Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 1, January 2024
https://doi.org/10.47709/cnahpc.v6i1.3374

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

Case Study: Gradient Boosting Machine vs Light GBM in Potential Landslide Detection

Djarot Hindarto[1)]*
Prodi Informatika, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional Jakarta
djarot.hindarto@civitas.unas.ac.id

ABSTRACT

An increasing demand for precise forecasts concerning the likelihood of landslides served as the impetus for this investigation. Human life, infrastructure, and the environment are all profoundly affected by this natural occasion. Constructing models capable of discerning intricate patterns among diverse factors that impact the likelihood of landslide occurrences constitutes the primary obstacle in landslide detection. Predicting potential landslides requires algorithms that are both accurate and efficient in their processing of vast quantities of data encompassing a variety of geographical, environmental, and ecological characteristics. An evaluation of the efficacy of both Gradient Boosting Machine and Light Gradient Boosting Machine in identifying patterns associated with landslides is accomplished by comparing their performance on a large and complex dataset. In the realm of potential landslide detection, the primary aim of this research endeavor is to assess the predictive precision, computation duration, and generalizability of Gradient Boosting Machine and Light Gradient Boosting Machine. This research aims to enhance comprehension regarding the comparative benefits of these two approaches in surmounting the obstacles associated with risk assessment and modeling pertaining to potential landslides, with a specific emphasis on efficiency and precision. The research findings are anticipated to serve as a valuable reference in the identification of more efficient approaches to reduce the likelihood of landslide-induced natural catastrophes. The accuracy of the GBM experiment reached 82% and LGBM reached 81%.

Keywords: Gradient Boosting Machine; Light Gradient Boosting Machine; Landslides Detection; Geographical;

INTRODUCTION

Landslides all gravely endanger human life, infrastructure, and the environment. These calamities can occur abruptly and result in catastrophic consequences, such as property damage, loss of life, and disruption to the social and economic activities of individuals in different regions of the globe. The risk of landslides is escalating, particularly in areas with delicate topography, environmental fluctuations, severe weather conditions, and significant human intervention. Prompt identification of potential landslides is essential to minimize their impact. The identification of potential landslides is hindered by the intricate interplay of geographic, geological, and environmental factors, which presents difficulties in creating precise and effective predictive models. Hence, the primary objective is to enhance the precision and efficiency of forecasting potential landslides by employing technology and machine learning techniques in the analysis of geospatial data. Within this framework, it is pertinent to assess and comprehend the merits and constraints of different methods, such as Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (Oluwatosin et al., 2023), to effectively detect potential landslides. Gradient boosting machines are boosting algorithms that are versatile enough to be used for both regression and classification; they rely on regression trees as their base learners. A classifier was iteratively trained using the base learners and the negative gradient of a differentiable loss function for a classification problem (Sunaryono et al., 2022). The training data was used to expect the classifier to minimize the loss function (Friedman, 2001). By acquiring a comprehensive comprehension of these methodologies, the aim is to develop more efficient and adaptable approaches to reducing risks and gaining a better understanding of the peril posed by landslides in pursuit of guaranteeing the enduring sustainability of both the environment and society.

The challenge in detecting potential landslides lies in the intricacy of identifying patterns that have the potential

* Corresponding author

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

to initiate the catastrophe. A range of environmental, geographic, and geological factors influences the probability of landslides. However, analysing data related to these factors is a complex task. Efficient algorithms that can handle vast amounts of data containing intricate characteristics are crucial, particularly for accurately forecasting potential landslides. Furthermore, the computational time needed for this process is an essential factor, as landslide detection necessitates a swift response to implement preventive or mitigation measures promptly. Hence, the primary objective is to discover effective and precise methods for analysing intricate data to forecast possible landslides.

The objective of this study is to assess and contrast the efficacy of Light Gradient Boosting Machine (Oram et al., 2021) and Gradient Boosting Machine (Yin et al., 2024) as machine learning approaches for the purpose of identifying potential landslides. The impending necessity for precise forecasts concerning the natural occurrence of landslides, which severely affect human life, infrastructure, and the environment, provides the impetus for this research. Confronting the intricate patterns that govern the occurrence of landslides, which are influenced by a multitude of environmental, geographic, and geological factors, represents the primary obstacle. To tackle this challenge, the main goal of this research is to develop a computational algorithm that is both accurate and efficient in handling large datasets with complex characteristics. In the context of potential landslide detection, the primary objective of this study is to assess the prediction accuracy, computation time, and generalization capabilities of both Light Gradient Boosting Machine and Gradient Boosting Machine (Deng et al., 2024) by comparing their performance on a large and diverse dataset.

The processing of exceedingly complex data primarily involves the application of machine learning (Hindarto & Djajadi, 2023) techniques to overcome the obstacles encountered in landslide detection. Gradient Boosting Machine (Thongthammachart et al., 2022), (Hindarto & Santoso, 2022) and Light Gradient Boosting Machine, two techniques selected for comparison, are iterations of ensemble learning algorithms that have demonstrated efficacy across a range of applications. Gradient Boosting Machine, despite its prowess in generating precise predictions, frequently encounters computational time challenges attributable to its sequential boosting methodology. As an alternative, Light Gradient Boosting Machine provides enhanced velocity through the utilization of a histogram-based method, which speeds up model construction through the parallel division of nodes. Hence, conducting a comparative analysis of Gradient Boosting Machine and Light Gradient Boosting Machine (Mishra et al., 2023) regarding the detection of potential landslides will yield more profound insights into the comparative merits of each methodology, particularly when confronted with the intricacies of data encompassing diverse environmental variables.

Research inquiries pertaining to the identification of potential landslides are as follows:
What is the performance comparison between Gradient Boosting Machine and Light Gradient Boosting Machine predictive models in forecasting the likelihood of landslides using different environmental, geographic, and geological characteristics? Research Question 1. Can the enhanced computational speed provided by Light Gradient Boosting Machine compensate for or improve the accuracy of predictions concerning landslide susceptibility compared to Gradient Boosting Machine when dealing with extensive and intricate datasets? Research Question 2.

LITERATURE REVIEW

This study presents a transparent machine-learning model for forecasting carbon price trends. It utilizes five techniques: CEEMDAN, TFS, LightGBM, BOA, and SHAP. The model demonstrates superior performance compared to other benchmark models in making predictions for multiple steps. The most crucial features are the high-frequency intrinsic mode function components. This model is a highly efficient tool for accurately predicting carbon prices (Deng et al., 2024). In this study, we compared the effects of three different machine learning algorithms on the classification of sleep stages in healthy participants and those with major depressive disorder: Support Vector Machines (Hindarto & Santoso, 2022), Random Forest, and LightGBM. The results indicated that LightGBM outperformed SVM and Random Forest in terms of accuracy. The inclusion of demographic features and HDRS score was found to be essential for distinguishing between DD and DD classifications (Tai et al., 2024). Machine learning models use subjective and random grading conditioning factors, according to a Wenzhou, China, study on landslide susceptibility modeling. The study examined non-grading, equal intervals, and natural breaks. The Support Vector Machine (SVM) model worked best with 8-level grading and natural breaks, while decision trees and ensemble models worked better without grading. GRU and DNN models benefit from equidistant grading over 12 levels, while LSTM models thrive with over 16 (Zeng et al., 2024) (Guo et al., 2023). Decisionmakers need landslide susceptibility maps to plan for future disasters. Landslide prediction model uncertainty was quantified using eight machine-learning techniques in a new framework. Four ensemble models were tested, and the weighted mean of probability was the best. A confident map shows that

* Corresponding author

74% of past landslides have higher susceptibility and low uncertainty. MLT-based micro-level zonation may improve landslide susceptibility maps and identify future landslide-prone areas (Achu et al., 2023). The study evaluated machine learning models like ANN, GBM, RF, and SVM in rainfall-induced landslide susceptibility mapping in Turkey's highest rainfall areas. The study used a landslide inventory with 533 polygons, with 70% for training and 30% for validation. The models' prediction rates were 93.8%, 94.8%, 96.1%, and 97%, with GBM outperforming other models (Achu et al., 2023). This paper proposes Bayesian optimization to improve landslide susceptibility assessment machine learning models. The study uses China's Anhua County, where many landslides have occurred. Random forest, SVM, and gradient boost decision tree are used. Considering training/test set ratios, the Bayesian optimization algorithm finds the optimal P/N sample ratio. SVM (Hindarto, 2022), RF, and GBDT performed better, but RF and GBDT are better for sample imbalance (Yang et al., 2023).

Research on carbon price modeling, sleep stage classification, and landslide susceptibility gives us much information about how machine learning can be used. There are, however, several gaps that can be seen. Cross-domain integration in the use of machine learning techniques still needs to be fully realized, which limits the model exploration that can be done on different problems. When people focus on how well a model works, they forget to think about how easy it is to understand. This can affect how much people trust and use the model, especially when it comes to predicting natural disasters. Getting rid of sample imbalance is another issue, and different studies give different advice on how to do this. In landslide susceptibility modeling or carbon price forecasting, the best way to use demographic variables that have been shown to be essential for sleep classification has yet to be found. Getting these holes filled can make machine learning applications stronger by making them easier to understand, allowing integration across domains, fixing uneven sample management, and using more data to make more accurate predictions.

METHOD

The research approach encompasses the methods and strategies implemented during the stages of data collection, analysis, and model development. By conducting a comprehensive literature review and preparing the dataset, this study establishes a solid groundwork. The implementation of training with the Gradient Boosting Machine and Light Gradient Boosting Machine yields a predictive model that is reasonably accurate. The validity of the model is confirmed through an assessment of its performance on distinct datasets. Methodical procedures are implemented to generate dependable and pertinent outcomes consistent with the declared research goals.

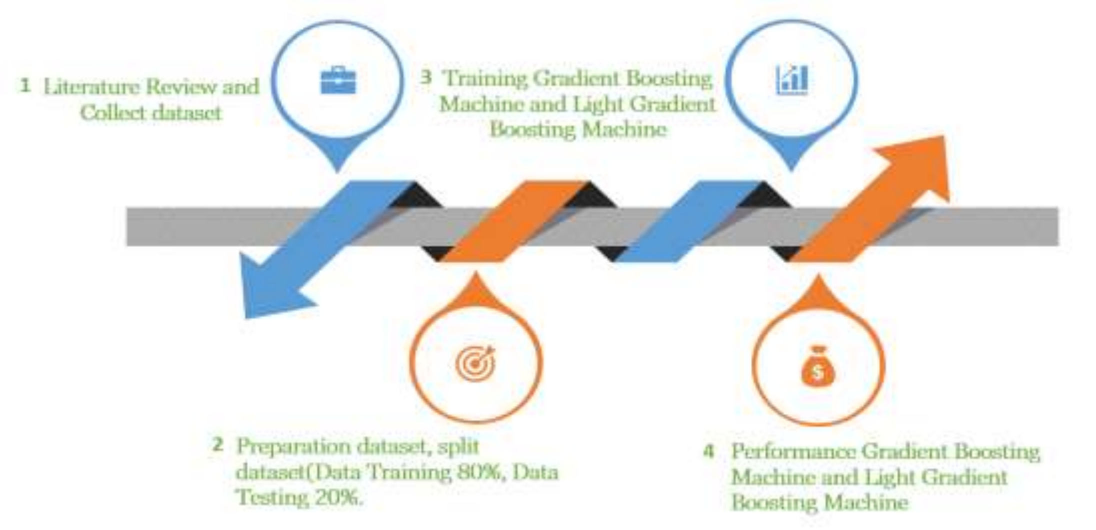

1 Literature Review and Collect dataset
3 Training Gradient Boosting Machine and Light Gradient Boosting Machine
2 Preparation dataset, split dataset(Data Training 80%, Data Testing 20%.
4 Performance Gradient Boosting Machine and Light Gradient Boosting Machine

Figure 1. Proposed research methods
Source: Researcher Property

Figure 1 highlights the crucial significance of the initial stage in research, which serves as a fundamental basis for subsequent progress. The initial stage of this methodological research centers on conducting a comprehensive review

* Corresponding author

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

of existing literature and gathering relevant datasets. At this stage, the initial step is to perform a comprehensive examination of the literature pertaining to the chosen research subject. This process facilitates comprehension of the context, theories, prior research, and methodologies employed in the relevant field. In addition to that, this stage also encompasses the acquisition of datasets that will serve as the foundation for subsequent analysis. Following the completion of the preliminary stage, the subsequent step is stage 2, specifically the preparation of the dataset. In this step, the gathered dataset will be processed to facilitate subsequent analysis.

One crucial aspect of this stage involves partitioning the dataset into two primary components: Training Data (80%) and Testing Data (20%). The division is crucial to ensure that the developed model can be adequately tested on data that was not utilized during training, thus reducing the likelihood of overfitting, and promoting improved generalization. Stage 3 is dedicated to training the model using two distinct methods: Gradient Boosting Machine and Light Gradient Boosting Machine. The training process of this model entails utilizing a pre-existing dataset that has been prepared in advance. Light Gradient Boosting Machines and Gradient Boosting Machines, in order to construct trustworthy prediction models, are two widely utilized machine learning techniques used in the construction of dependable prediction models.

The objective of this training process is to optimize the model to discern patterns within the data and generate predictions with a high degree of accuracy. Stage 4 involves conducting a performance evaluation of the model that was trained earlier. This evaluation is conducted using pre-separated Testing Data. The primary objective is to assess the model's capacity to generate precise and dependable predictions using previously unseen data. The evaluation phase is vital to guarantee the dependability of the model prior to its implementation in real-life scenarios. This methodology comprises a systematic sequence of steps aimed at ensuring the meticulous execution of research, starting from the initial comprehension of existing literature and culminating in the final assessment of the developed model. The objective of this process is to generate predictive models that are dependable and valuable in applicable scenarios.

Gradient Boosting Machine

Landslide detection is a critical component of mitigation strategies for natural disasters, given their destructive potential. In this instance, machine learning techniques, specifically the Gradient Boosting Machine, were implemented. Gradient Boosting Machine is an ensemble technique that iteratively enhances a model through the identification and rectification of prior deficiencies. When considering landslide detection, the application of Gradient Boosting Machine can be advantageous in the examination of numerous determinants that impact the propensity for landslides to transpire. The initial step in this detection procedure is the collection of a dataset containing innumerable variables, including geological conditions, precipitation, soil type, elevation, and the occurrence of landslides in the past. A comprehensive literature review is necessary at this stage to ensure that the variables used for model construction are pertinent and representative. Following the collection of the dataset, data preparation is performed by separating the data into training and testing sets.

Using a dataset that was previously prepared, the Gradient Boosting Machine model is trained during the training phase. This procedure facilitates the model's acquisition of intricate relationships between established variables and the propensity for landslides to transpire. The Gradient Boosting Machine improves its predictive performance iteratively through the identification of errors in previous predictions and the modification of subsequent predictions. This enables the model to acquire knowledge from the data in an iterative fashion, thereby enhancing the precision of its forecasts throughout the procedure. Validating the performance of the trained model is crucial. In the evaluation phase, a distinct dataset is utilized, which was not used during the training phase. By evaluating the model's performance on this dataset, the accuracy with which it predicted landslide tendencies was determined. It is critical to conduct this evaluation to ascertain that the model possesses the capability to not only discern patterns within the training data but also generate accurate predictions on novel data. The utilization of Gradient Boosting Machine for the detection of landslides exemplifies the potential of machine learning technology to aid in the mitigation of risks associated with natural disasters. Through the implementation of this methodology, the expectation is that a predictive framework can be constructed to aid in the identification of landslide-prone regions, thereby facilitating more precise and efficient initiatives for prevention and mitigation. The Gradient Boosting Machine can be seen in Figure 2.

* Corresponding author

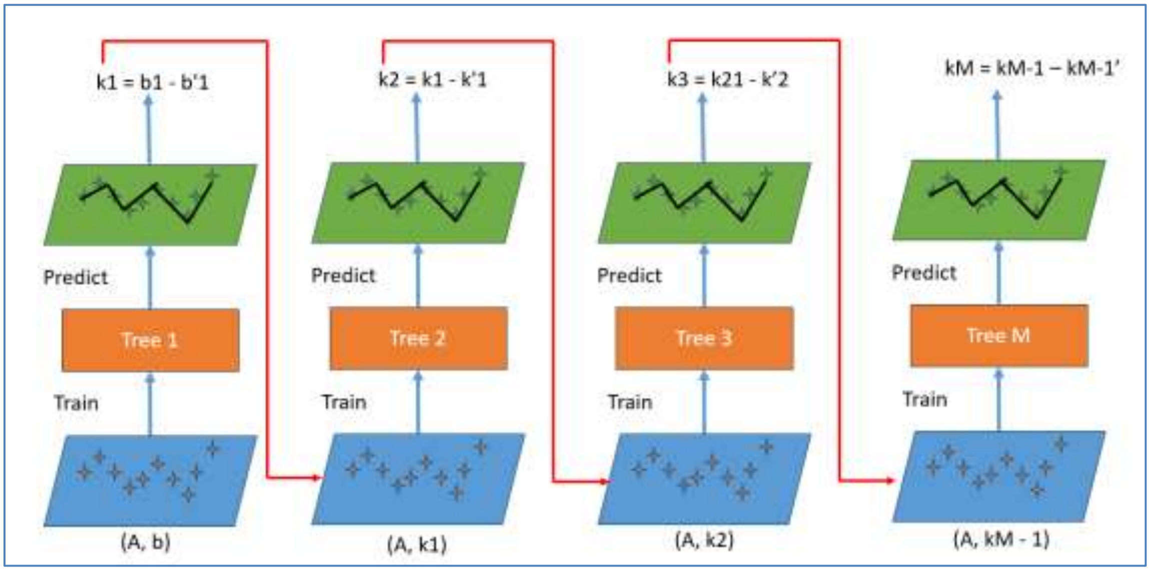


Figure 2. Gradient Boosting Machine
Source: Google Image

Light Gradient Boosting Machine

Landslide detection is possible with the aid of the practical and potent machine learning algorithm Light Gradient Boosting Machine (LightGBM). LightGBM is an excellent option for analyzing variables that affect the probability of landslides due to its streamlined data processing capabilities regarding large data volumes. Collecting a dataset comprising diverse variables—including geological attributes, climate data, soil composition, topographical features, and historical records of landslides—is the initial step in implementing LightGBM for landslide detection. To ensure that pertinent and representative variables are incorporated, this procedure necessitates the participation of domain specialists. Data preparation follows the collection of the dataset and consists of separating the data into sections for evaluation and training purposes.

During the training phase, the pre-prepared dataset will be processed by the LightGBM model. This algorithm operates similarly to a Gradient Boosting Machine in that it refines the model iteratively by concentrating on its previous flaws. In contrast, LightGBM exhibits notable benefits in terms of memory utilization efficiency and speed, which enables it to process sizable datasets with greater efficiency. Consequently, this renders it a favorable option for intricate analyses, including landslide detection. The performance of the models generated by LightGBM must be validated. A distinct testing dataset is utilized during the evaluation phase; this dataset was not utilized during training. This dataset will be used to evaluate the model's predictive capability regarding the probability of landslides. It is essential that the model not only comprehends patterns in the training data but also generates accurate predictions on never-before-seen data, as this evaluation verifies. In terms of comprehending the elements that contribute to landslides, the application of LightGBM in landslide detection exhibits tremendous promise. LightGBM can assist in the identification of landslide-prone regions by processing data quickly and effectively, thereby supplying critical information for risk reduction initiatives and the prevention of natural catastrophes.

RESULT

Dataset

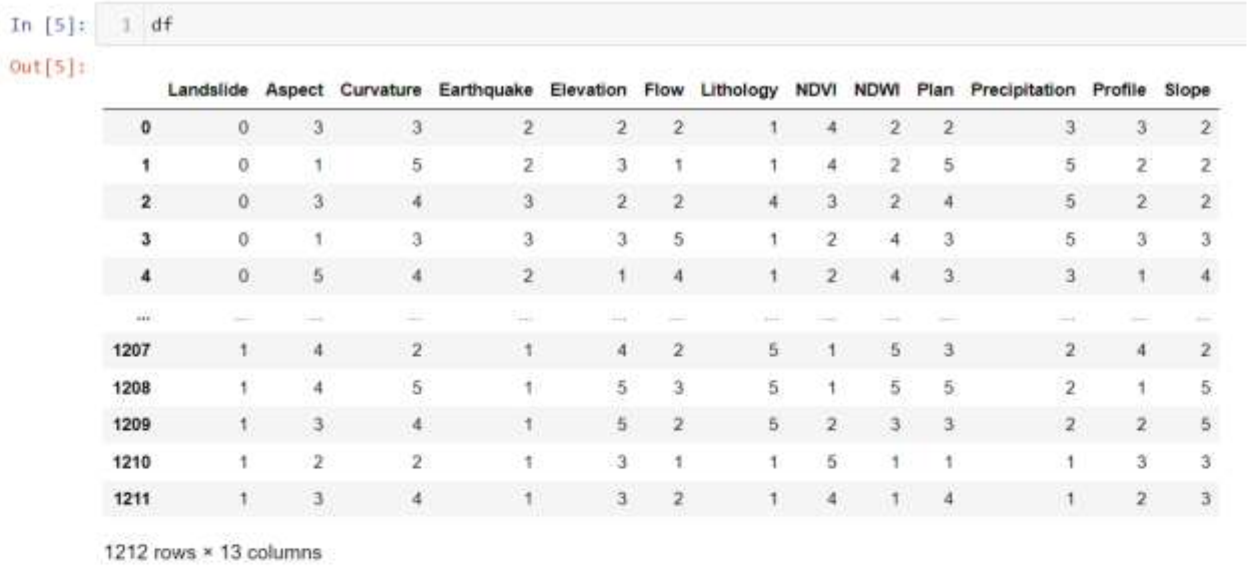
The dataset sourced from Kaggle comprises data on landslide disasters and associated factors. Each data point consists of 13 observed features, namely Landslide, Aspect (land aspect), Curvature (land stiffness), Earthquake, Elevation (height), Flow, Lithology, NDVI (Normalized Difference Vegetation Index), NDWI (Normalized Difference Water Index), Plan, Precipitation (rainfall), Profile, and Slope. This dataset comprises 1212 data points and offers an opportunity to examine the correlation between different geographic, environmental, and natural event factors and the likelihood of landslides. Every characteristic in this dataset plays a crucial role in analyzing and forecasting landslide

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 1, January 2024
https://doi.org/10.47709/cnahpc.v6i1.3374

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

propensities. For instance, elevation plays a pivotal role as more inclined terrains have an increased susceptibility to landslides.

Curvature, or the stiffness of the land, can also have a significant impact as it can indicate the likelihood of landslides in areas with extreme topographic features. Human factors, such as land use and rock type, can have a significant impact on slope stability, providing valuable insights. Furthermore, investigating the correlation between environmental factors, such as the Normalized Difference Water Index and the Normalized Difference Vegetation Index, and the likelihood of landslides could be a captivating research field. NDVI quantifies the density of vegetation, whereas NDWI quantifies the moisture content of soil. Both indices can offer insights into land conditions that are prone to landslides. Natural phenomena such as earthquakes and rainfall are significant factors to consider, as they have the potential to initiate or amplify the likelihood of landslides. Conducting a meticulous analysis of a dataset that encompasses a wide range of information will help you understand the patterns and factors that affect the likelihood of landslides better. Utilizing an appropriate data analysis methodology can aid in the prevention and mitigation of disasters by enabling the identification of susceptible regions and the formulation of more efficient strategies to minimize the likelihood of future landslides. The dataset is visible in Figure 3.



```
In [5]:  df
Out[5]:
```

	Landslide	Aspect	Curvature	Earthquake	Elevation	Flow	Lithology	NDVI	NDWI	Plan	Precipitation	Profile	Slope
0	0	3	3	2	2	2	1	4	2	2	3	3	2
1	0	1	5	2	3	1	1	4	2	5	5	2	2
2	0	3	4	3	2	2	4	3	2	4	5	2	2
3	0	1	3	3	3	5	1	2	4	3	5	3	3
4	0	5	4	2	1	4	1	2	4	3	3	1	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1207	1	4	2	1	4	2	5	1	5	3	2	4	2
1208	1	4	5	1	5	3	5	1	5	5	2	1	5
1209	1	3	4	1	5	2	5	2	3	3	2	2	5
1210	1	2	2	1	3	1	1	5	1	1	1	3	3
1211	1	3	4	1	3	2	1	4	1	4	1	2	3

1212 rows × 13 columns

Figure 3. Dataset
Source: Kaggle

Figure 4, a curve representing the performance of a classification model at various thresholds is referred to as the ROC (Receiver Operating Characteristic). When judging how well a classification model works, ROC is a crucial metric that defines the relationship between False Positive Rate (FPR) and True Positive Rate (TPR). As ensemble algorithms, they are frequently employed in predictive modeling and are highly effective in classification tasks when applied to Gradient Boosting Machines and Light Gradient Boosting Machines. By combining many weak models (weak learners) into one robust model, GBM is an ensemble algorithm that constructs predictive models. The assessment of the model's ability to distinguish between positive and negative classes is illustrated by the Area Under the Curve (AUC) value for GBM's ROC. Baseline GBM has an AUC of 0.80 in the given example; Model 1 exhibits a marginal enhancement to 0.81, suggesting a marginal improvement in the ability of the model to distinguish between classes.

* Corresponding author

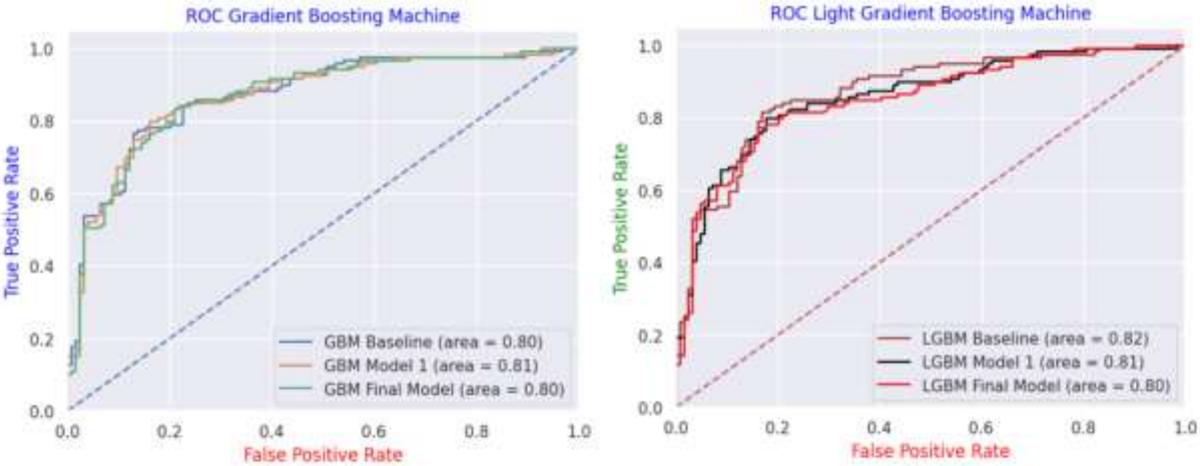


Figure 4 Receiver Operating Characteristic of GBM and LGBM
Source: Researcher Property

In contrast, LGBM is a refinement of GBM designed to enhance performance, speed, and efficiency. AUC is another metric used to assess ROC for LGBM. The illustration demonstrates that Baseline LGBM possesses an area under the curve (AUC) of 0.82. While Model 1 exhibits a marginally lower AUC value of 0.81, its efficacy on test data is enhanced to 0.830. This result indicates that the LGBM model improved its accuracy on the test data despite a slight decrease in the AUC. Understanding the model's ability to distinguish between positive and negative classes is the significance of ROC in both types of algorithms. In addition to accuracy, precision, recall, and F1-score, which offer a more comprehensive understanding of model performance, it is critical to incorporate additional metrics such as AUC or ROC values, which provide a broad indication of model performance. It is essential to acknowledge that while there may be minor variations in AUC values among models, these fluctuations may not necessarily indicate substantial disparities in their operational capability. On occasion, a more comprehensive understanding of the benefits between the developed GBM and LGBM models requires an additional emphasis on accuracy and other relevant metrics.

Table 1. Gradient Boosting Machine

	precision	recall	f1-score	support
0	0,83	0,84	0,82	124
1	0,82	0,78	0,82	119
Accuracy			0,82	243
Macro avg	0,82	0,81	0,82	243
Weighted avg	0,82	0,82	0,82	243

Table 1, the results of assessing the performance of the Gradient Boosting Machine (GBM) model in a classification task are presented in Table 1. The evaluation process incorporates various significant metrics, including precision, recall, f1-score, and accuracy values, to assess the model's overall performance. One measure of a model's accuracy is its precision, which is defined as the percentage of true positives out of all possible positives. In the given context, class 0 exhibits a precision value of 0.81, indicating that the model accurately predicts approximately 81% of instances. In contrast, the precision of class 1 is 0.82, which signifies that approximately 82% of predictions made by class 1 are valid. The recall metric evaluates the model's ability to recognize positive instances. With a recall of 0.84, the model is capable of accurately identifying approximately 84% of the actual cases of class 0. Class 1, on the other hand, exhibits a recall value of 0.78, signifying that the model can discern a mere 78% of the total instances belonging to class 1. The harmonic average value between recall and precision is the F1-score. The f1-score for class 0 is 0.82, which is the harmonic of the precision and recall for class 0. In the interim, the f1-score for class 1 is 0.82, suggesting that class 1 recall and precision are in equilibrium. The GBM model achieves an overall accuracy of 0.81, denoting the proportion of accurate predictions made across the complete evaluation dataset. The overall assessment outcomes demonstrate consistency in the performance of the model, as indicated by the weighted average (weighted average) and the mean value (macro average) of the precision, recall, and f1-score metrics. With an accuracy of 81%, it is possible to conclude that the GBM model exhibits satisfactory overall performance based on the findings of this evaluation. Despite a marginal disparity in precision and recall metrics between classes 0 and 1, the model demonstrates adequate classification performance for both courses, as indicated by the balanced f1-score value. To prevent the omission of actual positive examples in class 1, it may be necessary to place particular emphasis on improving recall for class 1.

Table 2. Light Gradient Boosting Machine

	precision	recall	f1-score	support
0	0,81	0,81	0,81	124
1	0,80	0,81	0,80	119
Accuracy			0,81	243
Macro avg	0,81	0,81	0,81	243
Weighted avg	0,81	0,81	0,81	243

Table 2 presents the performance evaluation outcomes of the Light Gradient Boosting Machine (LGBM) model in a classification task, utilizing multiple evaluation metrics such as precision, recall, f1-score, and overall accuracy. The precision of class 0, which represents the accuracy of optimistic predictions made by the model, is 0.81. This indicates that approximately 81% of the model's predictions for class 0 are correct. Class 1 has a precision of 0.80, meaning that approximately 80% of the forecasts for class 1 are accurate. The recall metric, which assesses the model's capacity to identify positive instances correctly, indicates that class 0 has a recall rate of 0.81. This implies that the model can recognize approximately 81% of all actual class 0 examples. The recall for class 1 is 0.81, which means that the model can correctly identify approximately 81% of all instances belonging to class 1. The F1 score is determined by summing the precision and recall scores harmonically. Both class 0 and class 1 exhibit a f1-score of 0.81 in this scenario, indicating a harmonious equilibrium between precision and recall in both classes.

The LGBM model achieved an overall accuracy of 0.81, representing the proportion of accurate predictions made across the entire evaluation dataset. The evaluation demonstrates consistency in model performance, as indicated by both the macro average and weighted average values of precision, recall, and f1-score. The results of the evaluation suggest that the LGBM model performs admirably in all respects, achieving an accuracy rate of 81%. Despite minor discrepancies in precision and recall between classes 0 and 1, the balanced f1-score accurately reflects the model's proficiency in classifying both classes. Importantly, LGBM performs very similarly to earlier GBM models with respect to recall, f1-score, and precision, suggesting that both exhibit similar performance in each classification task. While there is room for enhancement in class 1 precision and recall, the LGBM model has exhibited commendable performance in this classification task.

DISCUSSIONS

What is the performance comparison between Gradient Boosting Machine and Light Gradient Boosting Machine predictive models in forecasting the likelihood of landslides using different environmental, geographic, and geological characteristics?

Examining the performance comparison between Gradient Boosting Machine and Light Gradient Boosting Machine in forecasting the probability of landslides using different environmental, geographic, and geological features is crucial for comprehending the effectiveness of these models in tackling this significant problem. Accurate prediction models are necessary for the early detection and mitigation of landslides, which are complex natural disasters influenced by multiple factors. The comparison commences by comprehending the mechanisms that underlie the two models. GBM, an influential ensemble learning technique, builds a series of decision trees sequentially to rectify mistakes, leading to

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

a progressively robust predictive model. Light GBM, a refined version, utilizes a histogram-based technique that allows for concurrent node splitting during the construction of decision trees, resulting in improved computational speed.

When assessing their performance, the main emphasis is on their capacity to efficiently analyse diverse environmental, geographic, and geological data to estimate the likelihood of landslides. Sequential enhancement techniques in GBM may encounter computational limitations, particularly when handling large datasets with intricate features. Light GBM's novel method of parallelizing node splitting enables it to effectively process vast quantities of intricate data, potentially leading to expedited model training and predictions. Nevertheless, the crucial factor lies not solely in the computational velocity but also in the precision of the forecasts. The objective of the comparison is to evaluate whether the computational acceleration of Light GBM results in a decrease in prediction accuracy compared to GBM or if it can maintain or even enhance prediction accuracy while increasing processing speed. This study conducted experiments using comprehensive datasets encompassing various environmental, geographic, and geological factors that contribute to the occurrence of landslides. The data set comprises terrain attributes, rainfall patterns, soil characteristics, vegetation cover, and geological features. The GBM and Light GBM models underwent training, validation, and testing using these datasets, allowing for meticulous assessment of their predictive abilities. This evaluation incorporates metrics such as accuracy, precision, recall, and F1 score to gauge the model's effectiveness in detecting potential landslides thoroughly.

Furthermore, this study examines computational efficiency by quantifying the time needed for both models to process and generate predictions for specific data sets. This comparison facilitates the assessment of whether Light GBM's computational efficiency compromises the accuracy of landslide forecasting in comparison to GBM or if it successfully achieves a harmonious equilibrium between speed and precision. The primary objective of this comparison is to offer a deeper understanding of the balance between computational speed and prediction accuracy in landslide prediction models. This will empower stakeholders to make well-informed decisions when choosing the most suitable model for their requirements and limitations in managing landslide-related risks.

Can the enhanced computational speed provided by Light Gradient Boosting Machine compensate for or improve the accuracy of predictions concerning landslide susceptibility compared to Gradient Boosting Machine when dealing with extensive and intricate datasets? Research Question 2.

It is essential to investigate whether the increased computational speed provided by a Light Gradient Boosting Machine can enhance or improve the accuracy of landslide susceptibility predictions compared to a Gradient Boosting Machine when dealing with large and complex datasets. This exploration will help us understand the trade-off between efficiency and accuracy in predictive modelling. To predict landslide susceptibility, it is necessary to use strong models that can handle large and intricate datasets effectively while maintaining high accuracy. This analysis explores the distinct abilities of Light GBM and GBM in handling complex datasets, providing insight into their predictive accuracy when it comes to landslide susceptibility.

GBM, known for its ensemble learning methodology, sequentially builds decision trees, iteratively rectifying mistakes to generate a robust predictive model. Nevertheless, the sequential nature of the data may present difficulties when dealing with large datasets that have complex characteristics, potentially causing computational limitations. Conversely, Light GBM employs a novel histogram-based technique that concurrently splits nodes during decision tree building, resulting in improved computational efficiency. One question that arises is whether the increased speed in processing large datasets has a positive effect on the accuracy of landslide susceptibility predictions or if it potentially undermines the precision achieved by GBM.

The research aims to evaluate the predictive accuracy of both models in determining landslide susceptibility. This will be done by analysing comprehensive datasets that include various environmental, geographic, and geological factors known to affect the likelihood of landslides. The datasets include terrain attributes, rainfall patterns, soil properties, land cover, and geological features, which represent the complex factors that contribute to landslides. The GBM and Light GBM models are subjected to intensive training, validation, and testing using these datasets, with the goal of revealing their predictive capabilities. The evaluation metric encompasses accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the models' abilities to identify areas susceptible to landslides. The comparison focuses on whether the faster computational speed of Light GBM results in similar or better accuracy in predicting landslide susceptibility compared to GBM. It is crucial to determine whether Light GBM can maintain or improve predictive accuracy while also increasing computational efficiency. In addition, the study examines the trade-offs between computational speed and predictive accuracy by analysing the processing time of both models when

* Corresponding author

Submitted : Dec 28, 2023
Accepted : Dec 29, 2023
Published : Jan 1, 2024

handling large datasets and producing predictions of landslide susceptibility. This thorough examination helps determine if Light GBM's faster computation sacrifices accuracy or if it achieves a commendable equilibrium between efficiency and precision in predicting landslide susceptibility. The primary aim of this comparison is to enhance comprehension of the complex relationship between computational speed and predictive accuracy in landslide susceptibility models. The findings guide stakeholders in choosing the most appropriate model according to specific requirements and constraints, enabling the implementation of effective strategies to mitigate the risks associated with landslides.

CONCLUSION

Ensemble algorithms such as Gradient Boosting Machines and Light Gradient Boosting Machines rely heavily on the ROC to assess the efficacy of their classification models. Using a combination of weak models, GBM builds a more robust predictive model. This algorithm is part of an ensemble. The capacity of the model to differentiate between positive and negative classes is shown by the Area Under the Curve (AUC) value for GBM's ROC. Improved in speed, efficiency, and performance, LGBM is an upgrade from GBM. Several metrics are used to evaluate the model's overall performance, such as accuracy values, f1-score, recall, and precision. Overall, the GBM model performs satisfactorily, with an accuracy of 82%. Consistent model performance is demonstrated by the LGBM model, which attains an overall accuracy of 81%. Both models could improve upon class 1 precision and recall, but overall, they perform admirably on the given classification task.

REFERENCES

Achu, A. L., Aju, C. D., Di, M., Prakash, P., Gopinath, G., Shaji, E., & Chandra, V. (2023). *Geoscience Frontiers Machine-learning based landslide susceptibility modelling with emphasis on uncertainty analysis*. *14*.

Deng, S., Su, J., Zhu, Y., Yu, Y., & Xiao, C. (2024). *Forecasting carbon price trends based on an interpretable light gradient boosting machine and Bayesian optimization*. *242*(June 2023).

Friedman, B. J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *29*(5), 1189–1232.

Guo, Z., Guo, F., Zhang, Y., He, J., & Li, G. (2023). *A python system for regional landslide susceptibility assessment by integrating machine learning models and its application*. *9*(October).

Hindarto, D. (2022). Perbandingan Kinerja Akurasi Klasifikasi K-NN, NB dan DT pada APK Android. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *9*(1), 486–503. https://doi.org/10.35957/jatisi.v9i1.1542

Hindarto, D., & Djajadi, A. (2023). *Android-manifest extraction and labeling method for malware compilation and dataset creation*. *13*(6), 6568–6577. https://doi.org/10.11591/ijece.v13i6.pp6568-6577

Hindarto, D., & Santoso, H. (2022). PERFORMANCE COMPARISON OF SUPERVISED LEARNING USING NON-NEURAL NETWORK AND NEURAL NETWORK. *Janapati*, *11*, 49–62.

Mishra, D., Naik, B., Nayak, J., Souri, A., Byomakesha, P., & Vimal, S. (2023). *Light gradient boosting machine with optimized hyperparameters for identi fi cation of malicious access in IoT network*. *9*(October 2022), 125–137.

Oluwatosin, T., Opeoluwa, D., & Gbenga, E. (2023). *Healthcare Analytics A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis*. *4*(June).

Oram, E., Byomakesha, P., Naik, B., Nayak, J., & Vimal, S. (2021). *Light gradient boosting machine-based phishing webpage detection model using phisher website features of mimic URLs*. *152*, 100–106.

Sunaryono, D., Sarno, R., & Siswantoro, J. (2022). *Gradient boosting machines fusion for automatic epilepsy detection from EEG signals based on wavelet features*. *34*, 9591–9607.

Tai, C., Liao, T., Chen, S., & Chung, M. (2024). *Sleep stage classification using Light Gradient Boost Machine : Exploring feature impact in depressive and healthy participants*. *88*(October 2023).

Thongthammachart, T., Araki, S., Shimadera, H., Matsuo, T., & Kondo, A. (2022). *Incorporating Light Gradient Boosting Machine to land use regression model for estimating NO2 and PM2.5 levels in Kansai region, Japan*. *155*(May).

Yang, C., Liu, L., Huang, F., Huang, L., & Wang, X. (2023). *Machine learning-based landslide susceptibility assessment with optimized ratio of landslide to non-landslide samples*. *123*, 198–216.

Yin, H., Sharma, B., Hu, H., Liu, F., Kaur, M., Cohen, G., Mcconnell, R., & Eckel, S. P. (2024). *Predicting the climate impact of healthcare facilities using gradient boosting machines*. *12*(September 2023).

Zeng, T., Jin, B., Glade, T., Xie, Y., Li, Y., Zhu, Y., & Yin, K. (2024). *Assessing the imperative of conditioning factor grading in machine learning-based landslide susceptibility modeling : A critical inquiry*. *236*(November 2023).

* Corresponding author