OPEN ACCESS
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 9, No. 2, Dec 2023. pp. 137-149
doi: doi.org/10.21108/ijoict.v9i2.865

Predicting Forest Fire Hotspots with Carbon Emission Insights Using Random Forest and Gradient Boosting Regression

Irma Palupi [1*], Ayu Shabrina [2], Bambang Ari Wahyudi [3], Fauzi Arya Surya Abadi [4], Naila Al Mahmuda [5]

[1]*School of Computing,Telkom University, Indonesia*
[2]*Research Center for Computing, National Research and Innovation Agency, Bogor, Indonesia*
[5]*Department of Business Administration, Southeast University, Dhaka, Bangladesh*
[*] irmapalupi@telkomuniversity.ac.id

Abstract

This research paper focuses on predicting the dispersion of carbon emissions, a crucial indicator for identifying potential forest fire hotspots in the wooded regions of Sumatra Island, Indonesia. Forest fires, often triggered by extended periods of dry weather, result in significant environmental degradation, impacting both the ecosystem and the economy. Furthermore, health concerns arise from smoke inhalation, leading to respiratory problems. To achieve this predictive capability, we harnessed valuable datasets, including GFED4.1s for carbon emissions and ERA5 for historical climate indicators, spanning from 1998 to 2022. Employing supervised learning ensemble methods, specifically Random Forest Regression (RFR) and Gradient Boosting Regression (GBR), we sought to forecast carbon emissions. It is noteworthy that our predictions encompassed carbon emission values from 1998 to 2023, providing insights into recent trends. Our analysis showed that GBR did better than RFR in terms of evaluation metrics, with a root mean square error (RMSE) of 10.87 and a mean absolute error (MAE) of 2.91. This was done by carefully tuning the hyperparameters. Additionally, our study highlighted that precipitation, temperature, and humidity were the primary climate factors influencing carbon emission values. This research contributes to enhancing preparedness and mitigation strategies for forest fires in Sumatra Island, offering valuable insights for environmental and economic protection.

Keywords: Firespot Prediction, Carbon Emissions, Random Forest Regression, Gradient Boosting Regression

I. INTRODUCTION

FOREST fires are caused by a combination of natural and human factors. Among the natural elements, the onset of El Nino, which is marked by extended periods of dry weather, is critical. Forest flora becomes dry and prone to igniting during El Nino occurrences due to drought conditions. Forest fires can have an impact on living beings both directly and indirectly. Short-term changes are caused by the direct repercussions. Individual organisms are immediately exposed to flames, severe combustion, high-temperature gases, or confinement in soil and other settings where sufficient heat is conveyed to the organism's immediate surrounds in the event of direct effect [1]. This high temperature is enough to kill or hurt the creature. Forest fires can lead to detrimental ecological and economic consequences, including the propagation of carbon gas emissions and destruction of human property within the charred forest zones. In Indonesia, certain regions, particularly those on Sumatra

Received on 08 Nov 2023, Revised on 07 Dec 2023, Accepted and Published on 29 Dec 2023.

Island, frequently confront forest fires [2]. The initial step towards mitigating hotspot expansion involves analyzing data from hotspot predictions within the region. This predictive analysis employs the Ensemble Learning technique, incorporating models from Supervised Learning such as Random Forest Regression and Gradient Boosting Regression. These models are employed to anticipate firespot or hotspot values.

Machine learning, specifically gradient boosting and random forest, proves to be advantageous in various studies. In study [3] on COVID-19 patient recovery rates, an ensemble of machine learning models, including gradient boosting and random forest, efficiently identified high-risk patients with an impressive 98% accuracy, facilitating timely medical attention. Meanwhile study [4], machine learning models, including random forest, enhanced the precision of agricultural crop yield estimation by incorporating factors like temperature, precipitation, and soil conditions. These models outperform traditional regression methods, underscoring the effectiveness of random forests and machine learning in this context. Study [5] tackles dust forecasting using machine learning, with the gradient-boosting regressor (GBR) showing the highest accuracy. This highlights the advantage of gradient boosting, a machine learning technique, in addressing environmental challenges like dust forecasting and climate change impact mitigation. Furthermore, machine learning algorithms, particularly gradient boosting, excel in predicting myositis disease, showcasing their potential for early disease diagnosis and risk prediction based on clinical and demographic patient data. These results underscore the utility of gradient boosting in healthcare applications [6].

In order to provide better forecasting for forest fires, machine learning techniques or parameters still need to be developed. Several techniques, including Gradient Boosting Regression (GBR) and Random Forest Regression (RFR), can be used in forecasting. For this reason, several studies on those methods have recently been published. Weather prediction using Random Forest has less errors than Support Vector Machine (SVM) with values of 0.750 MSE [7]. This research underscores the challenges of using complex climate models for weather forecasting and highlights the value of machine learning. By employing a random forest machine learning model, the study achieves accurate predictions of solar radiation and wind speed in Tamil Nadu, India, without the need for expensive measuring instruments in various locations. One study proposes a multi-sensor system for ground-level fire detection, enhancing accuracy, especially in peatland areas where traditional satellite systems may fail due to adverse weather [8]. This approach integrates various sensors to measure factors like temperature, smoke, haze, and carbon dioxide, resulting in improved hotspot detection. Another researches from [9] and [10] focuses on forecasting forest fire hotspots in high-risk regions like Kalimantan and Sumatra, Indonesia, using the Long Short-Term Memory (LSTM) deep learning algorithm. The accuracy of hotspot prediction supports its potential for preventive measures. In a different study [11], machine learning algorithms were employed to estimate forest fire-affected areas, with SVM showing higher accuracy compared to K-Nearest Neighbor (KNN). This research emphasizes the applicability of machine learning in predicting and quantifying forest fire areas, notably in Riau Province. Study on [12] focuses on identifying carbon concentrations in haze from forest fires using carbon sensors. By distinguishing between different sources of fog, this approach aids in pinpointing forest fire emissions, particularly in fire-prone regions. Furthermore, [13] provided prediction models from various machine learning techniques, including SVM, k-Nearest Neighbor, Logistic Regression, Decision Tree, and Naïve Bayes, are explored to predict forest fire occurrences in South Kalimantan Province's peatland areas. SVM emerges as the most accurate method, shedding light on machine learning's potential in fire prediction. Moreover, researchers address imbalanced datasets in forest fire prediction by using the Synthetic Minority Oversampling Technique (SMOTE) to enhance classification performance [14]. The ensemble AdaBoost with SMOTE-balanced data consistently improves prediction accuracy. These studies collectively demonstrate the potential of machine learning and deep learning algorithms in enhancing forest fire prediction, monitoring, and mitigation efforts, not only in Indonesia but also in similar regions grappling with this environmental challenge. Accurate data analysis and innovative approaches play a crucial role in mitigating the destructive impact of forest fires on the environment and society [15].

Our previous research [16] delved into forecasting carbon emissions triggering forest fires using Random Forest Regression (RFR) and Artificial Neural Network methods. The temporal RFR model demonstrated heightened precision in projecting location and intensity for 2019, but its predictive accuracy diminished for 2020-2021. Spatially, both models capably identified fire locations, although RFR consistently overestimated

carbon emissions. Climate influences yielded distinct outcomes, with RFR emphasizing relative humidity as the dominant factor, while ANN indicated comparable contributions from average temperature, maximum temperature, and wind speed. This study diverges from [16] by integrating the Gradient Boosting Regression method into the existing Random Forest approach. Additionally, it introduces visualization outputs and performance outcomes via web-based interactions.

The main emphasis of this study is centered on the prediction of carbon emission distribution leading to hotspots on the island of Sumatra, Indonesia, using the RFR and GBR models. However, it is important to note that the study's scope primarily focuses on the implementation aspects of this prediction methodology, which is a key limitation. The objectives of the study are threefold, as fo¬¬¬llows: The primary goal is to ascertain the expected results of carbon emission data as indicators of hotspots between 2021 and 2023. This entails the utilization of random forest regression and gradient-boosting regression models for case studies focused on Sumatra Island. The second objective entails conducting a comparative analysis between the Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models. The main objective is to evaluate the precision of these models in forecasting carbon emissions that contribute to the formation of hotspot incidents, thus serving as indicators for potential fire hotspots. This comparison also encompasses an analysis of the implementation error rates linked to both models. The third objective aims to pinpoint climate indicators that exhibit the strongest correlation with the spread of carbon emissions, acting as indicators for the occurrence of hotspots. Furthermore, the aim is to quantify the percentage contribution of each climate indicator to the accuracy of the Random Forest Regression and Gradient Boosting Regression models. The dataset used for predicting hotspot values consists of monthly historical data from 1998 to 2022, obtained from the ERA5 climate indicator dataset. This dataset encompasses a multitude of climatic variables, encompassing rainfall, humidity, wind velocity, mean temperature, maximum temperature, and GFED carbon emissions.

II. METHOD

A. *Dataset*

The dataset employed for the purposes of training and testing comprises time series data obtained from GFED4.1s and ERA5, as referenced in sources [17] and [18]. The carbon emissions data obtained from GFED4.1s is structured in a grid format with a spatial resolution of 0.25° x 0.25°. This grid consists of 1440 columns and 720 rows. Each data file in the study reflects monthly emissions measured in grams of carbon per square meter [17]. The climate data utilized in this study encompasses various variables, including precipitation (tp), Humidity (d2m), windspeed (si10), avg. temperature (t2m), and maximum temperature (tmax). These data are sourced from ERA5, which is the fifth iteration of the European Centre Medium-Range Weather Forecasts (ECMWF) reanalysis dataset. The ERA5 dataset has been reformatted into a standardized grid system based on latitude and longitude, with a spatial resolution of 0.25° x 0.25° for the purpose of reanalysis and a resolution of 0.5° for uncertainty estimation (0.5° and 1° for ocean waves, respectively) [18].

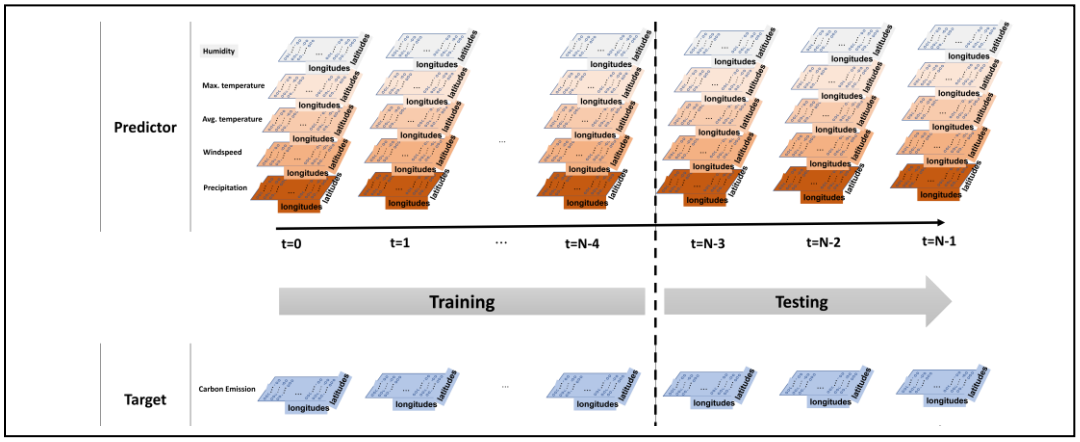


Fig. 1. Dataset description for monthly spatial predictors and response of the prediction model.

The main goal of data prediction, encompassing both training and testing phases, pertains to the carbon emission data. The aforementioned objective is accomplished by employing random forest regression and gradient-boosting regression models. In order to maintain uniformity of grid composition throughout the dataset, it is subjected to preprocessing with the cubic spline interpolation technique. In this scenario, the ERA5 data grid is used for computing because GFED's carbon emission data does not cover as much space. The dataset consists of 160,500 data entries and includes nine distinct features, consisiting time, latitude, longitude, precipitation, humidity, wind speed, average temperature, maximum temperature as previously mentioned. The dataset structure is illustrated in Fig. 1, while Table 1 provides an overview of the nine characteristics included in our experimental study. These features were utilized in the training of our model using the dataset variables. In addition to incorporating climate factors, it is imperative to include time indices (month and year) in the construction of the model. This is essential to capture seasonal patterns within our predictive model. Furthermore, the inclusion of location is necessary to minimize computational time, as the model is not developed for each individual location.

TABLE 1.
DATASET FEATURES/COMPONENT.

Features (X training)	Target (y training)
latitude, longitude, si10, d2m, t2m, tp, tmax, years_index, months_index	emissions

When we start our data recognition process, the first and most important step is to thoroughly examine and analyze the dataset. The main goal of this strategy is to gain important insights on the distribution of each component of the data. Fig. 2 illustrates this by using boxplots to display the various dataset components. The visualizations highlight the distinct distribution patterns that each component possesses, which are frequently distinguished by different units and value ranges. As a result, there is an urgent need for data scaling, which is an important strategy for standardizing feature sizes and mitigating the risk of overfitting in certain machine learning models.

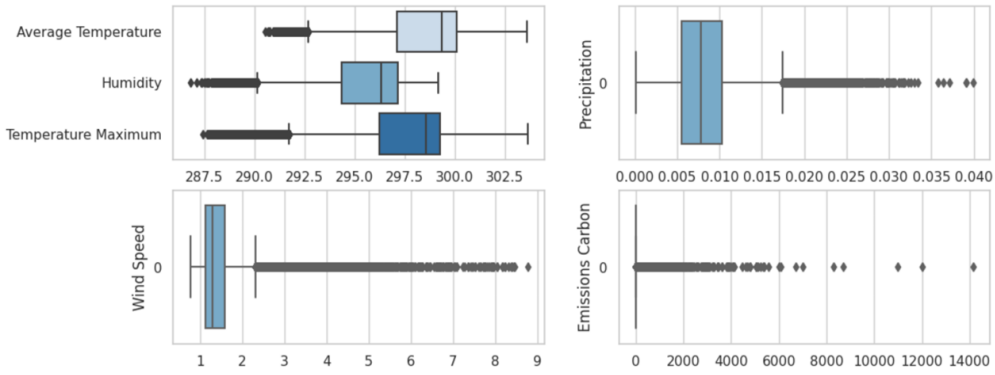


Fig. 2. Carbon Emission and climate factors distribution.

B. Regression Model

We use Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models to help with regression and data fitting problems. These models help anticipate carbon emissions by acting as markers of hotspot or fire spot occurrences. RFR aims to investigate the interaction between independent variables in climate data and their impact on dependent variables in carbon emissions. This approach aids in determining the importance of each element [19]. In this work, GBR emerges as a critical alternative to RFR in predicting carbon emissions, functioning as a signal for hotspots or fire spots once again. The fundamental distinction is

between bagging methods used in RFR and boosting techniques used in GBR [20]. The gradient boosting technique, which is the foundation of GBR, employs a regression function with decision trees designed within a gradient improvement framework. This framework results in more consistent improvements in carbon emission prediction estimates [21].

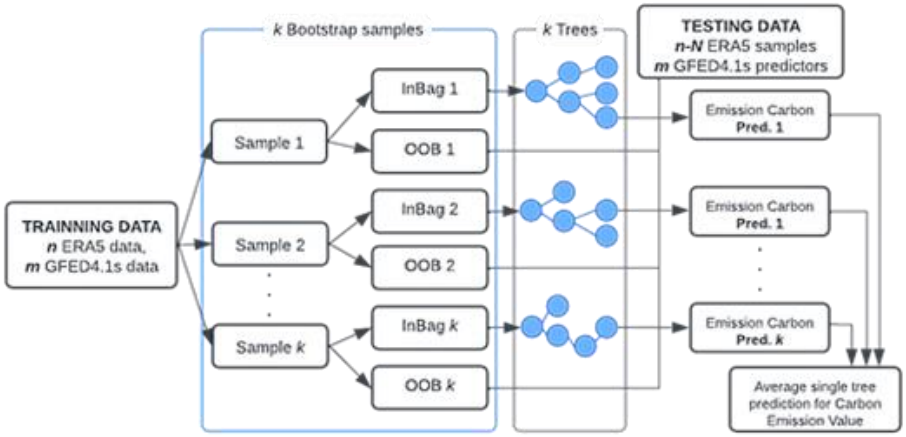


Fig. 3. Random Forest Regression predictors of carbon emissions, (Adapted from Rodriguez-Galiano [24]).

Let's have a look at how the RFR model predicts carbon emissions in Fig. 3. During training, we use feature data (n) that includes climate variables and target data (m) that represents carbon emissions. During the testing phase, we select N test data points from ERA5 spanning 2020–2022 (n–N). During training, the RFR algorithm generates a large number of decision trees using a resampling technique known as bagging or bootstrapping [22]. Each of these trees produces forecasts for carbon emission values, which are then aggregated to produce the best carbon emission predictions. The GBR algorithm, on the other hand, estimates carbon emission values through an iterative process of training and testing. Each GBR model iteration corrects earlier base model flaws in estimating carbon emission quantities as described in Fig. 4. By supplying the most relevant data to each successive model, resampling the training data enables this adjustment. The model iterates several times, building decision trees [20] until it reaches the nth tree, a parameter that conFig.s the model. Each iteration of the decision tree adds to the overall prediction strength, ending in a reliable carbon emission value prediction model. This method reduces the risk of overfitting that can occur with overly complex models [20].

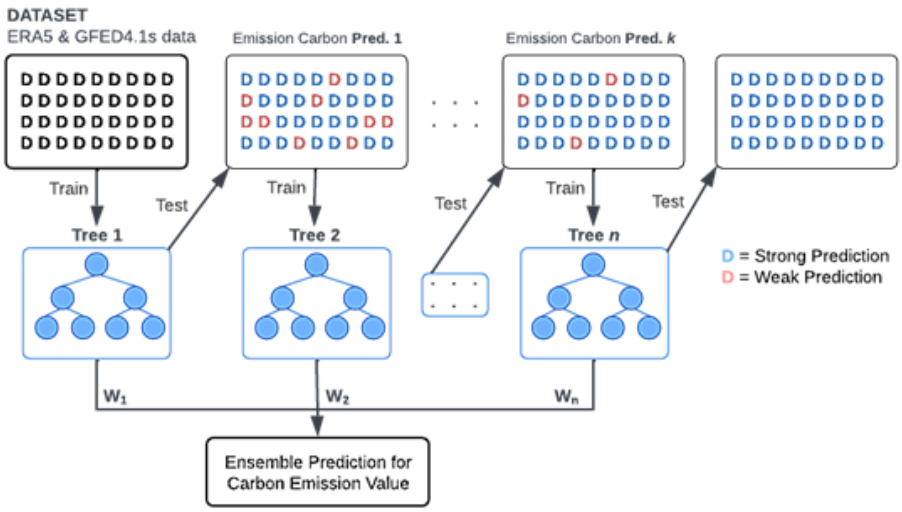


Fig. 4. Gradient Boosting Regression predictors of carbon emissions. (Adapted from Changming Zhao et al [25]).

C. Regression Model

In the implementation, the climate indicators and carbon emissions, as time series datasets, are carefully split into training and testing subsets. This is a very important part of the model development and evaluation processes. This splitting is exemplified in Fig. 5, which illustrates the integration process. In this process, the dataset is segmented into three distinct parts. The first part encompasses data spanning from 1998 to 2019, which serves as the foundation for training the model. The second part comprises data spanning from 2020 to 2022 and is earmarked for testing the model's performance. Notably, within this segment, data from 2020 is utilized for training models to predict carbon emissions in 2021, and data from 2021 is employed for training models to predict emissions in 2022. The third and final part of the dataset includes data from 2022, which acts as a critical component for projecting carbon emission predictions in 2023. Both the Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models undergo extensive training using the training data and subsequent testing using the testing data. The objective is to evaluate and ascertain the models' capabilities in predicting the distribution of carbon emissions from 2020 to 2022. Utilizing these effective models to forecast the emission of carbon in 2023 is the next step once both models have achieved satisfactory performance results during testing. Fig. 5 shows how the overall system design for this study organizes this process by putting the splitting of data, the training and testing of models, and the eventual projection of carbon emission values into a coherent and effective framework.

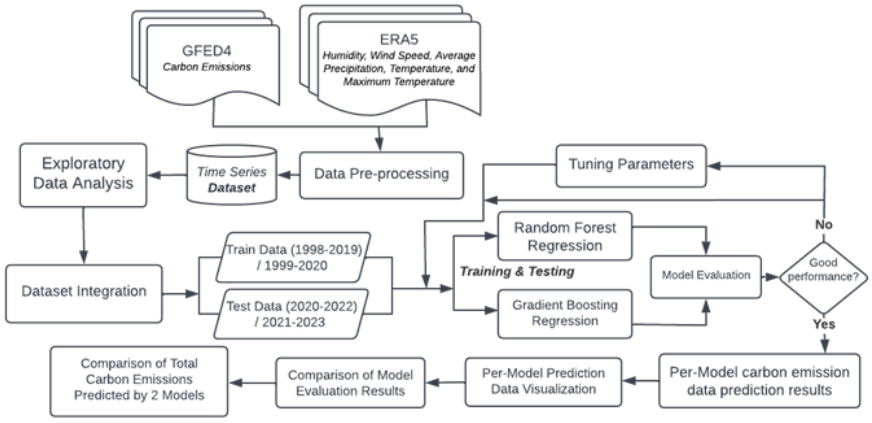


Fig.5. Flowchart System

To evaluate the predictive ability of the two proposed techniques, we utilize assessment scores that quantify errors, namely absolute error (MAE) and root mean squared error (RMSE). This study used RMSE and MSE (mean squared error) as the principal metrics for evaluating the performance of the model [23]. Root Mean Square Error (RMSE) is a significant statistic that is employed to quantify the typical magnitude of discrepancy between the projected values and the actual values. The metric calculates the mean vertical deviation between the observed values and their respective anticipated values along the regression line. In contrast, the mean absolute error (MAE) is concerned with assessing the average magnitude of absolute errors between anticipated values and actual values [23]. The metrics presented in formulas (1) and (2) are expressed mathematically, offering a quantitative approach to evaluating the precision and efficacy of our prediction models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{prediction}(i) - y_{actual}(i)\right)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{prediction}(i) - y_{actual}(i)\right| \quad (2)$$

In the analysis of our results, we present the experimental setup for each approach in order to evaluate the influence of hyperparameter adjustment on the accuracy of predictions. The aim of this research is to determine the most effective model for estimating carbon emissions from a subset of models that demonstrate high performance. In order to achieve this objective, we carried out three scenarios in order to assess each model. The initial scenario involves the development of models without adjusting parameters, wherein both models just utilize default configurations. In the subsequent scenario, we explore the process of hyperparameter tuning through the refinement of models using a subset of the dataset. The third scenario entails enhancing the outcomes of hyperparameter tweaking from the second scenario by employing the complete dataset for model construction. The purpose of these scenarios is to acquire a thorough comprehension of how hyperparameter adjustment improves forecast accuracy.

III. RESULTS AND DISCUSSION

We determined Pearson's correlation coefficient for each predictor variable to begin our investigation into how the variables under consideration influence carbon emissions, specifically as a hotspot indication for forest fires. Fig. 6 depicts the relationship between each climatic indicator and carbon emissions. This experiment was carried out in order to find the climate indicators that have the greatest influence on the distribution of carbon emissions. Based on Fig. 6, it is clear that the variable "wind speed" has the highest correlation value with carbon emissions, measuring 0.1611. It is crucial to highlight, however, that the linearity of any climate component and carbon emissions is not guaranteed. As a result, this correlation metric may not be totally appropriate for our situation. As a result, we use a metric known as "feature importance" to conduct a thorough factor analysis, analyzing the dependence of all components on carbon emissions as the target variable.

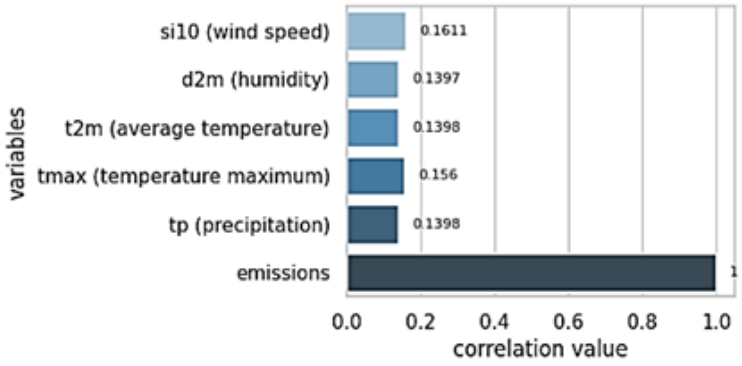


Fig.6. Cross-correlation of climate factors to carbon emission.

A. Model Evaluation

The findings of three scenarios of testing for both the RFR and GBR models are presented in Table 2. The purpose of these scenarios is to assess and contrast the efficacy of the models across different circumstances. The best model is found in scenario three, where both the MAE and the RMSE that are respectively written in formula (1) and (2), are the lowest for relevant RFR and GBR models. The first scenario in the RFR model, referred to as "without tuning," produces a MAE of 625.63 and a RMSE of 25.01. In the second scenario, labeled "With Hyperparameter Tuning," the model exhibits enhancement, as seen by a MAE value of 111.93 and a RMSE value of 12.57. The third scenario, which involves evaluating the adjustment of parameter values resulting from hyperparameter tuning, demonstrates superior performance with a MAE of 108.91 and a RMSE of 10.43. In the context of the GBR model, it is observed that the initial scenario yields a MAE value of 3.00 and a RMSE value of 15.26. In the second case, the MAE is observed to be 3.60, while the RMSE is calculated to be 13.75. The optimal performance is observed in the third scenario, with an MAE of 2.91 and a RMSE of 10.87. In general, the table presents a comprehensive analysis of the model's efficacy under various

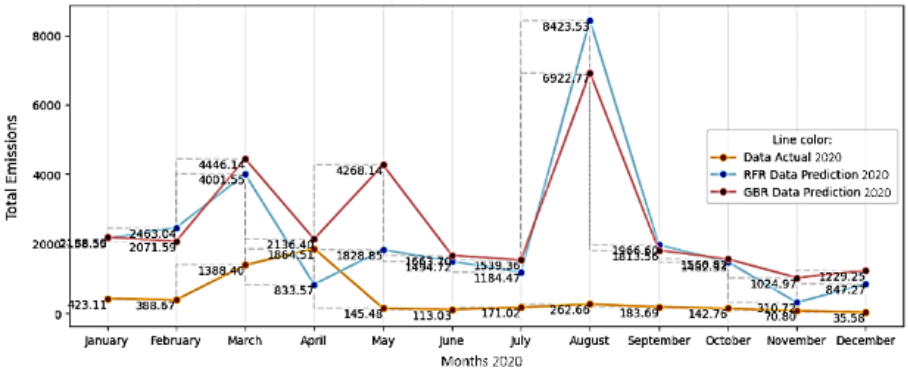
circumstances, enabling us to discern the optimal setup for forecasting carbon emissions within the framework of forest fire hotspots.

TABLE 2.
DATASET FEATURES/COMPONENT.

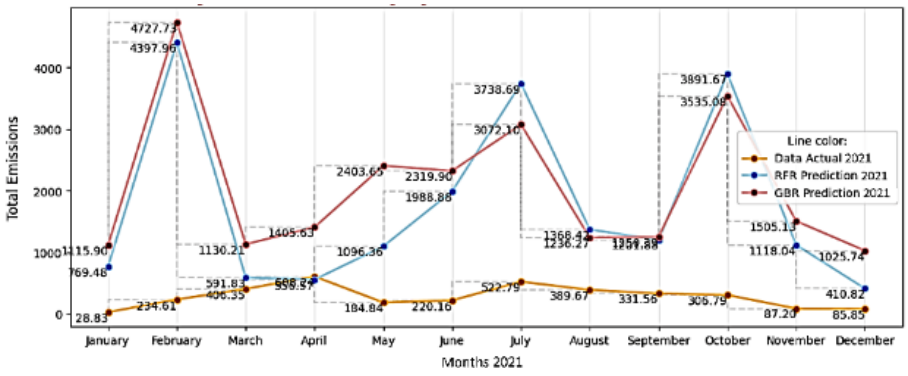
Scen.	Description	RFR			GBR		
		MAE	RMSE	Param value	MAE	RMSE	Param value
1	Without Tuning	625.63	25.01	n_estimators=150, max_depth=10, min_samples_split=10, random_state=42	3.00	15.26	n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42
2	With Hyperparameter Tuning	111.93	12.57	n_estimators=300, max_depth=30, max_features=0.5, random_state=42, min_samples_leaf=10, min_samples_split=30	3.60	13.75	n_estimators=500, max_depth=4, learning_rate=0.01, min_samples_split=10, random_state=42, criterion= 'friedman_mse'
3	Evaluation of adjustment of parameter values results of Hyperparameter Tuning	108.91	10.43	bootstrap=True, n_estimators=300, max_depth=30, max_features=0.3, min_samples_leaf=100 ,min_samples_split=30, random_state=42	2.91	10.87	n_estimators=300, max_depth=3, learning_rate= 0.01, min_samples_split= 120, random_state= 42 ,criterion= 'friedman_mse'

B. Carbon Emission Prediction in Sumatera

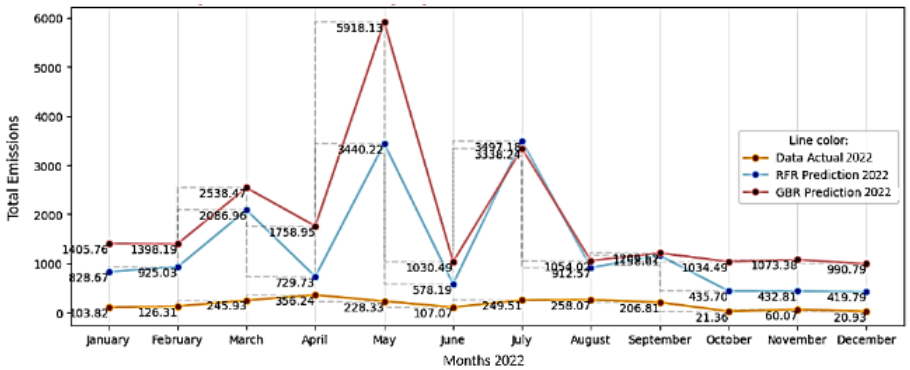
In scenario three, the optimal settings for both the Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models were identified. The optimal parameters were subsequently employed in constructing the predictive models. The objective of our testing phase was to predict the monthly carbon emissions from the years 2020 to 2022. The quantification of carbon emissions serves as a means to identify localized regions of heightened activity within the designated geographic area. The metric errors, as indicated in Table 2, are the mean values derived from all the sites encompassed within the selected grid. A comprehensive evaluation was carried out by consolidating total carbon emissions, as depicted in Fig. 7, in order to offer a comprehensive viewpoint on the degree of alignment between our projected values and the actual monthly data. This summary offers crucial insights into the overall precision and effectiveness of our prediction models in capturing trends and patterns of carbon emissions within the selected time period.



(a)



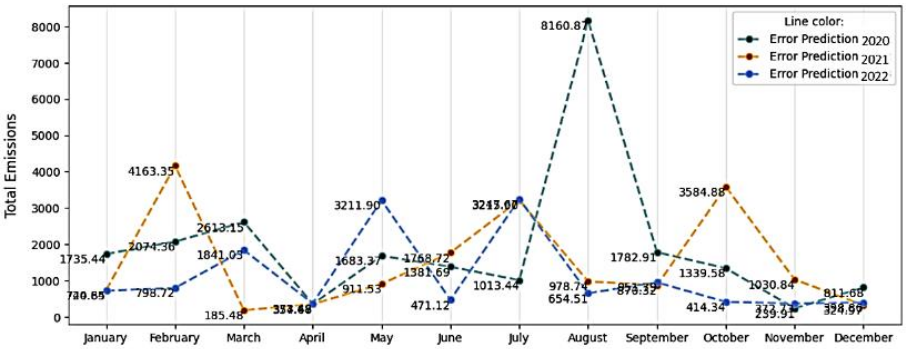
(b)



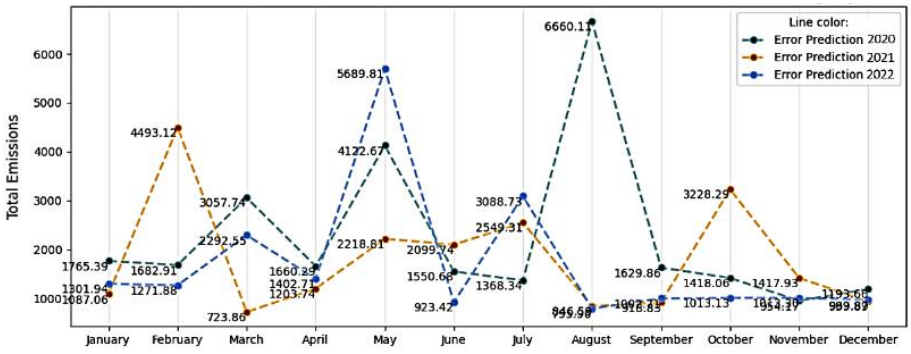
(c)

Fig.7. Carbon emissions prediction between RFR and GBR method (a) year 2020, (b) year 202, and (c) year 2022.

Fig. 7 demonstrates that both the RFR and GBR models' projections of carbon emissions for the year 2020 exhibit notable overestimations, particularly in the month of August. In contrast, the projections for carbon emissions in the year 2021 exhibit instances of overestimation throughout the months of February, July, October, and 2022, while displaying underestimations in May and July. It is worth mentioning that both the RFR and GBR models exhibit a tendency to overstate forecast values during some months. Notably, the predictions made for the year 2022 demonstrate a closer alignment with the actual GFED4.1s carbon emission data in comparison to the projections made for the years 2020 and 2021.



(a) RFR model's absolute error during 2020-2022.



(b) GBR model's absolute error during 2020-2022.

Fig. 8. Absolute error of RFR dan GBR during 3 years.

For a more comprehensive analysis, please refer to Fig. 8, which presents the monthly absolute error of both approaches in comparison to the actual data. According to Fig. 8, the RFR model demonstrates its highest level of carbon emission prediction error over the months of August 2020, February 2021, and July 2022. In contrast, the GBR model exhibits cumulative forecast inaccuracies throughout the months of August 2020, February 2021, and May 2022, as observed in the analysis of the prediction outcomes.

The Random Forest Regression (RFR) model, which was used to forecast carbon emissions, provided useful insights into the significance of various variables in explaining variation in emissions. Among the parameters analyzed, two stand out as having the most impact on carbon emission forecasts. The total precipitation (tp) proved to be the most relevant feature, accounting for roughly 25.54% of variable's contribution. This emphasizes the significant influence of rainfall intensity on carbon emissions, implying that greater precipitation may result in increased emissions. Furthermore, the maximum temperature (tmax) element was important, contributing around 22.11% to the prediction model. Meanwhile, total precipitation (tp) and geographical coordinates, or longitudes and latitudes, have scores of 42.8% and 16.1%, respectively, that show how much they affect the Gradient Boosting Regression (GBR) model. Notably, precipitation's contribution is dominant in both models, and maximum temperature also holds significant importance. However, according to Table 2, the GBR model exhibits superior accuracy compared to the Random Forest Regression (RFR) model, leading to the conclusion that only maximum temperature and precipitation significantly impact carbon emissions. Understanding the relative importance of these factors is crucial for guiding effective strategies to mitigate and regulate carbon emissions.

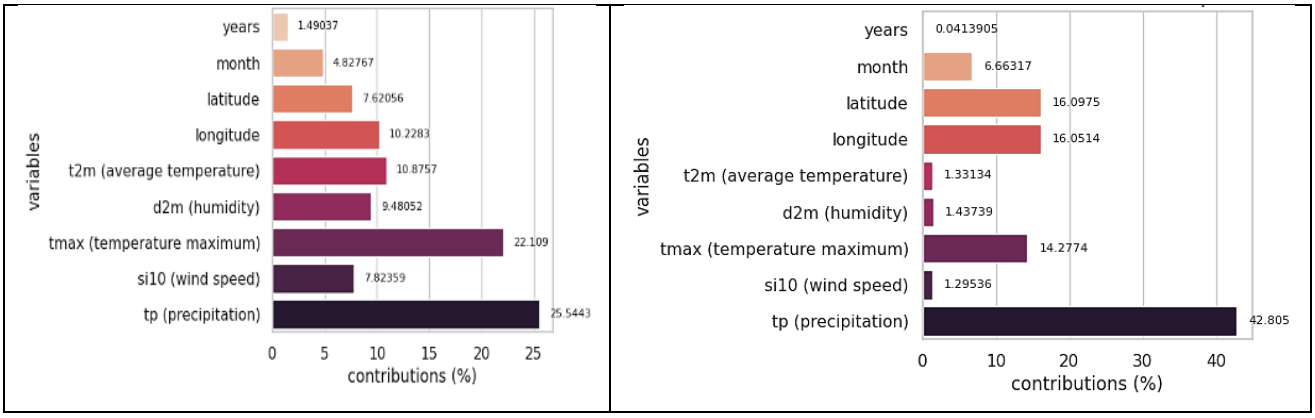


Fig. 9. Contribution of climate variables to RFR (left) and GBR (right) models.

C. Carbon Emission Distribution in Sumatera

The elevated carbon emission levels derived from the Global Fire Emissions Database (GFED) model serve as an indicator of heightened carbon production resulting from forest fire activities. Through the spatial representation of carbon emission values, it becomes feasible to make predictions regarding regions that are likely to experience high intensity burning based on specific climatic variables under consideration. As illustrated in Fig. 10, this predictive capability extends to the period spanning 2020 to 2022, with a focus on the months of September and October. Notably, our analysis highlights the model's adeptness in accurately forecasting the intensity of carbon emissions during the September to October 2019 period, as can be seen in Fig. 11, coinciding with the occurrence of the El Niño event, which is a plausible contributing factor to the observed elevated emissions.

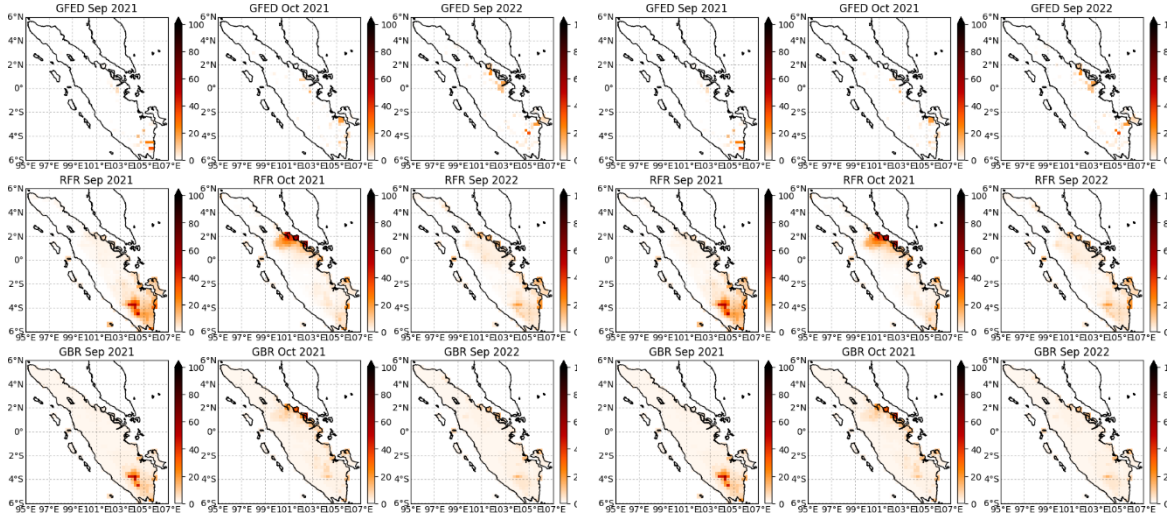


Fig. 10. Distribution of Carbon Emission Predictions 2021-2023 with RFR and GBR.

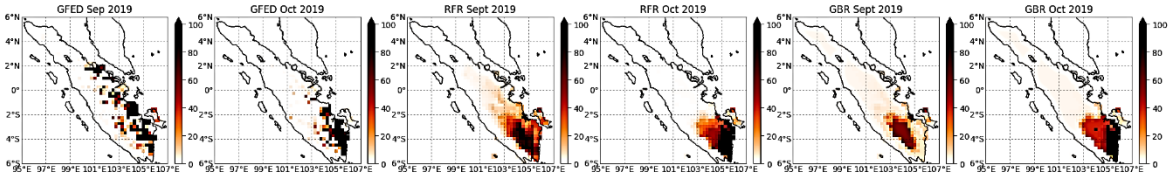


Fig. 11. Distribution of Carbon Emission Predictions 2019 with RFR and GBR.

IV. CONCLUSION

In conclusion, our study highlights the challenges associated with utilizing Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models for predicting carbon emissions as indicators of hotspots. Notably, these challenges pertain to issues of overestimation and inaccuracies, particularly in estimating the total amount of carbon emissions during specific months within the target area. Addressing these challenges necessitates thorough model evaluations and necessary adjustments to enhance accuracy and generalization capabilities, particularly during sensitive months. It is noteworthy that the RFR model exhibits a lower carbon emission error level compared to the prediction error rate score, indicating its relative effectiveness. Furthermore, our analysis underscores the significance of precipitation as a pivotal factor in predicting carbon dispersion, with both the RFR and GBR models highlighting its substantial contribution. It is worth mentioning that the GBR model outperforms the RFR model in terms of average error when predicting carbon emissions, underscoring its superior predictive capabilities. These insights provide valuable guidance for refining carbon emission prediction models and improving their reliability, especially in hotspot identification. Moreover, our findings reveal that precipitation is the dominant variable in terms of percentage contribution within the feature importance models for both RFR and GBR.

DATA AND COMPUTER PROGRAM AVAILABILITY

Data and result visualization used in this paper can be accessed in the following site https://fasa2297-visualization-of-carbon-emission-predictio-main-n6tuwd.streamlit.app/.

ACKNOWLEDGMENT

The first author would like to express our sincere gratitude to the Directorate of Research and Community Service at Telkom University for their invaluable support and funding. Their generous assistance has been instrumental in facilitating our research endeavors and has greatly contributed to the successful completion of this research article.

REFERENCES

[1] S. Verma and S. Jayakumar, "Impact of forest fire on physical, chemical and biological properties of soil: A," Proceedings of the International Academy of …, vol. 2, no. 3, pp. 168–176, 2012, [Online]. Available: http://www.iaees.org/publications/journals/piaees/articles/2012-2(3)/impact-of-forest-fire.pdf

[2] T. Tukiyat, F. H. Widodo, and R. D. Goenawan, "Disaster Risk Mitigation for Forest and Land Fire Prevention in Sumatera," Social, Humanities, and Educational Studies (SHEs): Conference Series, vol. 3, no. 1, pp. 54–60, 2020, doi: 10.20961/shes.v3i1.44980.

[3] A. Abdo, K. M. Elzalama, and A. E. Yakoub, "A machine learning model for predicting recovery rates of COVID-19 patients," vol. 31, no. 3, pp. 1656–1664, 2023, doi: 10.11591/ijeecs.v31.i3.pp1656-1664.

[4] Yuchen Sun; Cheng Chen; Xulei Shi; Xu Sun; Hong Fan, "Machine Learning Based Agricultural Crop Yield Estimation in Yingcheng District, Hubei Province," in 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Wuhan, China, 2023, pp. 1–4. doi: 10.1109/Agro-Geoinformatics59224.2023.10233541.

[5] A. Y. Hassan and M. H. Saleh, "Intelligence framework dust forecasting using regression algorithms models," vol. 32, no. 1, pp. 177–184, 2023, doi: 10.11591/ijeecs.v32.i1.pp177-184.

[6] S. R. K, Arunkarthick A and K, "Prediction of Myositis Disease using Machine Learning Algorithm," in 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), 2023, pp. 534–539. doi: 10.1109/ICIRCA57980.2023.10220885.

[7] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," Indonesian Journal of Electrical Engineering and Computer Science, vol. 22, no. 2, pp. 1208–1215, 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.

[8] E. A. Kadir, S. K. A. Rahim, and S. L. Rosa, "Multi-sensor system for land and forest fire detection application in Peatland Area," Indonesian Journal of Electrical Engineering and Informatics, vol. 7, no. 4, pp. 789–799, 2019, doi: 10.11591/ijeei.v7i4.1604.

[9] E. A. Kadir, H. T. Kung, A. A. AlMansour, H. Irie, S. L. Rosa, and S. S. M. Fauzi, "Wildfire Hotspots Forecasting and Mapping for Environmental Monitoring Based on the Long Short-Term Memory Networks Deep Learning Algorithm," Environments - MDPI, vol. 10, no. 7, 2023, doi: 10.3390/environments10070124.

[10] M. Listia Rosa, Sri and Abdul Kadir, Evizal and Syukur, Abdul and Irie, Hitoshi and Wandri, Rizky and Fikri Evizal, "Fire Hotspots Mapping and Forecasting in Indonesia Using Deep Learning Algorithm," in 2022 3rd International Conference on Electrical Engineering and Informatics (ICon EEI), Pekanbaru, Indonesia, 2022, pp. 190–194. doi: 10.1109/IConEEI55709.2022.9972281.

[11] Saruni Dwiasnati and Yudo Devianto, "Classification of forest fire areas using machine learning algorithm," World Journal of Advanced Engineering Technology and Sciences, vol. 3, no. 1, pp. 008–015, 2021, doi: 10.30574/wjaets.2021.3.1.0048.

[12] E. Abdul Kadir, S. Listia Rosa, A. Syukur, M. Othman, and H. Daud, "Forest fire spreading and carbon concentration identification in tropical region Indonesia," Alexandria Engineering Journal, vol. 61, no. 2, pp. 1551–1561, 2022, doi: 10.1016/j.aej.2021.06.064.

[13] D. Rosadi, Dedi and Andriyani, Widyastuti and Arisanty, Deasy and Agustina, "Prediction of Forest Fire Occurrence in Peatlands using Machine Learning Approaches," in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, pp. 48–51. doi: 10.1109/ISRITI51436.2020.9315359.

[14] Z. Rosadi, Dedi and Arisanty, Deasy and Andriyani, Widyastuti and Peiris, Shelton and Agustina, Dina and Dowe, David and Fang, "Improving Machine Learning Prediction of Peatlands Fire Occurrence for Unbalanced Data Using SMOTE Approach," in 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 2021, pp. 1660–163. doi: 10.1109/DATABIA53375.2021.9650084.

[15] S. Yang, M. Lupascu, and K. S. Meel, "Predicting Forest Fire Using Remote Sensing Data And Machine Learning," 35th AAAI Conference on Artificial Intelligence, AAAI 2021, vol. 17A, pp. 14983–14990, 2021, doi: 10.1609/aaai.v35i17.17758.

[16] 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," ACM International Conference Proceeding Series, pp. 194–198, 2022, doi: 10.1145/3575882.3575920.

[17] L. Giglio, J. T. Randerson, and G. R. Van Der Werf, "Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4)," J Geophys Res Biogeosci, vol. 118, no. 1, pp. 317–328, 2013, doi: 10.1002/jgrg.20042.

[18] H. Hersbach et al., "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.

[19] Y. Sun, F. Zhang, H. Lin, and S. Xu, "A Forest Fire Susceptibility Modeling Approach Based on Light Gradient Boosting Machine Algorithm," Remote Sens (Basel), vol. 14, no. 17, pp. 1–16, 2022, doi: 10.3390/rs14174362.

[20] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transp Res Part C Emerg Technol, vol. 58, pp. 308–324, 2015, doi: 10.1016/j.trc.2015.02.019.

[21] L. Gigović, H. R. Pourghasemi, S. Drobnjak, and S. Bai, "Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in Serbia's Tara National Park," Forests, vol. 10, no. 5, 2019, doi: 10.3390/f10050408.

[22] I. N. Wahyuni, A. Shabrina, and A. L. Latifah, "Investigating Multivariable Factors of the Southern Borneo Forest and Land Fire based on Random Forest Model," ACM International Conference Proceeding Series, pp. 71–75, 2021, doi: 10.1145/3489088.3489115.

[23] A. Jierula, S. Wang, T. M. Oh, and P. Wang, "Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data," Applied Sciences (Switzerland), vol. 11, no. 5, pp. 1–21, 2021, doi: 10.3390/app11052314.

[24] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[25] C. Zhao et al., "BoostTree and BoostForest for Ensemble Learning," IEEE Trans Pattern Anal Mach Intell, vol. 45, no. 7, pp. 8110–8126, 2023, doi: 10.1109/TPAMI.2022.3227370.