ASSESSMENT OF LIQUEFACTION POTENTIAL BASED ON THE LOGISTIC REGRESSION MACHINE LEARNING ALGORITHM

Md. Mahabub Rahman[1], Md. Belal Hossain[2], Abu Sayed[*3] and Sonu Thakur[4]

[1] *Lecturer, Department of Civil Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh, e-mail: mmr.civil@hstu.ac.bd*
[2] *Associate Professor, Department of Civil Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh, e-mail: mbh.civil@hstu.ac.bd*
[3] *Undergraduate student at Department of Civil Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh, e-mail: sayed55443138@gmail.com*
[4] *Graduate student at Department of Civil Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh, e-mail: sraazthakur076@gmail.com*

***Corresponding Author**

ABSTRACT

The assessment of liquefaction potential is a crucial aspect in the field of geotechnical engineering. It makes it possible to assess the soil's vulnerability to liquefaction in the event of a seismic event. The application of a machine learning algorithm to improve the precision of liquefaction potential assessment is the particular focus of this study. The Groundwater table (GWT), Effective stress (rav), Effective overburden stress (r'av), Fineness content (F<0.0075), Corrected SPT-N value (N1(60CS)), Depth (Z(m)), and Peak ground acceleration (PGA) are the input parameters that are used in logistic regression to predict the liquefaction potential. The input parameters used in this study were collected from the authors earlier studies. In order to build an estimator model, these parameters were gathered. This study intends to investigate the efficacy of logistic regression in precisely estimating the potential for liquefaction through an extensive analysis. The algorithm's performance is assessed through the use of metrics like f1-score, recall, accuracy, and precision. With an accuracy of 93.3% for testing data and 95% for training data, the experimental results show that the logistic regression algorithm performed perfectly on the dataset. The results of this study could greatly increase the precision of liquefaction assessment, leading to better decision-making in the domains of seismic hazard mitigation and geotechnical engineering. This research adds to the ongoing efforts to enhance comprehension of soil behavior under seismic conditions by utilizing machine learning capabilities.

Keywords: *Liquefaction, logistic regression, machine learning algorithm, geotechnical parameters*

1. INTRODUCTION

One of the most damaging effects of earthquakes is liquefaction, which occurs when saturated, loose sand deposits lose their shear strength. In the last five decades, scientists have carried out a great deal of research and put forth a number of strategies to forecast the occurrence of this catastrophic event. Initially, undrained cyclic loading laboratory tests have been used to assess a soil's liquefaction potential (Castro, 1975; Peck, 1979; Castro, 1987). However, many researchers have chosen to use in situ tests because it is more difficult to obtain undisturbed samples of loose sandy soils (Seed et al., 1983; Juang et al., 2014; Kayen & Mitchell, 2008). The researchers have developed a number of empirical formulas that are based on various in situ soil tests, including shear wave velocities (Vs), self-boring pressure meter tests (BPT), cone penetration test CPT, and standard penetration test SPT. The two essential components of empirical field-based procedures for liquefaction potential determination are: (i) an appropriate in situ index to represent soil liquefaction characteristics; and (ii) an analytical framework to organize historical experiences. However, because site investigation work and laboratory testing are required, these methods are expensive. It is therefore necessary to find a simpler method of computing soil liquefaction potential (SLP).

The application of machine learning (ML)-based techniques to complex geotechnical problems has been shown by geotechnical researchers (Karthikeyan & Samui, 2014; Fahim et al., 2022; Ghani & Kumari, 2022). A deep learning (DL) model for accurate soil classification in liquefaction determination was presented by Kumar et al., in 2021. With the use of emotional backpropagation neural networks (EMBP), the applicability of the DL model was examined. A computer model for calculating the potential for soil liquefaction using artificial neural networks is presented by Tung et al., (1993). It is claimed that the model, which was developed using data sets from previous occurrences, can be used to comprehend events that will occur in the future. Using geotechnical, geometric, and seismic load parameters, García et al., (2012) introduce a machine learning scheme to assess the liquefaction potential of soils. Field observations of the liquefaction performance of past earthquakes, along with a sizable database of CPT and vs measurements, are examined. Liquefaction can be predicted with neural networks and classification trees in a nonlinear environment created by this database. In order to assess the soil's potential for liquefaction in the event of an earthquake, Ahmad et al., (2021) examined the effectiveness of four machine learning (ML) algorithms using the cone penetration test (CPT) based on field case history records. Hu, (2021) developed two Bayesian network models to forecast soil liquefaction using shear wave velocity databases and dynamic penetration testing. When the created models were compared to the ones that already existed, it was found that they performed well. According to earlier research on the use of ML techniques for soil liquefaction potential (SLP), soils were correctly classified into liquefied and non-liquefied soil classes by the ML models (Li et al., 2020; Chen et al., 2018; Sharma & Singh, 2017).

The goal of this work is to develop an empirical machine learning (ML) approach for liquefaction potential assessment. Empirical formulas are used to evaluate liquefaction triggering based on standard penetration test (SPT) data from various government and private organizations in Dinajpur Sadar in order to meet research objectives. Subsequently, the acquired dataset's seismic liquefaction triggering is predicted using a machine learning algorithm. A machine learning algorithm of the linear classification model type is called logistic regression. In order to produce binary outputs, such as grid search cross-validation, it implements the sigmoid function. This study also emphasizes how various soil parameters are correlated to cause the liquefaction of the soil. Confusion matrices are then used to assess the developed models, and the results are subsequently utilized to determine the model is also assessed using the following metrics: AUC value, Cohen's kappa coefficient, F1 score, log loss, precision, recall, accuracy, Mathews Correlation Coefficient (MCC), Specificity, and Overall Accuracy.

2. METHODOLOGY

3. Study Area

This study used soft computing techniques to find the liquefaction potential index, representing an analytical procedure. This is carried out using the SPT test data that was gathered from various Dinajpur zones. Because Dinajpur is situated in the Terai basin, its soil has a significantly higher sand-to-silt ratio than clay. Because of the numerous fault lines in this area, including the Assam, Sub Dauki, Bogra, and Shillong faults, it is especially prone to earthquakes. Khansama, Ghoraghat, Nawabganj, Parbatipur, Fulbari, Biral, Birrampur, Birganj, Bochaganj, and Hakimpur are the 13 upazilas that make up the region. Dinajpur Sadar is expanding quickly as a result of industrialization and urbanization. Many infrastructure projects are under construction in order to meet the increasing demand.

4. Data Collection

The input parameters for this investigation were gathered from previous studies conducted by Hossain et al., (2022). In previous studies, data was collected from public and private organizations that have carried out subsurface investigations at different locations across the city. The previous study used data that was taken from the 150 soil test reports. The data sets collected cover the majority of the region. The SPT test results using the deterministic approach determined the soil layers' liquefaction state. These borehole data, along with the analysis results from previous studies, were then tabulated to produce the data sets required for the machine learning models. Groundwater table (GWT), Effective stress (rav), Effective overburden stress (r'av), Fineness content (F<0.0075), Corrected SPT-N value (N1(60CS)), Depth (Z(m)), and Peak ground acceleration (PGA) are the collected input parameters.

5. Working Principle

Nearly all complex geotechnical engineering problems have been solved in the past ten years using a variety of soft computing techniques. These studies are mostly based on Artificial Neural Network (ANN) models with various network architectures. This study attempts to estimate the soil liquefaction resulting from challenging and time-consuming laboratory studies using a logistic regression model based on grid search cross validation techniques. A machine learning algorithm's operation starts with data collection, which is the process of gathering pertinent data and ensuring its quality through pre-processing. Next, pre-processing is used to eliminate any outliers or inconsistencies from the data, and normalization is applied to guarantee that all variables have the same scale. Grid search cross validation techniques are used in conjunction with StandardScaler to determine the optimal hyperparameters for accurately predicting the target variable and achieve normalization. The optimal hyperparameter for this investigation is found to be c=10; penalty=12; solver=liblinear, utilizing the best estimator of grid search. A test dataset is used to train the logistic regression model and evaluate its capacity for generalization. Finally, appropriate metrics are used to assess the model's performance.

6. Logistic Regression

The logistic regression algorithm is a type of supervised learning model and also statistical model which is used to estimate the likelihood of a binary result, such as success or failure, true or false, or yes or no, depending on one or more independent variables. Logistic regression model usually uses a logistic function to model a binary dependent variable. The logistic function, also known as the sigmoid function, is represented by the equation:

$$1/(1 + e^{-z})$$

where 'z' is the input to the function, e is the base of natural logarithms, and the output is a value between 0 and 1. This output can be interpreted as the probability of the positive class.

To best predict the output, the algorithm determines the weights for each input feature. To do this, the difference between the values predicted by the algorithm and the actual values is measured by the cost function, which can be minimized. Given is the logistic regression cost function, which is:

$$-y\log(h) - (1-y)\log(1-h)$$

Here 'y' is the actual value, h is the predicted value, and log is the natural logarithm.

Gradient descent is used to update the weights iteratively until the cost function is minimized. The liquefaction potential of fresh data can be predicted using the model once it has been trained. Metrics like accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) are used to assess the model's performance. All things considered, the Logistic Regression Machine Learning Algorithm offers a reliable and effective way to determine liquefaction potential. It's an important tool for this task because it can handle big datasets and intricate relationships between variables.

7. Feature Scaling

Standardization is a technique used to scale features for machine learning models. Data standardization is a process that combines the structures of multiple datasets into a single, consistent data format. The alteration of datasets occurs subsequent to the collection of data from diverse sources, but precedes the ingestion of the data into the intended systems.

$$z = (x - \mu) / \sigma$$

Here 'z' represents the standard value, 'x' is the original value, the mean is denoted by 'μ' and standard deviation is represented by 'σ'.

Standardization preserves the distribution and form of the data while making it more suitable for algorithms that assume a normally distributed set of data. Easy-to-use tools for standardization are provided by Python modules like Scikit-Learn. To scale the input data, standardize the features in the dataset using the StandardScaler class.

8. Model Evaluation

Numerous techniques are employed to assess a model's performance in classification. The aforementioned metrics comprise precision, recall, accuracy, F1 score, log loss, Cohen's kappa coefficient, Mathews Correlation Coefficient (MCC), specificity, and AUC value. These metrics are designed to assess the degree of categorical accuracy that exists between the expected and actual outcomes. For the most part, a good model is anything above 0.8.

9. RESULTS AND DISCUSSIONS

10. Statistical Information

Descriptive statistics can also be used to characterize an entire dataset. Descriptive statistics, in short, help with the description and comprehension of the features of a given data set by providing concise summaries of the data set's samples and measurements. Measures of centre, such as the mean, median, and mode, are among the most widely used types of descriptive statistics. They are utilized in nearly all mathematics and statistics courses across all educational levels. To find the mean, also known as the average, add up all the numbers in the data set. Subtract this sum from the total number of figures in the dataset. Descriptive statistics (spread) are a statistical subset that includes measures of variability and central tendency. The mean, median, and mode are three frequently used metrics to assess central tendency. On the other hand, measures of variability include variance, standard deviation, minimum and maximum variables, and variance. The input dataset's statistical data is shown in Table 1.

Table 1: Statistical information of input data

Index	Z(m)	N1(60cs)	F<0.0075	G.W.T	rav	r'av	Pga
count	150	150	150	150	150	150	150
mean	3.76077	11.2483	72.6953	2.74667	70.5053	46.7281	0.2354
std	2.56634	7.12516	24.5857	0.79109	47.6095	24.288	0.64163
min	0.7621	2.2921	11	1.5	18.2427	1.1043	0.03
25%	1.82	7.0358	49.25	2.25	27.8089	24.073	0.12
50%	3.81	10.1059	81	2.5	67.0058	42.8637	0.2
75%	5.33	13.4512	96	3	94.5778	59.9119	0.23
max	10.67	70.1165	99	4.75	199.15	115.35	8

11. Model Validation

11.1.1 Correlation Matrix

A table that displays the pairwise correlations between a group of variables is called a correlation matrix. When working with multivariate data, it is an especially useful tool in statistics and data analysis. The degree and direction of a linear relationship between two variables are measured by correlation. In this study Pearson correlation coefficient matrix is used to represent the correlations between input variables. Figure 1 shows the correlation matrix for the dataset.

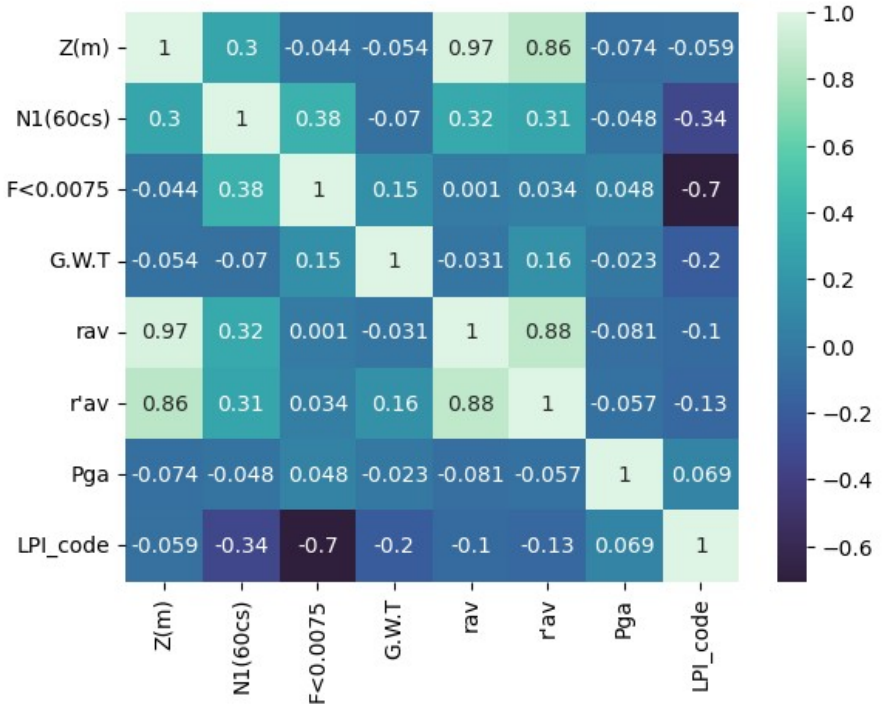

Figure 1: Correlation matrix

In this study, binary classification is done based on Logistic regression, where 0 means no liquefiable and 1 means liquefiable. Considering binary classification, the correlation matrix represents that Fineness content (F<0.0075) and N1(60CS) has a strong relationship (r=-0.7 and r=-0.34 respectively)

to predict Liquefaction Potential Index (LPI) and very lower or about to neutral relationship to predict Liquefaction Potential Index (LPI) for depth (r=-0.059) and PGA (r=0.069).

11.1.2 ROC Curve

Some conclusions are drawn from the evaluation, and these are then displayed using a Receiver Operating Characteristic (ROC) curve. The Area under the ROC Curve (AUC) represents a value between 0 and 1, with a value closer to 1 suggesting better model performance. The ROC curve is a plot of sensitivity versus false positive rate, where the line along the diagonal represents a pure 50% chance of accurate prediction of a model. To summarize, the ROC curve plots the false positive rate against the sensitivity (Park, Goo, & Jo, 2004). An AUC value of more than 0.7 is typically regarded as an acceptable value for the purposes of validating the model. An AUC value of more than 0.7 is typically regarded as an acceptable value for the purposes of validating the model. The study revealed that the Logistic Regression model had an AUC value of 0.9523 for testing data and an AUC value of 0.9509. These values suggest that the model is functioning well. The graphical representation of ROC curve for testing and training data is shown on Figure 2 and Figure 3.

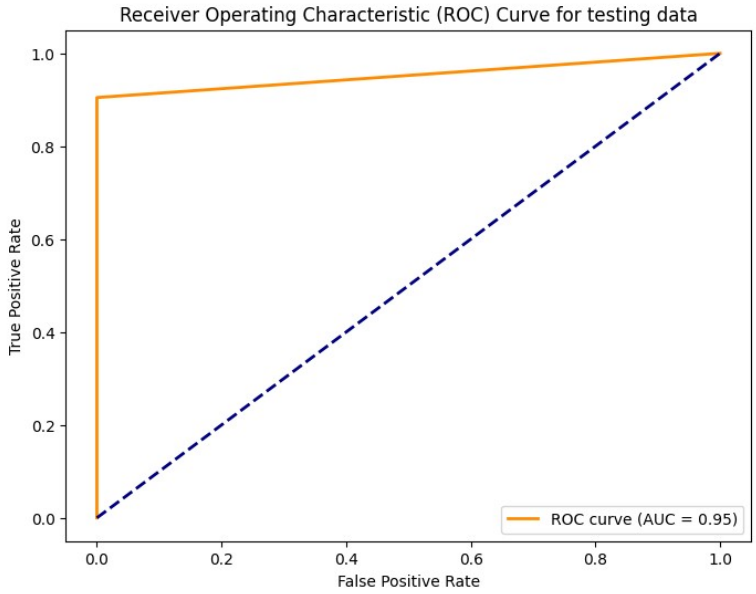

Figure 2: ROC curve for testing data

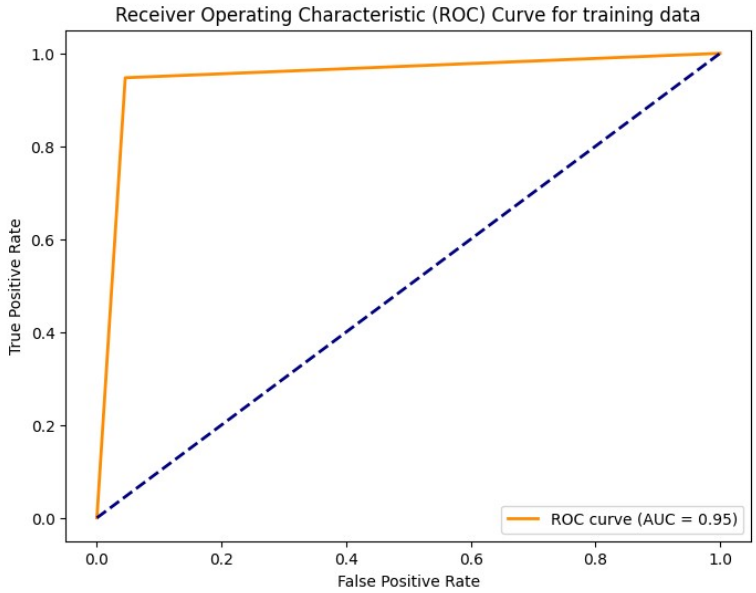

Figure 3: ROC curve for training data

11.1.3 Confusion Matrix

A confusion matrix is a technique used to analyse and summarize how well a classification technique performs. Relying only on classification accuracy may result in incorrect conclusions if dataset contains more than two classes or if the number of observations in each class is uneven. A better idea can be obtained about the classification model as it is working correctly and the errors that it is making by calculating a confusion matrix. So simply, an explanation of a classification model's performance when applied to a set of test data for which the true values are known is often provided by a confusion matrix, which is a table that compares actual true negative and positive data with predicted true negative and true positive data that is successfully achievable. Therefore, a graphical depiction of the confusion matrix can be used to understand the overall comparison between the actual and projected data. In order to assess the model's performance in terms of true positives, true negatives, false positives, and false negatives, it summarizes the predicted and actual classifications (Figure 4).

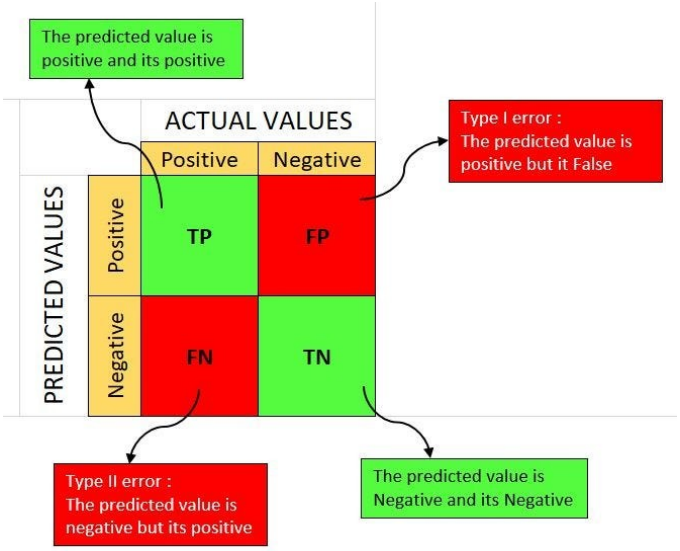

Figure 4: Confusion Matrix

The confusion matrix for testing data is graphically represented by the Figure 5.

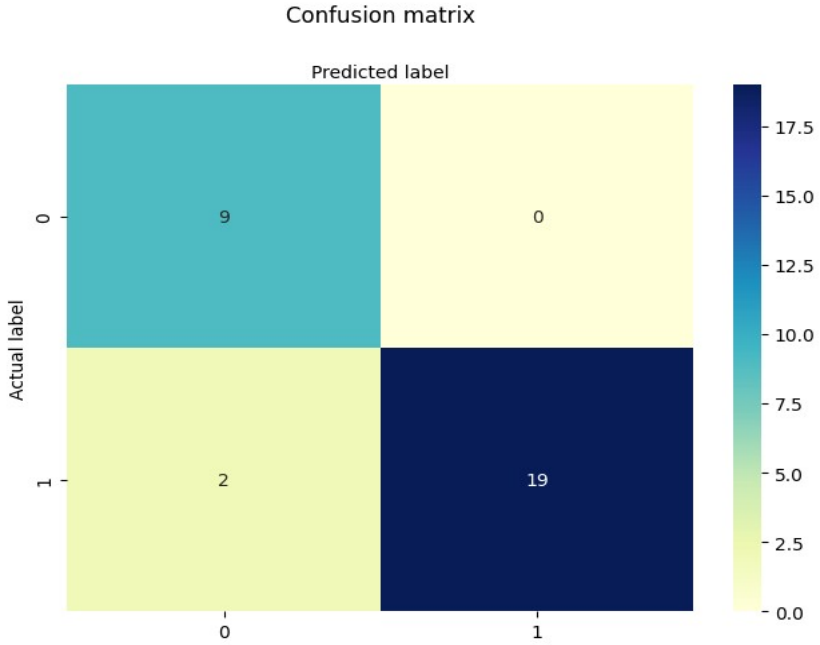

Figure 5: Confusion matrix for testing data

The Figure 5 represents that, 19 sample data is predicted as liquefiable and these are actually liquefiable, model predict 9 sample data as no liquefiable ad these are actually non liquefiable, no sample is predicted incorrectly as liquefiable but 2 sample data is predicted as no liquefiable but these 2 data is actually liquefiable. Figure 6 represents the graphical representation of confusion matrix in terms of training data.

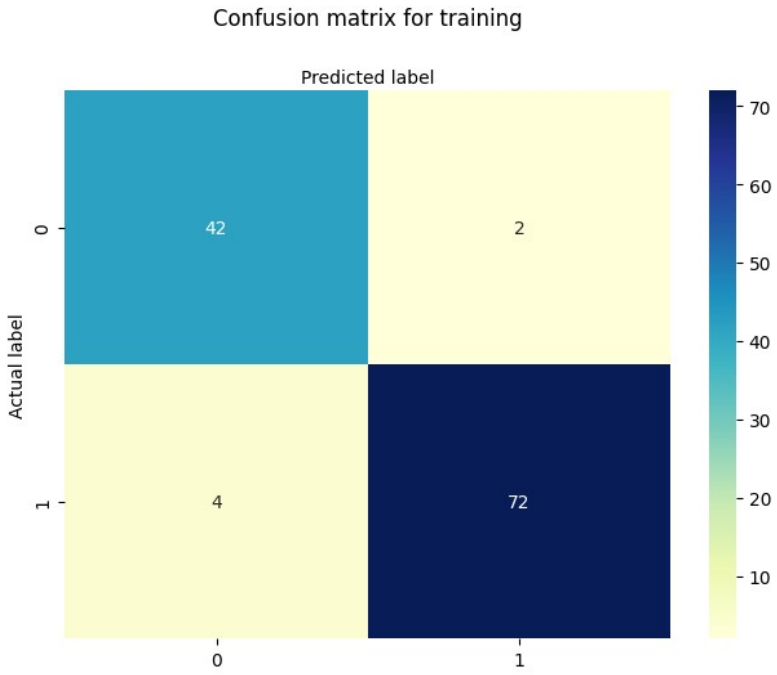

Figure 6: Confusion matrix for training data

Figure 6 shows that, the classification model predict 72 and 42 data as liquefiable and no liquefiable respectively and these data are actually liquefiable and no liquefiable. But the model predicts 2 data as liquefiable and is actually no liquefiable as well as 4 data is predicted as liquefiable but these are actually no liquefiable.

11.1.4 Performance Evaluation Indicators

A classification model's performance can be evaluated using a variety of metrics and performance indicators, depending on the situation. In this study, Precision, recall, accuracy, F1 score, Log loss, Cohen's kappa coefficient, Mathews Correlation Coefficient (MCC), Specificity and AUC value are used as performance evaluation indicators. The value obtained by the classification model is tabulated on the Table 2.

Table 2: Performance evaluation indicators

Indices	Testing	Training	Ideal Value
Precision	1	0.9729	1
Recall	0.9047	0.9474	1
Accuracy	0.933	0.95	1
F1 Score	0.95	0.9599	1
Log loss	2.4029	1.802	Lower
Cohen's kappa	0.8507	0.8934	1
MCC	0.8604	0.8939	1
Specificity	1	0.9545	1
AUC	0.95	0.95	1

From Table 2, It is proved that the logistic regression model shows an acceptable performance in terms of binary classification. That's why logistic regression is preferred as a classification machine learning algorithm.

11.1.5 Performance Comparison

In this study, the binary classification was done based on the logistic regression algorithm, where 0 means no liquefiable and 1 means liquefiable. In Table 3, the actual and predicted results for testing and training data are given.

Table 3: Comparison between analytical result and result predicted using ML

Actual and Predicted result for Training data							
Actual	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**
0	0	1	1	0	0	1	1
0	0	1	1	1	1	1	1
1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1
0	1	1	1	1	1	0	0
1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	0
0	0	1	1	1	1	1	1
1	1	1	0	1	1	0	0
1	1	0	0	1	1	1	1
1	1	1	1	0	0	1	1
1	1	0	0	0	0	1	1
1	1	1	1	0	0	1	1
0	0	0	0	1	1	0	0
1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	1
1	1	0	0	1	1	0	0
1	1	0	0	0	0	1	1
0	0	0	0	0	0	1	1
0	0	1	1	1	1	1	1
0	0	1	1	1	1	1	1
0	0	0	0	1	0	1	1
0	0	1	1	0	0	1	1
0	0	1	0	1	1	1	1
0	0	1	1	0	0	1	1
0	0	1	1	1	1	0	0
0	0	1	1	1	1	1	1
0	0	0	0	0	0	1	1
Actual and Predicted data For Testing data							
Actual	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**
1	1	1	1	1	1	1	1
0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0
1	1	0	0	1	0	0	0
0	0	1	1	1	1	1	1
1	1	0	0	1	0	0	0
1	1	1	1	1	1		
1	1	0	0	1	1		

Careful visualization proved that, among the testing dataset only 2 locations are falsely predicted (more details on figure 5) and among the training datasets, 6 locations predicted as wrong. The validation of such kind of binary classification study depends on the performance indices based on the comparison of result obtained by analytical method (actual) and by soft computing techniques (predicted). The performance indices for both testing data and training data are mentioned on the Table 2.

12.CONCLUSIONS

It is discovered in this study that performance indicators alone are insufficient for evaluating classification performance. There are also more effective ways to evaluate the model validation and classification capability, such as using ROC curves and confusion matrices. Thus, a number of performance indicators are employed in this study, and the results demonstrate that logistic regression performs well in both the training and testing datasets when it comes to classifying soil liquefaction. Additional results are enumerated below:

i. This study used logistic regression methods based on grid search cross validation procedures to accomplish its purpose of examining how effective logistic regression is in accurately forecasting the liquefaction potential.
ii. This study shows the logistic regression model has an accuracy to predict soil liquefaction of 93.3% for testing data and 95% for training data.
iii. This model has an AUC value of 0.95 for both training and testing data prediction. With an AUC of 0.95, the model appears to have a high degree of accuracy in differentiating between the two classes. It provides compelling evidence of the model's predictive power with minimal overlap, few misclassifications.
iv. The confusion matrix indicates that out of the 150-input data, the developed model correctly predicts 142 sample data, while only 8 sample data are incorrectly classified.

REFERENCES

Castro, G. 1975. Liquefaction and cyclic deformation of sands. *ASCE Journal of the Geotechnical Engineering Division,* 101(GT6): 551-569.

Peck, R.B. 1979. Liquefaction potential: science versus practice. *ASCE Journal of the Geotechnical Engineering Division,* 105(GT3): 393-398.

Castro, G. 1987. On the behavior of soils during earthquake liquefaction. *Proceedings, 3rd International Conference on Soil Dynamics and Earthquake Engineering,* St. Louis, Mo., pp. 169-205.

Seed, H.B., Idriss, I.M., & Arango, I. (1983). Evaluation of liquefaction potential using field performance data. *Journal of Geotechnical Engineering,* 109(3):458–482.

Juang, C.H., Jiang, T., & Andrus, R.D. (2014). Simplified SPT-based liquefaction analysis procedures using CPT measurements. *J. Geotechnical and Geo-environmental Engineering,* 140(2): 04013020. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000971.

Kayen, R., & Mitchell, J.K. (2008). Assessment of liquefaction potential during earthquakes by Aria's intensity. *Journal of Geotechnical and Geo-environmental Engineering,* 134(9):1285–1300. https://doi.org/10.1061/(ASCE)1090-0241(2008)134:9(1285).

Karthikeyan, J., & Samui, P. (2014). Application of statistical learning algorithms for prediction of liquefaction susceptibility of soil based on shear wave velocity. Geomatics, *Natural Hazards and Risk,* 5(1):7-25. https://doi.org/10.1080/19475705.2012.757252

Fahim, A.K.F., Rahman, M.Z., Hossain, M.S., & Kamal, A.M. (2022). Liquefaction resistance evaluation of soils using artificial neural network for Dhaka City, Bangladesh. *Natural Hazards,* 113(2): 933-963. https://doi.org/10.1007/s11069-022-05331-w

Ghani, S., & Kumari, S. (2022). Liquefaction behavior of Indo-Gangetic region using novel metaheuristic optimization algorithms coupled with artificial neural network. *Natural Hazards,* 111(3): 2995-3029. https://doi.org/10.1007/s11069-021-05165-y

Kumar, D., Samui, P., Kim, D., & Singh, A. (2021). A Novel Methodology to Classify Soil Liquefaction Using Deep Learning. *Geotech. Geol. Eng.,* 39(2):1049–1058. https://doi.org/10.1007/s10706-020-01544-7.

Tung, A.T.Y., Wang, Y.Y., & Wong, F.S. (1993). Assessment of liquefaction potential using neural networks. *Soil Dynamics Earthquake Engineering,* 12 (6), 325–335.

García, S., Ovando-Shelley, E., Gutierrez, J., & García, J. (2012). Liquefaction Assessment through Machine Learning Approach. *15th World Conference on Earthquake Engineering.*

Ahmad, M., Tang, X.W., Qiu, J.N., Ahmad, F., & Gu, W.J. (2021). Application of machine learning algorithms for the evaluation of seismic soil liquefaction potential. *Front. Struct. Civ. Eng.,* 15 (2):490–505. https://doi.org/10.1007/s11709-020-0669-5.

Hu, J. (2021). A new approach for constructing two Bayesian network models for predicting the liquefaction of gravelly soil. *Computational Geotechnics,* 137, 104304. https://doi.org/10.1016/j.compgeo.2021.104304.

Li, X., Wang, Z., & Li, L. (2020). Evaluation of liquefaction potential of sandy soils using machine learning algorithms. *Eng. Geol*., 269, 105569. https://doi.org/10.1016/j.enggeo.2020.105569.

Chen, Y., Zhang, Y., & Li, Z. (2018). Liquefaction potential assessment using artificial neural network: A case study in Tangshan, China. Geotech. Geol. Eng. 36(6):3671–3682. https://doi.org/10.1007/s10706-018-0620-2.

Sharma, L., & Singh, R. (2017). Liquefaction susceptibility mapping using artificial neural network model. *Natural Hazards,* 87 (1), 345–365. https://doi.org/10.1007/s11069-017-2759-6.

Hossain, M.B., Roknuzzaman, M., & Rahman, M.M. (2022). Liquefaction Potential Evaluation by Deterministic and Probabilistic Approaches. *Civil Engineering Journal,* 8(7):1459-1481. http://dx.doi.org/10.28991/CEJ-2022-08-07-010