Pol. J. Environ. Stud. Vol. XX, No. X (XXXX), 1-14
DOI: *10.15244/pjoes/217446* ONLINE PUBLICATION DATE: 2026-05-19

Original Research

Forest Fire Susceptibility Analysis with Remote Sensing Data and Machine Learning Algorithms using Region-based Datasets

Cihan Uysal[1]°*, Mustafa Mutlu Uysal[2], Murat Uysal[3]

[1]Çanakkale Onsekiz Mart University, Faculty of Science, Department of Space Sciences and Technologies, Çanakkale, Türkiye
[2]Balıkesir Regional Directorate of Forestry, Balıkesir, Türkiye
[3]Afyon Kocatepe University, Faculty of Engineering, Surveying Engineering, Afyonkarahisar, Türkiye

Received: 04 November 2025
Accepted: 27 January 2026

Abstract

Forests are vital elements of terrestrial ecosystems and ensure the integrity and sustainability of natural factors such as soil, water, and climate. Forest fires are one of the disasters that cause deterioration of the ecosystem as well as social and economic impacts. In combating disasters, determining pre-disaster risks and reducing disaster damage is the first stage of disaster management. In this study, a dataset consisting of 679×14 rows and columns derived from 312 forest fire points between 2017 and 2022 was prepared. For the dataset, 13 independent variables were mapped with remote sensing and geographic information systems techniques from satellite images and open-source data. Training and prediction datasets were created by extracting values for all variables in each pixel. Feature relevance was initially assessed using Mutual Information (MI), followed by model-specific interpretation using SHAP values. Model performance was evaluated using confusion matrices and ROC-AUC analysis. Machine learning algorithms, including Decision Trees, Multi-Layer Perceptron (MLP), Naive Bayes, Random Forest, and XGBoost, were trained using a dataset split into 80% training and 20% testing. At the end of the study, forest fire sensitivity maps were created with an accuracy rate of 96% using the Random Forest and XGBoost algorithms, which were the most powerful models. Susceptibility probabilities were rescaled to a percentage scale and classified into low, medium, and high categories using a quantile-based approach. Results indicate that distance to roads and population density are the most influential predictors, highlighting the dominant role of human activity in wildfire ignition.

Keywords: forest fire, machine learning, random forest, remote sensing, susceptibility mapping, XGBoost

*e-mail: cihan.uysal@comu.edu.tr
°ORCID iD: 0000-0001-6006-5672

Introduction

The effects of climate change have been increasing in the world, and it is predicted that the climate crisis will cause an increase in forest fires. In the Synthesis Report of the 6th Report of the Intergovernmental Panel on Climate Change, it is stated that the global surface temperature in the first two decades of the 21st century (2001-2020) is 0.99ºC (0.84 to 1.10ºC) higher than that in 1850-1900 [1]. Fires that occur during periods of high air temperatures, low relative humidity, and strong, dry winds are severe. In general, the potential fire hazard is considered to be very high when the temperature is very high, and the relative humidity is below 10%. In Türkiye, located in the Mediterranean climate zone, the Aegean and Mediterranean regions are the areas with the highest risk of forest fires, and an increase in disaster risks is expected due to the effects of climate change [2, 3]. As in other parts of the world, the Mediterranean Basin, where Türkiye is located, has seen an increase in extreme weather events in recent years due to the effects of climate change. This is expected to increase the frequency and intensity of forest fires. Success in fire management depends not only on interventions during a fire but also on preventive measures and planning taken before a fire occurs. Therefore, pre-fire risk analyses, sensitivity mapping, and preparedness plans are as important as the efforts made to combat fires. In this context, the potential fire risks of forest areas must be identified using dynamic models [4].

Firefighting efforts in Türkiye are carried out by the General Directorate of Forestry (GDF). The GDF classifies fires according to their cause as Intentional, Negligence, Natural, and Unknown Cause. In 2022, the cause of 41% of fires in Türkiye was unknown [5].

It is possible to minimize the number, severity, and damage of forest fires with pre-fire measures. For this reason, there is much academic and institutional research on the causes of forest fires, measures to be taken, and fire behavior. Numerous studies have been conducted on forests using remote sensing and geographic information systems techniques [6-9]. Especially with the information technologies that have developed in recent years, big data has been created with developments such as the increase in data producers, such as the Internet of Things, and the storage of this data. By integrating remote sensing and artificial intelligence techniques, risk analysis research is conducted for forest fires from big data [10-16]. Accurate wildfire susceptibility mapping is essential for prevention planning, resource allocation, and land-use management. Traditional statistical approaches often struggle to capture nonlinear relationships among wildfire drivers, motivating the adoption of machine learning techniques.

When examining test accuracies obtained using different machine learning algorithms in the literature, it is evident that the characteristics, size, and regional differences of the dataset used have a critical impact on accuracy. For example, in the study by [13], a decision tree algorithm was used with Remote Sensing data (NDVI, LST, TA) to achieve a test accuracy of approximately 0.93.

The objective of this study is to analyze remote sensing data and the coordinates of forest fires that occurred in the study area over 6 years, using a dependent variable (fire presence/absence) and 13 independent variables to analyze the meteorological effects on forest fires using machine learning algorithms in the cloud-based Google Colab interface, and to create susceptibility maps. Furthermore, it is aimed to contribute to forest fire susceptibility studies by comparing the effects of independent variables and models created with different algorithms.

Materials and Methods

Study Area

The study area shown in Fig. 1 covers a total area of 2262.65 km², including Ayvalık, Burhaniye, Edremit, Havran, and Gömeç districts of Balıkesir province of Türkiye, and a part of İvrindi district. This area, which has Mediterranean climate characteristics, is on the coast of the Aegean Sea.

Within the borders of the province where the study area is located, 445 fires occurred between 2017 and 2022, and the number of fires and the burned area by year are shown in Fig. 2. Although the amount of burned area varies according to year, the number of fires has been increasing from the past to the present.

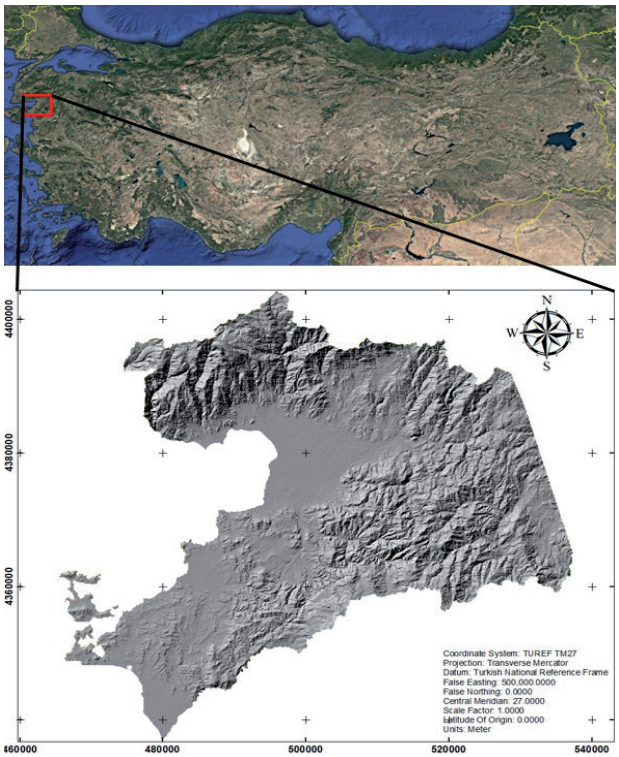

Fig. 1. Location map.

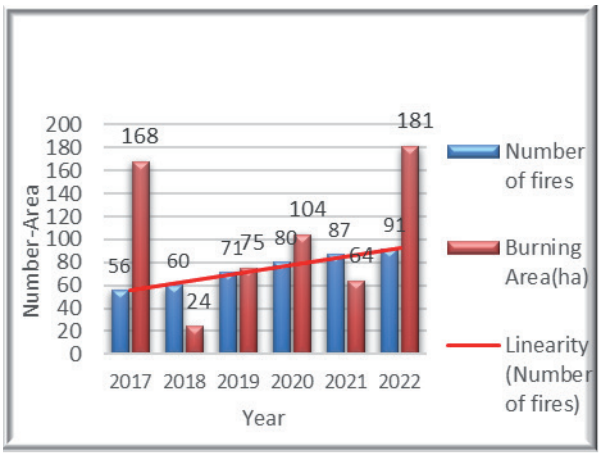

Fig. 2. Number and areas of forest fires in Balıkesir Province (2017-2022).

Dataset

Official fire data (312 points) provided by GDF were used as the dependent variable for fire occurrence points. The non-fire class was created using a random point sampling method to ensure that the points were entirely outside the fire polygons. However, due to missing and erroneous data in some variables, data cleaning (NaN removal) was performed, and the final sample size was determined as 679 rows.

For the dataset, 13 independent variables were identified as: Elevation, Slope, Aspect, Land Cover/ Land Use, Forest canopy density, NDVI, Distance from settlement, Distance from roads, Population density, Wind speed, Soil moisture, Precipitation, and Maximum temperature (Table 1). Independent variables were mapped with a spatial resolution of 30 m × 30 m using remote sensing and geographic information systems techniques from satellite imagery and open-source data.

A dataset was created by extracting values for each pixel. It consists of 13 independent variables, 1 dependent variable, and 679 × 14 rows and columns.

Topographic data such as elevation, slope, and aspect variables were generated from ASTER GDEM data. Elevation and slope map data are numeric and proportional, while Aspect data are categorical and nominal (Fig. 3(b-d)). Topographic data were generated as one map for each variable.

Vegetation data consist of Land Cover/Land Use, Forest Cover, and Normalized Difference Vegetation Index (NDVI) (Fig. 3a), h), i)). LU/LC and NDVI were created from Sentinel satellite imagery, and forest canopy was created from the GDF database, and 6 (year) × 1 (average of 5 months for each year) maps were produced with annual temporal resolution. Land use and forest cover variables are categorical data, and land use is scaled nominally, and forest cover is scaled ordinal. The NDVI variable was generated as numerical data. Vegetation data were produced for each year by taking the mean for the months of May, June, July, August, and September when fires are intense.

In forest canopy cover data, the 0 value indicated sparse canopy cover with 10% or less canopy cover, the 1 value indicated loose canopy cover between 11% and 40%, the 2 value indicated medium canopy cover between 41% and 70%, and the 3 value indicated full canopy cover between 71% and 100% [17].

Anthropogenic data consist of distance to settlement, distance to road, and population density variables (Fig. 3(e-g)).

OpenStreetMap was used as a source for distance to settlement and distance to road variables, while WorldPop data via Google Earth Engine was used as a source for population density.

The meteorological dataset consists of wind speed, soil moisture, precipitation, and maximum temperature variables (Fig. 3(j-m)). These variables were generated from the TerraClimate dataset via Google Earth Engine.

A total of 30 maps were produced for each variable by taking the average for each month in May, June, July, August, and September. The entire dataset was mapped in ArcMap software in the same coordinate system and at the same spatial resolution. A training set was prepared using the "Extract Multi Values to Point" tool with fire presence and absence points.

Aspect, land use/land cover (LULC), and canopy cover were encoded using dummy (one-hot) variables to avoid imposing artificial ordinal relationships among categories. Aspect was represented by 9 directional classes, LULC by 5 land cover categories, and canopy cover by 4 density classes. Although dummy encoding increases feature dimensionality, the total number of predictors remained modest relative to the sample size. Furthermore, Random Forest and XGBoost are robust to high-dimensional feature spaces, minimizing potential dimensionality issues.

Fire occurrence labels correspond only to observed fire events, while predictor variables represent the characteristic background environmental and climatic conditions of the fire season. This design ensures temporal consistency between predictor and response variables and prevents information leakage from post-event data into model training.

By extracting values for 13 independent variables in all pixels within the study area, the dataset to be used in the prediction phase, consisting of 113116 × 13 rows and columns, was prepared.

Methodology

A training and prediction dataset was created with remote sensing data of the 13 independent variables using the coordinates of forest fires that occurred within the borders of Balıkesir province, to which the study area is administratively connected. The data were classified at a monthly temporal resolution from May to September in 6 years between 2017 and 2022, and each point was checked from satellite images,

Table 1. Description of independent variables in the dataset.

VARIABLE	Data (~30 m)	Data Form	Unit	Source	Maps
Topographic	Elevation	Raster	meter	ASTER-GDEM (~30 m)	1
	Slope	Raster	degrees	ASTER-GDEM	1
	Aspect	Raster	degrees	ASTER-GDEM	1
Vegetation	LU/LC	Raster	meter	Sentinel (~10 m)	6
	Canopy	Raster	class	GDF-Database	6
	NDVI	Raster	ratio	Sentinel	6
Anthropogenic	Distance from Settlement	Raster	meter	Open Street Map	1
	Distance from Roads	Raster	meter	Open Street Map	1
	Population Density	Raster	ratio	Open Street Map	1
Meteorological	Wind speed	Raster	m/s	TerraClimate (~4 km)	30
	Soil moisture	Raster	mm	Terra Climate	30
	Precipitation	Raster	mm	Terra Climate	30
	Maximum Temperature	Raster	ºC	Terra Climate	30

and erroneous points were eliminated. For the burning areas in the finalized points, each pixel was assigned a fire presence value. Non-fire points were generated by random point generation and equal distribution control. By extracting values from 13 independent variables created with remote sensing data and techniques, a training dataset consisting of 679 × 13 rows and columns, and a prediction dataset consisting of 113116 × 13 rows and columns covering all pixels of the study area were prepared. The training dataset was visualized, and correlation analysis and Mutual Information (MI) were performed with Exploratory Data Analysis using Python language in the Google Colab interface. The visualization and applicability of the training dataset were analyzed with data science techniques using the cloud-based Google Colab interface and Python software language. A correlation heatmap and Mutual Information analysis showing the relationship between the independent variables and the dependent variable were created (Fig. 4, Table 2).

Dummy encoders were used for Aspect, Land Cover/Land Use, and Forest canopy density variables, which are categorical data in the training phase, so that the variables can be used in machine learning algorithms. The non-normal distribution of the variables in the dataset is a factor that affects the operation of some algorithms. If the data is skewed to the right or skewed to the left, model performance may be affected. In this case, standardization, which is a Feature Scaling method, was used to normalize the values and reduce dominance. The dataset was divided into 80% training and 20% testing. Next, the prediction phase was initiated, and training and predictions were performed using the most successful algorithms from among five different ones: Random Forest and XGBoost (Fig. 5).

Subsequently, model-specific interpretation was performed using SHAP values. Model performance was evaluated using confusion matrices and ROC-AUC analysis.

Feature Selection

Feature selection is the process of identifying and selecting a subset of input variables that are most relevant for model construction. Effective feature selection improves prediction accuracy, reduces computational cost, and enhances model interpretability [18].

In this study, Mutual Information (MI) was used as a model-independent measure to assess the relevance of predictor variables and capture potential nonlinear dependencies with forest fire occurrence, in addition to a correlation-based filtering approach for feature selection.

Pandas and NumPy libraries were used for data visualization. According to the correlation heatmap, the highest correlation with forest fires is negative for elevation in topographic variables, negative for NDVI in vegetation variables, negative for distance to road in anthropogenic variables, and positive for maximum temperature in meteorological variables (Fig. 4).

Mutual Information was employed as a model-independent measure to assess the relevance of predictor variables and to capture potential nonlinear dependencies with wildfire occurrence. MI was used as a preliminary screening step rather than as a strict elimination criterion. Since all predictors exhibited MI values greater than zero, no variables were excluded at this stage (Table 2).

Model-specific feature importance and interpretability were subsequently evaluated using SHAP analysis for both Random Forest and XGBoost models.

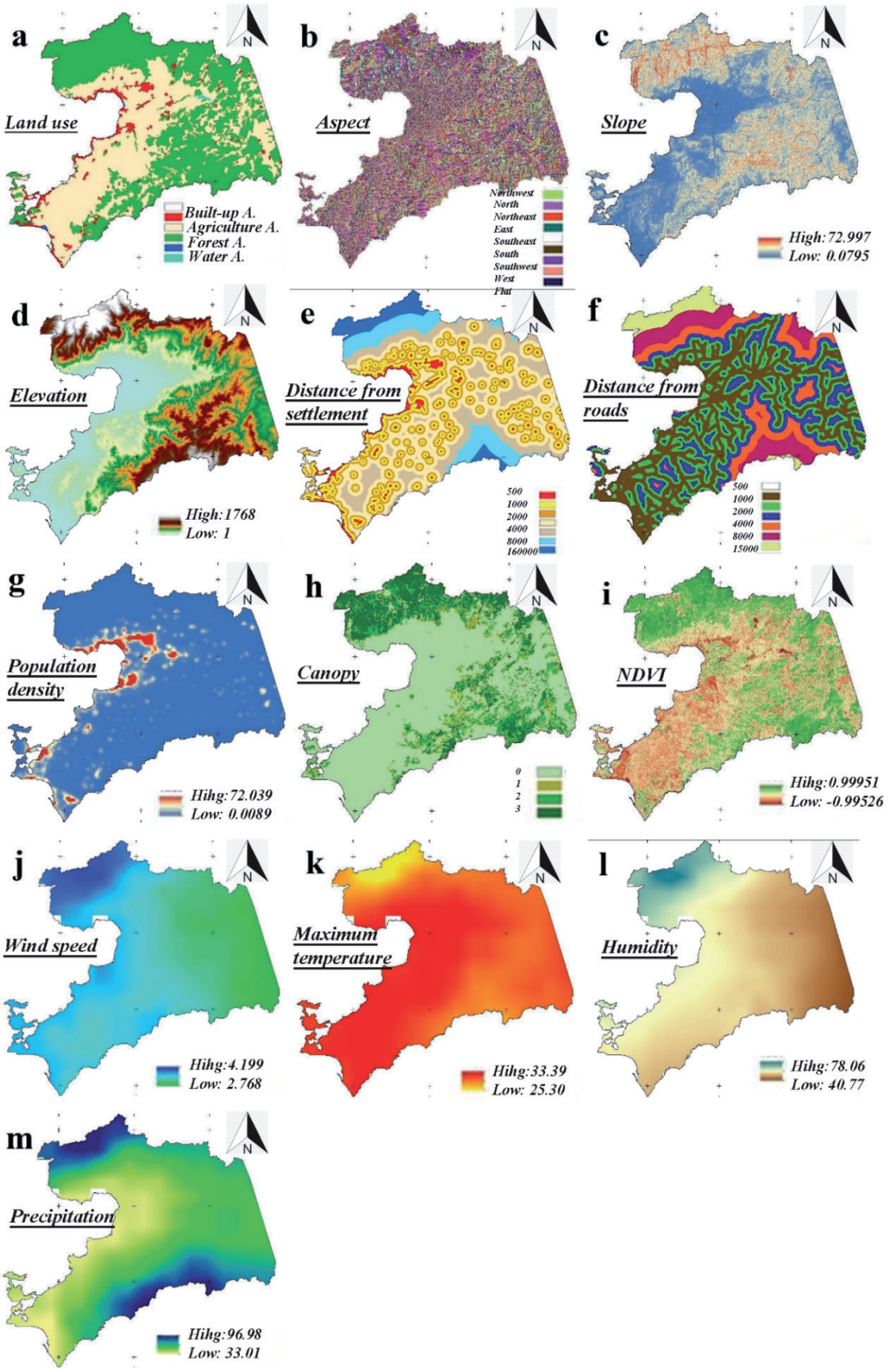

Fig. 3. Thirteen forest fire factors used in sensitivity analysis for forest fires.
a) Land use, b) Aspect, c) Slope, d) Elevation, e) Distance from settlement, f) Distance from roads, g) Population density, h) Canopy, i) Normalized Difference Vegetation Index (NDVI), j) Wind speed, k) Maximum temperature, l) Humidity, m) Precipitation.

	slope	elevation	aspect	lulc	NDVI	canopy	s_distance	r_distance	pop	wind speed	S_moisture	Precipitation	Max_temp	YES/NO
slope	1.000000	0.268010	-0.146143	0.116278	0.212532	0.172903	0.223711	0.321270	-0.231565	-0.021483	0.045177	0.062367	-0.086038	-0.154260
elevation	0.268010	1.000000	0.014835	0.209853	0.158911	0.229426	0.481149	0.442556	-0.580619	-0.180269	-0.098797	0.004473	-0.205196	-0.311876
aspect	-0.146143	0.014835	1.000000	-0.073276	-0.227091	-0.212218	-0.075821	-0.190675	0.045777	0.100649	-0.065799	-0.069607	0.086476	0.156232
lulc	0.116278	0.209853	-0.073276	1.000000	0.276771	0.306553	0.287602	0.293825	-0.358489	0.006736	0.105552	0.069212	-0.168591	-0.388419
NDVI	0.212532	0.158911	-0.227091	0.276771	1.000000	0.554581	0.352519	0.417558	-0.278124	0.017609	0.176427	0.146288	-0.212830	-0.389686
canopy	0.172903	0.229426	-0.212218	0.306553	0.554581	1.000000	0.403509	0.418565	-0.284030	0.034426	0.111451	0.100514	-0.173385	-0.322855
s_distance	0.223711	0.481149	-0.075821	0.287602	0.352519	0.403509	1.000000	0.781071	-0.558169	0.035687	0.164937	0.158214	-0.307416	-0.533148
r_distance	0.321270	0.442556	-0.190675	0.293825	0.417558	0.418565	0.781071	1.000000	-0.532209	-0.040169	0.245816	0.247772	-0.402875	-0.651207
pop	-0.231565	-0.580619	0.045777	-0.358489	-0.278124	-0.284030	-0.558169	-0.532209	1.000000	0.185050	-0.116208	-0.179164	0.329959	0.579076
wind speed	-0.021483	-0.180269	0.100649	0.006736	0.017609	0.034426	0.035687	-0.040169	0.185050	1.000000	0.005103	-0.334516	0.318139	0.106413
S_moisture	0.045177	-0.098797	-0.065799	0.105552	0.176427	0.111451	0.164937	0.245816	-0.116208	0.005103	1.000000	0.446192	-0.427865	-0.350996
Precipitation	0.062367	0.004473	-0.069607	0.069212	0.146288	0.100514	0.158214	0.247772	-0.179164	-0.334516	0.446192	1.000000	-0.482933	-0.313905
Max_temp	-0.086038	-0.205196	0.086476	-0.168591	-0.212830	-0.173385	-0.307416	-0.402875	0.329959	0.318139	-0.427865	-0.482933	1.000000	0.498467
YES/NO	-0.154260	-0.311876	0.156232	-0.388419	-0.389686	-0.322855	-0.533148	-0.651207	0.579076	0.106413	-0.350996	-0.313905	0.498467	1.000000

Fig. 4. Correlation heatmap.

Table 2. Mutual Information (MI) scores of predictor variables used in wildfire susceptibility modeling. Higher MI values indicate stronger dependency between the predictor and fire occurrence.

Rank	Predictor Variable	MI Score
1	r_distance	0.492
2	pop	0.420
3	s_distance	0.412
4	elevation	0.333
5	slope	0.285
6	NDVI	0.284
7	Max_temp	0.261
8	Precipitation	0.140
9	S_moisture	0.137
10	lulc	0.104
11	wind speed	0.068
12	canopy	0.053
13	aspect	0.040

Algorithms

Decision Tree (CART)

One of the algorithms used in this study is CART, which recursively splits the sample space into binary partitions; it selects the best splits based on the impurity measure (Gini/Entropy) [19]. The CART algorithm converts complex structures into simple decision structures by making a series of binary choices [20]. CART can be considered a subset of decision trees and generally has a more distinct and consistent methodology.

Multi-Layer Perceptron (MLP)

The other algorithm used in this study is a feedforward neural network composed of fully connected (dense) layers. Nonlinearity is provided by activation functions (ReLU, tanh, sigmoid); it learns through backpropagation and derivative-based optimization. It is the building block of classical pattern recognition and modern deep learning [21].

Naive Bayes (NB)

The other algorithm used in this study is based on Bayes' theorem and the assumption of conditional independence (between features). GaussianNB assumes that the class-conditional distribution of each feature is Gaussian; it provides closed-form parameter estimation and very fast learning [22].

Random Forest (RF)

Bagging trains multiple decision trees on random samples and feature subsets; it provides a collective vote/average prediction. It reduces variance and is generally robust and resilient [23]. Random Forest is a widely used machine learning algorithm developed by Leo Breiman and Adele Cutler that combines the outputs of multiple decision trees to reach a conclusion. It is one of the ensemble algorithms that stands out for its simplicity and flexibility, as it handles both classification and regression problems [20].

Extreme Gradient Boosting (XGBoost)

In XGBoost, successively added weak learners (trees) focus on previous errors; regularization (L1/L2) and second-order approaches (Newton/approx.) are efficient and powerful. It is optimized for big data and sparsity [24]. In machine learning, ensemble methods,

a subset of supervised learning algorithms, typically exhibit high performance in prediction and classification problems. One of the ensemble methods, Extreme Gradient Boosting, is a gradient boosting algorithm that works on decision trees and has high performance [20].

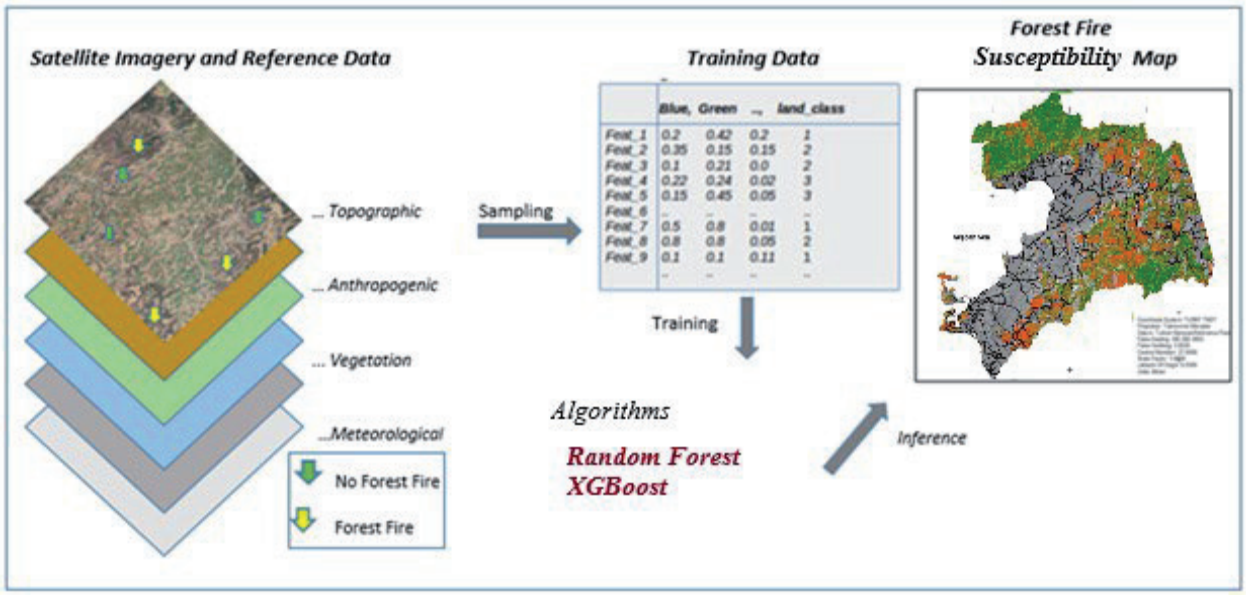

Fig. 5. Research methodology.

Table 3. The base metrics used for model evaluation.

Model	Accuracy	Precision	Recall	Specificity	F1 Score	K-Index
Decision Tree	0.904412	0.905385	0.904412	0.923077	0.904458	0.97862
MLP	0.926471	0.926935	0.926471	0.938462	0.926502	0.986673
Naive Bayes	0.647059	0.796981	0.647059	1	0.604534	0.686243
Random Forest	0.955882	0.957575	0.955882	0.984615	0.955901	0.971248
XGBoost	0.955882	0.957575	0.955882	0.984615	0.955901	0.971248

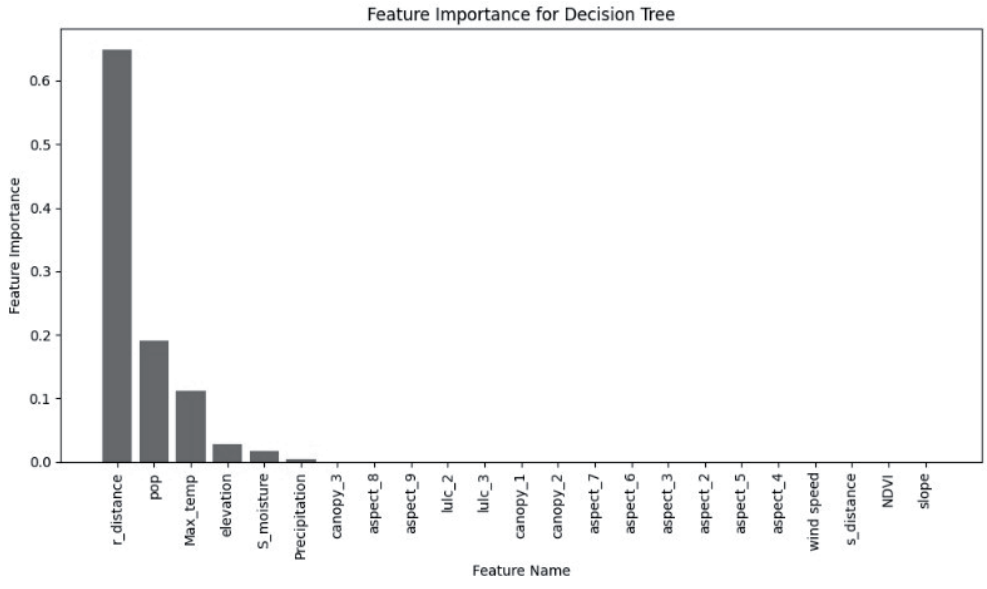

Fig. 6. Feature importance for the Decision Tree algorithm.

Results

After the data for the dependent and independent variables were collected, pre-processed, matched, and analyzed, the Decision Trees, Multi-Layer Perceptron (MLP), Naive Bayes, Random Forest, and XGBoost algorithms, which are used in the literature, were selected as models. In all algorithms, the input dataset was split into 80% training and 20% test data, and the accuracy metrics are shown in Table 3.

When analyzing the accuracy metrics of algorithms, we see that MLP, XGBoost, and Random Forest algorithms produce strong models, but there is a potential risk of overfitting. Therefore, the feature importance metrics of the models should be checked. Feature Importance for Tree-Based Algorithms and Permutation Importance for MLP and Naive Bayes (MLP & NB) are visualized.

In the Decision Tree algorithm, distance to the road, population, and maximum temperature were the three most influential independent variables, while in the MLP and Naive Bayes algorithms, distance to the road, land use, and population density were the most influential variables (Figs 6-8).

In the Random Forest algorithm, distance to road, population, and maximum temperature were the three most influential independent variables, while distance to road, land use, and aspect-3 (northwest) were the most influential variables in the XGBoost algorithm (Fig. 9 and Fig. 10).

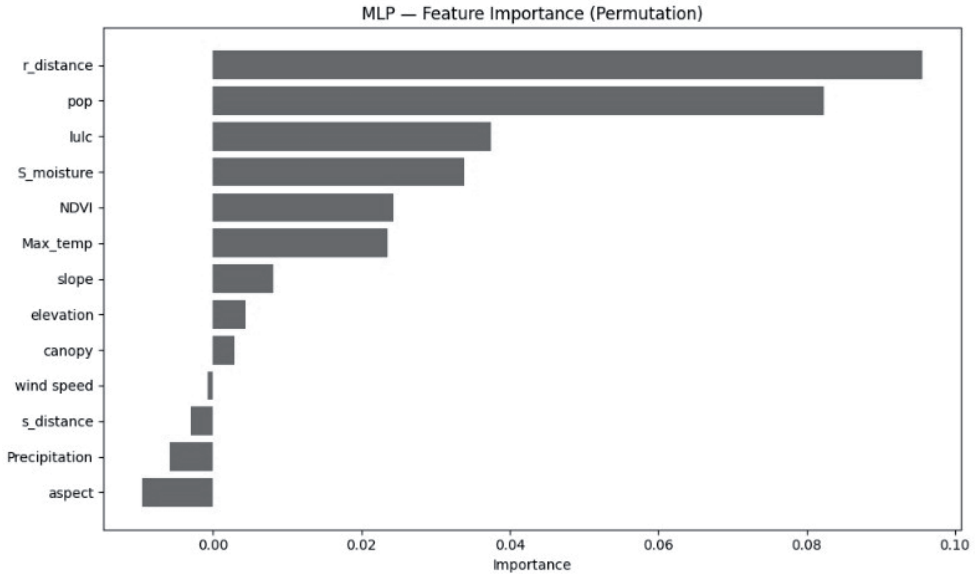

Fig. 7. Feature importance for the MLP algorithm.

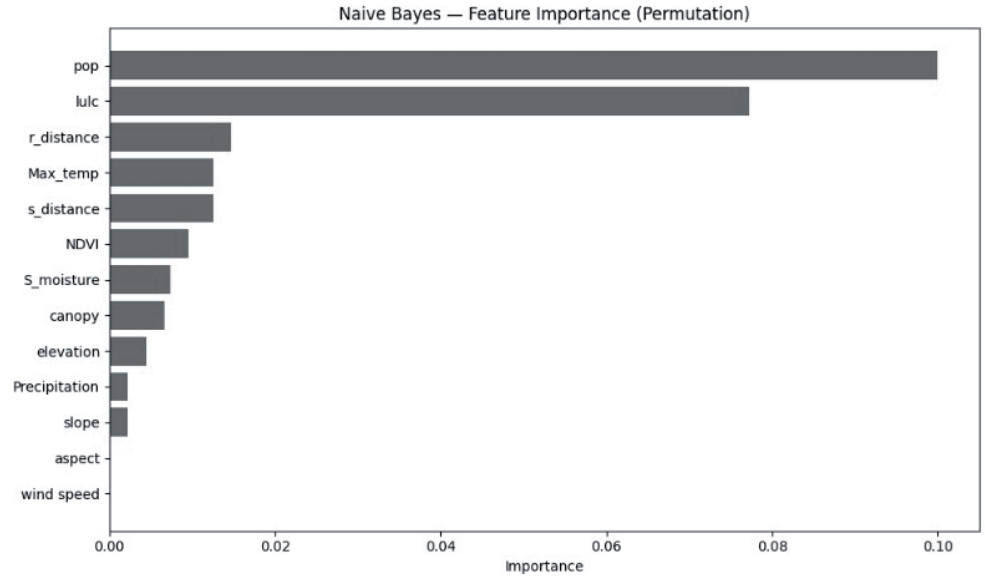

Fig. 8. Feature importance for the Naive Bayes algorithm.

After analyzing the dependent and independent variables, Decision Trees, Multi-Layer Perceptron (MLP), Naive Bayes, Random Forest, and XGBoost algorithms used in the literature were compared as models. Among these models, the Random Forest and XGBoost models, which achieved the highest accuracy metrics, were selected for prediction. The performance and reliability of the Random Forest and XGBoost models were evaluated using a set of complementary diagnostic tools. ROC-AUC analysis was applied to evaluate overall discriminative ability, confusion matrices were used to examine classification errors, and SHAP analysis was used to interpret model predictions and identify dominant forest fire factors.

The confusion matrices demonstrate strong classification performance for both Random Forest and XGBoost models. For the Random Forest model, 196 out of 204 test samples were correctly classified, with only 1 false positive and 7 false negatives. Similarly, the XGBoost model correctly classified 197 samples, also producing only 1 false positive while reducing the number of false negatives to 6 (Fig. 11). Both models therefore exhibit very low false alarm rates, which is particularly important for wildfire susceptibility mapping, where excessive false positives may lead to inefficient resource allocation.

The Receiver Operating Characteristic (ROC) curves further confirm the robustness of the models. Random

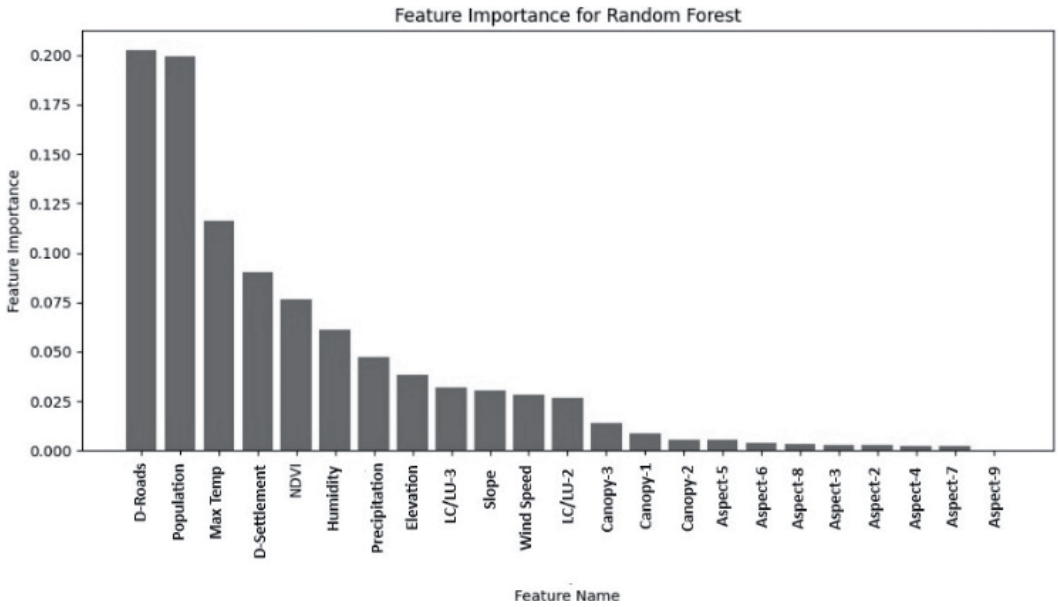

Fig. 9. Feature importance for the Random Forest algorithm.

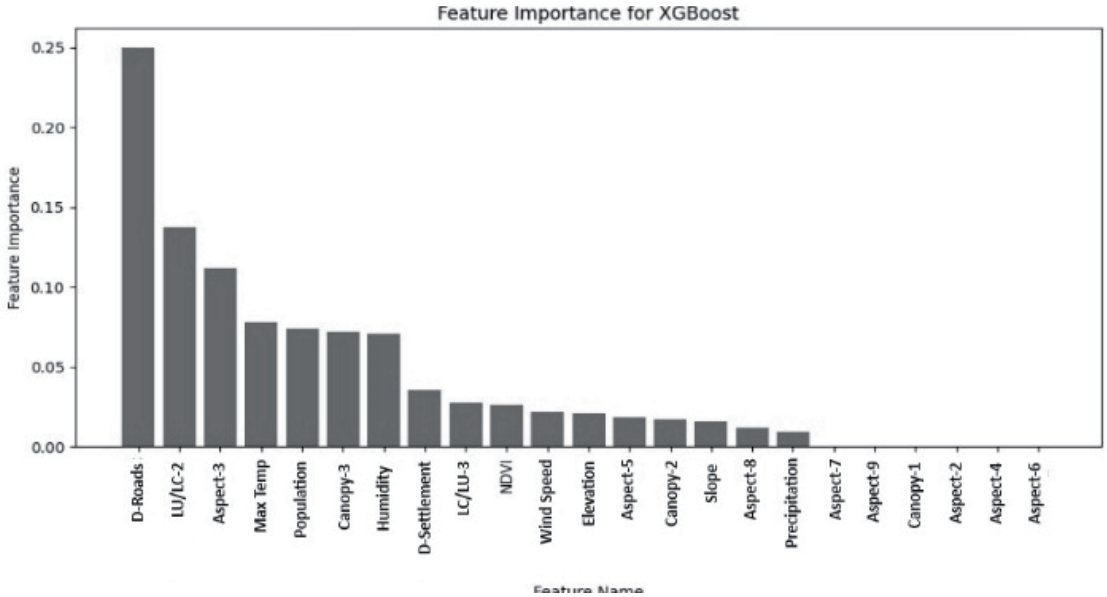

Fig. 10. Feature importance for the XGBoost algorithm.

Forest and XGBoost achieved AUC values of 0.997 and 0.996, respectively, indicating near-perfect discrimination between fire and non-fire classes. The ROC curves for both models are concentrated near the top-left corner, demonstrating high sensitivity at very low false positive rates. The marginal difference in AUC values suggests that both tree-based models possess comparable and highly reliable discriminative capabilities (Fig. 12).

SHAP analysis was applied to interpret the predictions of both models and to identify the most influential variables. For Random Forest, global SHAP bar plots based on mean absolute SHAP values indicate that r_distance and pop are the dominant predictors, followed by Max_temp, s_distance, and S_moisture. For XGBoost, SHAP summary plots reveal clear nonlinear relationships, showing that proximity to human activity (low r_distance and high pop) substantially increases wildfire probability, whereas higher soil moisture and precipitation tend to reduce fire risk. The consistency in feature importance rankings across both models highlights the robustness of the identified wildfire drivers.

Fig. 13 represents the mean absolute SHAP values, indicating the average magnitude of each variable's contribution to the model output. Proximity-related variables (r_distance and pop) exhibit the strongest influence on wildfire occurrence, highlighting the dominant role of anthropogenic factors, while climatic and vegetation-related variables contribute at secondary levels.

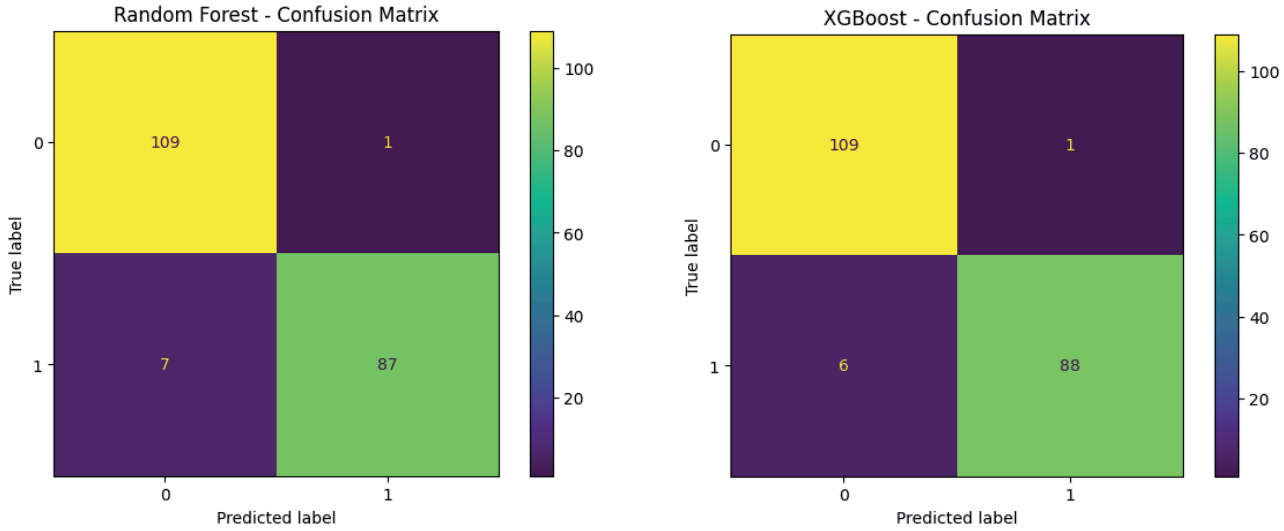

Fig. 11. Confusion matrices for the Random Forest and XGBoost models on the test dataset. Both models correctly classify the majority of fire and non-fire samples, with only one false positive in each case. XGBoost demonstrates slightly higher fire detection sensitivity by reducing the number of false negatives, while maintaining a similarly low false alarm rate.

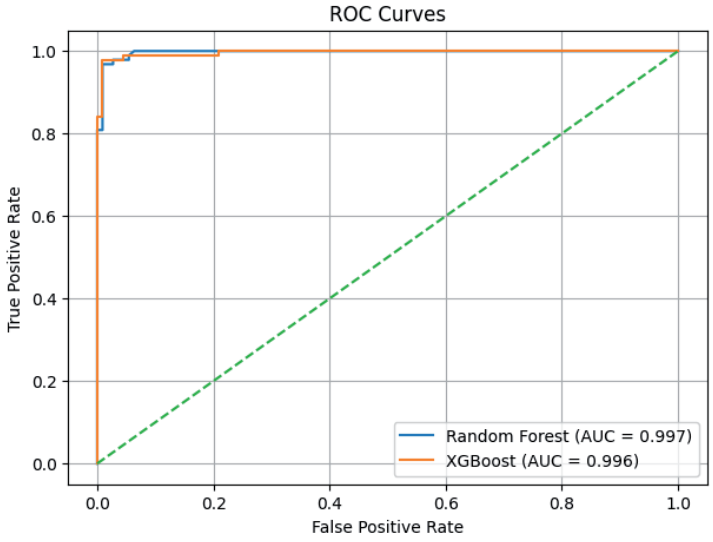

Fig. 12. Receiver operating characteristic (ROC) curves for the Random Forest and XGBoost models.

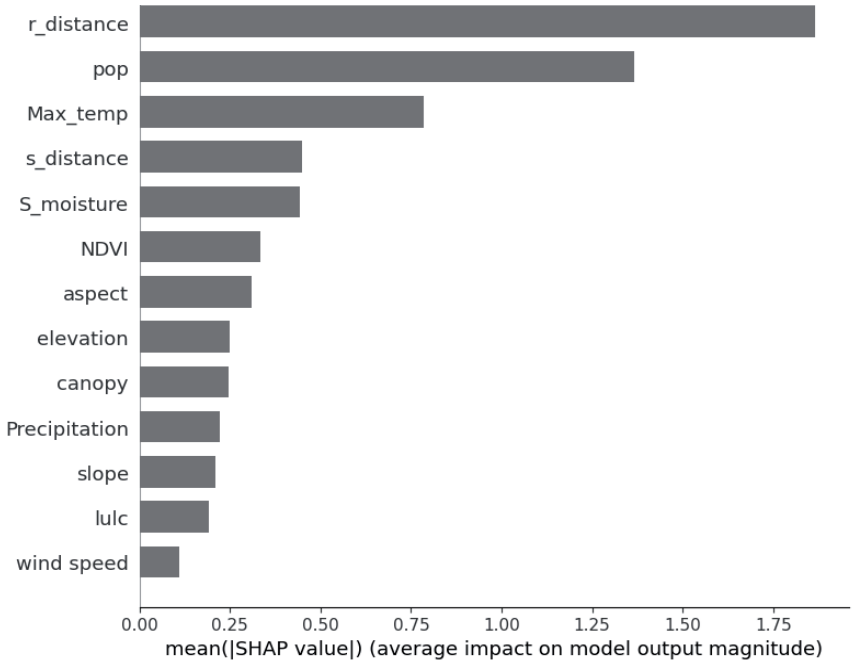

Fig. 13. Global feature importance derived from SHAP analysis for the Random Forest model.

Each point represents an individual observation, with color indicating the feature value (low to high). Positive SHAP values increase the predicted fire probability, whereas negative values decrease it. The results reveal clear nonlinear relationships, with r_distance and pop showing strong and consistent effects on wildfire susceptibility (Fig. 14).

After model selection and training, the forecasting phase was initiated. In the forecasting phase, the entire study area was mapped in the ArcMap software by making forecasts for the forecast dataset consisting of 113116 × 13 rows and columns for the entire study area. The study area, which has Mediterranean climate characteristics, generally carries high fire risks. In this study, the probability of fire was divided into three categories and classified as low, medium, and high, and mapped (Fig. 15 and Fig. 16).

The results obtained from experiments conducted with 13 independent variables in this study are as follows: The highest accuracy was achieved with

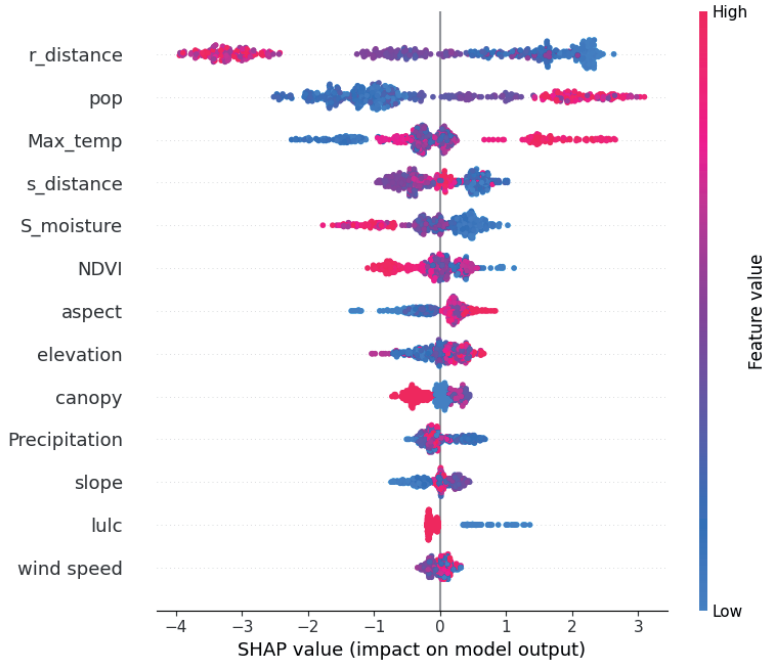

Fig. 14. SHAP summary plot for the XGBoost model illustrating the distribution of feature contributions across all samples.

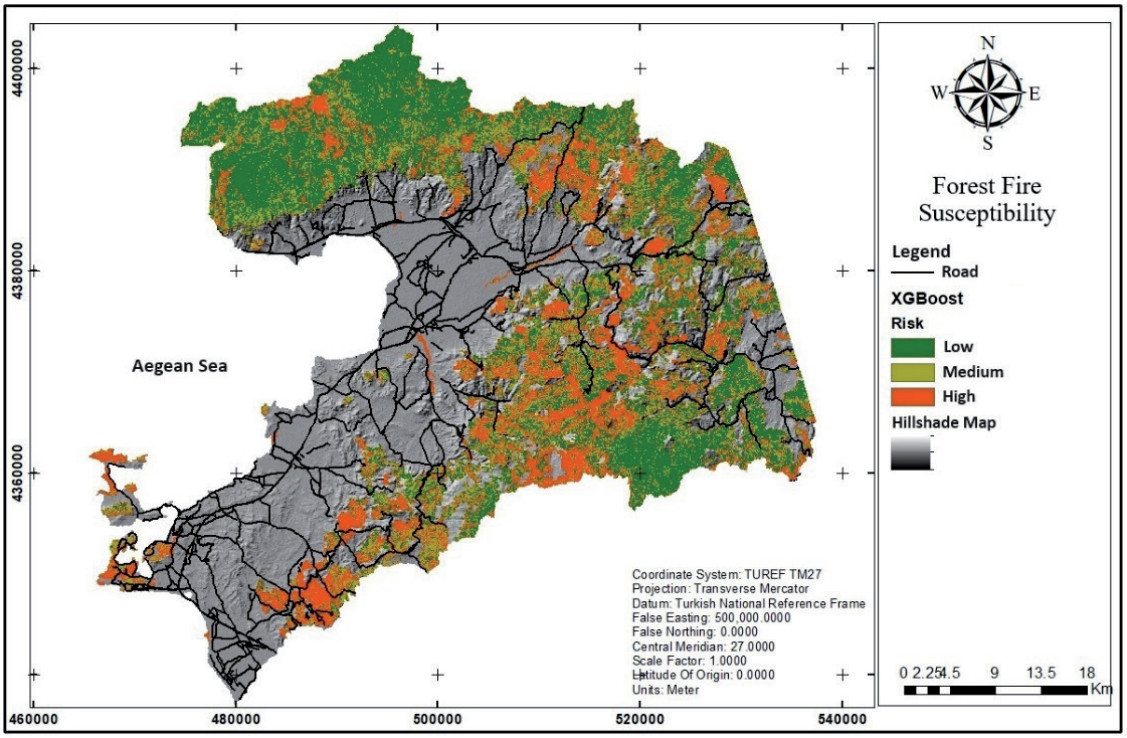

Fig. 15. Forest fire susceptibility map predicted with the Random Forest algorithm.

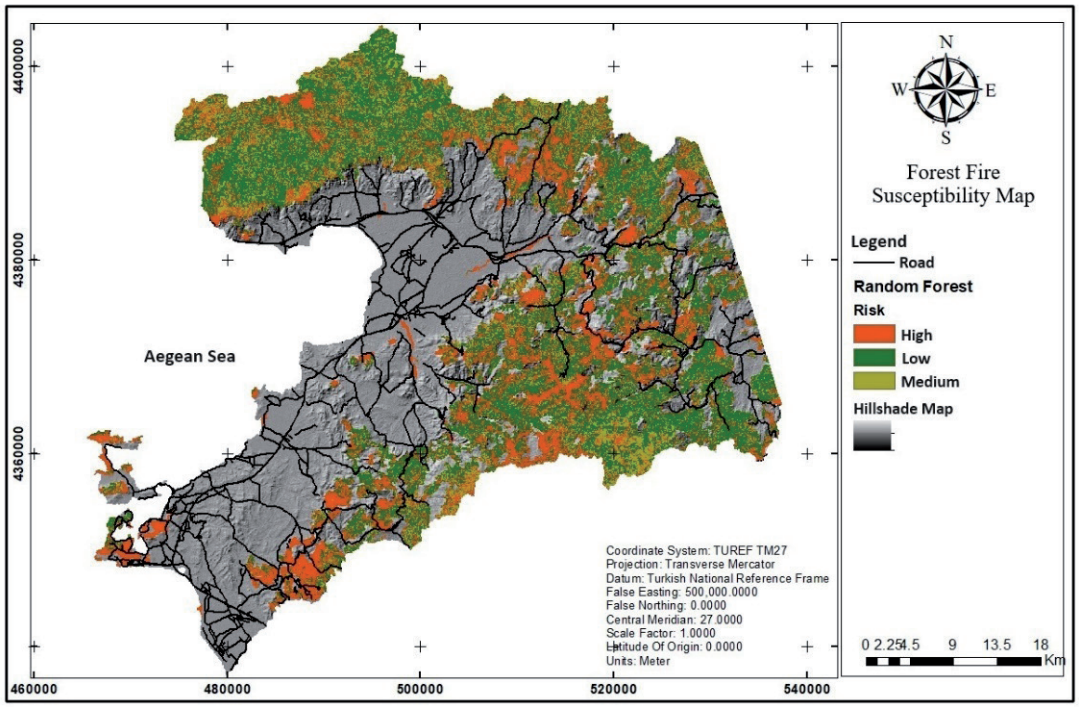

Fig. 16. Forest fire susceptibility map predicted with the XGBoost algorithm.

the Random Forest and Extreme Gradient Boosting (XGBoost) algorithms at 0.96, followed by the Multi-Layer Perceptron (MLP) at 0.93, Decision Trees at 0.90, and Naive Bayes test accuracy at 0.65. These results show parallelism with the literature for XGBoost, while other algorithms yielded results that differed from the literature. The most important reasons for this difference are the size of the datasets, class imbalances, the number of features, and the effect of regional differences.

Preparing region-specific datasets for forest fire risk prediction is expected to lead to more successful training and predictions using these datasets. In this study, the

Random Forest and XGBoost algorithms are proposed within the framework of the algorithms used, and these algorithms are consistent with the literature [25, 26].

In addition to the machine learning algorithms used in this study, various risk predictions have also been produced using deep learning algorithms in the literature. These algorithms use more data than machine learning algorithms. It appears that techniques based on deep learning networks can achieve more successful results with ever-growing datasets [27-29].

Discussion

The results of this study demonstrate that wildfire occurrence is strongly influenced by anthropogenic factors, particularly proximity to roads or settlements and population density. The dominance of r_distance and pop across MI, SHAP, and model performance analyses is consistent with previous wildfire susceptibility studies, which have emphasized the critical role of human activity in fire ignition processes [30].

Climatic variables such as maximum temperature and soil moisture also exhibit substantial contributions, supporting the well-established link between fuel dryness, atmospheric conditions, and wildfire ignition potential [31]. Vegetation-related variables, including NDVI and canopy cover, contribute at secondary levels, reflecting their indirect influence through fuel availability rather than direct ignition.

The near-perfect ROC-AUC values obtained for both Random Forest and XGBoost indicate strong generalization capability and confirm that the selected predictors effectively separate fire and non-fire conditions. Importantly, the low false positive rates observed in the confusion matrices suggest that the models avoid excessive overprediction, a common limitation in wildfire susceptibility mapping.

Although Random Forest achieved a slightly higher AUC, XGBoost demonstrated marginally better sensitivity by reducing the number of false negatives. This trade-off suggests that XGBoost may be preferable in early-warning or prevention-oriented applications, where missing a fire event may be more critical than generating a small number of additional alerts.

Overall, the combined use of Mutual Information for feature relevance, SHAP for model interpretability, and confusion matrix and ROC-AUC metrics for performance evaluation provides a comprehensive and reliable framework for nonlinear wildfire susceptibility modeling. The agreement between model outputs and established wildfire ecology literature further supports the validity of the proposed approach.

Conclusions

This study presents a robust and interpretable wildfire susceptibility modeling framework by integrating remote sensing data, environmental variables, and anthropogenic indicators with nonlinear machine learning algorithms.

Overall, the combined use of Mutual Information for feature relevance, SHAP for explainable artificial intelligence, and robust performance diagnostics provides a transparent and reliable framework for wildfire susceptibility mapping. The findings are consistent with established fire behavior literature and offer valuable insights for wildfire risk management, land-use planning, and prevention strategies in fire-prone regions.

Future studies may further extend this framework by incorporating spatiotemporal cross-validation strategies, dynamic climate indicators, and higher-resolution socio-environmental datasets to enhance model generalizability. In addition, integrating finer-grained anthropogenic variables such as road networks classified into provincial, district, rural, and forest road categories, as well as crop-type-specific agricultural data, could enable models to better capture real-world human-environment interactions influencing wildfire occurrence.

Acknowledgments

This research received no specific grant from any donor agency in the public, commercial, or nonprofit sectors, and these organizations have had no involvement in the analysis and interpretation of data.

Conflict of Interest

The authors declare no conflict of interest.

References

1. The Intergovernmental panel on climate change (IPCC). Available online: https://www.ipcc.ch/report/ar6/syr/summary-for-policymakers/ (accessed on 2nd January 2025).
2. KÜÇÜK Ö., SAĞLAM B. Forest fires and weather conditions. Kastamonu University Journal of Forestry Facult. **4** (2), 220, **2004** [In Turkish].
3. SATIR O., BERBEROĞLU S. A methodological overview of risk mapping approaches used in prevention of forest fires from past to present. In book: Forest fires: causes, effects, monitoring, precautions to be taken and rehabilitation activities. Editor: KAVZAOĞLU T. Turkish Academy of Sciences, Ankara, Türkiye, **33**, 137, **2021**.
4. BİLGİLİ E., KÜÇÜK Ö., SAĞLAM B., COŞKUNER A. Mega forest fires: causes, organization and management. In book: Forest fires: causes, effects, monitoring, precautions to be taken and rehabilitation activities. Editor: KAVZAOĞLU T. Turkish Academy of Sciences, Ankara, Türkiye, **33**, 1, **2021** [In Turkish].
5. General Directorate of Forestry (GDF), Available online: https://www.ogm.gov.tr/tr/ekutuphane/resmi-istatistikler (accessed on 2nd February 2025).

6. ÇOLAK E., SUNAR F. Evaluation of forest fire risk in the Mediterranean Turkish forests: A case study of Menderes region, Izmir. International Journal of Disaster Risk Reduction. **45**, 101479, **2020**.
7. SABUNCU A., ÖZENER H. Detection of burned areas using remote sensing techniques: İzmir Seferihisar forest fire example. Journal of Natural Hazards and Environment. **5** (2), 317, **2019**.
8. KAVLAK M.Ö., KURTIPEK A., ÇABUK S.N. Creating forest fire risk map with Geographic Information Systems: Ören example. Resilience. **4** (1), 33, **2020** [In Turkish].
9. ALKAYIŞ M.H., KARSLIOĞLU A., ONUR M.İ. Determination of forest fire risk potential map of Menteşe region of Muğla province using geographic information systems. Geomatik. **7** (1), 16, **2022** [In Turkish].
10. THACH N.N., NGO D.B.T., XUAN-CANH P., HONG-THI N., THI B.H., NHAT-DUC H., DIEU T.B. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. Ecological Informatics. **46**, 74, **2018**.
11. HE Q., JIANG Z., WANG M., LIU K. Landslide and wildfire susceptibility assessment in southeast asia using ensemble machine learning methods. Remote Sensing. **13** (8), 1572, **2021**.
12. SAYAD Y.O., MOUSANNIF H., AL MOATASSIME H. Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Safety Journal. **104**, 146, **2019**.
13. BEŞLİ N., TENEKECİ E. Forest fire prediction using decision trees from satellite data. Dicle University Faculty of Engineering, Engineering Journal. **11** (3), 906, **2020** [In Turkish].
14. SEVINÇ V., KÜÇÜK O., GÖLTAŞ M. A Bayesian network model for prediction and analysis of possible forest fire causes. Forest Ecology and Managemen. **457**, 117723, **2020**.
15. NADERPOUR M., RIZEEI H.M., RAMEZANI F. Forest fire risk prediction: a spatial deep neural network-based framework. Remote Sensing. **13** (13), 2513, **2021**.
16. MPAKAIRI K.S., TAGWIREYI P., NDAIMANI H., MADIRI H.T. Distribution of wildland fires and possible hotspots for the Zimbabwean component of Kavango-Zambezi Transfrontier Conservation Area. South African Geographical Journal/Suid-Afrikaanse Geografiese Tydskrif. **101** (1), 110, **2019**.
17. AVCI Z.D.U., KUŞAK B., KUŞAK L. Evaluation of different texture criteria in determining stand types using satellite data. Proceedings of the XVI Academic Informatics Conference. Mersin University, 121, **2014** [In Turkish].
18. KUHN M., JOHNSON K. Feature engineering and selection: A practical approach for predictive models. Chapman and Hall/CRC, Florida, USA, **74** (3), 308, **2019**.
19. BREIMAN L., FRIEDMAN J., OLSHEN R.A., STONE, C.J. Classification and regression trees. Chapman and Hall/CRC, 1st ed., New York, USA, pp. 368, **2017**.
20. DEPERLIOĞLU Ö., KÖSE U. Machine Learning Basic Concepts with Python: Classification – Regression – Clustering, 1st ed.; Seçkin Publishing, Ankara, Türkiye, pp. 315-337, **2024** [In Turkish].
21. GOODFELLOW I., BENGIO A., COURVILLE A. Deep Learning. The MIT Press, **2016**.
22. MITCHELL T.M. Does machine learning really work? AI Magazine. **18** (3), 11, **1997**.
23. BREIMAN L. Random forests. Machine Learning, **45** (1), 5, **2001**.
24. CHEN T., GUESTRIN C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. **2016**.
25. YOUSSEF A.M., POURGHASEMI H.R. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geoscience Frontiers. **12**, 639, **2021**.
26. MERGHADI A., YUNUS A.P., DOU J., WHITELEY J., THAIPHAM B., BUI D.T., AVTAR R., ABDERRAHMANE B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews. **207**, 103225, **2020**.
27. TRUONG T.X., NHU V.H., PHUONG D.T.N., NGHI L.T., HUNG N.N., HOA P.V., BUI D.T. A new approach based on Tensorflow deep neural networks with adam optimizer and GIS for spatial prediction of forest fire danger in tropical areas. Remote Sensing. **15**, 3458, **2023**.
28. SUN J., LIU Y., CUI J., HE H. Deep learning-based methods for natural hazard named entity recognition. Scientific Reports. **12**, 4598, **2022**.
29. LIU Z., ZHANG K., WANG C., HUANG S. Research on the identification method for the forest fire based on deep learning. Optik. **223**, 165491, **2020**.
30. CHUVIECO E., PETTINARI M.L., KOUTSIAS N., FORKEL M., HANTSON S., TURCO M. Human and climate drivers of global biomass burning variability, Science of The Total Environment. **779**, 146361, **2021**.
31. ABATZOGLOU J.T., WILLIAMS A.P. Impact of anthropogenic climate change on wildfire across western US forests. The Proceedings of the National Academy of Sciences. **113** (42), **2016**.