JURNAL INFOTEL
VOL. 16, NO. 3, AUGUST 2024, PP. 642–659. DOI:10.20895/INFOTEL.V16I3.1180

RESEARCH ARTICLE

The Development of a Prediction Model for Potential Forest and Land Fires using Machine Learning Algorithms Based on Patrol Data

Angga Bayu Santoso[1,*], Imas Sukaesih Sitanggang[2], and Medria Kusuma Dewi Hardhienata[3]

[1,2,3]Department of Computer Science, IPB University, Bogor 16680, Indonesia
[3]Department, Institution, City Zip Code, Country

*Corresponding email: anggabayu@apps.ipb.ac.id

Received: May 10, 2024; Revised: July 22, 2024; Accepted: August 2, 2024.

Abstract: Approximately 64 percent of the land area in Indonesia is classified as forest. Deforestation in Indonesia occurs due to forest and land fires. Forest and land fires are prevented through integrated patrols. Integrated patrols utilize the Forest and Land Fire Prevention Patrol Information System to manage patrol data. Patrol data is used for data observation and simple spatial analysis. However, patrol data has not been used for further forest and land fire prevention studies. This research aims to use SVM, Random Forest, AdaBoost, and XGBoost algorithms to develop a prediction model for potential forest and land fires. The performance of these models will be compared to determine the prediction model with the best performance. The preprocessing stage combines the SMOTE-ENN method to handle data class imbalance and the Random Search method for hyperparameter tuning. In this study, the best performance was obtained using the XGBoost model, which had an accuracy of 95.5%. On the other hand, the accuracies of Random Forest, AdaBoost, SVM-Linear, SVM-Polynomial, SVM-RBF, and SVM-Sigmoid are 94.95%, 91.24%, 79.65%, 76.82%, 76.75%, and 33.83% respectively. It is implied that applying SMOTE-ENN and Random Search methods can improve the accuracy of the XGBoost model. In addition, the results show that the XGBoost model can employ boosting techniques to minimize residuals. This study also found that the variable with the highest correlation was the condition of dry vegetation.

Keywords: forest and land fires, prediction, Random Forest, SVM, XGBoost

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Indonesia has tropical forests that are rich in resources and biodiversity [1]. Approximately 64 percent of the total land area in Indonesia, or 120 million hectares, is designated as forest [2]. Indonesia has forests consisting of conservation areas (22.109 million ha), protected areas (29.680 million ha), and production areas (29.247 million ha) [1]. Forests as national development assets provide tangible benefits for the lives and livelihoods of the Indonesian people. Although Indonesia has a vast forest area, the reality is that forests in Indonesia continue to experience deforestation. Indonesia had its highest deforestation rate during 2018-2019, reaching 844.72 hectares per year [3]. Some activities identified as causing deforestation include agricultural expansion, illegal logging, and forest fires [2]. Increased deforestation is primarily attributed to forest and land fires, which can arise from natural and human causes [4].

Land and forest fires are burning forests and land caused by human or other natural causes. Land and forest fires cause losses and environmental damage in terms of ecology, economy, socio-culture, and politics [5]. In 2015, Indonesia experienced a sharp increase in land and forest fires, with 2.6 million hectares affected [6]. Based on the data from the Forest and Land Fire Early Detection and Warning Information System (SiPongi), 3,295,804 hectares of forest were burned in Kalimantan and Sumatra between 2013 and 2022.

The Ministry of Environment and Forestry continues conducting integrated patrol operations to prevent forest and land fires early in Indonesia [5]. The integrated patrol team conducts patrols to confirm information on forest and land fires. The Land and Forest Fire Prevention Patrol Information System facilitates integrated patrols in managing patrol data. The system was developed in collaboration with the Department of Computer Science at IPB University and the Ministry of Environment and Forestry [7]. Patrol data is currently used for data observation and simple spatial analyses in the spatial module. However, patrol data obtained from the Forest and Land Fire Prevention Patrol Information System has not been used for further forest and land fire prevention studies. Patrol data can be utilized to predict land and forest fire potential in Kalimantan and Sumatra, reducing the severity of this problem.

Previous research has successfully built a prediction model of forest and land fire potential using a combination of machine learning algorithms and methods for handling data class imbalance. One of them is a study that built a model of suitable conditions for forest fires in Southeast China using the SVM algorithm and the SMOTE method with an average accuracy of 62.7% [8]. Another study identified forest fires using the SVM algorithm and the Random Search method, achieving an accuracy of 90.9% [9]. In addition, a study predicted the causes of forest fires in Southern France using the Random Forest algorithm and the SMOTE method, achieving 70% accuracy [10]. A previous study detected forest fire scars in South Korea using the Random Forest algorithm and Grid Search methods, achieving 88% accuracy [11]. Furthermore, a study used the XGBoost algorithm and SMOTE method to achieve 88.8% accuracy in modeling the initial success rate of forest fire suppression in Liangshan [12]. Moreover, previous research classified forest cover and mapped forest fire vulnerability using the XGBoost algorithm and the Artificial Bee Colony-Adaptive Neuro-Fuzzy Inference System (ABC-ANFIS), achieving an accuracy rate of 81.44% [13]. A study assessed landslide and forest fire vulnerability in Southeast Asia using the AdaBoost algorithm and the Adaptive Resampling method, achieving 74% ac-

curacy [14]. Previous research predicted fire ignition events from lightning forecasts using the AdaBoost algorithm and Grid Search method, achieving 72.95% accuracy [15].

Based on several previous studies mentioned above, two studies show relatively robust performance with high accuracy: research [9] using the SVM algorithm and the Randomized Search method and research [12] using the XGBoost algorithm and the SMOTE method. Although previous studies have shown relatively good performance, there is potential to further improve the accuracy of machine learning algorithms in predicting forest and land fires. In addition to the aforementioned challenges, other significant issues in prediction are data imbalance and hyperparameter tuning. According to a related study [16], the SMOTE-ENN method outperforms the single SMOTE method in handling the problem of data class imbalance. Research [17] indicates that the Random Search method is more effective than Grid Search for hyperparameter optimization across different learning algorithms and datasets. Thus, this study suggests integrating the SMOTE-ENN and Random Search approaches to enhance the accuracy of predicting forest and land fire potential.

The SVM, Random Forest, and XGBoost algorithms have been widely applied in previous research to build prediction models due to their respective advantages. SVM can effectively classify both linear and non-linear data [18]. The Random Forest algorithm can produce relatively low error with high performance [19]. Moreover, the XGBoost algorithm can reduce model complexity and prevent over-fitting [20]. Meanwhile, the AdaBoost algorithm is faster, easy to operate, and simple to implement [21]. Based on these advantages, SVM, Random Forest, XGBoost, and AdaBoost algorithms should be compared for predicting potential forest and land fires based on patrol data.

To address the above problems, this study aims to develop a prediction model for potential forest and land fires using SVM, Random Forest, AdaBoost, and XGBoost algorithms while employing the SMOTE-ENN and Random Search methods. This study will analyze how the application of these methods affects the accuracy of the prediction model based on patrol data to obtain the best prediction model. Furthermore, the results of this study will provide information about areas with potential for forest and land fires. It is expected that the findings of this study will help create more efficient strategies for preventing forest and land fires in Kalimantan and Sumatra.

2 Research Method

This section presents the steps for developing a prediction model for potential forest and land fires using machine learning algorithms. The study began by collecting sample patrol data in August 2023 from the Land and Forest Fire Prevention Patrol Information System. The August sample patrol data was chosen because it contains the highest amount of patrol data in 2023. After the data collection stage, the next step is data exploration to describe the amount of potential class data and the number of forest and land fire occurrences. The patrol data ready for use in this study results from the data preprocessing stage, which includes data transformation, addressing class imbalance, handling missing values, and handling outliers.

Meanwhile, the method used in this study to address data class imbalance is the Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN). Furthermore, patrol data will be divided into 80% training data and 20% testing data at the data partition stage. Then, 80% of the training data is used in the k-fold cross-validation

stage to reduce biases when developing a prediction model for potential forest and land fires. The k-fold cross-validation uses a stratification method to ensure balanced class distribution in each subset. This study uses SVM, Random Forest, AdaBoost, and XGBoost algorithms to develop a prediction model for potential forest and land fires. The SVM prediction models also include the Linear kernel, Polynomial kernel, Radial Basis Function (RBF) kernel, and Sigmoid kernel. The prediction model that has been constructed is then optimized using the best hyperparameters generated from the hyperparameter tuning process. The hyperparameter tuning in this study uses the Random Search method to determine the best hyperparameter combination randomly. After the prediction model is successfully optimized, it is tested using test data to predict the potential for forest and land fires. In addition, the prediction model for forest and land fire potential is evaluated using the Confusion Matrix method. Furthermore, this study employs several comparison scenarios to assess the impact of applying the SMOTE-ENN and Random Search methods on the accuracy of each prediction model. The final stage of this study involves visualizing the best model prediction results through spatial plots, displaying the amount of data per prediction class, and exploring variable correlations. Figure 1 illustrates the phases of this study.

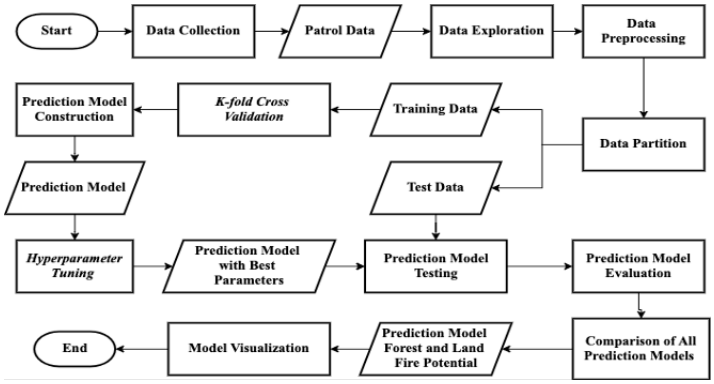


Figure 1: Stages of building a prediction model of forest and land fire potential.

2.1 Data Collection

The Ministry of Environment and Forestry provided closed-source secondary patrol data for this study. Patrol data was obtained from the Land and Forest Fire Prevention Patrol Information System website (sipongi.menlhk.go.id/sipp-kebakaran-hutan-and-lahan). Forest and land fire patrol data covers Sumatra and Kalimantan. Patrol data consists of climatic and environmental factors. Then, for patrol data attributes used as predictors consist of rainfall (mm), temperature (°C), humidity (%), wind speed (km/h), soil type, soil condition, peat depth (m), land slope (degrees), morning weather, daytime weather, afternoon weather, vegetation type, and vegetation condition. The potential for land and forest fires, classified into low, medium, high, and extreme classes, is also the target attribute. Then, the attributes used to visualize the prediction results are latitude, longitude, patrol date, province, district, sub-district, and village.

2.2 Data Exploration

Data exploration is done after data collection. Data exploration aims to comprehend the properties of the data. Part of the exploration process is to provide information on the frequency of forest and land fires by presenting data on the potential classes of forest and land fires and the potential for daily occurrences of forest and land fires.

2.3 Data Preprocessing

Data preprocessing consists of several activities, such as identifying and handling outliers and identifying and handling missing values. Identifying data items that deviate from expected or typical behavior is known as outlier detection [22]. Missing value handling is performed using imputation techniques for numeric and categorical attributes. Furthermore, data transformation is performed to convert categorical data into numerical data. Additionally, the Synthetic Minority Oversampling Technique with the Edited Nearest Neighbor approach (SMOTE-ENN), which combines under-sampling and oversampling to improve model performance, addresses data imbalances [23].

2.4 Data Partition and K-fold Cross Validation

The data partition stage uses resampled datasets with 80% training and 20% test data. This ratio was selected because it has been extensively used in previous research and has shown positive outcomes. After that, 80% of the training data from the earlier data partition is used in the k-fold cross-validation stage. K-fold cross-validation can provide a more stable estimation of model performance by using various training and test data combinations. This research employs a stratified k-fold cross-validation method that maintains the distribution of target classes, with k = 5 folds.

2.5 Prediction Model Construction

The model construction stage utilizes SVM algorithms (Linear, Polynomial, RBF, and Sigmoid), Random Forest, AdaBoost, and XGBoost. The model construction utilizes training and validation data from the stratified k-fold cross-validation process. Subsequently, the average accuracy of all models is computed in each iteration. The following is an explanation of these algorithms:

2.5.1 Support Vector Machine (SVM)

The SVM algorithm aims to identify the hyperplane with the largest margin. A hyperplane is a line that separates data between classes or categories [24]. Then, the SVM can be seen in Equation 1 and Equation 2.

$$(w \cdot x_i + b) \leq 1, y_i = -1 \tag{1}$$

$$(w \cdot x_i + b) \geq 1, y_i = 1 \tag{2}$$

Description:
x_i : The i-th data of the dataset

$w \cdot x_i$: Weight value for the i-th data class
b : Bias value
y_i : The class of the i-th data

2.5.2 Random Forest (RF)

When several decision trees are combined using random forests, the decision tree models are executed in parallel. This results in a prediction that is the target class mode or mean prediction for the regression problem [25]. Random Forest through forest formation using Equation 3 [26].

$$\text{forest} = \{h(x, \Theta_k), k = 1, \cdots\} \tag{3}$$

Description:
h : Hypothesis or classifier
x : Input vector
Θ_k : Independent and identically distributed (IID) random vectors

2.5.3 XGBoost (XGB)

XGBoost uses gradient-enhanced decision trees for regression analysis and classification [27]. XGB reduces model complexity to avoid over-fitting [28]. Equation 4 gives the overall value for XGBoost.

$$L^(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

Description:
n : Number of models to be used
l : Function to measure the difference between target prediction y_i and $\hat{y}_i f_t(x_i)$
Ω : Function to create a spared model of overfitting

2.5.4 AdaBoost

AdaBoost is an algorithm that improves weak classification algorithms by enhancing the classification ability of data through continuous training [29]. Equation 5 gives the overall value for AdaBoost.

$$F_n(x) = F_{m-1}(x) + \text{argmin} \sum_{i=1}^{n} L(y_i F_{m-1}(x_i) + h(x_i)) \tag{5}$$

Description:
$F_n(x)$: Overall model
$F_{m-1}(x)$: Overall obtained in the previous round
y_i : Prediction result of the i-th tree
$h(x_i)$: Newly added trees

2.6 Hyperparameter Tuning

The prediction model that has been built previously will be optimized using the best hyperparameter combination obtained from the hyperparameter tuning process. Hyperparameter tuning in this study uses the Random Search method. Random Search is used to combine random samples of parameter values. Random Search can help explore the hyperparameter space more efficiently [30]. Hyperparameter combinations are tested based on the hyperparameters of each prediction model.

2.7 Prediction Model Testing

The prediction model testing stage is performed on SVM models (Linear, Polynomial, RBF, and Sigmoid), Random Forest, AdaBoost, and XGBoost. Each prediction model utilizes test data to estimate the potential of land and forest fires. The process begins with predicting the potential for land and forest fires, combining test data and prediction findings, and integrating characteristics identified in the outcomes of each prediction model.

2.8 Evaluation and Comparison of All Prediction Models

The confusion matrix method is used for the prediction model evaluation. At this point, accuracy values will be computed. The precision value indicates the accuracy of the model for each class. The maximum number of cases in each class can be found using the recall value. The F1 score shows how well the model balances precision and recall [31]. The confusion matrix method includes TP (True Positive) for correctly predicted positives, TN (True Negative) for correctly predicted negatives, FP (False Positive) for incorrectly predicted positives, and FN (False Negative) for incorrectly predicted negatives. In this study, the evaluation stage was conducted for five classes: low, medium, high, and extreme. Equation 6, Equation 7, Equation 8, and Equation 9 also compute accuracy, precision, recall, and F1 Score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

The next stage involves comparing the accuracy of SVM (Linear, Polynomial, RBF, and Sigmoid), Random Forest, AdaBoost, and XGBoost models using SMOTE-ENN and Random Search methods. The results of the comparison are to get the best prediction model.

2.9 Model Visualization

This stage aims to display the spatial plot of the best model prediction results as an interactive map using the Folium library (Python) and leveraging the Leaflet library (JavaScript). Marker dots in the spatial representation of the best model prediction outcomes indicate

the potential of land and forest fires in Sumatra and Kalimantan. Plotting the amount of prediction data per class simultaneously shows the quantity of data from the best prediction model for each class.

3 Results

3.1 Data Collection

The data collected are patrol data from January to September 2023, totaling 18,202 data. The amount of patrol data in January (10 data), February (117 data), March (1,269 data), April (1,668 data), May (2,030 data), June (2,914 data), July (2,474 data), August (6,032 data), and September (1,688 data). Based on this data, the most significant amount of patrol data was recorded in August 2023, with 6,032 data. Therefore, this study uses the patrol data from August as a sample. The forest and land fire patrol data used covers the islands of Kalimantan and Sumatra. Table 1 shows an example of patrol data from the Land and Forest Fire Prevention Patrol Information System.

Table 1: Example of patrol data

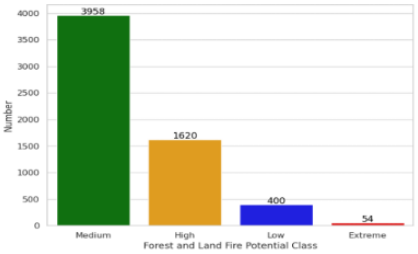
Id	Lat.	Long.	Province	Rainfall	Temperature	···	Potential
1	-2.523	112.93	Central Kalimantan	0.2967	31.98	...	Medium
2	0.445	111.35	West Kalimantan	26.9	32	...	Low
...	...	...	...	...	...	...	...
6031	-2.705	104.41	South Sumatra	0	33	...	High
6032	-1.465	113.85	Central Kalimantan	0	30.54	...	Extreme

3.2 Data Exploration

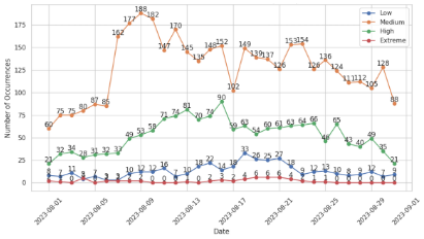
The exploration stage illustrates the amount of forest and land fire data for each class through visual representation. Based on the visualization results, there is an imbalance in the data classes, with the medium class being dominant, totaling 3,958 data. Then, visualize the number of occurrences of potential forest and land fire classes per day to show the daily fluctuations of potential forest and land fires in August 2023. The visualization results show the highest number of daily forest and land fire events for low (33 data), medium (188 data), high (128 data), and extreme (6 data) classes. Figure 2 displays the data exploration outcomes.

3.3 Data Preprocessing

Data preprocessing begins with outlier identification using box plots on the numeric variables. The results show the number of outliers above the upper limit for the variables rainfall (816), temperature (29), humidity (3), wind speed (132), land slope (67), and peat depth (60). The number of outliers is below the lower limit for the temperature (195) and humidity (15) variables. The IQR (Interquartile Range) method is used to handle outliers by replacing them with predetermined upper or lower limit values. If the lower limit value is negative, it will be replaced with the minimum value of the numerical variable. The results of outlier identification and handling can be seen in Figure 3.

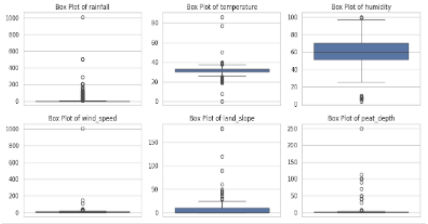


(a) Number of potential class data

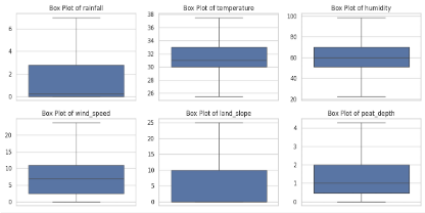


(b) Number of potential class occurrences per day

Figure 2: Results of data exploration using visualization.



(a) Outlier identification results



(b) Outlier handling results

Figure 3: Results of identifying and handling outliers.

Furthermore, this study identifies and handles variables with missing values using a specific method. Missing value handling for numerical variables uses the average value (mean), while categorical variables use the most frequently occurring value (mode). Meanwhile, geographical distance is calculated to handle outliers in sub-district and village variables. Table 2 presents the results of the identification of the missing value.

Table 2: Missing value identification results

Variables	Missing values
District	56
Village	605
Rainfall	601
Soil Type	4,013
Soil Condition	4,073
Land Slope	5,309
Peat Depth	5,064
Vegetation Type	2,040
Vegetation Condition	2,219

After handling missing values, data transformation is performed to convert categorical data into numerical data. This stage uses the one-hot encode method for nominal data and label encode for ordinal data. The result of label encoding for the land and forest fire potential variable involves reclassifying the extreme category as class 3, the high category as class 2, the medium category as class 1, and the low category as class 0. In addition,

this research utilizes the SMOTE-ENN method to address class imbalance. ENN removes outlier data (undersampling), and SMOTE adds synthesized data (oversampling). In this study, the percentage of the SMOTE-ENN method used is based on experiments conducted. This percentage indicates the amount of synthetic data added or removed, which can affect model accuracy. Table 3 shows the percentage, number of original data, and number of data resulting from applying SMOTE-ENN for each class.

Table 3: Result of handling data class imbalance

Class	SMOTE Percentage	ENN Percentage	Original Data	SMOTE-ENN Result Data
Low	3.90%	-	400	1,808
Medium	-	33.61%	3,958	1,937
High	0.68%	-	1,620	1,869
Extreme	33.8%	-	54	1,802
Total data			6,032	7,416

3.4 Data Partition and K-fold Cross Validation

In the data partition stage, resampling data was used (7,416 data points), with 80% of the data used for training and 20% for testing. The result is 5,932 training data and 1,484 testing data. Next, a stratified k-fold cross-validation method is employed to ensure that each fold maintains a balanced distribution of target classes throughout the cross-validation process [32]. This stage uses 5,932 data (80% of the previous training data) divided into k = 5 equal-sized parts (folds). The process resulted in 4,745 training data and 1,187 validation data.

3.5 Prediction Model Construction

The construction phase of all prediction models used training data (4,745 data) and was tested using validation data (1,187 data) generated from the stratified k-fold cross-validation process. All prediction models were tested using the fit and predict met hod. Then, the accuracy of each iteration is calculated, and the average accuracy of each prediction model is estimated. The XGBoost model achieved the highest accuracy at 93.80%. On the other hand, the accuracies of the Random Forest, AdaBoost, SVM-Linear, SVM-Polynomial, SVM-RBF, and SVM-Sigmoid models are 92.87%, 85.39%, 78.24%, 55.85%, 52.71%, and 27.32%, respectively. However, the prediction model has not used the best hyperparameter combination, so it needs to undergo optimization in the hyperparameter tuning stage. A plot of the average accuracy of each model can be seen in Figure 4.

3.6 Hyperparameter Tuning

The prediction model that has been previously built is optimized using the best hyperparameters identified through the Random Search method. The Random Search method is used to try random combinations to obtain the best hyperparameters. The results of hyperparameter tuning identified the best hyperparameters for SVM models (Linear, Polynomial, and RBF), specifically the C-parameter (12) and gamma (scale). In contrast, the

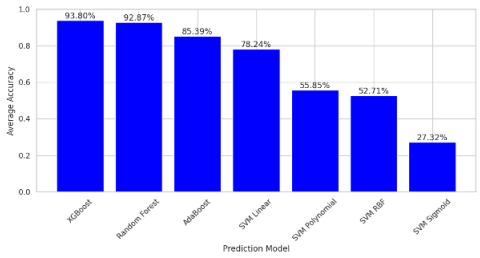


Figure 4: Plot of the average accuracy of the prediction model.

SVM-Sigmoid model has C-parameters (3) and gamma (scale). Then, the best hyperparameters for the Random Forest model are n-estimators (200), max depth (20), max features (log2), min samples leaf (1), and min samples split (5). The best hyperparameters for the AdaBoost model are n-estimators (100), max depth (6), learning rate (0.1), and random state (42). Furthermore, the best hyperparameters for the XGBoost model are n-estimators (400), max depth (30), learning rate (0.05), subsample (0.6), scale pos weight (5), and gamma (0.2).

3.7 Prediction Model Testing

During the prediction model testing stage, test data (1,484 data) are utilized with the prediction method. The predictions from each model will be stored in the "predictions" column and saved as a prediction result dataframe. Next, the prediction results and test data are integrated based on the index of the data. Then, the integration of latitude, longitude, province, district, sub-district, village, and patrol date features is performed on the data frame to facilitate visualization of the prediction results. Additionally, data containing NaN (missing values) was removed, resulting in 1,218 data. For example, the test results on the XGBoost prediction model can be seen in Table 4.

Table 4: XGBoost prediction model test results

Id	Lat.	Long.	Province	Rainfall	Temperature	···	Potential	Prediction
1	2.452	112.93	North Sumatra	6.97	27	...	0	0
2	-2.232	102.93	Jambi	0.20	33	...	1	1
...	...	...	...	...	...	...	...	...
1217	0.962	104.41	Riau	3.12	32	...	2	2
1218	2.777	113.85	South Kalimantan	0	33	...	3	3

3.8 Evaluation and Comparison of All Prediction Models

The confusion matrix method is a performance metric for evaluating SVM (Linear, Polynomial, RBF, and Sigmoid), Random Forest, AdaBoost, and XGBoost prediction models. The confusion matrix method is used to calculate the precision, recall, f1 score, and accuracy values for each prediction model. A confusion matrix is calculated using test data and prediction results, paying attention to the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. The accuracy values for each prediction model are presented in Table 5.

Table 5: Prediction model accuracy values (%)

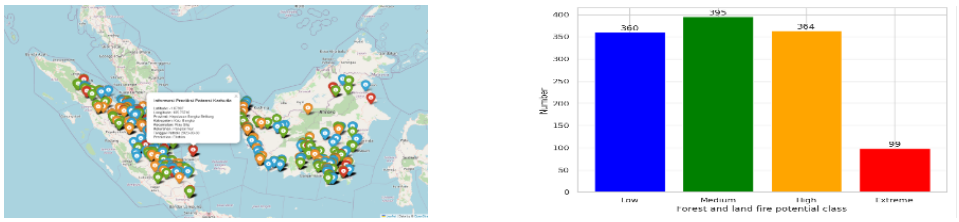
Prediction model	Prediction Model Accuracy		
	Without SMOTE-ENN and Random Search	Using SMOTE-ENN Without Random Search	Using SMOTE-ENN and Random Search
SVM-Linear	69.87%	78.24%	79.65%
SVM-Polynomial	65.62%	55.85%	76.82%
SVM-RBF	65.62%	52.71%	76.75%
SVM-Sigmoid	64.64%	27.32%	33.83%
AdaBoost	67.92%	85.39%	91.24%
Random Forest	77.60%	92.87%	94.95%
XGBoost	79.71%	93.80%	95.55%

The prediction model comparison stage aims to evaluate the accuracy of SVM (Linear, Polynomial, RBF, and Sigmoid), AdaBoost, Random Forest, and XGBoost models using SMOTE-ENN and Random Search methods. The model comparison involves three scenarios: modeling without SMOTE-ENN and Random Search, modeling with SMOTE-ENN but without Random Search, and modeling with a combination of SMOTE-ENN and Random Search. Based on Table 5, the model before applying SMOTE-ENN and Random Search exhibits reasonably high accuracy, but this accuracy metric only reflects the performance of the majority class. After applying SMOTE-ENN and Random Search, the XGBoost algorithm achieved the highest accuracy compared to other models, reaching 95.55%. The high accuracy of the XGBoost prediction model indicates its excellent performance. The results show that the SMOTE and ENN methods are successfully used to generate synthetic samples of minority classes and clean the dataset from noise when dealing with data class imbalance. The Random Search method is successfully employed to identify the best hyperparameter combination for optimizing the prediction model based on the hyperparameter space. The accuracy results of other prediction models after applying SMOTE-ENN and Random Search, such as the accuracy of Random Forest, AdaBoost, SVM-Linear, SVM-Polynomial, SVM-RBF, and SVM-Sigmoid, are 94.95%, 91.24%, 79.65%, 76.82%, 76.75%, and 33.83% respectively. The low accuracy of the model suggests its complexity, rendering it ineffective for identifying patterns in patrol data.

Based on these results, the XGBoost model is the best model for predicting potential forest and land fires on the islands of Sumatra and Kalimantan based on patrol data. The findings show that applying SMOTE-ENN and Random Search methods can improve the accuracy of the XGBoost model. Another factor that makes the XGBoost model perform better than other models is its use of boosting techniques. The boosting technique allows the XGBoost model to continuously improve its performance by minimizing previous errors, thus achieving better accuracy.

3.9 Model Visualization

At the model evaluation and comparison stage, XGBoost proves to be the best model. The spatial plot of the XGBoost model prediction results is used to display marker points that indicate the potential for land and forest fires in Sumatra and Kalimantan. The marker points represent different potential classes: blue for low, green for medium, orange for high, and red for extreme. The marker points are used to display information about latitude,



(a) Spatial plot of prediction results (b) Plot of the number of data per predicted class

Figure 5: Visualization of best model prediction results.

longitude, province, district, sub-district, village, patrol dates, and predictions of potential land and forest fires. The data obtained from predicting potential forest and land fires is as follows: 360 instances in the low class, 395 in the medium class, 364 in the high class, and 99 in the extreme class. Figure 5 shows the spatial plot and data plot per class of the XGBoost model.

The spatial plot visualization of XGBoost model predictions successfully displayed the potential for forest and land fires on the islands of Sumatra and Kalimantan. Information on the potential for forest and land fires in each region can be used to help prevent, monitor, and suppress forest and land fires in Sumatra and Kalimantan. Then, it can help determine the appropriate areas for resource management and assist in making decisions for planning forest and land fire mitigation policies in Sumatra and Kalimantan. Areas with the most potential points can be seen in Table 6 and Table 7.

Table 6: Regions with the most potential points on the island of Sumatra

Potential Class	Province	District	Number of Occurrences
Low	Riau	Siak	18
Middle	Riau	Dumai	14
High	Riau	Rokan Hulu	15
Extreme	Riau	Dumai	7

Table 7: Regions with the most potential points on the island of Kalimantan

Potential Class	Province	District	Number of Occurrences
Low	West Kalimantan	Sambas	20
Middle	West Kalimantan	Sambas	16
High	South Kalimantan	Banjar	16
Extreme	West Kalimantan	Kubu Raya	5

Furthermore, the variable correlation plot can be used to analyze the correlation between each predictor variable and the target variable (potential prediction result) in the XGBoost model. Figure 6 shows the ten variables most correlated with predicting the potential for forest and land fires. The variable correlations are as follows: dry vegetation condition (10.69%), soil conditions associated with forest and land fires (8.88%), moist vegetation condition (7.95%), rubber vegetation type (5.77%), cloudy afternoon weather (4.77%), peat depth (4.09%), peat soil type (3.37%), cloudy sunny afternoon weather (3.14%), soil con-

ditions prone to forest and land fires (3.00%), and sunny afternoon weather (2.98%). This study also identified dry vegetation condition as the variable with the highest correlation, although the strength of this correlation was relatively low at 10.69%.

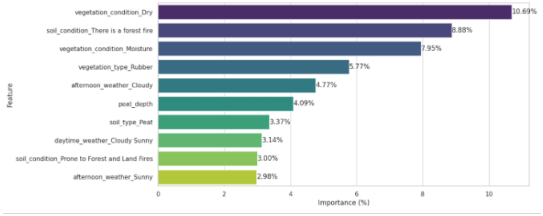


Figure 6: Correlation plot of predictor variables with the target variable.

4 Discussion

This research successfully utilizes patrol data from the Forest and Land Fire Prevention Patrol Information System to build an optimal forest and land fire potential prediction model. Then, this research successfully constructed a prediction model for potential forest and land fires using SVM, Random Forest, AdaBoost, and XGBoost algorithms, applying the SMOTE-ENN and Random Search methods. The results show that the SMOTE-ENN method has been successfully used to handle data class imbalance. The Random Search method was successfully used to obtain the best hyperparameter combination for model optimization and improve the accuracy of the prediction model. The prediction model comparison stage also benefits by revealing the performance of each model in predicting the potential for forest and land fires, thus identifying the best prediction model, namely the XGBoost model. Then, the most influential variable is dry vegetation conditions, indicating a need to prioritize these variables in managing forest and land fires. The results of this study will also provide information about areas with the potential for forest and land fires, which can be used to develop more effective strategies for early prevention of such fires on the islands of Kalimantan and Sumatra.

This research uses climate and environmental factors to build a prediction model. Future research could incorporate social factors, such as community activities, as contributors to forest and land fires. In addition, future research could use other methods, such as SMOTE-Tomek Link with Grid Search, to improve the performance of the prediction model. The current model could also be compared with different algorithms, such as LightGBM or CatBoost, for better prediction model performance. Further research is expected to implement the best prediction model in the Forest and Land Fire Prevention Patrol Information System to enable real-time predictions.

There is a difference between this research and previous studies in that this research not only compares SVM, Random Forest, and XGBoost algorithms. This research also examines the impact of applying the SMOTE-ENN and Random Search methods on the accuracy of each prediction model. Based on Table 8, combining the SMOTE-ENN and Random Search methods in prediction provides higher accuracy than previous studies. Another difference is that the output of this research is not only in the form of prediction results but also visualized as spatial plots used to display information related to potential forest and land

fire locations. The spatial plot can be used as a reference to implement in the Land and Forest Fire Prevention Patrol Information System as a prediction module for potential land and forest fires.

Table 8: Results comparison with previous research

Previous Study	Methods	Scope of Study	Accuracy
Shirazi, Wang and Bondur [8]	SVM + SMOTE	Forest and Land Fires	62.7%
Davis [9]	SVM + Random Search		90.9%
Bountzouklis, Fox and Bernardino [10]	Random Forest + SMOTE		70%
Lee et al. [11]	Random Forest + Grid Search		88%
Xu, Zhou and Zhang [12]	XGBoost + SMOTE		88.8%
Pham et al. [13]	XGBoost + ABC-ANFIS		81.44%
He et al. [14]	AdaBoost + Adaptive Resampling		74%
Coughlan et al. [15]	AdaBoost + Grid Search		72.95%
Proposed Method	XGBoost + SMOTE-ENN + Random Search		95.55%

5 Conclusion

This research successfully built a prediction model for predicting potential forest and land fires using SVM, Random Forest, AdaBoost, and XGBoost algorithms based on patrol data. This research utilizes the Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) method, effectively overcoming the imbalance of data classes. Additionally, the Random Search method was effectively employed to identify the optimal hyperparameter combination, enhancing the accuracy of the prediction model.

The performance of the prediction model can be observed by comparing the accuracy before and after applying the SMOTE-ENN and Random Search methods to each prediction model. The best model accuracy is the XGBoost model, with an accuracy of 95.55% after applying the SMOTE-ENN and Random Search methods. While other prediction models, such as Random Forest, AdaBoost, SVM-Linear, SVM-Polynomial, SVM-RBF, and SVM-Sigmoid, have accuracies of 94.95%, 91.24%, 79.65%, 76.82%, 76.75%, and 33.83%, respectively. The low accuracy of the model suggests its complexity, rendering it ineffective for identifying patterns in patrol data. Based on these results, the XGBoost model provides the best predictions for potential forest and land fires in Sumatra and Kalimantan based on patrol data. The results of this study suggest that the combined use of SMOTE-ENN and Random Search methods effectively improved the accuracy of the XGBoost model. Another factor contributing to the XGBoost model outperforming other prediction models is its boosting techniques. Boosting techniques can be used to improve model performance by minimizing the difference between actual data and predicted results (residuals). This

study also identified dry vegetation condition as the variable with the highest correlation, although the strength of this correlation was relatively low at 10.69%. In addition, climatic and environmental factors also affect the prediction results of this study, so the results may differ in different regions.

Acknowledgments

Thanks to the Ministry of Education, Culture, Research, and Technology for funding this study and research through the Indonesian Education Scholarship research fund Number 03108/J5.2.3./BPI.06/10/2022.

References

[1] R. M. Sukarna, N. Hidayat, and M. S. Tambunan, "Kondisi hutan tropis lahan kering berdasarkan struktur dan komposisi jenis tegakan (studi kasus pada pt. sindo lumber provinsi kalimantan tengah, indonesia)," *Journal of Environment and Management*, vol. 3, no. 1, pp. 80–88, 2022.

[2] http://journal.ummgl.ac.id/index.php/urecol/article/view/719/804%0Ahttp://www.forestprogramme.com/files/2011/05/FOREST-Standard-Guide_V04_UK.pdf. [Accessed 23-09-2024].

[3] H. Nugraheni, "Deforestasi dan peran kesatuan pengelolaan hutan (kph) sivia patuju untuk mengatasinya," *Jurnal Planoearth*, vol. 5, no. 2, pp. 62–68, 2020.

[4] H. A. Adrianto, D. V. Spracklen, S. R. Arnold, I. S. Sitanggang, and L. Syaufina, "Forest and land fires are mainly associated with deforestation in riau province, indonesia," *Remote Sensing*, vol. 12, no. 1, p. 3, 2019.

[5] hantoro.aji@gmail.com, "Peraturan Lingkungan Hidup dan Kehutanan - [jdih.menlhk.go.id] — jdih.menlhk.go.id." https://jdih.menlhk.go.id/new2/home/portfolioDetails/32/2016/4. [Accessed 23-09-2024].

[6] T. A. Pratiwi, M. Irsyad, and R. Kurniawan, "Klasifikasi kebakaran hutan dan lahan menggunakan algoritma naïve bayes (studi kasus: Provinsi riau)," *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, vol. 9, no. 2, pp. 101–107, 2021.

[7] I. S. Sitanggang, L. Syaufina, R. Trisminingsih, D. Ramdhany, E. Nuradi, M. F. A. Hidayat, H. Rahmawan, Wulandari, F. Ardiansyah, I. Albar, *et al.*, "Indonesian forest and land fire prevention patrol system," *Fire*, vol. 5, no. 5, p. 136, 2022.

[8] Z. Shirazi, L. Wang, and V. G. Bondur, "Modeling conditions appropriate for wildfire in south east china–a machine learning approach," *Frontiers in Earth Science*, vol. 9, p. 622307, 2021.

[9] M. Davis and M. Shekaramiz, "Desert/forest fire detection using machine/deep learning techniques," *Fire*, vol. 6, no. 11, p. 418, 2023.

[10] C. Bountzouklis, D. M. Fox, and E. Di Bernardino, "Predicting wildfire ignition causes in southern france using explainable artificial intelligence (xai) methods," *Environmental Research Letters*, vol. 18, no. 4, p. 044038, 2023.

[11] C. Lee, S. Park, T. Kim, S. Liu, M. N. Md Reba, J. Oh, and Y. Han, "Machine learning-based forest burned area detection with various input variables: A case study of south korea," *Applied Sciences*, vol. 12, no. 19, p. 10077, 2022.

[12] Y. Xu, K. Zhou, and F. Zhang, "Modeling wildfire initial attack success rate based on machine learning in liangshan, china," *Forests*, vol. 14, no. 4, p. 740, 2023.

[13] Van The Pham, T. A. T. Do, H. D. Tran, and A. N. T. Do, "Classifying forest cover and mapping forest fire susceptibility in dak nong province, vietnam utilizing remote sensing and machine learning," *Ecol. Inform.*, vol. 79, p. 102392, Mar. 2024.

[14] Q. He, Z. Jiang, M. Wang, and K. Liu, "Landslide and wildfire susceptibility assessment in southeast asia using ensemble machine learning methods," *Remote Sensing*, vol. 13, no. 8, p. 1572, 2021.

[15] R. Coughlan, F. Di Giuseppe, C. Vitolo, C. Barnard, P. Lopez, and M. Drusch, "Using machine learning to predict fire-ignition occurrences from lightning forecasts," *Meteorological applications*, vol. 28, no. 1, p. e1973, 2021.

[16] M. Prabha and Sasikala, "Data analytics for imbalanced dataset," *J. Comput. Sci.*, vol. 20, pp. 207–217, Feb. 2024.

[17] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[18] J. Mase, M. T. Furqon, and B. Rahayudi, "Penerapan algoritme support vector machine (svm) pada pengklasifikasian penyakit kucing," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3648–3654, 2018.

[19] F. Y. Pamuji and V. P. Ramadhan, "Komparasi algoritma random forest dan decision tree untuk memprediksi keberhasilan immunotheraphy," *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 1, pp. 46–50, 2021.

[20] S. E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan metode extreme gradient boosting (xgboost) pada klasifikasi nasabah kartu kredit," *Journal of Mathematics: Theory and Applications*, pp. 21–26, 2022.

[21] T. Chengsheng, L. Huacheng, and X. Bing, "Adaboost typical algorithm and its application research," in *MATEC Web of Conferences*, vol. 139, p. 00222, EDP Sciences, 2017.

[22] A. F. Hassan, S. Barakat, and A. Rezk, "Towards a deep learning-based outlier detection approach in the context of streaming data," *Journal of Big Data*, vol. 9, no. 1, p. 120, 2022.

[23] Z. Darojah, R. Susetyoko, and N. Ramadijanti, "Strategi penanganan imbalance class pada model klasifikasi penerima kartu indonesia pintar kuliah berbasis neural network menggunakan kombinasi smote dan enn," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 457–466, 2023.

[24] C. Cortes, "Support-vector networks," *Machine Learning*, 1995.

[25] J. M. Rudd *et al.*, "An empirical study of downstream analysis effects of model preprocessing choices," *Open journal of statistics*, vol. 10, no. 5, pp. 735–809, 2020.

[26] M. Mara, A. Nursyahid, T. Setyawan, A. Sriyanto, *et al.*, "Adjustment pattern of ph using random forest regressor for crop modelling of nft hydroponic lettuce," in *Journal of Physics: Conference Series*, vol. 1863, p. 012075, IOP Publishing, 2021.

[27] N. R. Octavianto and A. Wibowo, "Stacking classifier method for prediction of human body performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 34, p. 1832, June 2024.

[28] C. Qin, Y. Zhang, F. Bao, C. Zhang, P. Liu, and P. Liu, "Xgboost optimized by adaptive particle swarm optimization for credit scoring," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 6655510, 2021.

[29] C. Wang, S. Xu, and J. Yang, "Adaboost algorithm in artificial intelligence for optimizing the iri prediction accuracy of asphalt concrete pavement," *Sensors*, vol. 21, no. 17, p. 5682, 2021.

[30] C. Arnold, L. Biedebach, A. Küpfer, and M. Neunhoeffer, "The role of hyperparameters in machine learning models and how to tune them," *Political Science Research and Methods*, pp. 1–8, 2023.

[31] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Oxford, England: Morgan Kaufmann, 3 ed., June 2012.

[32] S. Szeghalmy and A. Fazekas, "A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning," *Sensors*, vol. 23, no. 4, p. 2333, 2023.