

Sensitivity and feature importance of climate factors for predicting fire hotspots using machine learning methods

Endar Hasafah Nugrahani, Sri Nurdyati, Fahren Bukhari, Mohamad Khoirun Najib, Denny Muliawan
Sebastian, Putri Afia Nur Fallahi

Department of Mathematics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Jul 29, 2023

Revised Oct 29, 2023

Accepted Jan 6, 2024

Keywords:

Bayesian regression

Feature importance

Machine learning

Sensitivity analysis

Wildfire

ABSTRACT

Every year, Indonesia experiences a national crisis due to forest fires because the resulting impacts and losses are enormous. Hotspots as indicators of forest fires capable of quickly monitoring large areas are often predicted using various machine learning methods. However, there is still few research that analyzes the sensitivity and feature importance of each predictor that forms a machine learning prediction model. This study evaluates and compares machine learning methods to predict hotspots in Kalimantan based on local and global climate factors in 2001-2020. Using the most accurate machine learning model, each climate factor used as a predictor is analyzed for its sensitivity and feature importance. Four methods used include random forest, gradient boosting, Bayesian regression, and artificial neural networks. Meanwhile, measures of sensitivity and feature importance used are variance, density, and distribution-based sensitivity indices, as well as permutation and Shapley feature importance. Evaluation of the machine learning model concluded that the Bayesian linear regression model outperformed other models with an RMSE of 750 hotspots and an explained variance score of 68.96% on testing data. Meanwhile, tree-based models show signs of overfitting, including gradient boosting and random forest. Based on the results of sensitivity analysis and feature importance of the Bayesian linear regression model, the number of dry days is the most important feature in predicting fire hotspots in Kalimantan.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sri Nurdyati

Department of Mathematics, Faculty of Mathematics and Natural Sciences, IPB University

Meranti Kampus Road, Babakan, Dramaga, Bogor Regency, West Java 16680, Indonesia

Email: nurdiati@apps.ipb.ac.id

1. INTRODUCTION

In the last three decades, Indonesia is heavily affected by land and forest fires. There have been three remarkable land and forest fires reported in 1997–1998, 2015, and 2019. Nonetheless, Indonesia always experiences land and forest fires, particularly in Sumatra and Kalimantan [1]. Forest fires are a major environmental problem with significant impacts on the atmosphere, carbon cycle, and various ecosystem benefits. The haze caused by forest fires causes short to long-term health problems, as well as causing economic losses, and even affects neighboring countries [2]. Therefore, it is vital to know the indications of forest fires in order to reduce their impact.

One of the factors causing forest fires is climatic conditions such as temperature, humidity, and rainfall, which can affect surface dryness [1]. Climate is the average condition of temperature, rainfall, pressure,

humidity, wind direction, and other climate parameters over a long period. Meanwhile, climate change is the term used to describe shifts in the climate that are caused, either directly or indirectly, by human activity, causing changes in the composition of the atmosphere and increasing climate variability over a long period. Indonesia is included in the category of countries that are very vulnerable to climate change can be seen from Indonesia's location, which lies between the Pacific and Indian oceans. As a result, the climate on land is influenced by ocean phenomena such as the Indian Ocean Dipole (IOD) and the El Nino Southern Oscillation (ENSO).

ENSO is defined by sea surface temperatures that are higher or lower than normal in the eastern Pacific Ocean [3]. El Nino, or rising temperatures and humidity in the Pacific Ocean, can lead to abnormally low rainfall and a protracted dry season in a number of Indonesian regions. Previous studies have shown that El Nino affects fires in Kalimantan [4], such as the great fires in 1997 and 2015 [5]. Meanwhile, IOD, an atmospheric-oceanic phenomenon in the equatorial region of the Indian Ocean, can have an impact on the climate of Indonesia and other nations surrounding the Indian Ocean. IOD is important to the condition of Indonesia's seasons, along with the ENSO phenomenon [6].

The need for a forest fire prediction model is considered necessary to reduce its impact on society, such as death of flora and fauna, haze which affects the health of local residents, and deforestation which has long-term impacts. Researchers have developed models of forest fires, including the development of a probabilistic multilayer perceptron model utilizing fifth-generation seasonal forecasting system (SEAS5) from ECMWF [7], modeling of carbon emissions based on climate indicators in Sumatra with random forests and artificial neural networks [8], and modeling of hotspots in Kalimantan using Bayesian inference based on precipitation, relative dry spells, ENSO and IOD [9]. However, of the various models offered, not many have conducted a deeper analysis of the models obtained, such as analysis of the sensitivity and feature importance of each predictor or climate indicator used. Thus, the effect or influence of the predictors mentioned above is not seen in more detail on forest fires. Analysis of sensitivity and feature importance has been carried out [10] to examine how sub-basins affect the hydrological response of catchments.

This article focuses on the analysis of the sensitivity and feature importance of each climatic factor for forest fires in Kalimantan using four machine learning techniques: random forests, gradient boosting, Bayesian regression, and artificial neural networks. The results provide a comparison of the accuracy of the four machine learning models used. In addition, a summary of each climatic factor's sensitivity and feature importance is given in this article, based on the fittest machine learning (ML) model, such as variance, density, and distribution-based sensitivity indices, as well as permutation and Shapley feature importance.

The main contribution of this article is to disseminate sensitivity analysis in supervised learning which can be used as a way to select explanatory variables that influence response variables, which is still rarely used. This article also compares the results of sensitivity analysis with feature importance analysis, which is also widely used to select explanatory variables. Selection of explanatory variables using sensitivity analysis is more effective because it can be done without the ML model training process like feature importance analysis which sometimes also causes misunderstanding. Apart from that, the ML model formed can explain the connection between hotspot density and climate variables; and become the initial basis for further modeling of hotspots.

2. STUDY AREA

The world's largest tropical peatlands are found in Indonesia, covering a total of 13.43 million hectares across three major islands: Papua, Kalimantan, and Sumatra. This study is concentrated in Kalimantan, which is comprised of five provinces: West, East, Central, South, and North Kalimantan and contributes to 33.8% of Indonesia's peatlands [11]. The provinces of Central and West Kalimantan had the most hotspots during the 2019 fire event, followed by Jambi, Riau, and South Sumatra provinces [12]. Every year, forest fires in Kalimantan become a national concern that receives major attention from the government and researchers.

The Indonesian part of Borneo, the largest island in Asia and the third largest in the world, is called Kalimantan. Kalimantan is renowned for its rich biodiversity, vast rainforests and unique geography. Kalimantan experiences year-round high temperatures and high humidity due to its tropical climate. There are two distinct seasons in the region: November to March is the rainy season and April to October is the dry season. The rainfall patterns in Kalimantan are classified as equatorial and monsoonal. Equatorial rainfall patterns are found in majority parts of East, West, and North Kalimantan, according to fast Fourier transform and empirical orthogonal function analysis [13]. In the meantime, the majority of South and Central Kalimantan experiences

monsoonal rainfall patterns. The climate is essential in supporting the island's lush rainforests and diverse ecosystems. On the other hand, these climatic conditions greatly influence forest fire events in Kalimantan, especially when accompanied by a strong El Nino event.

3. DATASETS

This research uses data on global and local climate factors and the number of hotspots. Hotspots are the outcome of land and forest fires detected at particular pixel sizes using a specific algorithm [14]. The local climate factors used include total precipitation, precipitation anomaly, and the number of dry days (dry spells, i.e., daily precipitation less than one millimeter per day). Meanwhile, the global climate factors used include indices for the ENSO and IOD phenomena. Table 1 describes the source of each variable in the datasets.

Table 1. Source of each variable in the datasets

No.	Name	Description
1	Total precipitation	Extracted from CMORPH (https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH.CRT/).
2	Precipitation anomaly	Extracted from CMORPH (https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH.CRT/).
3	Number of dry days	Extracted from CMORPH (https://ftp.cpc.ncep.noaa.gov/precip/PORT/SEMDP/CMORPH.CRT/).
4	Number of hotspots	Agency for Meteorology, Climatology, and Geophysics (BMKG) Indonesia.
5	ENSO index	produced by NOAA and can be downloaded at https://psl.noaa.gov/gcos_wgsp/Timeseries/Nino34/ .
6	IOD index	produced by NOAA and can be downloaded at https://psl.noaa.gov/gcos_wgsp/Timeseries/DMI/ .

The data used (local climate factors and the number of hotspots) in this study has been processed in fire-prone areas in Kalimantan [15]. There are two main seasonal rainfall patterns in Kalimantan: equatorial and monsoonal. Using the clustering method, hotspot data in Kalimantan is grouped into clusters to find areas that are vulnerable to forest fires. Most of these areas are located in central, western and southern Kalimantan, which has a monsoonal rainfall pattern. In these selected areas, the data is aggregated to retrieve the general characteristics of rainfall, dry spells, and hotspots data in fire-prone areas in Kalimantan. The data were then analyzed for dependency on monthly hotspot data and it was found that the two-month average of total precipitation, the monthly precipitation anomaly, and the three-month accumulative of the number of dry days provided the strongest dependency on monthly hotspots. All data were obtained in 2001-2020.

4. METHOD

There are three stages to this study. The first stage is analyzing the sensitivity of climatic factors to hotspots data. Then, the second stage is training and testing ML models to predict hotspots data based on climatic factors. The final stage is analyzing the feature importance of climate factors based on the fittest ML models. Figure 1 shows the research flow in this article and the following details each step.

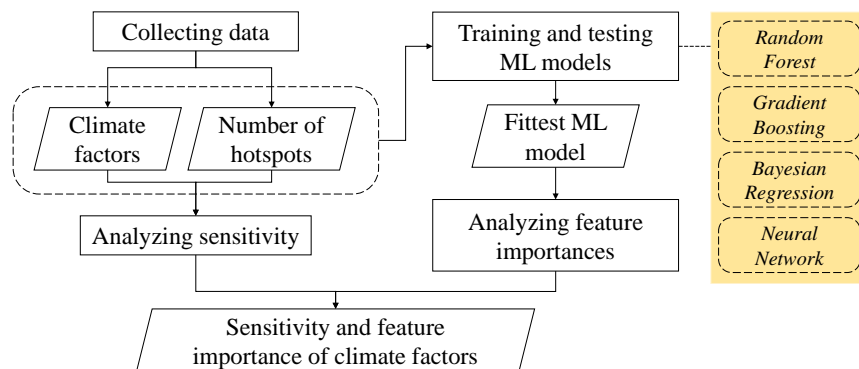


Figure 1. Research flow in this article

4.1. Feature importance and sensitivity analysis for supervised learning

In this subsection, three sensitivity measures of sensitivity analysis and described: variance, density, and distribution-based approaches. Additionally, we implement two pertinent approaches to the ML model-agnostic feature importance. Sensitivity analysis is used initially before modeling, while feature importance analysis is carried out after the ML model is selected.

4.1.1. Sensitivity analysis

There are three bases that are used to measure the sensitivity of each feature in the sensitivity analysis: variance, density, and distribution-based approaches. Variance-based sensitivity index [16], [17]:

$$\eta_j^2 = \frac{\mathbb{V}[Y] - \mathbb{E}_{\mathbf{x}_{-j}}[\mathbb{V}_{X_j}[Y|X_j]]}{\mathbb{V}[Y]} \quad (1)$$

Density-based sensitivity index [18]:

$$\delta_j = \frac{1}{2} \mathbb{E}_{X_j} \left[\int_{\mathcal{Y}} |p_Y(y) - p_{Y|X_j}(y)| dy \right] \quad (2)$$

Distribution-based (CDF) sensitivity index [19]:

$$\beta_j^{KS} = \mathbb{E}_{X_j} \left[\sup_{\mathcal{Y}} |\mathbb{P}_Y(y) - \mathbb{P}_{Y|X_j}(y)| dy \right] \quad (3)$$

Where $p_{Y|X_j}$ and p_Y represent the conditional density and marginal output density via the L_1 -norm, respectively, with $\mathbb{P}_{Y|X_j}$ and \mathbb{P}_Y are the corresponding cumulative distribution functions. From the same features-forecast realizations dataset, In (1) to (3) can possibly be computed. Using the given-data (or one-sample) approach described in [20], the computation is carried out.

4.1.2. Feature importance

Here, we present importance measures designed for ML use cases. The most common measure is called permutation feature importance (PFI) defined by [21], which can be estimated in (4):

$$\text{PFI}_j \approx \frac{1}{N} \sum_{i=1}^N \mathcal{L} \left(y^{(i)}, \hat{f}(X_j^{\pi,i}, \mathbf{X}_{-j}^{(i)}) \right) - \frac{1}{N} \sum_{i=1}^N \mathcal{L} \left(y^{(i)}, \hat{f}(X_j^{(i)}, \mathbf{X}_{-j}^{(i)}) \right) \quad (4)$$

Where X_j^{π} is the distribution of feature X_j . A high PFI_j value indicates that when a permutation of X_j breaks the dependency between Y and X_j , the performance of the prediction model dramatically declines. However, when features have a significant statistical reliance on one another, PFI measurements may produce deceptive results [22].

The second feature importance measure is the Shapley additive explanations (SHAP) method [23]. The SHAP approach uses the optimal Shapley values from game theory to explain individual predictions. In ML, the Shapley value indicates how the feature contributed to a prediction at the given query point. Moreover, Shapley values can be combined to create global explanations. A matrix of Shapley values is obtained by running SHAP for each query point. Each row in this matrix corresponds to a query point, and each column to a feature. We are able to analyze the complete model by examining the Shapley values in this matrix.

The idea behind SHAP feature importance is simple: features are important or relevant if their absolute Shapley values are high. To determine the global importance, we take the average of the absolute Shapley values for each feature throughout the data:

$$\text{SFI}_j \approx \frac{1}{N} \sum_{i=1}^N |\phi_j^{(i)}| \quad (5)$$

Where $\phi_j^{(i)}$ is the Shapley value of the j -th feature for the i -th query point. SHAP is an alternative to PFI. Both importance metrics have significant differences. Whereas SHAP depends on the quantity of feature attributions, PFI is based on the model's performance declining [24].

4.2. Supervised machine learning

One area of artificial intelligence called ML was created to enable a machine to learn a problem and find a solution on its own without human assistance. Supervised learning is one type where the algorithm of this type begins with a training process that objectives to acquire knowledge about the relationship of features or predictors to a specified target (output). Thus, if there is a new input outside of the training data, the supervised learning algorithm can predict the appropriate target. There are many types of methods in supervised learning. Here, we employ four models, i.e., random forest, gradient boosting, Bayesian regression, and artificial neural network.

4.2.1. Random forest and gradient boosting

An ensemble method employs multiple learning algorithms simultaneously and then combines them to obtain more accurate modeling results. Ensemble models that leverage tree-based models include random forests [21] and gradient boosting [25] machines. These tree-based ensemble models can handle nonlinear and complicated feature connections. Furthermore, multicollinearity has little or no impact on random forest model [26].

Decision trees are developed into random forest by applying bootstrap aggregating and random feature selection methods [27]. Bootstrap is a random subset sampling process from a data set with a certain number of iterations and variables. The sample is returned to the data set so that it can be re-selected in the next process. In a random forest, every tree receives independent predictions after being trained on a random selection of features. By averaging the decision trees' projections, the response variable's final estimation is determined [10]. Figure 2(a) displays an example of the random forest model. There are hyperparameters that need to be tuned in a random forest model including criterion (the function for evaluating a split's quality), n-estimators (the number of trees), and max-depth (the tree's maximum depth).

Gradient boosting builds a strong learner (ensemble model) iteratively using weak learner models, typically decision trees. This algorithm's primary concept is to build models one after the other, with each new model attempting to minimize the mistakes of the preceding model. We train a decision tree at each step using the residuals from the preceding tree series. The additive model described by each tree's contribution is used to build the resulting ensemble model [10]. An illustration of the gradient boosting model is shown in Figure 2(b). There are hyperparameters that need to be tuned in a random forest model including loss function to be optimized, n-estimators (the number of trees), and max-depth (the tree's maximum depth). Implementation of random forest and gradient boosting models using MLJ.jl package in julia.

4.2.2. Bayesian linear regression

For a multiple linear regression (MLR) model, $y_i = \beta \mathbf{x}_i + \varepsilon_i$, there are two approaches for estimating its parameters. The least squares and maximum likelihood approaches are examples of the classical approach, which handles the parameters as fixed but the quantities are unknown. As an alternative, Bayesian approach treats the parameters as random variables [28].

The goal of Bayesian analysis is to update the parameters' probability [29], from prior distributions (the parameter distribution assumed before observing the data) into posterior distributions, when more evidence or data becomes available. Priors can have a significant effect on estimation and inference. Many Bayesian regression methods have been proposed to fit different situations for various prior distributions, including the hierarchical linear model [30] and the Bayesian lasso model [31]. Table 2 shows the prior distribution options that can be used. The normal-inverse-gamma conjugate model is a frequently selected option [32]. Using MATLAB's econometrics toolbox, the Bayesian linear regression model is implemented.

Using Bayes' theorem, the conditional probability as the posterior density is given in (6):

$$P_{posterior}(\beta|y) = P_{prior}(\beta) \times \frac{P_{sample}(y|\beta)}{P_{pred}(y)} \quad (6)$$

or can be simplified to 'posterior \propto likelihood \times prior' where *prior* is the parameter distribution we assume, allowing us to include knowledge about the model before data are imported, and likelihood is the information about the parameters provided by the sample response [28]. The marginal distribution, denoted by $P_{pred}(y)$, is the likelihood averaged across all possible values of the parameters concerning the prior density:

$$P_{pred}(y) = \int P_{prior}(\beta) P_{sample}(y|\beta) d\beta$$

The density of $P_{sample}(y|\beta)$, or the probability of a parameter value given a particular outcome, is the likelihood function. $P_{prior}(\beta)$ stands for the arbitrary opinions regarding the parameter values before to measurement. Then, a posterior distribution $P_{posterior}(\beta|y)$ could be interpreted as a higher degree of belief attained through the use of experimental data [9].

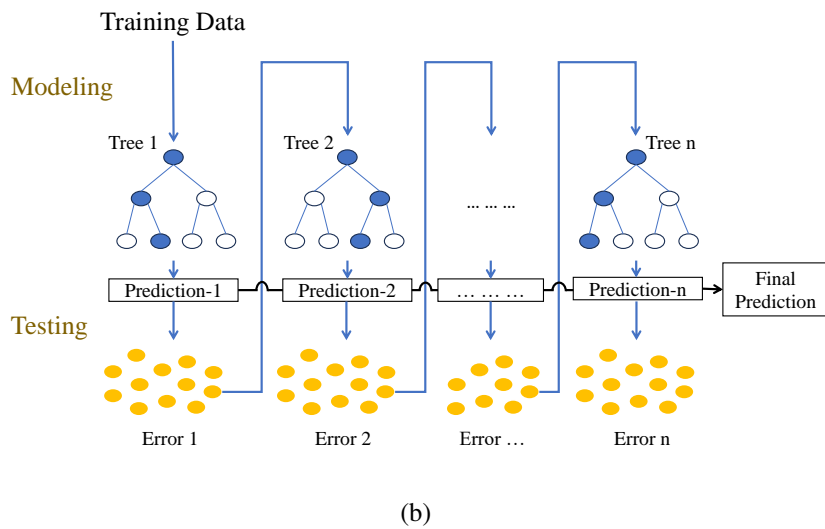
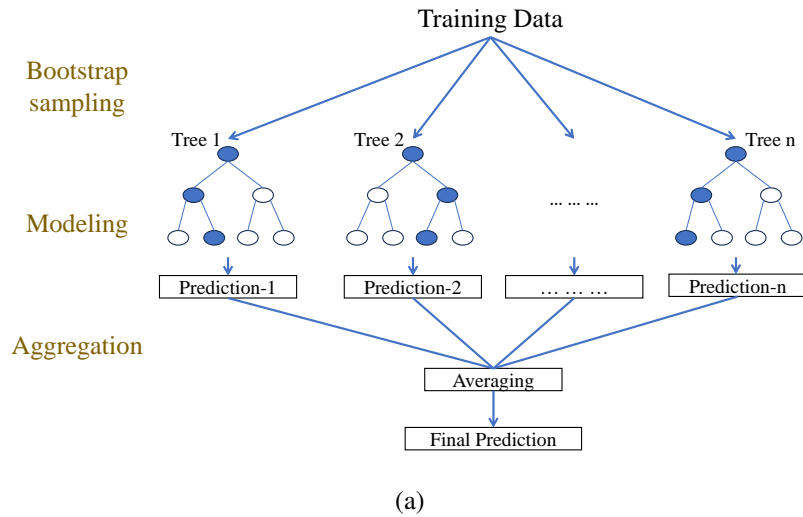


Figure 2. Illustration of tree-based ML algorithms (a) random forest (b) gradient boosting

Table 2. Prior distribution options and its descriptions

Prior Model	Description
Conjugate	A normal-inverse-gamma conjugate model, where β and σ^2 are independent. $\beta \sigma^2 \sim N_{p+1}(\mu, \sigma^2 V)$ and $\sigma^2 \sim IG(A, B)$
Semi-conjugate	Same as conjugate model, but β and σ^2 are dependent.
Diffuse	The joint prior distribution of (β, σ^2) is proportional to $1/\sigma^2$
Mix conjugate	Implementing stochastic search variable selection (SSVS) assuming β and σ^2 are dependent random variables, given γ_k and σ^2 , $\beta_k = \gamma_k c_1 Z + (1-\gamma_k) c_2 Z$, where $c_j = \sigma^2 V_j, j = 1, 2$.
Mix semi-conjugate	Same as conjugate model, but $c_j = V_j, j = 1, 2$.
Lasso	Implementing Bayesian lasso regression $\beta \sigma^2, \lambda \sim Laplace(0, \sigma/\lambda)$ and $\sigma^2 \sim IG(A, B)$ where λ is the shrinkage parameter.

4.2.3. Artificial neural network

Artificial neural network is implemented as a software simulation of the properties of human neural networks due to their high ability to process information [33]. The adaptability of artificial neural network models is well known. Artificial neural network is made up of several processing components that process input and produce output in response to an activation function. These processing elements are called units, or nodes which represent a neuron in a human neural network [34]. Broadly speaking, there are four building blocks of artificial neural network architecture, including nodes, layers, activation functions, and training methods (optimizers). In this study, an artificial neural network model is focused on a network structure with single hidden layer H_q , several input neurons X_p and an output layer with the observed outcome Y . An illustration of the artificial neural network model is shown in Figure 3.

Determining the weight value for each signal in a multi-layer architecture so that the model has good accuracy is not easy. Therefore, a backpropagation algorithm is introduced which allows to determine the error value at the hidden layer's node, so that the weight value can be adjusted. Adjustment of this weight value is done by a training method. A number of training methods commonly used [35], including gradient descent, momentum, nesterov accelerated gradient descent (NAG), adaptive moment estimation (Adam), and nesterov-adam (Nadam). There are hyperparameters that need to be tuned in an artificial neural network model including the number of neurons in the hidden layer, optimizer, learning rate, and loss function. Implementation of the artificial neural network model employs the Flux.jl package in Julia.

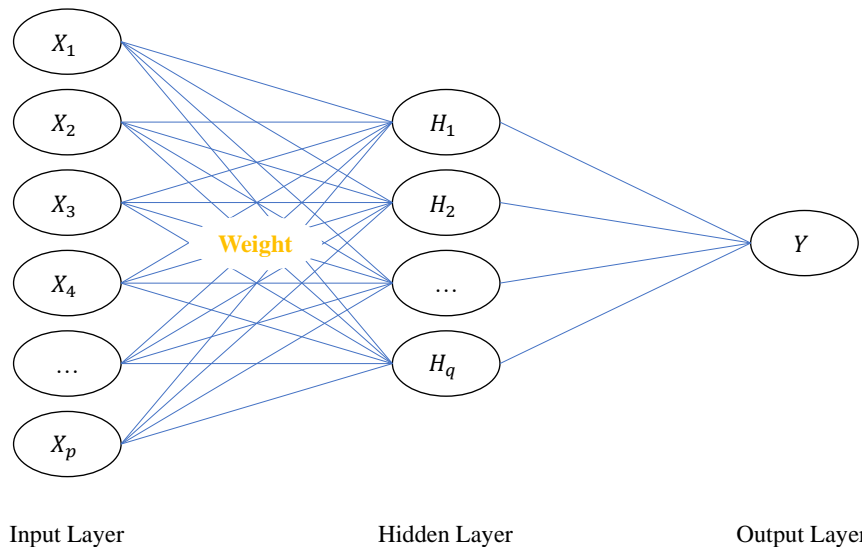


Figure 3. Illustration of a single hidden layer artificial neural network architecture

4.3. Performance assessment

This study uses two metrics to assess the performance of ML models to predict hotspots in the testing data, i.e., root mean squared error (RMSE) and explained variance score (EVS). The RMSE is defined as (7):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (7)$$

where y is the actual value and \hat{y} is the predicted value. This performance measure ranges in $[0, \infty)$, with 0 indicates a perfect match. Meanwhile, the EVS is estimated in (8):

$$\text{EVS} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \quad (8)$$

EVS simply shows the degree of variation in the actual value that can be explained by a model. Scores near 1.0 are extremely desirable, suggesting lower squares of standard deviations of errors [36].

5. RESULTS AND DISCUSSION

In this section, the sensitivity analysis of climate factors is presented in subsection 4.1. The results of training and testing processes for each ML model are presented in subsection 4.2. Subsection 4.3 presents the importance feature of each climate factor by the most suitable ML model.

5.1. Sensitivity analysis of climate factors

Looking back at subsection 4.1, we apply the three sensitivity measures in (1)-(3) on the climate factors data to the fire hotspots data in 2001-2020. The sensitivity value of each climate factor is shown in Figure 4. From the three sensitivity indices, the number of dry days and total precipitation have the highest sensitivity values. According to variance-based, the number of dry days has the highest sensitivity to fire hotspots data compared to other climatic factors. Meanwhile, total precipitation has the highest sensitivity to fire hotspots data based on other sensitivity indices. After the two climatic factors, the month is the factor that has the third highest sensitivity index.

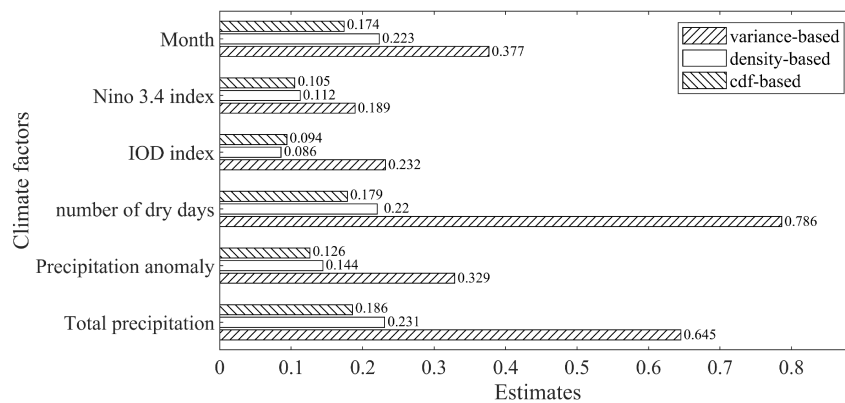


Figure 4. Sensitivity indices of climate factors respect to fire hotspots in Kalimantan, Indonesia

The lowest sensitivity is shown by the IOD and Nino 3.4 indices, indicating that these climatic factors do not have a direct effect on fire hotspots, although many studies have looked at the influence of the two indices on fire hotspots in Indonesia [37]. Even though extreme hotspots coincide with strong El Nino and positive IOD phenomena in 1997 and 2015, in fact these two phenomena affect rain and drought conditions in Indonesia which indirectly affect the emergence of hotspots that trigger forest fires. Thus, it can be concluded that IOD and Nino 3.4 indices have an indirect effect on forest fires in Indonesia.

5.2. Hyperparameter tuning and performance of ML methods

Based on the data described in Table 1, there are six predictors used from X_1, X_2, \dots, X_6 respectively: total precipitation, precipitation anomaly, number of dry days, ENSO index, IOD index, and month. Meanwhile, the response variable Y is the number of hotspots. There are two divisions to the data: 80% training (2001-2016) and 20% testing (2017-2020). Here, the hyperparameter tuning results on the training data will be presented for each ML model.

There are hyperparameters for tree-based models. There are four criteria (squared error, absolute error, Friedman MSE, and Poisson), number of trees (1-20) and maximum depth (2-40) to be tuned for random forest. Meanwhile, four loss functions (least square, least absolute deviation, huber, and quantile), number of trees (1-60) and maximum depth (1-10) are tuned for gradient boosting. These are the hyperparameter values that we acquire after training the models:

- Random forest: criterion = squared error, n-estimator = 5, max-depth = 18
- Gradient boosting: loss function = huber, criterion = friedman-mse, learning-rate = 0.1, n-estimator = 47, max-depth = 16

Bayesian linear regression has a hyperparameter, i.e., its prior distribution. Based on Table 2, there are six prior distributions that were tried and found that the diffuse prior distribution is the most fit with the regression equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_3^2 + \beta_8 x_3^3 + \beta_9 x_5^2 \quad (9)$$

where each parameter coefficient is shown in Table 3. Since the predicted value of \hat{y} allows for negative values, we take $\max(0, \hat{y})$ as the predicted value for hotspots based on Bayesian linear regression.

Meanwhile, there are the number of neurons in the hidden layer, optimizer, learning rate, and loss function which are tuned for the artificial neural network model. From the training process up to 1000 epochs, the most suitable artificial neural network structures are obtained: 6 neurons in the input layer, batch normalization layer, 5 neurons in the hidden layer, and an output. Meanwhile, the most appropriate optimizer is Nadam with a learning rate of 0.01 and an MAE loss function.

Table 3. Fittest parameter coefficient of the Bayesian linear regression model

Coefficient	Mean	Standard Deviation	95% Conf. Interval	Positive	Distribution
Intercept	-4695.56	2369.58	[-9345.187, -45.931]	0.024	t (-4695.56, 2356.53 ² , 1.8e+02)
β_1	-38.03	47.37	[-130.970, 54.915]	0.210	t (-38.03, 47.11 ² , 1.8e+02)
β_2	32.17	30.08	[-26.849, 91.185]	0.858	t (32.17, 29.91 ² , 1.8e+02)
β_3	498.97	159.65	[185.699, 812.247]	0.999	t (498.97, 158.77 ² , 1.8e+02)
β_4	10.68	77.62	[-141.620, 162.979]	0.555	t (10.68, 77.19 ² , 1.8e+02)
β_5	-76.70	236.87	[-541.490, 388.094]	0.373	t (-76.70, 235.57 ² , 1.8e+02)
β_6	16.40	17.61	[-18.155, 50.947]	0.825	t (16.40, 17.51 ² , 1.8e+02)
β_7	-15.71	3.69	[-22.954, -8.464]	0.000	t (-15.71, 3.67 ² , 1.8e+02)
β_8	0.16	0.03	[0.105, 0.213]	1.000	t (0.16, 0.03 ² , 1.8e+02)
β_9	-183.46	604.16	[-1368.947, 1002.027]	0.380	t (-183.46, 600.83 ² , 1.8e+02)
σ^2	528980.00	56071.76	[430355.487, 649785.797]	1.000	IG(91.00, 2.1e-08)

In Table 4, the ML models' performance metrics are displayed. The training data can be effectively used to train the random forest and gradient boosting models, which is indicated by the low RMSE value and high explained variance score. The explained variance score for both models exceeds 90%, and is almost perfect for the gradient boosting model. However, the evaluation results on the testing data show that both models are overfit, due to the high RMSE values and low explained variance scores, especially gradient boosting.

Table 4. Performance measures of the models

ML model	Training		Testing	
	RMSE	Explained variance	RMSE	Explained variance
Random Forest	412.63	93.25%	995.44	41.97%
Gradient Boosting	97.39	99.63%	1085.45	26.64%
Bayesian Linear Regression	702.15	80.44%	750.60	68.96%
Artificial Neural Network	655.97	83.20%	827.65	57.53%

Performance improvements are seen in the artificial neural network model. Although the training process is not as fit as the tree-based ensemble model, the artificial neural network model gives a better explained variance score of more than 50% and an RMSE of 827 hotspots on the testing data. However, the Bayesian linear regression model outperforms the four ML models. By maintaining the explained variance score above 80% during training, the Bayesian linear regression model gives the best performance on data testing, i.e., an RMSE of 750 hotspots and an explained variance score of 69%. Therefore, Bayesian linear regression model is the best performing model compared to all other models. As a result, we decide to use this ML model for the hotspots predicting analysis. Figures 5(a) to 5(d) displays the predictions of hotspots using four machine learning models.

Figures 5(a) and (b) show very fit training results for the random forest and gradient boosting models, but the prediction results on the testing data are unsatisfactory, especially the predictions for 2018 and 2019. In addition, the predicted number of hotspots in 2016 and 2020 is higher than the actual number of hotspots which is almost zero. The artificial neural network model in Figure 5(d) is slightly better than the previous two models. The prediction results for 2016 and 2020 are very low according to their actual values, while predictions for 2019 have increased accuracy towards their actual values compared to both tree-based models. Moreover, the most satisfactory results are shown by the Bayesian linear regression model Figure 5(c). Even though the performance on the training data is not as fit as the other models, the predictions on the testing data are the most accurate compared to other ML models that have been tried.

Interesting results are shown in the predictions for 2018, where all ML models show overestimated prediction results. That is, actually, based on existing climatic conditions, the number of hotspots that should

have occurred is higher than the actual number of hotspots at that time. This is due to preventive actions from the Indonesian government to reduce the number of hotspots, in order to make the ASIAN Games 2018 successful in Indonesia [38]. This is the second time in a row that Indonesia has been able to reduce its deforestation rate. As a consequence of decreases in deforestation in 2017 and 2018, Indonesia received the first installment of REDD+ payments, a program that compensates developing countries that successfully reduce emissions by maintaining their forests [39]. This shows that the existence of an appropriate early warning system model can assist the government in making policies as a preventive action to reduce deforestation and minimize the impact and losses due to forest fires in Indonesia.

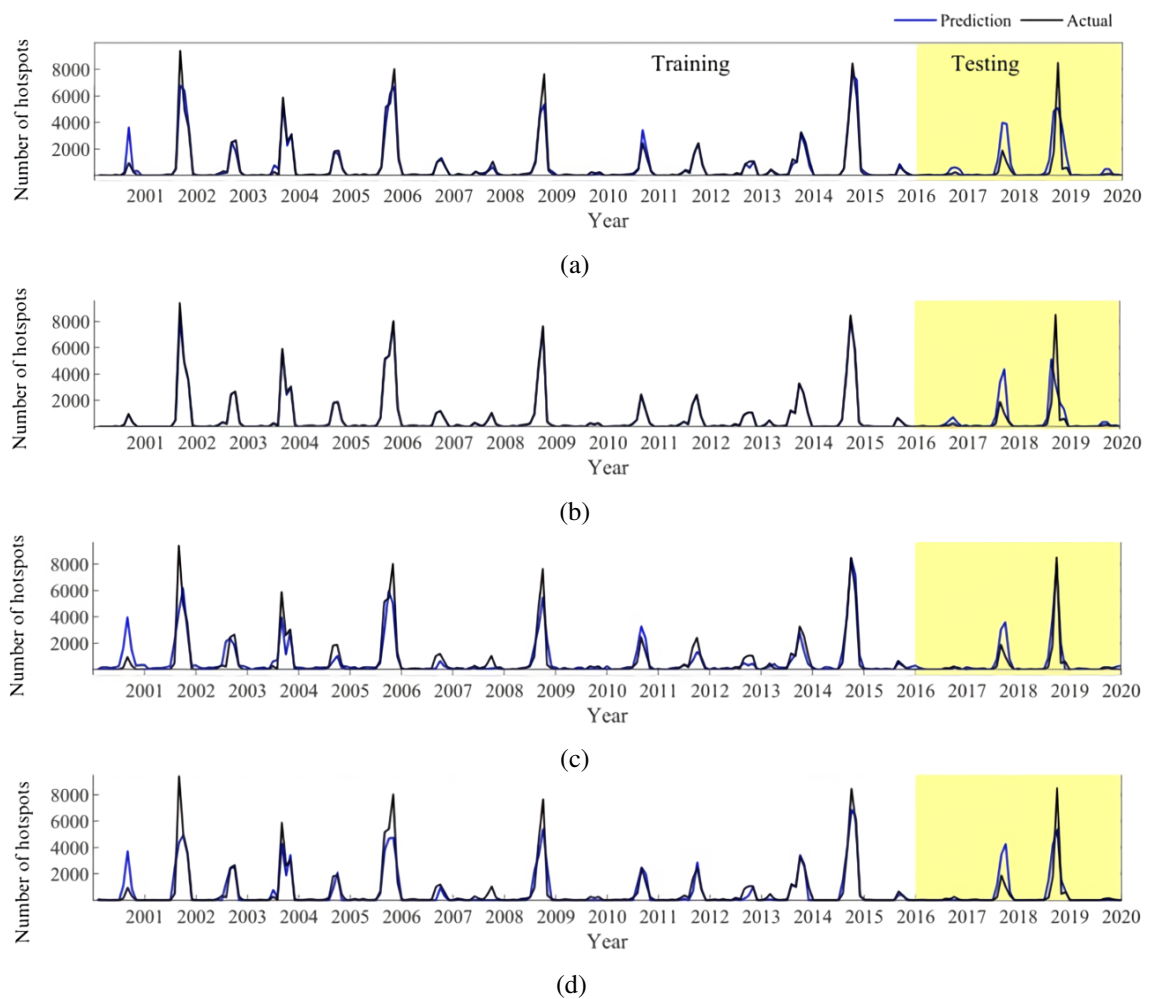


Figure 5. Comparison of the predictions of hotspots on the training and testing data of the four ML models: (a) random forest, (b) gradient boosting, (c) Bayesian linear regression, and (d) artificial neural network

5.3. Feature importance analysis

In contrast to the sensitivity measures, which are determined directly from the data, the feature importance measures in (4) and (5) are calculated using the predictions of the optimum ML model. The permutation feature importance is calculated using the implementation of the algorithm by [40] using data testing. RMSE is used as a loss function in the computation of performance-based measures. Meanwhile, SHAP feature importance is calculated using the implementation of the algorithm by [23].

The estimations of the feature importance measures employed in the case study are shown in Figure 6. Permutation feature importance is obtained from the absolute mean of 100 repetitions of the permutations of the observed features. Meanwhile, Shapley feature importance is the absolute mean of the Shapley values for each query point on the observed features. Recalling that feature importance analysis will be misleading if there is

multicollinearity between the variables, so here multicollinearity is detected using the variance inflation factor (VIF) value. It can be said that there is multicollinearity that must be handled appropriately, if the $VIF \geq 10$ [41]. The VIF values of each feature are 5.88, 1.44, 6.96 1.43, 1.17 and 1.29, respectively. This shows that there is no multicollinearity in each climate indicator.

The feature importance of dry-spells or the number of dry days far outperforms the other features. This is different from the sensitivity of each feature where all features have sensitivity values that are almost close to one another. This shows that based on the sensitivity value, all features have an impact on hotspots because they have a correlation [10]. Thus, a small feature importance value does not mean that the feature has no effect at all on the hotspots in Kalimantan. To better understand the results presented in Figures 4 and 6, Table 5 presents the rankings deriving from the set of important measures [42].

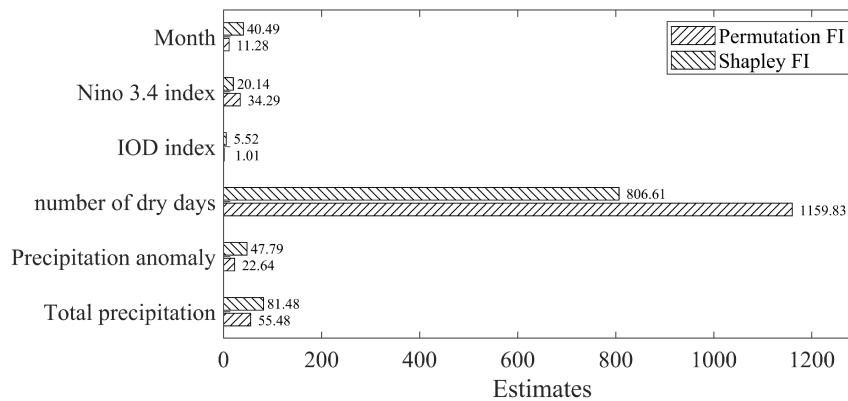


Figure 6. Feature importance of climate factors as predictors of Bayesian linear regression models to predict hotspots in Kalimantan, Indonesia

Table 5. Ranking for each feature importance measure and the mean ranking

Features	Variance SA	Density SA	Distribution SA	Permutation FI	Shapley FI	Mean ranking
Number of dry days	1	3	2	1	1	1
Total precipitation	2	1	1	2	2	2
Month	3	2	3	5	4	3
Precipitation anomaly	4	4	4	4	3	4
Nino index	6	5	5	3	5	5
IOD index	5	6	6	6	6	6

In general, the main and most important feature of the regression model for hotspots is the number of dry days. Even though the ENSO and IOD indices are in the last ranking, this does not mean they do not have an effect on hotspots. Both indices still have an effect on hotspots through their influence on decreasing rainfall and extending the dry season in Kalimantan. Moreover, the Nino index has more influence on hotspots in Kalimantan than the IOD index, in line with studies [4] and [43].

6. CONCLUSION

This article analyzes the sensitivity and feature importance of climatic factors for forest fires in Kalimantan using four machine learning techniques: random forests, gradient boosting, bayesian regression, and artificial neural networks. Three sensitivity measures are used such as variance-based, density-based, and distribution-based, as well as feature importance such as permutation and Shapley feature importances. Evaluation of the ML model concluded that the Bayesian linear regression model outperformed other ML models, which was presented by the best evaluation of data testing based on RMSE and explained variance score. Meanwhile, tree-based models, such as random forest and gradient boosting, are indicative of overfit, which is shown by the very good evaluation results on the training data but poor evaluation on the testing data. On the other hand, the artificial neural network model gives quite good results, although not as good as the Bayesian linear regression model. Based on the results of sensitivity analysis and feature importances, the number of dry days

is the most important feature for the Bayesian linear regression model in predicting the number of hotspots in Kalimantan. Followed by total precipitation and month features. The two features of least importance are the IOD and ENSO indices. Even so, the two features still have an indirect influence on hotspots in Kalimantan based on sensitivity analyses.




REFERENCES

- [1] B. H. Suharjo and W. A. Velicia, "The role of rainfall towards forest and land fires hotspot reduction in four districts in Indonesia on 2015-2016," *Journal of Tropical Silviculture*, vol. 9, no. 1, pp. 24–30, 2018, doi: 10.29244/j-siltrop.9.1.24-30.
- [2] S. Yang, M. Lupascu, and K. S. Meel, "Predicting forest fire using remote sensing data and machine learning," in *35th AAAI Conference on Artificial Intelligence*, May 2021, vol. 35, no. 17, pp. 14983–14990, doi: 10.1609/aaai.v35i17.17758.
- [3] C. Wang, and P. C. Fiedler, "ENSO variability and the eastern tropical Pacific: A review," *Progress in oceanography*, vol. 69, no. 2-4, pp. 239-266, 2006, doi: 10.1016/j.pocean.2006.03.004
- [4] S. Nurdianti *et al.*, "The impact of El Niño southern oscillation and Indian Ocean Dipole on the burned area in Indonesia," *Terrestrial, Atmospheric and Oceanic Sciences*, vol. 33, no. 1, pp. 1-17, 2022, doi: 10.1007/S44195-022-00016-0.
- [5] T. Fanin and G. R. Van Der Werf, "Precipitation-fire linkages in Indonesia (1997-2015)," *Biogeosciences*, vol. 14, no. 18, pp. 3995–4008, 2017, doi: 10.5194/bg-14-3995-2017.
- [6] M. N. Nur'utami and R. Hidayat, "Influences of IOD and ENSO to Indonesian rainfall variability: role of atmosphere-ocean interaction in the Indo-Pacific sector," *Procedia Environmental Sciences*, vol. 33, pp. 196-203, 2016, doi: 10.1016/j.proenv.2016.03.070
- [7] T. Nikonovas, A. Spessa, S. H. Doerr, G. D. Clay, and S. Mezbahuddin, "ProbFire: A probabilistic fire early warning system for Indonesia," *Natural Hazards and Earth System Sciences*, vol. 22, no. 2, pp. 303–322, 2022, doi: 10.5194/nhess-22-303-2022.
- [8] A. Shabrina, I. Palupi, B. A. Wahyudi, I. N. Wahyuni, M. D. Murti, and A. L. Latifah, "Modelling the climate factors affecting forest fire in Sumatra using Random Forest and Artificial Neural Network," in *ACM International Conference Proceeding Series*, 2022, pp. 194–198, doi: 10.1145/3575882.3575920.
- [9] E. Ardiyani, S. Nurdianti, A. Sopaheluwakan, P. Septiawan, and M. K. Najib, "Probabilistic hotspot prediction model based on bayesian inference using precipitation, relative dry spells, ENSO and IOD," *Atmosphere (Basel)*, vol. 14, no. 2, pp. 1-20, 2023, doi: 10.3390/atmos14020286.
- [10] F. Cappelli, F. Tauro, C. Apollonio, A. Petroselli, E. Borgonovo, and S. Grimaldi, "Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: a proof of concept," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 4, pp. 1247–1264, 2023, doi: 10.1007/s00477-022-02332-w.
- [11] T. W. Yuwati *et al.*, "Restoration of degraded tropical peatland in Indonesia: A review," *Land*, vol. 10, no. 11, pp. 1-31, 2021, doi: 10.3390/land10111170.
- [12] A. S. Thoha *et al.*, "Spatial distribution of 2019 forest and land fires in Indonesia," *Journal of Physics: Conference Series*, vol. 2421, no. 1, pp. 1-9, 2023, doi: 10.1088/1742-6596/2421/1/012035.
- [13] S. Nurdianti, E. Khatizah, M. K. Najib, and R. R. Hidayah, "Analysis of rainfall patterns in Kalimantan using fast fourier transform (FFT) and empirical orthogonal function (EOF)," *Journal of Physics: Conference Series*, vol. 1796, no. 1, pp. 1-10, 2021, doi: 10.1088/1742-6596/1796/1/012053.
- [14] H. A. Nainggolan, D. P. O. Veanti, and D. Akbar, "Utilisation of Nasa - Gfwd and Firms Satellite Data in Determining the Probability of Hotspots Using the Fire Weather Index (Fwi) in Ogan Komering Ilir Regency, South Sumatra," *International Journal of Remote Sensing and Earth Sciences (IJReSES)*, vol. 17, no. 1, pp. 85-98, 2020, doi: 10.30536/ij.ijreses.2020.v17.a3202.
- [15] M. K. Najib, S. Nurdianti, and A. Sopaheluwakan, "Copula-based joint distribution analysis of the ENSO effect on the drought indicators over Borneo fire-prone areas," *Modeling Earth Systems and Environment*, vol. 8, no. 2, pp. 2817–2826, 2022, doi: 10.1007/s40808-021-01267-5.
- [16] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of nonlinear models," *Reliability Engineering & System Safety*, vol. 52, no. 1, pp. 1–17, 1996, doi: 10.1016/0951-8320(96)00002-6.
- [17] R. L. Iman and S. C. Hora, "A Robust Measure of Uncertainty Importance for Use in Fault Tree System Analysis," *Risk analysis*, vol. 10, no. 3, pp. 401–406, 1990, doi: 10.1111/j.1539-6924.1990.tb00523.x.
- [18] E. Borgonovo, "A new uncertainty importance measure," *Reliability Engineering & System Safety*, vol. 92, no. 6, pp. 771–784, 2007, doi: 10.1016/j.ress.2006.04.015.
- [19] E. Borgonovo, S. Tarantola, E. Plischke, and M. D. Morris, "Transformations and invariance in the sensitivity analysis of computer experiments," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 76, no. 5, pp. 925–947, 2014, doi: 10.1111/rssb.12052.
- [20] E. Plischke, E. Borgonovo, and C. L. Smith, "Global sensitivity measures from given data," *European Journal of Operational Research*, vol. 226, no. 3, pp. 536–550, 2013, doi: 10.1016/j.ejor.2012.11.047.
- [21] L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [22] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," *Statistics and Computing*, vol. 31, no. 6, pp. 1-16, 2021, doi: 10.1007/s11222-021-10057-z.
- [23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in neural information processing systems*, 2017, vol. 30, pp. 4768–4777.
- [24] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. 2023, Ferndale, USA: Lean Publishing.
- [25] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.




- [26] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199-231, 2002, doi: 10.1214/ss/1009213726.
- [27] D. Chutia, D. K. Bhattacharyya, J. Sarma, and P. N. L. Raju, "An effective ensemble classification framework using random forests and a correlation based feature selection technique," *Transactions in GIS*, vol. 21, no. 6, pp. 1165-1178, 2017, doi: 10.1111/tgis.12268
- [28] Y. Xue, Y. Liu, C. Ji, and G. Xue, "Hydrodynamic parameter identification for ship manoeuvring mathematical models using a Bayesian approach," *Ocean Engineering*, vol. 195, 2020, doi: 10.1016/j.oceaneng.2019.106612.
- [29] M. Movaghar and S. Mohammadzadeh, "Bayesian Monte Carlo approach for developing stochastic railway track degradation model using expert-based priors," *Structure and Infrastructure Engineering*, vol. 18, no. 2, pp. 145-166, 2022, doi: 10.1080/15732479.2020.1836001.
- [30] H. Woltman, A. Feldstain, J. C. MacKay, and M. Rocchi, "An introduction to hierarchical linear modeling," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, pp. 52-69, 2012, doi: 10.20982/tqmp.08.1.p052.
- [31] T. Park and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681-686, 2008, doi: 10.1198/016214508000000337.
- [32] C. Robert, "Machine Learning, a Probabilistic Perspective," *Chance*, vol. 27, pp. 62-63, 2014, doi: 10.1080/09332480.2014.914768.
- [33] Y. Safi and A. Bouroumi, "Prediction of forest fires using artificial neural networks," *Applied Mathematical Sciences*, vol. 7, no. 5-8, pp. 271-286, 2013, doi: 10.12988/ams.2013.13025.
- [34] L. V. Fausett, *Fundamentals of Neural Network, Architectures, Algorithms, Applications*. New York: John Wiley & Sons, 2018.
- [35] S. Nurdianti, M. K. Najib, F. Bukhari, R. Revina, and F. N. Salsabila, "Performance Comparison of Gradient-Based Convolutional Neural Network Optimizers for Facial Expression Recognition," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 3, pp. 927-938, 2022, doi: 10.30598/barekengvol16iss3pp927-938.
- [36] A. A. Oyedele, A. O. Ajayi, L. O. Oyedele, S. A. Bello, and K. O. Jimoh, "Performance evaluation of deep learning and boosted trees for cryptocurrency closing price prediction," *Expert Systems with Applications*, vol. 213, pp. 927-938, 2023, doi: 10.1016/j.eswa.2022.119233.
- [37] X. Pan, M. Chin, C. M. Ichoku, and R. D. Field, "Connecting Indonesian Fires and Drought With the Type of El Niño and Phase of the Indian Ocean Dipole During 1979-2016," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 15, pp. 7974-7988, 2018, doi: 10.1029/2018JD028402.
- [38] A. Gunadi, G. Gunardi, and M. Martono, "The Law of forest in Indonesia: Prevention and suppression of forest fires," *Bina Hukum Lingkungan*, vol. 4, no. 1, pp. 113-134, 2019, doi: 10.24970/bhl.v4i1.86
- [39] S. Ruiz and A. Putraditama, "Will the Start of Forest Fires Season Hamper Indonesia's Progress in Reducing Deforestation?," *World Resources Institute*, 2019. [Online]. Available: <https://www.wri.org/insights/will-start-forest-fires-season-hamper-indonesias-progress-reducing-deforestation> (accessed Jul. 26, 2023).
- [40] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, pp. 1-81, 2019.
- [41] M. O. Akinwande, H. G. Dikko, and A. Samson, "Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis," *Open Journal of Statistics*, vol. 5, no. 7, pp. 754-767, 2015, doi: 10.4236/ojs.2015.57075.
- [42] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*. New Jersey: John Wiley & Sons, Inc., 2004.
- [43] A. Kurniadi, E. Weller, S. K. Min, and M. G. Seong, "Independent ENSO and IOD impacts on rainfall extremes over Indonesia," *International Journal of Climatology*, vol. 41, no. 6, pp. 3640-3656, 2021, doi: 10.1002/joc.7040.

BIOGRAPHIES OF AUTHORS







Ender Hasafah Nugrahani    is a lecturer and researcher at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University. Bogor, Indonesia. She earned her Bachelor of Statistics and Master of Science in Applied Statistics from IPB University in 1987 and 1993 respectively. In 2003, she received her Doctorate in Applied Mathematics from the University of Saarland, Germany. She is currently head of the Department of Mathematics, IPB University. Her research areas include mathematical modeling and financial mathematics. She has published research articles in reputable national and international journals. She can be contacted at email: e_nugrahani@apps.ipb.ac.id.







Sri Nurdianti    received her Bachelor of Statistics and Master of Applied Statistics degrees from IPB University in 1984 and 1987, respectively. She earned her Masters in Computer Science from Western Ontario, Canada in 1991. In 2005, she received her Ph.D. in Applied Mathematics from Twente University, The Netherlands. She is currently a professor at Department of Mathematics, IPB University, Bogor, Indonesia. She is also a lecturer at the Department of Computer Science, IPB University. Her research area includes computational mathematics, natural language processing, fuzzy logic, singular value decomposition, machine learning, and data science. She has published many research papers in international conferences and reputable international journals. She can be contacted at email: nurdiati@apps.ipb.ac.id.







Fahren Bukhari     received his Bachelor of Statistics and Master of Applied Statistics degrees from IPB University in 1984 and 1987, respectively. He earned his Masters in Computer Science from Western Ontario, Canada. In 2012, he received his Ph.D. in Computing Science from Newcastle University, UK. He currently heads the division of Computational Mathematics at the Department of Mathematics, IPB University, Bogor, Indonesia. His research area includes parallel computing, computational mathematics, machine learning, and data science. He has published many research papers in international conferences and reputable international journals. He can be contacted at email: fahrenheit@apps.ipb.ac.id.







Mohamad Khoirun Najib     holds a Bachelor of Science in Mathematics and Master of Science in Applied Mathematics from IPB University, Indonesia in 2019 and 2022, respectively. He currently works as a research assistant in the division of Computational Mathematics, Department of Mathematics at IPB University, Bogor, Indonesia. His research area is applied mathematics and statistics in the field of climatology, including applied probability, statistical bias correction and downscaling, quantile mapping, empirical orthogonal function, fast Fourier transform, copula, and machine learning. He has published various research papers in international journals and conferences indexed in Scopus and Web of Science. He can be contacted at email: mkhoirun.najib@apps.ipb.ac.id or mohknajib@gmail.com.



Denny Muliawan Sebastian     is a fresh graduate with a bachelor of science in Mathematics at IPB University in 2023 with a thesis entitled "construction of the artificial neural networks for modeling the number of hotspots based on the climate indicators". He joined a computational mathematics research group with an interest in machine learning applications in geoscience research. He can be contacted at email: dennym111@gmail.com.



Putri Afia Nur Fallahi     is a fresh graduate with a bachelor of science in Mathematics at IPB University in 2023 with a thesis entitled "machine learning model using random forest and gradient boosting regression to estimates the number of hotspot in Kalimantan". She joined a computational mathematics research group with an interest in machine learning applications in geoscience research. She can be contacted at email: putriafianurfallahi@gmail.com.