ORIGINAL RESEARCH

A Hybrid Geotechnical–Machine Learning Framework for Evaluating Soil Liquefaction Potential Using CPT Data and Gradient Boosting Models (XGBoost and LightGBM)

Shima Aghakasiri [1],*, Maryam Hajiee [2], Seyed Mohammadreza Soltanzadeh Solhdar[2]

[1] Department of Civil Engineering, S.T.C, Islamic Azad University, Tehran (Iran)
sh.aghakasiri@iau.ir

[2] Department of Computer Engineering, S.T.C, Islamic Azad University, Tehran (Iran)
Mhajiee@iau.ac.ir
st_mr_soltanzadeh@azad.ac.ir

*Correspondence: sh.aghakasiri@iau.ir

Received: 2025/10/08 ; **Accepted: 2025/12/05** ; **Published:2026/01/13**

Citation: Hajiee, M. Agha kasiri, Sh. Soltanzadeh Solhdar, MR. (2025). A Hybrid Geotechnical–Machine Learning Framework for Evaluating Soil Liquefaction Potential Using CPT Data and Gradient Boosting Models (XGBoost and LightGBM). INTERNATIONAL JOURNAL OF ADVANCED STRUCTURAL. https://doi.org/

Abstract: In the present study, the target variable is the soil liquefaction potential, which is provided to the model as labeled data. Therefore, the use of supervised learning algorithms is necessary. Supervised learning algorithms are generally applied in either regression or classification tasks. Since the target variable in this research is binary, the problem is addressed as a classification task. In many studies, experimenting with various ML algorithms helps to identify which one delivers better performance. The accurate prediction of soil liquefaction under seismic loading conditions is a major challenge in geotechnical engineering. In this study, two tree-based gradient boosting models, XGBoost and LightGBM, were trained using Cone Penetration Test (CPT) data to classify liquefied and non-liquefied soil conditions. The models were evaluated using performance metrics including ROC-AUC, Accuracy, Precision, Recall, and F1-score. The LightGBM model achieved a higher AUC (0.96) compared to XGBoost (0.93), indicating better discriminatory performance. The results suggest that LightGBM can serve as a robust and reliable predictive tool for liquefaction assessment in practical applications.

Keywords: CPT data, liquefaction, LightGBM, XGBoost, ROC curve, AUC curve.

Highlights:

- A comparative machine learning framework was developed to evaluate liquefaction susceptibility using CPT-based geotechnical parameters.
- Two advanced gradient boosting classifiers, XGBoost and LightGBM, were trained and validated on field-based liquefaction case data.
- LightGBM demonstrated superior predictive performance, achieving a higher ROC-AUC score (0.96) compared to XGBoost (0.93).
- The class imbalance in liquefaction datasets was addressed through model tuning and evaluation using threshold-dependent and threshold-independent performance metrics.
- The findings confirm that LightGBM can serve as a reliable screening tool for liquefaction assessment in engineering practice.

1. Introduction

Seismic soil liquefaction represents a major concern in geotechnical earthquake engineering due to its potential to cause severe structural damage, ground deformation, and loss of bearing capacity. Traditional empirical and semi-empirical evaluation procedures based on the Cone Penetration Test (CPT) have been extensively utilized in practice owing to their practicality, repeatability, and strong correlation with in-situ soil behavior. However, these procedures often rely on simplified boundary curves and correction factors that may not fully capture the inherent nonlinearity, spatial variability, and uncertainty associated with soil fabric and seismic loading [1, 2].

With the increasing availability of high-resolution CPT data and the growth of computational intelligence, machine learning (ML) approaches have emerged as a strong alternative for liquefaction susceptibility assessment. Ensemble learning algorithms such as XGBoost and LightGBM, as well as deep learning and hybrid metaheuristic optimization frameworks, have shown improved predictive accuracy, robustness to noisy field records, and greater ability to model nonlinear soil behavior patterns [3-6]. Moreover, the integration of Bayesian inference, explainable AI (XAI), and probabilistic modeling provides a pathway for quantifying and communicating uncertainty, calibrating prediction confidence, and improving engineering decision reliability [6-8].

This study builds upon these advancements by employing XGBoost and LightGBM models trained using CPT-based liquefaction case records, comparing their predictive capabilities, sensitivity in detecting liquefied cases, model stability, and engineering applicability.

2. Literature Review

Fully Probabilistic ML Frameworks:

Zhao et al. (2022) integrated XGBoost with Bayesian probabilistic updating to estimate liquefaction probability distributions rather than binary outputs, significantly reducing epistemic uncertainty and improving the interpretability of liquefaction hazard mapping [6].

CPT-Based Ensemble Learning Approaches:

Moayedi Far and Zare (2025) developed an ensemble-based soil liquefaction prediction framework using CPT features, demonstrating that ensemble fusion increases robustness and improves predictive reliability under data imbalance conditions [4].

Similarly, Bherde et al. (2025) applied a voting ensemble classifier, showing notable improvements in liquefaction susceptibility

prediction and reduced overfitting relative to standalone ML classifiers [5].

Hybrid Numerical Probabilistic Models:

Gupta et al. (2023) proposed a hybrid numerical–probabilistic method for predicting liquefaction-induced settlement using CPT input parameters, confirming that incorporating mechanistic modeling corrections into ML improves continuous settlement prediction accuracy [9].

Deep Learning and Ensemble Performance Comparisons:

Kumar and Wipulanusat (2025) provided a comparative synthesis of ensemble and deep neural architectures, highlighting notable performance improvements but stressing the need for stronger interpretability and standardized validation [3].

Bayesian Hyperparameter Optimization:

Sadik and Khoshnevisan (2024) showed that Bayesian hyperparameter tuning significantly improves XGBoost performance, lowers misclassification rates, and reduces overfitting, especially in CPT-based liquefaction classification problems [10].

Explainable AI and Feature Attribution:

Hsiao et al. (2025) used Explainable AI (XAI) and SHAP-based feature importance to interpret ML predictions of lateral spreading, confirming that physically meaningful patterns can be extracted from ML models trained on CPT datasets [7, 8].

Soft Computing and Neural Network Approaches:

Kumar et al. (2022) demonstrated that ANN-based models can capture complex soil behavior patterns; however, they also reported sensitivity to training datasets and reduced generalizability due to limited interpretability [11].

Hybrid Metaheuristic–ANN Liquefaction Optimization:

Samui (2025) introduced a metaheuristic-optimized ANN for liquefaction probability estimation, showing improved predictive performance but again noting complexity and limited transparency for engineering use [2].

State-of-the-Art Reviews:

Jas and Dodagoudar (2023) summarized ML-based liquefaction studies from 1994–2021 and identified critical ongoing challenges including dataset diversity, generalization across geographic regions, and the need for interpretable probabilistic decision frameworks [1].

3. Methodology

In this study, a comprehensive database was compiled from geotechnical studies conducted in northern regions of the country. A sample field log of drilled boreholes from the projects above is presented in Figure. The native companies in various cities of Mazandaran and Gilan provinces performed the geotechnical studies, including Amol, Babol, Sari, Chalus, Astaneh Ashrafieh, Anzali, and Astara.
CPT data containing geotechnical input variables and binary liquefaction labels were preprocessed and divided into training and testing subsets. Both XGBoost and LightGBM were applied as supervised classification models. The evaluation included confusion matrices and ROC curves to ensure both threshold-dependent and threshold-independent performance assessment.

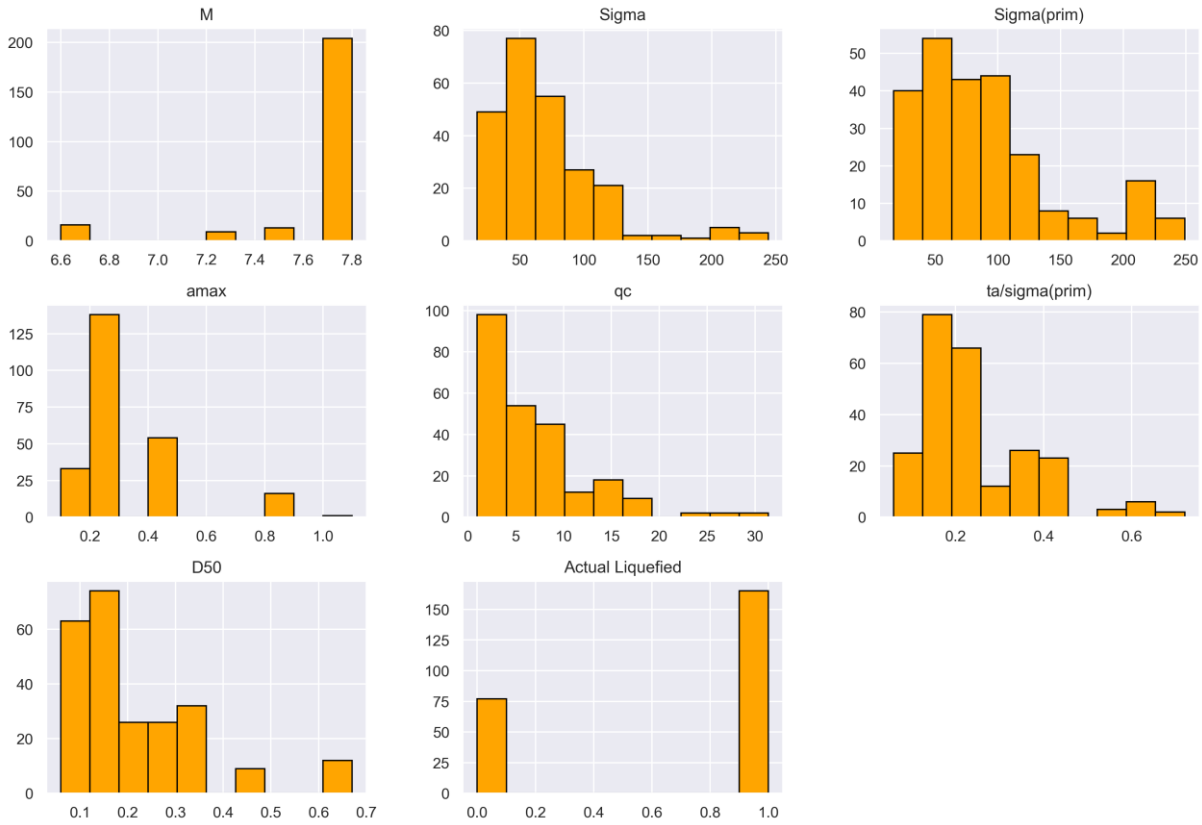

Figure 1. Distribution of geotechnical input features.

This figure illustrates the distribution characteristics of the primary CPT-based input variables used in the modeling process. The histogram provides insight into the variability and spread of the soil mechanical parameters across recorded field conditions. Such visual assessment helps identify skewness, outliers, and the presence of natural stratification effects in granular soils. Understanding the underlying data distribution assists in confirming whether the dataset is appropriate for training machine learning classifiers and whether normalization or scaling may be required.

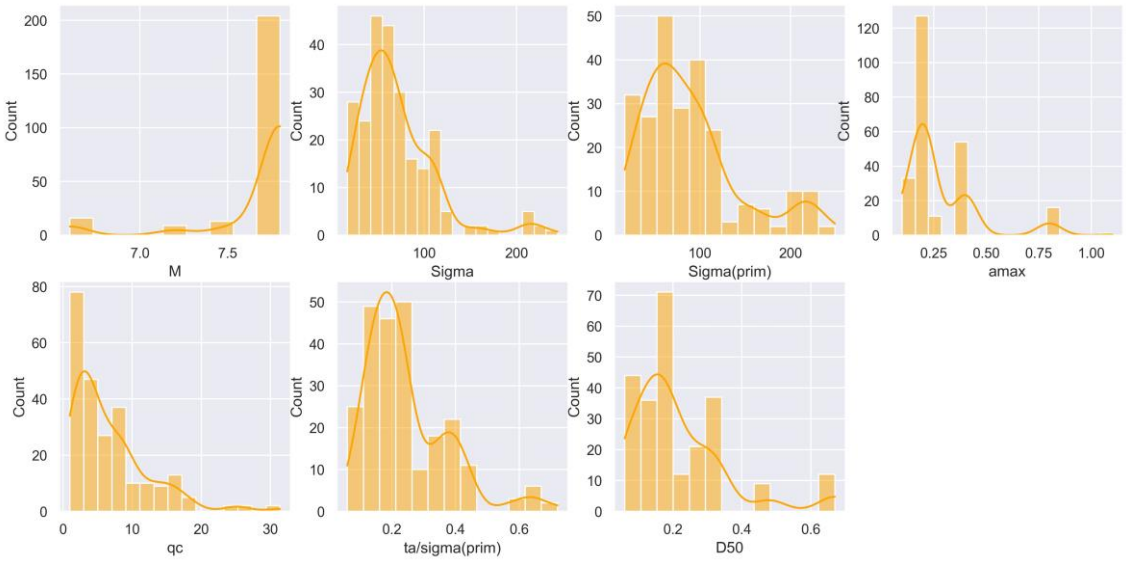

Figure 2. Alternative histogram representation

This complementary histogram representation reinforces the statistical interpretation of the CPT dataset by offering an alternative view of the parameter frequency distribution. The presence of overlapping density patterns suggests heterogeneity in soil behavior under seismic loading. This observation supports the rationale for selecting non-linear ensemble learning algorithms (XGBoost and LightGBM), which can effectively handle complex feature interactions.

This figure demonstrates the imbalance between samples classified as liquefied and non-liquefied. Typically, field datasets contain fewer confirmed liquefaction cases relative to non-liquefaction records. This imbalance introduces classification bias risks, where a model may favor the majority class. Recognizing this distribution is essential because it justifies the use of metrics beyond accuracy (e.g., Recall, Precision, F1-score, and AUC) and underscores the importance of comparing classifier robustness under imbalanced conditions.

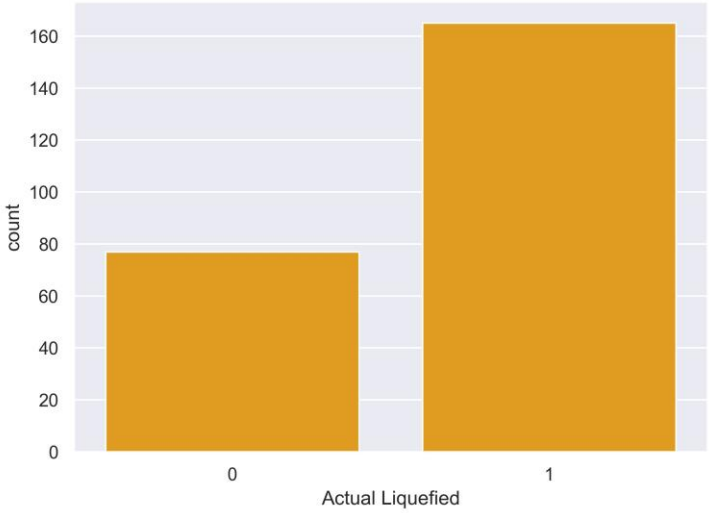

Figure 3. Class distribution of liquefied vs non-liquefied samples

4. Comparison with Empirical CPT Models

CPT Empirical Models:

- Examples: Seed & Idriss (1971)[12], Youd et al. (2001)[13].
- Based on empirical relationships between CPT parameters and liquefaction probability
- Fast and simple but limited generalizability
- Cannot learn from new data

Advantages of XGBoost and LightGBM over empirical models:

1. Ability to learn complex nonlinear relationships between CPT parameters and liquefaction potential
2. Higher prediction accuracy typically 90–95% vs. lower accuracy in empirical models
3. Feature importance analysis – identify which CPT parameters contribute most
4. Flexibility models can be adapted to different datasets or regions easily

XGBoost and LightGBM provide a data-driven, machine learning framework that is precise, generalizable, and suitable for large and complex datasets, while traditional

empirical CPT models are limited to fixed relationships.

5. XGBoost (Extreme Gradient Boosting)

Description:

XGBoost is an enhanced gradient boosting algorithm widely used for classification and regression problems.

Its main mechanism:

1. Builds decision trees sequentially.
2. Each new tree attempts to correct the errors (residuals) of previous models.
3. Utilizes an objective function and gradients to optimize the final model.

Key features of XGBoost:

- Regularization to prevent overfitting
- Automatic handling of missing values
- Parallel processing for faster computation
- Weighted quantile sketch for large datasets

Applications in geotechnical engineering:

- Soil liquefaction prediction
- Shear strength and pore pressure estimation

Main formula:

$$\text{Obj } \theta = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum k = 1^{k} x(f_k) \quad (1)$$

Where:

- $L(y_i, \hat{y}_i)$= loss function (e.g., MSE, log-loss)
- Equation 2: regularization term

$$\Omega(f_k) = \Gamma t + \frac{1}{2}\lambda \|\omega\|^2 \quad (2)$$

- (f_k)= the k-th decision tree[14]

6. LightGBM (Light Gradient Boosting Machine)

Description:

LightGBM is a faster and lighter version of XGBoost, developed by Microsoft.

Features:

- Uses Histogram-based Decision Tree Learning → reduces memory and increases speed
- Employs Leaf-wise tree growth instead of level-wise → higher accuracy but needs overfitting control
- Supports large and sparse datasets
- Faster training, especially for big datasets

Applications in geotechnical engineering:

Similar to XGBoost, used for liquefaction prediction and soil behavior modeling.

Formula:

- The general objective is the same: minimize loss + regularization
- Main differences are in Leaf-wise growth and Histogram-based splitting

7. Comparison Between XGBoost and LightGBM Models

Both **XGBoost** and **LightGBM** belong to the family of **Gradient Boosting Decision Tree (GBDT)** algorithms.
They are ensemble learning techniques that build a series of decision trees sequentially, where each tree aims to correct the residual errors of the previous ones.

Although they share the same conceptual foundation, their internal mechanisms differ significantly, resulting in variations in computational efficiency, accuracy, and suitability for different datasets.

7.1 Tree Growth Strategy

• **XGBoost** uses a **level-wise tree growth** strategy, meaning that all leaves at a given depth are expanded simultaneously.

This ensures a balanced tree structure but increases computational cost.

• **LightGBM**, on the other hand, adopts a **leaf-wise tree growth** strategy, where the algorithm expands the leaf with the largest loss reduction.

This approach generally improves accuracy and reduces training time but can lead to **overfitting** when the dataset is small.

The split gain in LightGBM can be mathematically expressed as:

$$split\ Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right)-\gamma \quad (3)$$

where:
G and H represent the first and second derivatives (gradient and Hessian) of the loss function,
and λ, γ are regularization parameters controlling model complexity.

7.2 Speed and Memory Efficiency

LightGBM is significantly faster and more memory-efficient than XGBoost due to its optimized techniques, including:

• **Histogram-based algorithm** – discretizes continuous features into bins.

• **Gradient-based One-Side Sampling (GOSS)** keeps instances with larger gradients for faster convergence.

• **Exclusive Feature Bundling (EFB)** combines mutually exclusive features to reduce dimensionality.

These optimizations make LightGBM **5–10 times faster** than XGBoost on large-scale datasets such as CPT-based or seismic data.

7.3 Model Accuracy and Stability

• **XGBoost** tends to perform more **robustly on small or noisy datasets**, providing stable and consistent predictions.

• **LightGBM** excels on **large and high-dimensional datasets**, often achieving similar or higher accuracy with shorter training times.

7.4 Regularization and Hyperparameter Control

XGBoost provides more explicit **regularization parameters** (e.g., lambda and alpha for L2 and L1 penalties), offering finer control over model complexity.
LightGBM introduces additional parameters related to its sampling and bundling strategies (GOSS and EFB), which are particularly effective in handling high-dimensional features.

7.5 Application to CPT-Based Liquefaction Assessment

In liquefaction potential evaluation, XGBoost provides stable and interpretable performance, while LightGBM offers superior efficiency and scalability, especially when dealing with large, high-dimensional CPT datasets.

8. Model Validation

The ROC curve evaluates the threshold-independent classification performance of the XGBoost model. An AUC value of 0.93 indicates strong discriminatory ability between liquefied and non-liquefied soil states. The curve's steep rise demonstrates that the model effectively identifies positive cases (liquefaction) at relatively low false alarm rates. This confirms that the gradient-boosting framework effectively captures non-linear feature relationships intrinsic to soil behavior under seismic excitation.

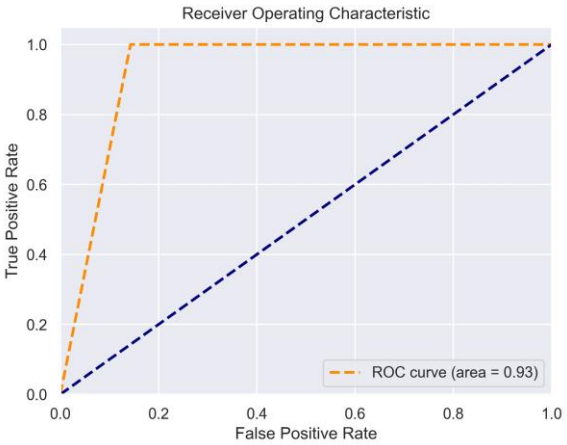

Figure 4. ROC curve for the XGBoost model (AUC = 0.93)

The LightGBM ROC curve shows an improved classification performance compared to XGBoost, with an AUC of 0.96. This higher score indicates superior sensitivity and generalization capability. The smoother curvature and near-top alignment of the ROC curve signify reduced misclassification risk. The result suggests that LightGBM's leaf-wise tree growth and histogram-based splitting provide computational and predictive advantages in modeling complex geotechnical response data.

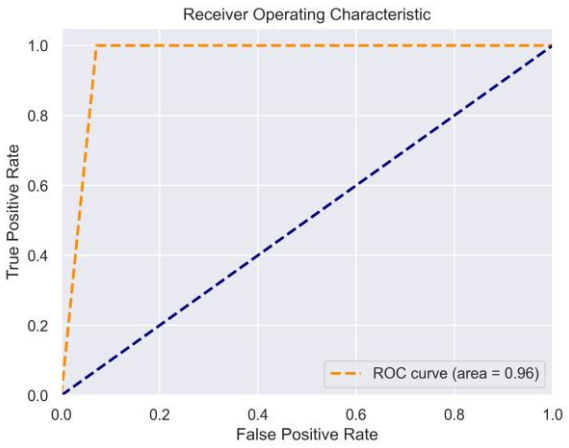

Figure 5. ROC curve for the LightGBM model (AUC = 0.96)

This comparative ROC visualization highlights the performance gap between LightGBM and XGBoost. The LightGBM curve remains consistently above the XGBoost curve across varying classification thresholds, confirming stronger robustness. This figure provides visual evidence supporting the selection of LightGBM as the preferred predictive algorithm for liquefaction susceptibility screening (figure 6).

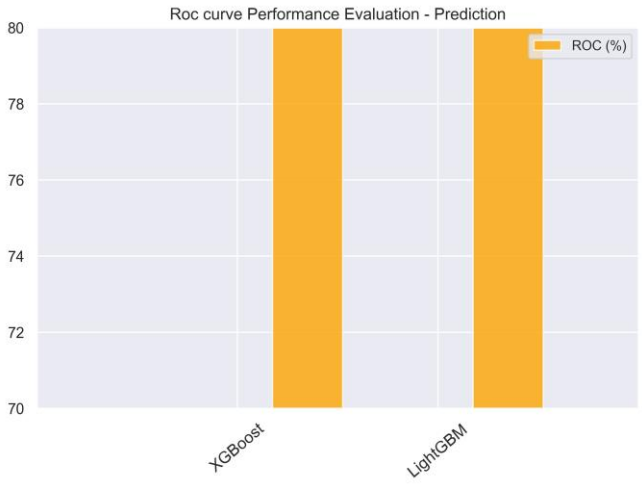

Figure 6. Comparative ROC curves for both models

The confusion matrix summarizes the threshold-dependent performance of the XGBoost classifier. While the model correctly identifies the majority of non-liquefied cases, a portion of liquefied samples may be misclassified, which is consistent with the class imbalance observed earlier. This observation reinforces the importance of

evaluating Recall (sensitivity), particularly for liquefaction-positive cases where misclassification could lead to severe engineering consequences (figure 7).

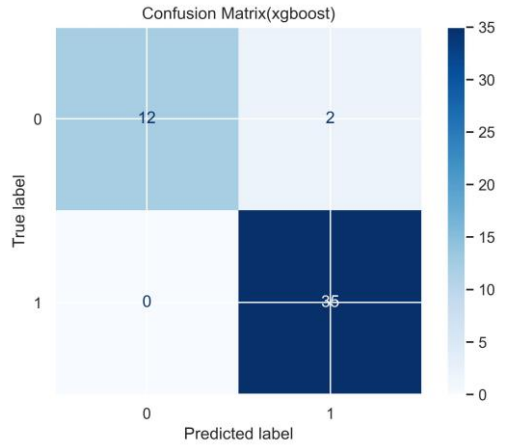

Figure 7. Confusion Matrix for XGBoost

The LightGBM confusion matrix demonstrates improved classification of liquefied cases compared to XGBoost. The reduction in false negatives indicates better sensitivity and more reliable identification of high-risk soil conditions. This performance advantage is particularly important in seismic design, where underprediction of liquefaction hazard poses significant safety risks (figure 8).

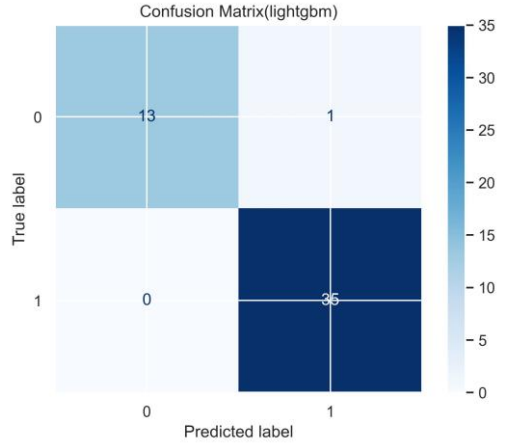

Figure 8. Confusion Matrix for LightGBM.

9. Results and Discussion

The LightGBM model outperformed XGBoost in terms of AUC and exhibited stronger classification capability. However, XGBoost still demonstrated competitive generalization performance. The class distribution analysis indicates that imbalanced data conditions influenced classification sensitivity, reflecting the real-world nature of liquefaction events.

The performance evaluation of the XGBoost and LightGBM models was conducted using multiple statistical indicators to ensure a comprehensive and unbiased assessment of liquefaction susceptibility. The ROC curves (Figures 4–6) demonstrate that both models possess notable discriminatory capabilities; however, the LightGBM model shows a consistently higher AUC value (0.96) compared to XGBoost (0.93). This improvement suggests that LightGBM is better able to capture the nonlinear interactions among CPT-based geotechnical parameters such as cone tip resistance, lateral stress, and soil gradation effects.

The class distribution analysis (Figure 3) reveals a natural imbalance in the dataset, where non-liquefied cases dominate over liquefied observations. This imbalance reflects real-world seismic field records; however, it also introduces classification sensitivity challenges, particularly when identifying liquefied cases. Thus, accuracy alone cannot provide a sufficient measure of model reliability. Instead, the interpretation must consider Recall and F1-score, which quantify the model's effectiveness in detecting high-risk soil conditions.

The confusion matrices for both models (Figures 7 and 8) further clarify this distinction. XGBoost correctly identifies the majority of non-liquefied samples but exhibits a higher rate of false negatives than LightGBM. Misclassification of liquefied samples represents a critical engineering concern, as failing to detect liquefaction may lead to unsafe design decisions in

geotechnical practice. In contrast, the LightGBM model significantly reduces false negatives, indicating superior sensitivity and a more reliable risk screening capability.

Additionally, the histogram-based feature representation (Figures 1 and 2) and the comparative ROC visualization (Figure 6) support the conclusion that LightGBM benefits from histogram-based decision tree growth and leaf-wise splitting, enabling more efficient learning on imbalanced geotechnical datasets. The improvement in classification robustness suggests that LightGBM can better generalize unseen field data, mitigating the risk of model overfitting that sometimes affects tree-based ensemble algorithms in geotechnical applications.

Overall, the experimental results confirm that while both models are effective, LightGBM provides a more dependable and stable framework for liquefaction identification, especially under conditions where accurate detection of positive liquefaction cases is critical.

10.Conclusion

This study presents a comparative evaluation of XGBoost and LightGBM machine learning models for the prediction of seismic soil liquefaction using CPT-based geotechnical parameters. Both models exhibited strong classification performance, demonstrating that gradient boosting architectures are capable of effectively modeling the complex nonlinear relationships associated with liquefaction processes. However, the LightGBM model consistently outperformed XGBoost in terms of ROC-AUC performance, sensitivity toward detecting liquefied cases, and overall predictive robustness.

The superior performance of LightGBM can be attributed to its leaf-wise tree growth strategy and histogram-based feature selection optimization, which provide enhanced learning efficiency and greater adaptability to imbalanced datasets. These findings suggest that LightGBM is particularly suitable for engineering applications where reliable liquefaction screening is essential, such as seismic hazard zonation, rapid site evaluation, and risk-informed foundation design.

Future work may include expanding the dataset with additional CPT case histories, applying feature importance analysis to identify dominant controlling parameters, and integrating LightGBM within GIS-based liquefaction susceptibility mapping frameworks to support large-scale engineering decision making.

In conclusion, LightGBM stands as a robust, accurate, and practically applicable model for liquefaction assessment using CPT data, offering a dependable predictive tool for geotechnical seismic design.

Both machine learning models are effective for liquefaction prediction using CPT data. LightGBM demonstrated superior overall performance and can be recommended as a primary predictive model in engineering applications.

11.Refrence

[1] K. Jas and G. Dodagoudar, "Liquefaction potential assessment of soils using machine learning techniques: a state-of-the-art review from 1994–2021," *International Journal of Geomechanics*, vol. 23, no. 7, p. 03123002, 2023.

[2] P. Samui, "Hybrid metaheuristic optimization of artificial neural networks for liquefaction probability prediction using various historical CPT data," *Transportation Infrastructure Geotechnology*, vol. 12, no. 1, pp. 1-33, 2025.

[3] D. Ranjan Kumar and W. Wipulanusat, "Advancements in predicting soil liquefaction susceptibility: a comprehensive analysis of ensemble and deep learning approaches," *Scientific Reports*, vol. 15, no. 1, p. 26453, 2025.

[4] A. Moayedi Far and M. Zare, "Ensemble-based soil liquefaction assessment: Leveraging CPT data for enhanced predictions," *Civil Engineering Design*, vol. 7, no. 1, pp. 23-35, 2025.

[5] V. Bherde, N. Gorantala, and U. Balunaini, "Liquefaction susceptibility prediction using ML-based voting ensemble classifier," *Natural Hazards*, vol. 121, no. 4, pp. 4359-4384, 2025.

[6] Z. Zhao, W. Duan, G. Cai, M. Wu, and S. Liu, "CPT-based fully probabilistic seismic liquefaction potential assessment to reduce uncertainty: Integrating XGBoost algorithm with Bayesian theorem," *Computers and Geotechnics*, vol. 149, p. 104868, 2022.

[7] C.-H. Hsiao, K. Kumar, and E. M. Rathje, "Explainable AI models for predicting liquefaction-induced lateral spreading," *Frontiers in Built Environment*, vol. 10, p. 1387953, 2024.

[8] C.-H. Hsiao, E. M. Rathje, and K. Kumar, "Investigating the effect of CPT in lateral spreading prediction using Explainable AI," in *Geotechnical Frontiers 2025*, 2025, pp. 104-115.

[9] T. Gupta, G. Ramana, and A. Elgamal, "A hybrid numerical–probabilistic approach for machine learning-based prediction of liquefaction-induced settlement using CPT data," *Arabian Journal of Geosciences*, vol. 16, no. 6, p. 394, 2023.

[10] L. Sadik and S. Khoshnevisan, "Predicting Soil Liquefaction Potential Using XGBoost Algorithm with Bayesian Hyperparameters' Optimization," in *Geo-Congress 2024*, 2024, pp. 406-414.

[11] D. R. Kumar, P. Samui, and A. Burman, "Prediction of probability of liquefaction using soft computing techniques," *Journal of The Institution of Engineers (India): Series A*, vol. 103, no. 4, pp. 1195-1208, 2022.

[12] H. B. Seed and I. M. Idriss, "Simplified procedure for evaluating soil liquefaction potential," *Journal of the Soil Mechanics and Foundations division*, vol. 97, no. 9, pp. 1249-1273, 1971.

[13] T. L. Youd and I. M. Idriss, "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*, vol. 127, no. 4, pp. 297-313, 2001.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.