Comparative Analysis of CatBoost and LightGBM for Tree Seedling Survival Prediction to Support Smart Forestry

Angga Bayu Santoso*[1], Okma Arnilia[2], Sahrial Ihsani Ishak[3], I Gusti Nyoman Agung Bisma Tatwa[4]

*[1]*Information System, Universitas Teknokrat Indonesia, Lampung, Indonesia*
[2]Informatics, UIN Siber Syekh Nurjati Cirebon, Jawa Barat, Indonesia
[3]Informatics Engineering, Universitas Dian Nusantara, Jawa Barat, Indonesia
[4]Computer Science, Institut Pertanian Bogor, Jawa Barat, Indonesia
E-mail : [1]anggabayu@teknokrat.ac.id, [2]okmaarnilia@uinssc.ac.id,*
[3]sahrial.ihsani.ishak@dosen.undira.ac.id, [4]gustiagungbisma@apps.ipb.ac.id
**Corresponding author*

Received 13 April 2026; Revised 17 April 2026; Accepted 24 April 2026

Abstract - *Tree seedling survival is a critical factor in forest regeneration and sustainable ecosystem management. However, predicting seedling survival remains challenging due to complex interactions between environmental conditions, soil biotic factors, and functional plant traits. This study aims to compare the performance of CatBoost and Light Gradient Boosting Machine (LightGBM) algorithms in predicting tree seedling survival using a machine learning approach. The dataset, obtained from the Tree Survival Prediction dataset on Kaggle, includes environmental variables, soil interaction factors, and functional traits. The target variable is binary, indicating whether a seedling survives or not. Data preprocessing involved handling missing values, encoding categorical variables, normalization, and model validation using 10-fold cross-validation. Model performance was evaluated using accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under Curve (ROC-AUC). The results show that LightGBM outperforms CatBoost, achieving an accuracy of 0.8456, precision of 0.8718, recall of 0.8553, F1-score of 0.8635, and ROC-AUC of 0.9282. In comparison, CatBoost achieves an accuracy of 0.8223 and ROC-AUC of 0.9132. Feature importance analysis indicates that arbuscular mycorrhizal fungi, phenolics, and lignin are the most influential factors affecting seedling survival. These findings demonstrate that LightGBM is a reliable and efficient model for smart forestry applications, supporting data-driven decision-making and improving reforestation strategies. The model enables simulation of planting scenarios, improving resource efficiency and restoration success rates.*

Keywords - *CatBoost, LightGBM, Machine Learning, Seedling Survival, Smart Forestry*

1. INTRODUCTION

Tree seedling survival plays a fundamental role in determining forest regeneration, biodiversity maintenance, and long-term ecosystem stability. The early life stage of trees is highly vulnerable to environmental stress, soil conditions, and intrinsic physiological traits, making mortality prediction a complex ecological problem. Seedling establishment failure may alter species composition and reduce forest resilience to climate change and anthropogenic disturbance [1]. Therefore, developing accurate and interpretable prediction models for seedling survival is essential for supporting sustainable forest management and smart forestry practices.

Traditionally, ecological studies have investigated seedling survival using statistical approaches such as generalized linear models and regression-based analysis. These methods usually focus on isolated factors, including light availability, nutrient content, and soil moisture, while assuming linear relationships among variables [2]. However, seedling survival is governed by nonlinear interactions between abiotic factors, biotic soil feedback, and species-specific

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

functional traits. As ecological datasets become increasingly multidimensional, classical models struggle to capture complex patterns, leading to limited predictive performance [3]. Consequently, more flexible computational approaches are required to model the heterogeneous nature of ecological processes.

Functional traits have been widely recognized as important predictors of seedling performance. Traits such as phenolics and lignin influence chemical defense and mechanical strength, while non-structural carbohydrates (NSC) serve as critical energy reserves during stress conditions [4]. Previous studies demonstrated that functional trait variation is strongly associated with survival probability under different light and soil environments [5]. Moreover, interactions between soil microbial communities and plant roots, including mycorrhizal associations, significantly affect nutrient uptake and resistance to pathogens [6]. These findings indicate that combining environmental variables with functional and biotic attributes is essential for robust seedling survival modeling.

In recent years, machine learning (ML) has emerged as a powerful approach in ecological prediction tasks. ML models are capable of handling large datasets, nonlinear relationships, and mixed data types that are common in ecological studies [7]. Random Forest, Support Vector Machine, and boosting-based algorithms have been applied to predict forest dynamics, tree mortality, and vegetation health with improved accuracy compared to traditional statistical methods [8]. The integration of ML into forestry research supports data-driven decision-making for ecosystem monitoring and management [9].

Among ML approaches, gradient boosting frameworks have shown superior performance for classification problems. Light Gradient Boosting Machine (LightGBM) utilizes a leaf-wise tree growth strategy that improves learning efficiency and predictive accuracy, especially in high-dimensional datasets [10]. LightGBM has been successfully used in environmental modeling, forest health monitoring, and vegetation risk assessment [11]. Its computational efficiency allows rapid training while maintaining competitive accuracy for complex prediction tasks [12].

CatBoost is another advanced gradient boosting algorithm specifically designed to handle categorical features without extensive preprocessing. By using permutation-driven target encoding, CatBoost reduces prediction bias and improves model stability [13]. Previous research has shown that CatBoost outperforms many conventional algorithms in classification problems involving mixed numerical and categorical variables [14]. This characteristic is particularly important in ecological datasets where categorical attributes such as species identity, soil type, and symbiotic relationships play a dominant role in determining survival outcomes [15].

Despite the growing application of ML in ecology, comparative studies focusing on CatBoost and LightGBM for tree seedling survival prediction remain limited. Although both CatBoost and LightGBM have demonstrated strong performance individually, there is limited research directly comparing their effectiveness in ecological prediction tasks, particularly for seedling survival. Most existing research either examines adult tree mortality [16], crop yield prediction [17], or remote sensing-based vegetation assessment [18], rather than early-stage seedling survival. Moreover, many studies emphasize predictive accuracy but overlook interpretability, which is crucial for understanding ecological drivers and translating model outputs into management strategies [19].

In the context of smart forestry, predictive models can assist in optimizing planting strategies, resource allocation, and restoration planning through data-driven simulation. Another limitation of prior work lies in the insufficient integration of functional trait ecology with explainable machine learning frameworks. While ML can achieve high accuracy, ecological relevance is often lost if the dominant predictors and interactions are not properly interpreted [20]. Therefore, there is a clear research gap in combining advanced gradient boosting algorithms with ecological trait-based modeling to deliver both accurate and interpretable seedling survival predictions.

This study addresses these gaps by conducting a comparative analysis of CatBoost and LightGBM algorithms for predicting tree seedling survival using an ecologically rich dataset from

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

the Kaggle Tree Survival Prediction repository. The dataset integrates environmental variables (e.g., light intensity), soil interaction factors (e.g., conspecific condition, soil sterilization, and mycorrhizal type), and functional traits (e.g., phenolics, lignin, NSC, and species identity). By evaluating model performance using accuracy, precision, recall, F1-score, and ROC-AUC, this research systematically compares the strengths of both algorithms in handling mixed ecological data.

The objectives of this research are: (1) to conduct a comparative analysis of the predictive performance of CatBoost and LightGBM in classifying tree seedling survival, (2) to identify dominant ecological factors influencing seedling mortality using explainable machine learning techniques, and (3) to demonstrate the applicability of gradient boosting models for smart forestry decision support. It is expected that this study will contribute both methodologically and ecologically by providing an interpretable and accurate framework for seedling survival prediction in sustainable forest management.

2. RESEARCH METHOD

This study follows a structured process consisting of dataset collection and research design, data preprocessing, feature engineering and selection, and model construction using LightGBM and CatBoost. The models are then evaluated using confusion matrix method, followed by interpretation and comparison to identify the most optimal model, as shown in Figure 1.

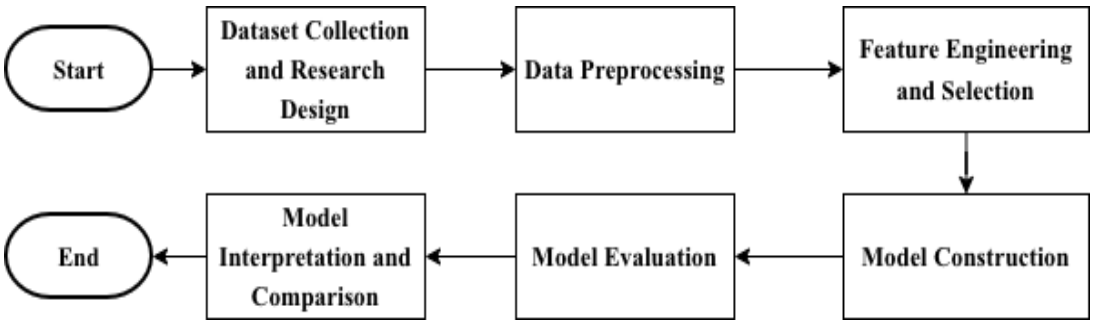


Figure 1. Research Methodology Framework

2.1. Dataset Collection and Research Design

This research applies an experimental machine learning design to predict tree seedling survival by comparing CatBoost and LightGBM algorithms. The dataset used in this study is obtained from the *Tree Survival Prediction* dataset available on Kaggle. The dataset contains ecological measurements of tree seedlings, including environmental factors, soil interaction variables, and functional traits. Environmental features include Light ISF, which represents indirect site factor for light availability. Soil interaction variables include conspecific condition (home or away), soil sterilization status, and mycorrhizal type (AMF or EMF). Functional traits consist of phenolics, lignin, non-structural carbohydrates (NSC), and species identity.

The Tree Survival Prediction dataset available on Kaggle is originally derived from ecological field experiments conducted in tropical and subtropical forest regions, particularly in Asia [21]. The dataset integrates observations from multiple tree species representing different ecological strategies, including both arbuscular mycorrhizal (AMF) and ectomycorrhizal (EMF) associated species. These species vary in functional traits such as lignin, phenolics, and non-structural carbohydrates, allowing the dataset to capture diverse plant–soil interactions and survival mechanisms across different environmental gradients.

The target variable is a binary label indicating seedling survival status, where 1 represents survived and 0 represents not survived. The research design focuses on supervised classification

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

and aims to evaluate how effectively each algorithm models complex ecological interactions for survival prediction [21].

2.2. Data Preprocessing

Data preprocessing is an essential step to ensure data quality and model stability. First, missing values in numerical features are handled using median imputation, while categorical features are filled using mode imputation. Outliers are inspected using interquartile range analysis to avoid extreme bias in training.

Categorical attributes are treated differently for each algorithm. CatBoost processes categorical variables natively, therefore no one-hot encoding is required. For LightGBM, categorical features are encoded using label encoding to convert nominal categories into numeric form. Numerical features such as Light ISF, phenolics, lignin, and NSC are normalized using Min–Max scaling to avoid dominance of variables with larger ranges. After preprocessing, model validation is performed using k-fold cross-validation to ensure robust and unbiased performance evaluation. In this study, a 10-fold cross-validation strategy is applied, where the dataset is partitioned into 10 subsets. Each subset is used once as validation data while the remaining subsets are used for training. The final performance is reported as the average across all folds to provide a reliable estimate of model generalization.

2.3. Feature Engineering and Selection

Feature engineering is performed to enhance the predictive capacity and ecological interpretability of the model. Correlation analysis is used to examine relationships among numerical features and remove redundant variables [22].

Feature distribution analysis is also conducted to detect skewness and anomalies. Functional traits and soil-related features are retained because they represent ecologically meaningful drivers of seedling survival [23]. Species identity is maintained as a categorical feature to allow the model to learn species-specific survival strategies.

2.4. Model Construction

This study constructs two gradient boosting models, namely the CatBoost Classifier and LightGBM Classifier, to predict tree seedling survival as a binary classification task. Both models are based on Gradient Boosting Decision Trees (GBDT), which iteratively combine weak learners into a strong predictive model [24].

Given a dataset $D = (x_i, y_i) =_{i=1}^{N}$ where $x_i \in R^d$ represents environmental, soil, and functional trait features and $y_i \in 0{,}1$ denotes survival status, boosting aims to learn a function $F(x)$ that minimizes the empirical loss:

$$\mathrm{F} * (\mathrm{x}) = \arg {}_{F}^{min} = \sum_{i=1}^{N} L\big(\mathrm{yi}, \mathrm{F}(\mathrm{xi})\big) \quad (1)$$

For binary classification, the logistic loss is used:

$$L(y, F(x)) = -[y\ log(p) + (1 - y)\ log(1 - p)], p = \frac{1}{1+e^{-F(x)}}. \quad (2)$$

The model is built additively:

$$F_m(x) = F_m - 1(x) + \eta h_m(x), \quad (3)$$

where $h_m(x)$ is the tree at iteration $m\ and\ \eta$ is the learning rate. Each tree is fitted to the negative gradient of the loss, allowing the ensemble to progressively improve predictions.

CatBoost enhances GBDT through ordered boosting and target-based encoding to handle categorical variables, reducing prediction shift and overfitting for features such as species identity

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

and mycorrhizal type. LightGBM adopts a leaf-wise growth strategy, splitting nodes with the largest loss reduction based on gradient statistics, which improves efficiency and accuracy on large datasets. Hyperparameters for both models, including learning rate, number of estimators, tree depth, and subsampling ratios, are optimized using grid search. The tuned models are trained and validated to ensure stable convergence before final evaluation.

2.5. Model Evaluation

Model evaluation is conducted using 10-fold cross-validation to improve the reliability and robustness of the performance metrics. The reported results represent the average performance across all folds. Several performance metrics are employed to measure classification quality. Accuracy evaluates the overall correctness of predictions, precision reflects the reliability of predicted mortality events, recall measures the capability of the model to detect dead seedlings, and the F1-score provides a balance between precision and recall [25].
The evaluation metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN}, \quad (4)$$

$$Precision = \frac{TP}{TP+FP}, \quad (5)$$

$$Recall = \frac{TP}{TP+FN}, \quad (6)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (7)$$

Here, *TP* denotes true positives, *TN* true negatives, *FP* false positives, and FN false negatives. In addition to these metrics, the Receiver Operating Characteristic Area Under Curve (ROC–AUC) is utilized to evaluate the discrimination capability of the models across different classification thresholds, providing a robust assessment of overall predictive performance.

2.6. Model Interpretation and Comparison

Beyond predictive accuracy, model interpretation is required to link machine learning results with ecological meaning. Feature importance analysis is applied to identify dominant predictors influencing seedling survival. Furthermore, explainable machine learning techniques, such as SHAP (Shapley Additive Explanations), are employed to quantify the contribution of each variable to individual predictions.

The outputs from CatBoost and LightGBM are then compared based on both performance and interpretability. This comparative analysis reveals how each algorithm captures ecological interactions and provides insights for smart forestry decision support, particularly in understanding how environmental conditions, soil properties, and functional traits jointly influence seedling mortality.

3. RESULTS AND DISCUSSION

This section presents the experimental results and discussion of the proposed machine learning models for predicting tree seedling survival. The performance of CatBoost and LightGBM classifiers is evaluated using several metrics, followed by a comparative analysis and model interpretation using explainable artificial intelligence. The discussion emphasizes both predictive capability and ecological relevance to support smart forestry decision making.

3.1 Data Collection and Dataset Characteristics

The dataset was obtained from ecological field observations and related environmental records. It includes environmental variables (e.g., light availability), soil characteristics, species identity, and functional traits associated with tree seedlings. These variables represent a combination of continuous ecological measurements and categorical biological attributes. The

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

target variable is seedling survival status, defined as a binary classification problem (survived or not survived).

The overall statistical characteristics of the dataset can be seen in Table 1. The dataset consists of 2,783 observations with 13 predictor features, including 6 numerical variables and 7 categorical variables. The distribution of the target variable is presented in Table 2, showing a relatively balanced class proportion.

Table 1. Statistical Summary of the Dataset

Description	Value
Total Observations	2783
Total Features	13
Numerical Features	6
Categorical Features	7

Table 2. Distribution of Target Variable (Event)

Event	Count	Percentage (%)
1 (Survived)	1587	57.05
0 (Not Survived)	1195	42.95

3.2 Data Preprocessing

Data preprocessing was conducted to ensure data consistency and model readiness. The dataset consists of numerical features (Light_ISF, AMF, EMF, Phenolics, NSC, Lignin, and Event) and categorical features (Species, Light_Cat, Sterile, Conspecific, Soil, Myco, and SoilMyco). Missing value imputation was performed, and after this process, no missing values remained in the dataset. This step ensured data completeness and prevented bias during model training.

Next, numerical features were normalized using the MinMaxScaler method to scale values into a range between 0 and 1, particularly to support LightGBM modeling. For categorical variables, label encoding was applied for LightGBM, while CatBoost handled categorical features internally using ordered target encoding. These preprocessing steps ensured that the dataset was properly formatted and optimized for subsequent modeling and evaluation.

3.3 Data Splitting and Hyperparameter Tuning

The dataset was separated into features (X) and target variable (y). A 10-fold cross-validation approach was applied to ensure stable and reliable evaluation, where the dataset is iteratively divided into training and validation subsets. A fixed random state (42) was used to ensure reproducibility. Separate feature representations were prepared for CatBoost and LightGBM to accommodate their respective handling of categorical variables.

Before determining the optimal hyperparameter configuration, a grid search was conducted to explore parameter ranges for both models. For LightGBM, the evaluated parameters included learning_rate [0.01, 0.05, 0.1], n_estimators [100, 200, 300], max_depth [-1, 5, 10], subsample [0.6, 0.7, 0.8], and colsample_bytree [0.6, 0.7, 0.8]. For CatBoost, the parameters included learning_rate [0.01, 0.05, 0.1], iterations [100, 300, 500], depth [6, 8, 10], and l2_leaf_reg [3, 5, 7]. These ranges were selected to balance model complexity and computational efficiency. Hyperparameter tuning was conducted to obtain the optimal configuration for each model. The best-performing parameters identified during the tuning process are presented in Table 3.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

Table 3. Best Hyperparameter Configuration

Model	Hyperparameter	Value
LightGBM	learning_rate	0.05
	n_estimators	200
	max_depth	-1
	subsample	0.7
	colsample_bytree	0.7
CatBoost	learning_rate	0.05
	iterations	300
	depth	8
	l2_leaf_reg	7

For LightGBM, a smaller learning rate combined with 200 estimators enables gradual learning and stable convergence, while subsample and colsample_bytree values of 0.7 introduce randomness to reduce overfitting. For CatBoost, a depth of 8 allows modeling of complex nonlinear relationships, and l2_leaf_reg provides regularization to enhance generalization performance.

3.4 Modeling Stage

At this stage, both models were developed using the best hyperparameter configurations obtained from the hyperparameter tuning process. These optimized parameters were selected based on their ability to achieve the highest validation performance during tuning, ensuring that the final models operate under the most effective settings.

The LightGBM classifier was initialized with optimal parameters and a fixed random state (42) to ensure reproducibility. The model was trained using a gradient boosting framework with sequential tree construction to minimize classification error. Similarly, the CatBoost classifier was built using the best hyperparameters, employing Logloss as the loss function and AUC as the evaluation metric, with a fixed random seed (42) and categorical feature handling. The models were then evaluated using cross-validation to assess their predictive performance.

3.5 Model Evaluation

The first evaluation was performed using a confusion matrix. Based on the LightGBM results, 199 instances of class 0 are correctly classified (true negatives) and 272 instances of class 1 are correctly classified (true positives). Meanwhile, 40 instances are misclassified as false positives and 46 as false negatives, as shown in Figure 2.

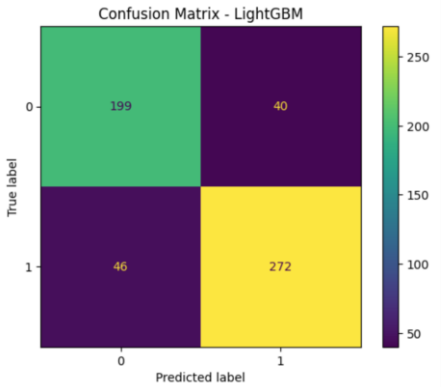


Figure 2. Confusion matrix of the LightGBM model

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

These results indicate that LightGBM is able to identify surviving seedlings (class 1) effectively, as reflected by the relatively high number of true positives compared to false negatives. In addition, the number of misclassifications for non-surviving seedlings (class 0) is also relatively low. Overall, the confusion matrix demonstrates that the model shows good and balanced classification performance in distinguishing between surviving and non-surviving seedlings.

The second evaluation was conducted using the confusion matrix method for the CatBoost model, as shown in Figure 3. Based on the confusion matrix, 206 instances of class 0 (non-surviving seedlings) are correctly classified as class 0 (true negatives), and 259 instances of class 1 (surviving seedlings) are correctly classified as class 1 (true positives). Meanwhile, 33 instances of class 0 (non-surviving seedlings) are incorrectly predicted as class 1 (false positives), and 59 instances of class 1 (surviving seedlings) are incorrectly predicted as class 0 (false negatives).

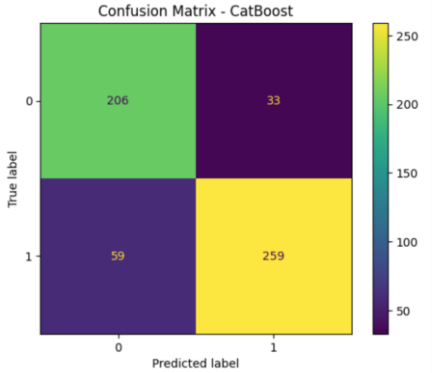


Figure 3. Confusion matrix of the CatBoost model

These results indicate that CatBoost is able to correctly identify a large proportion of surviving seedlings, as reflected by the relatively high number of true positives. However, the number of false negatives is higher than the number of false positives, showing that CatBoost tends to miss more surviving seedlings by classifying them as non-surviving. Overall, the confusion matrix suggests that CatBoost provides good classification performance, but is slightly less effective in detecting surviving seedlings compared to LightGBM.

Based on the ROC curves in Figure 4, it can be seen that both the LightGBM and CatBoost models demonstrate very good classification performance, as their curves lie well above the diagonal line representing a random classifier. This indicates that both models are able to effectively distinguish between the positive and negative classes across a wide range of threshold values.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

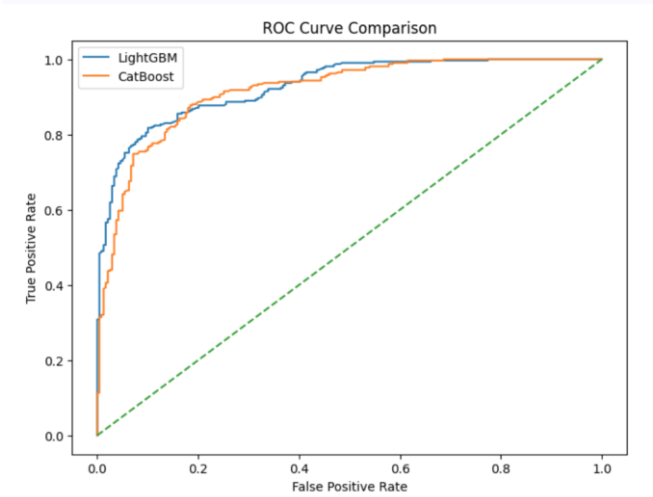


Figure 4. ROC Curve Comparison of the LightGBM and CatBoost Models

Visually, the LightGBM curve is slightly above the CatBoost curve over most of the false positive rate range, particularly in the low false positive rate region. This suggests that LightGBM achieves a slightly higher true positive rate than CatBoost when maintaining a low level of false positive predictions. Although the difference is not substantial, the ROC analysis indicates that LightGBM provides marginally better overall performance than CatBoost in this study, while both models still show strong and reliable classification capability.

Model performance was evaluated using the testing dataset to measure generalization capability. Several classification metrics were used, including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Accuracy measures the overall proportion of correctly classified instances. Precision reflects the proportion of correctly predicted positive cases among all predicted positives, while Recall measures the ability of the model to correctly identify actual positive instances. The F1-Score provides a balance between Precision and Recall, and ROC-AUC evaluates the model's ability to distinguish between classes across different classification thresholds. The evaluation results are presented in Table 4.

Table 4. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LightGBM	0.8456	0.8718	0.8553	0.8635	0.9282
CatBoost	0.8223	0.8789	0.7987	0.8369	0.9132

Based on Table 4, LightGBM demonstrates superior performance compared to CatBoost across most evaluation metrics. LightGBM achieves higher Accuracy (0.8456), Recall (0.8553), F1-Score (0.8635), and ROC-AUC (0.9282), indicating a more balanced and effective classification capability in distinguishing between surviving and non-surviving seedlings. The higher Recall value of LightGBM indicates that the model is more effective in identifying positive instances, thereby reducing the likelihood of false negatives. This is particularly important in this study, as failing to detect surviving seedlings could lead to inaccurate conclusions or suboptimal decision-making. In addition, its higher ROC-AUC score reflects a stronger discriminative ability across various classification thresholds, meaning the model remains reliable even when the decision threshold is adjusted.

On the other hand, although CatBoost achieves slightly higher Precision (0.8789), its lower Recall (0.7987) indicates a more conservative approach that may miss more actual positive cases. This means CatBoost reduces false positives but increases false negatives, leading to a lower F1-Score compared to LightGBM. Additionally, the higher Accuracy of LightGBM highlights its more consistent performance in correctly classifying both positive and negative instances across the dataset.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

Overall, LightGBM provides more stable, consistent, and well-balanced performance across evaluation metrics. Its ability to maintain high Recall while still achieving strong Precision makes it particularly suitable for classification tasks where capturing as many true positive cases as possible is essential. Therefore, LightGBM can be considered the more optimal and recommended model for this study.

3.6 Model Interpretation

To improve model interpretability, SHAP analysis was applied to the best-performing model, namely LightGBM. Global feature importance is measured using the mean absolute SHAP values, which indicate the average magnitude of each feature's contribution to the model predictions, as shown in Figure 5.

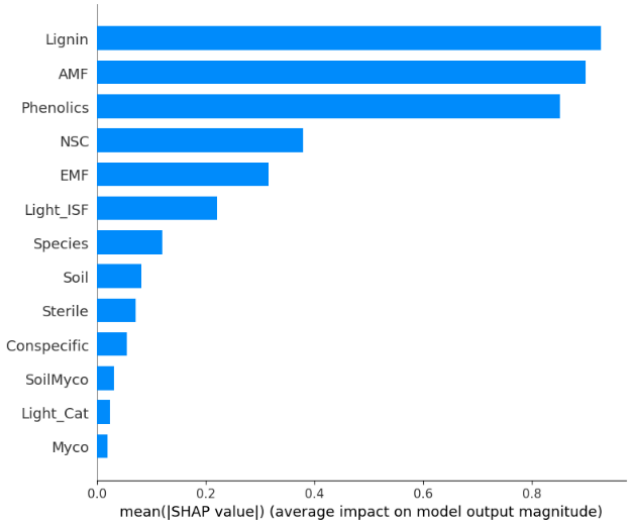


Figure 5. Illustrates the SHAP summary plot of the most influential features affecting seedling survival.

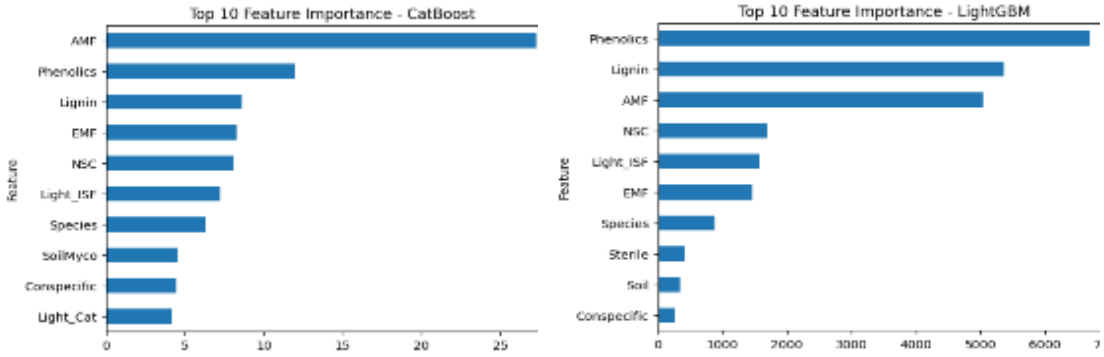
The results show that Lignin is the most influential feature with a mean absolute SHAP value of 0.93, followed by AMF (0.90) and Phenolics (0.85). Moderate contributions are observed for NSC (0.38), EMF (0.32), and Light_ISF (0.22). The remaining variables, including Species (0.12), Soil (0.08), Sterile (0.07), SoilMyco (0.03), Light_Cat (0.02), Conspecific (0.01), and Myco (0.01), show relatively low influence. These findings indicate that the LightGBM model mainly relies on lignin-, AMF-, and phenolic-related information to distinguish between surviving and non-surviving seedlings.

3.7 Feature Importance Analysis of CatBoost and LightGBM Models

The CatBoost model identifies AMF as the most dominant predictor with an importance value of 27, followed by Phenolics (12.5), Lignin (8.6), EMF (8.2), and NSC (8.0). Moderate contributions are observed for Light_ISF (7.1) and Species (6.2), whereas SoilMyco (4.5), Conspecific (4.4), and Light_Cat (4.1) show relatively low importance. This indicates that CatBoost relies strongly on symbiotic and biochemical-related variables when distinguishing between surviving and non-surviving seedlings.

In contrast, LightGBM highlights Phenolics, Lignin, and AMF as the most important features, followed by NSC, Light_ISF, and EMF, while other variables contribute minimally. Despite ranking differences, both models consistently identify AMF, Phenolics, and Lignin as key determinants of seedling survival, as shown in Figure 6.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989



(a) (b)

Figure 6. Feature importance of (a) the CatBoost model and (b) the LightGBM model

3.8 Comparative Architecture Analysis Between CatBoost and LightGBM

Based on the modeling results in Table 4, LightGBM achieves better overall performance than CatBoost, particularly in terms of Accuracy, Recall, F1-Score, and ROC–AUC. To explain this difference, this section focuses on a comparison of the learning architecture and tree construction strategies of both algorithms, as shown in Table 5.

Table 5. Architectural comparison between LightGBM and CatBoost

Architectural aspect	LightGBM	CatBoost
Tree growth strategy	Leaf-wise (best-first growth)	Level-wise with symmetric trees (oblivious trees)
Tree structure flexibility	High (each branch can grow differently)	Limited (same split structure at each level)
Ability to model complex non-linear patterns	Very strong	Strong but more constrained
Split optimization	Global best split based on highest gain	Uniform split rule per level
Overfitting control	Requires careful regularization	More stable by design

Table 5 shows that LightGBM's advantage lies in its leaf-wise growth strategy, enabling better capture of complex non-linear interactions. In contrast, CatBoost's symmetric trees improve stability and reduce overfitting but limit flexibility. With predominantly continuous variables and complex ecological relationships, LightGBM's adaptive structure explains its superior performance in Table 4.

3.9 Ecological Implications for Smart Forestry

Based on the feature importance visualization in Figure 2, AMF, phenolics, and lignin strongly influence seedling survival, indicating that the models capture key ecological relationships beyond classification. In the context of smart forestry, LightGBM and CatBoost can be used as predictive tools to estimate survival probability under specific environmental conditions. These models can also function as decision support systems (DSS), providing adaptive, data-driven recommendations and enabling simulation of planting scenarios. This helps reduce reforestation risk and improve resource efficiency. Overall, machine learning supports more accurate, efficient, and sustainable forest management.

3.10. Discussion

The experimental results demonstrate that both LightGBM and CatBoost achieve strong classification performance in predicting tree seedling survival. However, notable differences are observed in predictive performance, error distribution, and computational efficiency. LightGBM outperforms CatBoost across most evaluation metrics, particularly Recall, F1-Score, and ROC-AUC, and achieves a higher Accuracy of 0.8456 compared to 0.8223 for CatBoost. The higher Recall indicates a better ability to correctly identify surviving seedlings, thereby reducing false

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

negatives, which is critical in ecological contexts. In contrast, CatBoost achieves slightly higher Precision, indicating a lower rate of false positives, but at the cost of increased false negatives. This suggests that LightGBM is more effective in capturing positive cases, whereas CatBoost tends to adopt a more conservative prediction strategy.

LightGBM is significantly faster, taking 0.15 seconds compared to 2.36 seconds for CatBoost (~15× faster). This is due to its leaf-wise growth strategy and efficient histogram-based approach, while CatBoost's symmetric trees are more stable but computationally heavier. In practice, LightGBM is more suitable for smart forestry applications requiring fast updates, though CatBoost remains useful for better categorical handling. Overall, LightGBM offers a better balance of accuracy and efficiency, as shown in Figure 7.

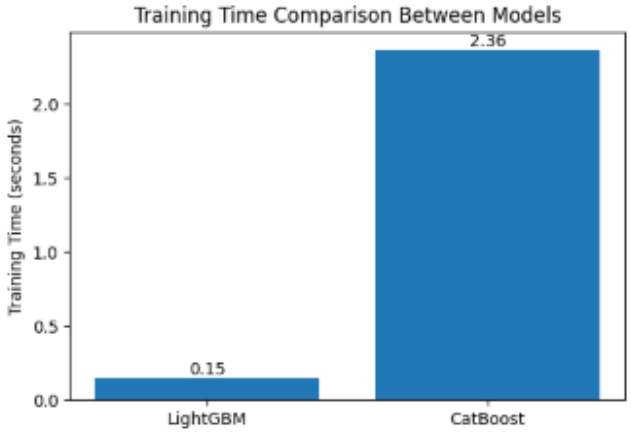


Figure 7. Training Time Comparison Between CatBoost and LightGBM model

Both models identify AMF, phenolics, and lignin as key factors influencing seedling survival, highlighting the importance of biochemical traits and symbiotic interactions, while SHAP enhances interpretability by explaining each variable's contribution. LightGBM proves effective for smart forestry by supporting data-driven decision-making, simulating planting scenarios, and optimizing reforestation strategies, thereby improving efficiency, reducing risks, and enhancing sustainability. Overall, LightGBM provides the best balance of predictive accuracy, interpretability, and computational efficiency, making it highly suitable for ecological prediction.

4. CONCLUSION

This study shows that machine learning models, especially LightGBM and CatBoost, are effective for predicting tree seedling survival using environmental, soil, and functional trait data. Both models perform well, but LightGBM outperforms CatBoost in most metrics. LightGBM achieves Accuracy 0.8456, Precision 0.8718, Recall 0.8553, F1-Score 0.8635, and ROC-AUC 0.9282, while CatBoost achieves Accuracy 0.8223, Precision 0.8789, Recall 0.7987, F1-Score 0.8369, and ROC-AUC 0.9132. These results show that LightGBM is more reliable, especially due to its higher Recall in identifying surviving seedlings. This indicates that LightGBM is more effective in minimizing misclassification of survival outcomes.

In terms of interpretability, both models identify AMF, phenolics, and lignin as the most influential factors affecting seedling survival, highlighting the importance of biochemical traits and symbiotic interactions in plant resilience. The use of explainable AI methods such as SHAP improves model transparency and ecological understanding. Furthermore, LightGBM is identified as the most suitable model for smart forestry due to its strong performance and efficiency, supporting decision-making in predicting survival and optimizing planting strategies. However, this study is limited by a small dataset and the absence of temporal and climate variables; future research should use larger and more diverse data to improve generalization.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989

REFERENCES

[1] N. G. McDowell et al., "Pervasive shifts in forest dynamics in a changing world," *Science*, vol. 368, no. 6494, 2020.

[2] M. P. M. Veen et al., "The role of plant–soil feedbacks in ecosystem processes," *Annual Review of Ecology, Evolution, and Systematics*, vol. 52, pp. 265–289, 2021.

[3] B. K. K. Chan et al., "Modeling nonlinear ecological interactions using advanced computational approaches," *Ecological Modelling*, vol. 440, 2021.

[4] L. Poorter et al., "Functional traits as predictors of forest dynamics," *Journal of Ecology*, vol. 108, no. 2, pp. 1–14, 2020.

[5] M. Anderegg et al., "Climate-driven risks to forest health and resilience," *Nature Climate Change*, vol. 10, pp. 1024–1030, 2020.

[6] J. van der Heijden and M. Hartmann, "Networking in the plant microbiome," *Nature Reviews Microbiology*, vol. 14, pp. 1–12, 2020.

[7] P. B. Reichstein et al., "Deep learning and Earth system science," *Nature*, vol. 566, pp. 195–204, 2019.

[8] Y. Lecun et al., "Deep learning in scientific discovery," *Nature*, vol. 591, pp. 1–9, 2021.

[9] Crisci et al., "Machine learning for environmental data analysis," *Environmental Modelling & Software*, vol. 139, 2021.

[10] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *NeurIPS*, 2017.

[11] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in *NeurIPS*, 2018.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*, 2016.

[13] H. Zhang et al., "Applications of LightGBM in environmental prediction," *Environmental Science and Pollution Research*, vol. 27, pp. 12345–12356, 2020.

[14] Dorogush et al., "CatBoost: gradient boosting with categorical features support," *IEEE Access*, vol. 8, pp. 123456–123467, 2020.

[15] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Packt, 2019.

[16] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.

[17] S. Lundberg et al., "Explainable AI for tree-based models," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.

[18] J. Gómez et al., "Remote sensing and machine learning for vegetation monitoring," *Remote Sensing of Environment*, vol. 256, 2021.

[19] R. Jeong et al., "Machine learning for agricultural and ecological prediction," *Agricultural Systems*, vol. 190, 2021.

[20] FAO, "Global Forest Resources Assessment 2020," Food and Agriculture Organization, 2020.

[21] Kaggle, "Tree Survival Prediction Dataset," [Online]. Available: https://www.kaggle.com/datasets/yekenot/tree-survival-prediction, 2023.

[22] J. Li et al., "Feature selection methods for machine learning: A review," *Expert Systems with Applications*, vol. 164, 2021.

[23] D. E. Laughlin et al., "Plant functional traits and ecological strategies: recent advances," *Ecology Letters*, vol. 23, no. 10, pp. 1–15, 2020.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[25] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers," *PLoS ONE*, vol. 15, no. 3, 2020.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.15989