Dinajpur City: A Case Study on Ward No. 06. *Proc. Int. Conf. Planning, Archit. Civ. Eng.*, 2019, p. 1–6.
- [31] Oldham RD. Report on the great earthquake of 12th June 1897. 1899.
- [32] Stuart M. The Srimangal earthquake of 8th July 1918. *Mem Geo Surv India* 1920;46:1–70.
- [33] Hossain MS, Kamal ASMM, Rahman MZ, Farazi AH, Mondal DR, Mahmud T, et al. Assessment of soil liquefaction potential: a case study for Moulvibazar town, Sylhet, Bangladesh. *SN Appl Sci* 2020;2:777. <https://doi.org/10.1007/s42452-020-2582-x>.
- [34] Hossain MB, Roknuzzaman M, Rahman MM. Liquefaction Potential Evaluation by Deterministic and Probabilistic Approaches. *Civ Eng J* 2022;8:1459–81. <https://doi.org/10.28991/CEJ-2022-08-07-010>.
- [35] Morino M, Maksud Kamal ASM, Muslim D, Ekram Ali RM, Kamal MA, Zillur Rahman M, et al. Seismic event of the Dauki Fault in 16th century confirmed by trench investigation at Gabrakhari Village, Haluaghat, Mymensingh, Bangladesh. *J Asian Earth Sci* 2011;42:492–8. <https://doi.org/10.1016/j.jseaes.2011.05.002>.
- [36] Morino M, Kamal ASMM, Akhter SH, Rahman MZ, Ali RME, Talukder A, et al. A paleo-seismological study of the Dauki fault at Jaflong, Sylhet, Bangladesh: Historical seismic events and an attempted rupture segmentation model. *J Asian Earth Sci* 2014;91:218–26. <https://doi.org/10.1016/j.jseaes.2014.06.002>.

- [37] Steckler MS, Mondal DR, Akhter SH, Seeber L, Feng L, Gale J, et al. Locked and loading megathrust linked to active subduction beneath the Indo-Burman Ranges. *Nat Geosci* 2016;9:615–8. <https://doi.org/10.1038/ngeo2760>.
- [38] Idriss IM, Boulanger RW. Semi-empirical procedures for evaluating liquefaction potential during earthquakes. *Soil Dyn Earthq Eng* 2006;26:115–30. <https://doi.org/10.1016/j.soildyn.2004.11.023>.
- [39] Jahangiri A, Rakha HA. Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Trans Intell Transp Syst* 2015;16:2406–17. <https://doi.org/10.1109/TITS.2015.2405759>.
- [40] Zhang D, Tsai J. *Advances in Machine Learning Applications in Software Engineering*. Idea Group Inc; 2007.
- [41] Quinlan J. Learning with continuous classes. *Proc. 5th Aust. Jt. Conf. Artif. Intell., WORLD SCIENTIFIC*; 1992, p. 1–410. <https://doi.org/10.1142/9789814536271>.
- [42] Wang Y, Witten I. Induction of model trees for predicting continuous lasses. *Proc. ninth Eur. Conf. Mach. Learn.*, 1997, p. 128–37.