International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at :www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

A Machine Learning Framework for AI-Based Wildfire Risk Assessment

Paril Ghori

Email: parilghori@gmail.com

ARTICLEINFO

Article History:

Accepted: 01 Nov 2023

Published: 30 Nov 2023

Publication Issue

Volume 9, Issue 6

November-December-2023

Page Number

422-434

ABSTRACT

Forest fires pose a severe risk to both the environment and human life, often causing extensive damage to ecosystems and property. With the growing impact of climate change, including rising temperatures and prolonged droughts, the frequency and intensity of forest fires have increased significantly. Effective early detection and intervention are crucial to minimizing these impacts. In recent years, machine learning (ML) techniques have been applied to forest fire prediction to improve early warning systems. This study aims to reduce the risk and impact of forest fires by predicting their occurrence ahead of time using machine learning. A dataset, sourced from NASA's Oak Ridge National Laboratory (ORNL), containing detailed environmental and forest-related factors, was used for this purpose. Preprocessing steps were applied to prepare the data for classification, including feature selection techniques that narrowed down the dataset from 35 to the most relevant features. Various machine learning algorithms—Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbor (K-NN), and Naive Bayes (NB)—were employed for classification. The performance of the models was evaluated, and hyperparameter optimization was performed to select the best parameters. The Random Forest model achieved an impressive accuracy of 97%, closely followed by Naive Bayes at 96%, highlighting the effectiveness of these algorithms in predicting forest fires.

Keywords – AI, Decision Tree, Gradient Boosting, K-Nearest Neighbor, Machine Learning, Naive Bayes, ORNL, Random Forest, Support Vector Machine.

I. INTRODUCTION

Forests are not just groups of trees and shrubs. Forests are one of the most valuable ecological values of living beings in every aspect of life. Forests are the most important source of basic living conditions such as clean air, drinkable water resources, shelter and food that are necessary for our lives. Forest fires that occur due to natural and human reasons are at the top of the list of natural disasters and cause great environmental damage by destroying the ecosystem where many living beings live together. In addition to other damages, these disasters most importantly threaten human life. With higher than normal temperatures, low relative humidity and strong

winds, forest fires sometimes reach disaster dimensions that the entire country or even almost the entire world struggles with. Especially in recent years, with global warming due to climate change, increasing population, industrialization and agricultural developments are causing forest areas to disappear rapidly. The Food and Agriculture Organization of the United Nations (FAO) announced in its "2020 Global Forest Resources Assessment Report (FRA) that there are 4.06 billion hectares of forest in the world and that this forest area corresponds to one-third of the world's land area. The report stated that the world's forests have decreased due to various reasons such as forest fires, environmental disasters, forest pests, and that a total of 178 million hectares of forest area were destroyed between 1990 and 2020 [1]. When the official statistical data of the General Directorate of Forestry, affiliated to the Ministry of Agriculture and Forestry, are examined, it is seen that 29% of the forest fires occur due to negligence and carelessness, 4% due to intent, 6.4% due to accidents, 13% due to natural causes, and 47.6% due to unknown causes [2]. Knowing the factors that threaten forest ecosystems and the causes of forest fires will enable more effective prevention activities. The main factors affecting forest fires include weather conditions (temperature, relative humidity, precipitation and wind speed), time factor (season, month, certain hours of the day), topographic structure (aspect, altitude, terrain slope, landform, etc.) and human-induced factors [3].

II. LITERATURE REVIEW

Early warning, rapid, and effective intervention are very important in firefighting to prevent forest fires and minimize their effects. When looking at the literature, many studies have been conducted to predict forest fires. The authors of [4] used the Bayesian network model to estimate and analyze possible forest fire causes. With this model, rates of fire outbreaks such as hunting, stubble burning, picnics, and shepherd fires were found and compared. In another study [5], spatial fire distribution was estimated with machine learning algorithms Maximum Entropy (MaxEnt) and Random Forest algorithms using environmental datasets and historical fire data. As a result of the study, it was seen that the two factors affecting spatial fire distribution were population density and climate. The authors of [6] developed a forest fire prediction model using the deep learning method TensorFlow and Keras library as a neural network using data obtained from forest fires that occurred between 2013-2019 in various regions of Turkey and created a regional fire risk map. The authors of [7] applied backpropagation neural network (BPNN), recurrent neural network (RNN), and long short-term memory (LSTM) methods to estimate the scale of fires (5 levels according to fire severity) in Alberta forests in Canada by taking meteorological factors as input. As a result of the study, it was shown that it is possible to estimate the scale of a forest fire and the fire in its initial stage using meteorological information. It was stated that the scale of a fire is determined by the combination of the duration of the fire and the size of the area burned. The authors of [8] estimated the burned area size with machine learning algorithms using various features such as temperature, precipitation, wind, and relative humidity from 512 forest fire data obtained as a result of a forest fire in a national park in Portugal.

In another study, a forest fire early warning system was created using fuzzy logic and machine learning techniques, taking into account weather and geographic data in the state of Acre, Brazil. In the study, the forest fire warning map was combined with two indices as forest fire risk and forest fire danger, and it was stated that the fire risk index measures the probability of a fire breaking out at a certain point, and the forest fire risk was obtained from past data with fuzzy KNN [9]. Touching on the issues that predicting forest fires will contribute to fire prevention and elimination risk, the authors of [10] created a forest risk map for China. The results of various metrics obtained from models established with machine learning algorithms were compared, showing that there were significant seasonal and regional differences in forest fire risks in China. Stating that forest fire prediction is the most important component in controlling forest fires, the authors of [11] developed a prediction model by using meteorological parameters such as temperature, humidity, wind, and rain to predict forest fire formation, and by performing hyperparameter adjustment with random forest regression from machine learning algorithms. The authors of [12] in their study determined the forest fire area in order to take rapid precautions before forest fires spread to a wide area and created the best regression model by comparing

machine learning techniques. In a study conducted by the authors of [13], data were clustered with k-means clustering and a label was made with five class labels as very low risk, low risk, medium risk, high risk, and very high risk. Using the labeled data obtained as a result of clustering, it was subjected to classification, and they evaluated that the best classifier was the random forest algorithm. In a similar study, the authors of [14] created the fire tendency levels of a certain area according to the amount of burned area with machine learning classification algorithms. The authors of [15] developed a prediction model for the burned area of forest fires and the occurrence of large-scale forest fires using ensemble learning approaches and stated that the adjusted random forest approach performed better than other regression models in predicting the burned area. The authors of [16] developed an intelligent system based on genetic programming to predict the area that will burn during a possible forest fire using the relationships between meteorological and forest-related data and the amount of burned area. In another study conducted to predict the burned area, the fire was classified as large or small with machine learning algorithms, and it was aimed to help the fire management team allocate appropriate resources during the fire [17]. One of the causes of forest fires is fires caused by lightning. For this reason, the authors of [18] analyzed the relationship between lightning and ignition to predict forest fires and created a forest fire prediction model based on lightning estimates and environmental conditions with a machine learning approach. The authors of [19] tried to predict the main causes of forest fires in China. In the study, the causes of fire outbreaks were predicted using machine learning algorithms such as artificial neural networks, radial basis function networks, support vector machines, and random forest algorithms, and spatial distribution maps of forest fire-prone areas from high to low were generated. In another study, the authors of [20], who stated that precipitation is an important factor affecting the probability of future forest fires, estimated the forest fire risk in Central and Northern China with a time-decreasing precipitation model in a study they conducted. In the study, a time-decreasing precipitation algorithm was used to calculate the comprehensive precipitation index. This method showed that the effect of precipitation was better represented in predicting the occurrence of forest fires. The authors of [21] proposed a new data balancing procedure for unbalanced data distribution in cases where large fires are fewer than small fires and forest fire prediction using deep neural networks. They stated that large-scale forest fires can be predicted more accurately with the proposed method, which will better benefit the advance management of forest fires and the prevention of serious fire accidents.

Forest fire prediction is an important component of forest fire management. Predicting forest fires in advance and intervening early in the fire, the earlier the intervention, the more the forest is saved. Being able to predict possible forest fires in advance can provide important benefits such as taking precautions before the fire and enabling the fire management team to intervene in the fire to make sound decisions. Thus, using the vehicle and personnel resources in the right place and on time, faster and more effective intervention in the fire is provided. The aim of this study is to control the fires before they break out or at the beginning stage and minimize their effects by predicting possible forest fires in advance in order to combat forest fires. The method proposed in this study, unlike other methods, is to process a data set with a pre-determined class (fire/no fire) with machine learning methods, which is a sub-branch of artificial intelligence, and to predict forest fires with the binary classification method. In other studies in the literature, regression models such as estimating the size of the burned area and creating fire risk maps have been created or the data has been analyzed by creating a new target variable using a certain threshold value on the data. Another difference of the proposed method from other studies is the use of dimensionality reduction techniques. In this study, the dataset used for the forest fire prediction model was taken from the official website of ORNL, which belongs to NASA. The data was subjected to various processes such as deletion of unnecessary data, completion of missing data, and normalization by passing through preprocessing steps. Thus, since unnecessary input features were not included in the training while training the model, the processing time and processing load were significantly reduced. With the completion of the preprocessing steps, the data was processed with machine learning methods and the forest fire prediction model was created. Since the class label in the dataset consists of binary values (fire-1/fire-0), binary classification was performed. The performance was examined by applying improvement strategies to the model.

The classification process was performed by determining the most important and most useful features in finding the target variable by reducing the dimensionality with feature selection techniques. As a result of feature selection, the model was trained with the new dataset and the accuracy rate was close to the rate obtained with the entire dataset. In this study, model performances were compared using different classification algorithms of machine learning algorithms. Confusion matrix was applied to evaluate binary classification models, validation process was applied to eliminate overfitting problem and hyperparameter optimization was applied to select the best parameter for the model.

III. PROPOSED METHODOLOGY

This section includes information about the forest fire dataset, data preprocessing steps, and classification algorithms used in the classification process.

3.1 Dataset

The dataset used in this study was taken from the official DAAC website of NASA [22]. This dataset consists of data collected from burned and unburned areas in Alaska and Canada between 1983 and 2016. This dataset, consisting of 1171 sample data, is divided into two classes, 1019 burned and 152 unburned. The dataset contains 49 features along with system components including fire weather indices (FWI) such as relative humidity, temperature, wind speed, precipitation, drought codes for each location, topographic structure such as aspect, elevation, slope, and forest information such as tree age and tree density. In the dataset, unnecessary features (project id, name, etc.) and features with too many empty values were deleted for model training and reduced to 35.2 of these features are categorical and 33 of them contain numerical data. In the dataset used in the study, the data labeled with the class variable zero represents the class "no fire", and the data labeled with one represents the class "fire". Table below shows 15 rows, 11 features and their values as an example from the dataset. As can be seen, there is both numerical and categorical data in the dataset. Since machine learning works with numerical data, the data belonging to categorical features must be converted to numerical data.

Table 1: Randomly taken samples from the dataset [22]

S. No.	Humidity (%)	Temperature (°C)	Wind Speed (m/s)	Precipitation (mm)	Elevation (m)	Slope (%)	Aspect (degrees)	Tree Age (years)	Tree Density (trees/ha)	Fire Weather Index (FWI)	Fire Status (0/1)
1	75	20	2	5	300	10	45	100	250	8.5	0
2	60	25	1.5	3	400	15	90	120	270	7.3	1
3	85	18	3	10	250	5	180	80	300	10.2	0
4	50	30	2.5	0	350	8	60	110	230	6.7	1
5	65	22	1.2	4	500	12	135	130	220	5.9	0
6	70	24	3.1	7	600	20	160	90	280	9.1	1
7	78	21	2.2	6	380	14	75	105	240	7.8	0
8	60	26	1.8	2	320	10	30	115	210	6.5	1
9	68	19	2.7	8	450	17	120	100	260	7.0	0
10	55	28	1.9	1	330	11	150	125	220	6.8	1
11	80	23	2.3	9	270	9	180	95	290	8.3	0
12	72	27	3	6	200	13	70	140	240	8.7	1
13	90	17	1	12	400	6	110	80	300	10.5	0
14	77	20	1.6	5	500	18	130	115	230	7.2	1
15	62	24	2.4	4	600	16	90	105	270	6.9	0

The distribution of the number of burned and unburned forest fires by region in the dataset used in the study is shown in Table 1, and it was seen that the region with the most fires was Taiga Plains.

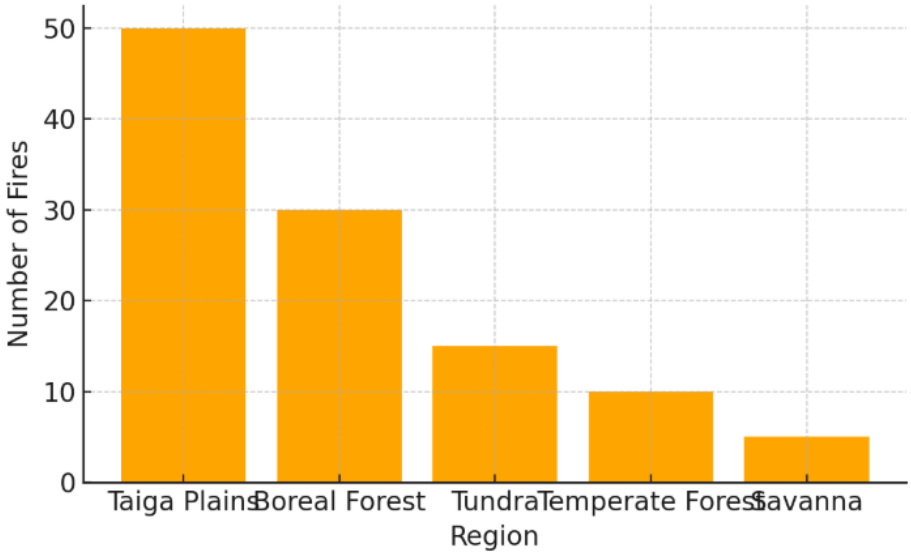

Figure 1: Distribution of fire status by region

In the study, the presence of forest fire was predicted using 6 different machine learning techniques on the forest fire dataset. The architectural structure of the model proposed for the classification process is given in Figure 2.

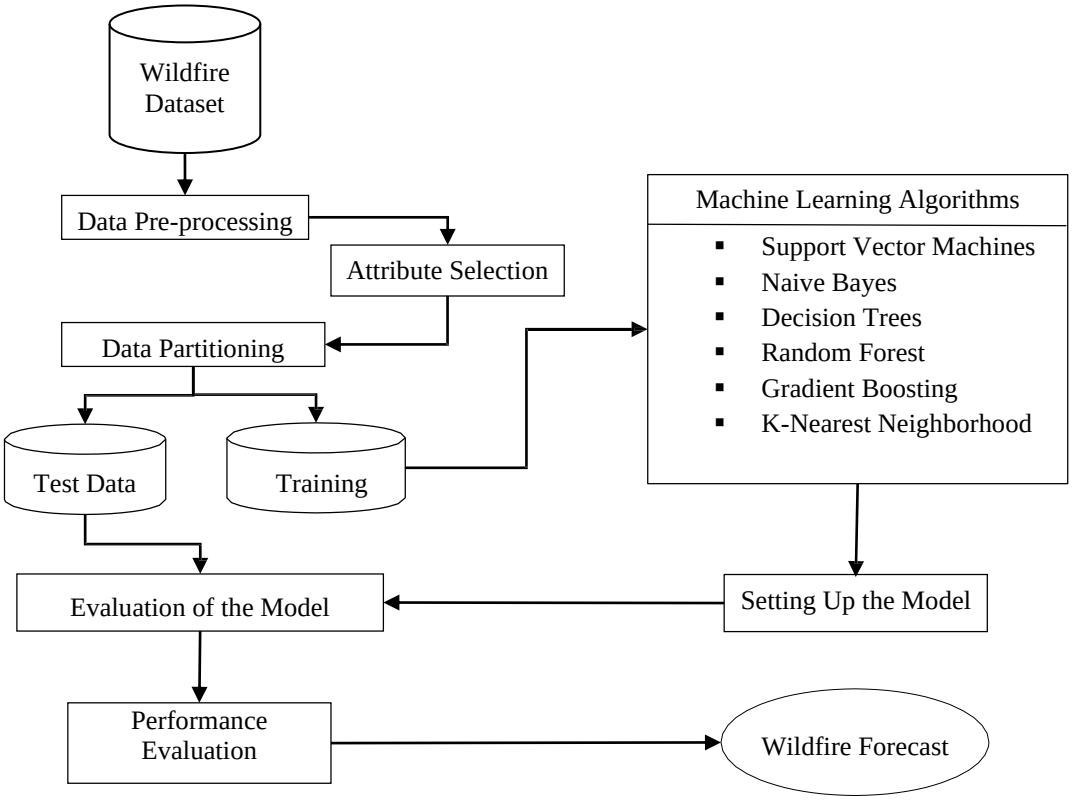

Figure 2: Proposed architecture to compare different models in forest fire prediction

3.2 Data Preprocessing

One of the most important and necessary steps in the process of establishing machine learning models is to prepare the data. Data preprocessing is the process of completing missing data on the data set, cleaning data, transforming, normalizing and reducing dimensions, and making the data suitable for the model. These steps are applied so that the machine learning model can make reliable, accurate and successful predictions. Some attributes (project id, name etc.) that will not contribute to the model were deleted from the forest fire data set used in the study. There are quite a few empty values in some attributes in the data set. Of the attributes in question, 50% of which were empty were removed from the data set. The remaining attributes with missing values were completed by filling them with the mean value. The SimpleImputer method of the sklearn impute library was applied in Python for this process. The boxplot (IQR) method was used to detect outliers. Numerical attributes containing outliers are set equal to the lower and upper limit values. Figure 3 shows the outliers of some attributes.

To establish machine learning models, both input and output variables must be in numerical form. In this study, the categorical dependent variable fire status (fire/no fire) was converted to binary numerical form. Similarly, categorical attributes in the input variables were digitized. The LabelEncoder class from Python's sklearn preprocessing library was used for this process.

The fact that the data is not normally distributed significantly affects the accuracy, speed and performance in distance-based machine learning algorithms such as support vector machines and k-nearest neighbors.

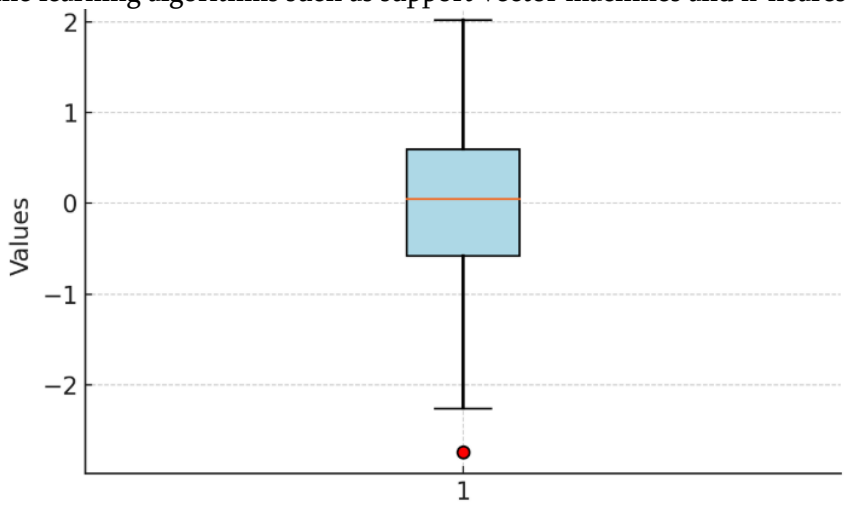

Figure 3: Use of box plots to detect outliers

For this reason, it is necessary to convert the distances between the values to a common scale without changing the distribution of the data set. In this preprocessing step, called the feature scaling process, only numerical features are processed. In this study, statistical and Min-Max normalization processes were performed to eliminate the distances between the data, and it was aimed for all values to contribute equally to the success of the algorithms used.

The first and most important step in creating a model design is data cleaning on the raw data set and selecting the best feature (feature, attribute or variable selection). In cases where the number of input features in the data set is too high, the effect on the output feature decreases. Increasing the complexity of the established model will reduce its interpretability and applicability. The data set to be analyzed may contain various features and a high cardinality. The data reduction technique aims to create a reduced representation of the data set to increase performance during model training. Some of the features may be irrelevant or even unnecessary in model training [23]. Therefore, the purpose of feature selection is to prepare the most important features in the data set for model training. The selected features should be a data subset that can replace the original data set. A successful application of feature selection is to create the model with a minimum subset without decreasing the performance of the model [24]. In addition, feature selection reduces the possibility of overfitting and prevents

unnecessary resource consumption during the training phase of the model. The features to be used for the model to be created in this study were determined by applying the Principal Component Analysis (PCA) and Analysis of Variance (ANOVA) methods from the Feature Selection methods in dimensionality reduction techniques. The data set was reduced from 35 features to 15 with PCA, and the model was established by selecting the best 11 features from 35 features with ANOVA. Classification processes were performed with both all features of the data set and the features determined after dimension reduction.

ANOVA is one of the filtering methods that select according to the relationship between the input variable and the output variable. This method, which is used when one of the input or output variables is categorical, statistically measures the relationship between a categorical variable and a numerical variable. This approach ranks the features by calculating the variance ratio between the groups [25]. ANOVA, which is used to determine whether the feature between two groups is significantly different, is removed from the model if the groups in the input variable do not differ at a statistically significant level according to the output variable.

PCA, on the other hand, is obtained by calculating the eigenvectors and eigenvalues of the covariance matrix of the input vector. It linearly transforms a high-dimensional input vector into a low-dimensional input whose components are unrelated [26]. Thus, the data set is represented with fewer variables while preserving the main features of the multivariate data. When PCA is applied, most of the variation in the data is concentrated in the first few components, only the components with significant differences are preserved and the rest are ignored. As a result, the number of variables is reduced by obtaining new variables consisting of the combination of existing variables with dimension reduction. Thus, the prepared models are provided to work in optimum time and with the best performance. In this study, the dataset with 35 features was reduced to 15 features with PCA and it was seen that 99.9% of the total variance of the data could be preserved.

3.3 Classification Algorithms

A sub-branch of artificial intelligence, machine learning uses programmed algorithms that learn and optimize their operations by analyzing input data to make predictions within an acceptable range. As new data is included in the data set, algorithms tend to make more accurate predictions [27]. Machine learning algorithms use various statistical, probability, and optimization methods to learn from past experiences and detect useful patterns from large, unstructured, and complex data sets [28]. In short, machine learning is the modeling of systems that make inferences and predictions on data with statistical and mathematical operations using computers. Classification is one of the supervised learning problems of machine learning. The classification process is divided into two according to the number of dependent variables. If the class label in the data set consists of binary values (female/male, positive/negative, etc.), the binary classification method is used. If the number of classes consists of three or more values (e.g., determination of plant species), the multi-class classification method is used.

In this study, models were created with six different supervised machine learning algorithms, namely SVM, DT, RF, GB, K-NN, and NB, using the forest fire data set. Since the number of classes in the data set consists of two values (no fire = 0, fire = 1), the binary classification method was used.

1) 3.3.1 K-Nearest Neighbor Algorithm (K-NN)

The K-NN algorithm is a classification method. The unknown class is compared with other data in the training set, and a distance measurement result is calculated. The closest k examples to the relevant example from the training set are determined according to the distance, and the class label of the new example is assigned according to the majority of the class labels of the k nearest neighbors. It is a suitable method for multimodal classes as well as applications where the object may have many labels. It is low-efficiency, and its performance depends on the selection of a good 'k' value [29]. It is negatively affected by noise and is sensitive to irrelevant features. Since all data must be revisited, performance also varies according to dimension [30].

2) *3.3.2 Support Vector Machines (SVM)*

SVM falls within the scope of supervised learning methods in classification problems. It is based on the principle of classifying points on the plane by separating them with a line or hyperplane. It is a learning method used for solving classification and curve fitting problems based on statistical learning theory and the principle of minimizing structural risk. It is robust to high-dimensional data and has good generalization ability. However, the training speed is low, and its performance varies depending on the parameter selection [31].

3) *3.3.3 Decision Trees (DT)*

DT, a tree-based learning algorithm, is a structure used to divide a dataset containing a large number of records into smaller clusters by applying a set of decision rules. The clustering process continues in depth until all elements of the cluster have the same class label. Decisions are divided at the leaves and the data at the nodes. In a classification tree, the decision variable is categorical, and in a regression tree, the decision variable is continuous. Decision Trees are easy to interpret, handle categorical and quantitative values, and have the ability to fill in missing values in the attributes with the most probable value [32].

4) *3.3.4 Random Forest (RF)*

RF is one of the supervised classification algorithms. It is an ensemble method that works by training a set of decision trees and returning the class with the majority vote across all trees in the ensemble [33]. Random Forest is the process of combining many decision trees working independently of each other and selecting the value with the highest score among them. Its main difference from the decision tree algorithm is that the process of finding the root node and dividing the nodes is random. It is a very useful classifier due to its efficiency and accuracy [34]. Most of the time, it can give good results without using hyperparameters. It can provide fast and reliable results even in complex and noisy data sets.

5) *3.3.5 Gradient Boosting (GB)*

The GB model is an ensemble machine learning model introduced by Friedman [35]. The term boosting refers to a specific type of algorithm in which weak predictor trees are combined to produce a stronger prediction [36]. It is a machine learning technique that creates prediction models similar to decision trees. Gradient boosting and random forest methods both use core decision tree algorithms. The main idea of the gradient boosting model is to combine multiple weak learners to improve the accuracy and robustness of the final model. To avoid overfitting problems, the gradient boosting model uses a learning rate to scale the contribution from the new tree [37]. This approach does not include dividing the dataset into multiple sub-datasets as in random forest. Instead, it creates a decision tree using the dataset as is and creates a new decision tree based on its errors.

6) *3.3.6 Naive Bayes (NB)*

The basis of the NB classification algorithm is Bayes' theorem. This technique, based on conditional probability, has a probability table that is updated with training data. This table is based on the feature values that need to be looked at for class probabilities to predict a new observation [32]. In the working method of the algorithm, the probability of each situation for an element is calculated and the classification process is performed according to the one with the highest probability value. This classifier assumes that a certain feature in a class is not directly related to any other feature and is based on the principle that only the features of that class can have mutual dependence among themselves [38].

IV. RESULTS AND ANALYSIS

The classification performances of the binary classification models created in this study were evaluated with the confusion matrix. There are 4 different values in the confusion matrix that contain the combination of the predicted and real values. Of these; True Positive value is the situation where the test result prediction is positive while the actual situation is positive, False Positive value is the situation where the test result prediction

value is positive while the actual situation is negative, True Negative value is the situation where the test result prediction value is negative while the actual situation is negative, and False Negative value is the situation where the test result prediction value is negative while the actual situation is positive. True Positive (DP) and True Negative (DN) show the correct predictions of the model, False Positive (YP) and False Negative (FN) show the wrong predictions of the model. The representation of these values in the confusion matrix and some measurements made using the values in the confusion matrix are given in Figure 4.

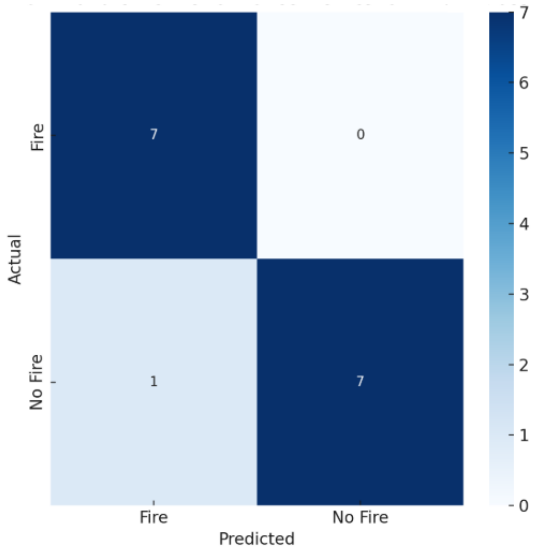

Figure 4: Confusion matrix and other performance metrics for binary classification

Comparison of Different Classification Models And Success Evaluation: After the preprocessing steps, the data set was divided into two, 75% as training data and 25% as test data. Since the dependent variable (class label) consists of binary values (1 is fire, 0 is no fire), binary classification was performed on this data set. Classification was performed with all 35 features without dimensionality reduction. The cross-validation coefficient of 5 was selected from the model validation methods, and the accuracy rates and running times are shown in Table 2. By analyzing the performance of the models in question, it was seen that the fastest working algorithm was DT, and the algorithm with the highest accuracy rate was GB.

Table 2: Performance evaluation for classification models without dimensionality reduction

Classification Algorithms	Accuracy	Cross-Validation Accuracy	Precision	Sensitivity	F1 Score	Execution Time (s)
KNN	98%	94%	99%	98%	99%	0.046
SVM	97%	89.5%	100%	97%	98%	0.25
DT	98%	93.5%	100%	98%	99%	0.015
RF	98%	96%	100%	98%	99%	0.140
GB	99%	93%	100%	100%	100%	0.656
NB	98%	98.5%	100%	98%	99%	0.985

In this study, the new features obtained by applying feature selection methods on the dataset were classified with the help of cross-validation. The most important features in finding the target variable were determined by reducing the dimensionality with PCA and ANOVA techniques. Testing operations were carried out until the smallest dimensional feature subset with the highest accuracy rate selected for the learning algorithm was obtained. As a result of these operations, the best 15 features were selected from 35 features with PCA and 11 features were selected with ANOVA for the best subset selection representing the original dataset. For

classification, the dataset was divided in the same way as 75% training and 25% test data. The algorithms in Table 3 and their performance values were compared.

Table 3: Performance evaluation for classification models after dimensionality reduction

Classification Algorithms	Accuracy	Cross-Validation Accuracy	Precision	Sensitivity	F1 Score	Execution Time (s)
KNN (PCA)	97%	94%	99%	97%	98%	0.031
KNN (ANOVA)	96%	84%	98%	97%	98%	0.062
SVM (PCA)	92.5%	89.5%	92%	100%	96%	0.125
SVM (ANOVA)	94%	87%	97%	95%	96%	0.046
DT (PCA)	96%	89%	99%	96%	98%	0.015
DT (ANOVA)	97%	95%	100%	97%	98%	0.010
RF (PCA)	97%	93%	100%	97%	98%	0.218
RF (ANOVA)	98%	97%	100%	98%	99%	0.14
GB (PCA)	96%	90%	98%	97%	98%	1.140
GB (ANOVA)	97.6%	95%	100%	98%	99%	0.5
NB (PCA)	90%	87.5%	98%	91%	94%	0.015
NB (ANOVA)	96%	96%	100%	95%	98%	0.010

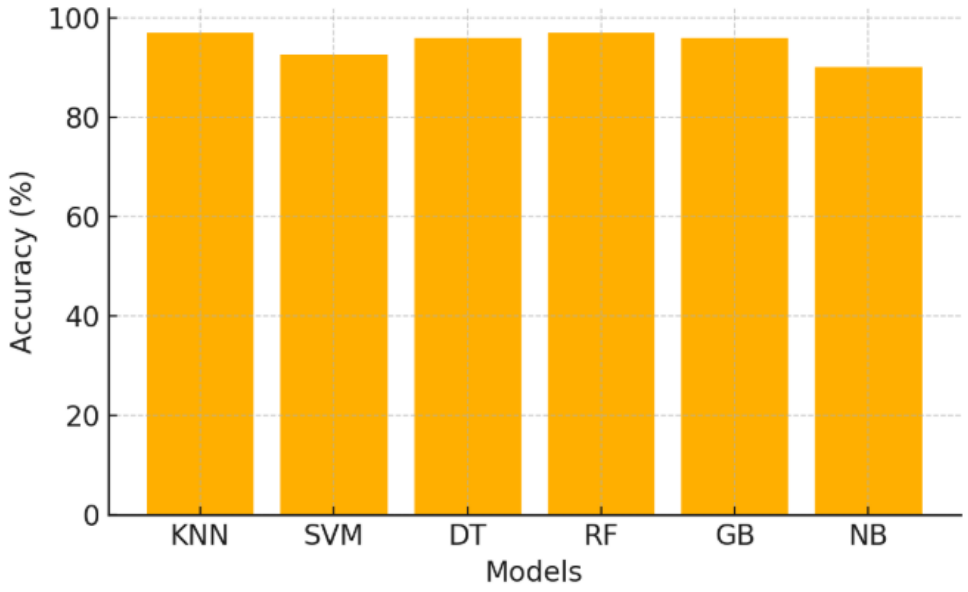

Figure 5: Accuracy rates after dimension reduction

In order to eliminate the overfitting problem of the models, the coefficient was taken as 5 in the validation process and the data set was divided into 5 equal parts and the classification process was performed. Accordingly, different parts of the data set were used for both testing and training purposes each time. The average accuracy values of the validation process are shown in Table 3. It was observed that the ANOVA method gave better results in terms of both the accuracy rate and the running times of the algorithms compared to the PCA method in the feature selection process. The highest accuracy rate as a result of feature selection with ANOVA was obtained with the RO algorithm as 98% and again, a higher success was observed compared to the other algorithms with a success rate of 97% as a result of the validation process of this algorithm (Figure 5). When the algorithms were analyzed in terms of running times, the KA and NB algorithms completed the model training in the shortest time as 0.010 s. When the results are compared in Table 2 and Table 3, since the number of features

has decreased significantly, the training time of the model has also decreased accordingly. The accuracy rate has decreased by approximately 1.5%. If the number of samples in the data set is too high, the feature selection can affect the speed and performance of the model more significantly.
In this study, Chi-Square, Recursive Feature Elimination and Decision Trees and Variable Selection (Feature Importances) methods from the feature selection methods were also tested and no significant changes were observed in the model duration and success rate.

V. CONCLUSION

This study developed a predictive model for forest fires to enable quicker intervention and more effective control. By using a comprehensive dataset from NASA, we applied a series of preprocessing techniques, including the reduction of feature dimensionality using Principal Component Analysis (PCA) and Analysis of Variance (ANOVA). The binary classification models were trained and tested using six different machine learning algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), and Naive Bayes (NB). Our results show that, with the reduced feature set, the models achieved high accuracy, with Random Forest and Naive Bayes reaching 97% and 96% accuracy, respectively. The use of feature selection not only improved the accuracy but also reduced the training time and mitigated the risk of overfitting. This study demonstrates the potential of machine learning techniques for forest fire prediction. In future research, the model can be enhanced by incorporating additional features such as the causes of forest fires and proximity to residential areas, which may provide further insights into fire risk factors. The inclusion of these variables could help create more targeted fire prevention strategies. Additionally, employing deep learning techniques and meta-optimization methods could lead to even more accurate predictions.

REFERENCES

[1] FAO, "Global Forest Resources Assessment 2020 – Key findings," Rome, 2020. [Online]. Available: https://www.fao.org/3/ca8753en/ca8753en.pdf. [Accessed: 22-Oct-2023].

[2] Yeşilyurt, E.N. and Türker, M.F., 2019. Economic Rationality Analysis of Forestry Sector with Econometric Methods (The General Directorate of Forestry Case). Bartın Orman Fakültesi Dergisi, 21(3), pp.893-898.

[3] M. Arif, K. K. Alghamdi, S. A. Sahel, S. O. Alosaimi, M. E. Alsahaft, M. A. Alharthi, and M. Arif, "Role of machine learning algorithms in forest fire management: A literature review," J. Robot. Autom., vol. 5, pp. 212–226, 2021.

[4] V. Sevinc, O. Kucuk, and M. Goltas, "A Bayesian network model for prediction and analysis of possible forest fire causes," Forest Ecol. Manage., vol. 457, p. 117723, 2020.

[5] A. Arpaci, B. Malowerschnig, O. Sass, and H. Vacik, "Using multivariate data mining techniques for estimating fire susceptibility of Tyrolean forests," Appl. Geogr., vol. 53, pp. 258–270, 2014.

[6] M. Fidanboy, N. A. Nihat, and S. Okyay, "Development of a deep learning-based forest fire prediction model and creation of the Turkey fire risk map," Ormancılık Araştırma Dergisi, vol. 9, no. 2, pp. 206–218, 2022.

[7] H. Liang, M. Zhang, and H. Wang, "A neural network model for wildfire scale prediction using meteorological factors," IEEE Access, vol. 7, pp. 176746–176755, 2019.

[8] G. Bayat and K. Yıldız, "Comparison of the Machine Learning Methods to Predict Wildfire Areas," Turkish J. Sci. Technol., vol. 17, no. 2, pp. 241–250, 2022.

[9] I. D. B. Silva, M. E. Valle, L. C. Barros, and J. F. C. Meyer, "A wildfire warning system applied to the state of Acre in the Brazilian Amazon," Appl. Soft Comput., vol. 89, p. 106075, 2020.

[10] Y. Shao, Z. Feng, L. Sun, X. Yang, Y. Li, B. Xu, and Y. Chen, "Mapping China's forest fire risks with machine learning," Forests, vol. 13, no. 6, p. 856, 2022.

[11] T. Preeti, S. Kanakaraddi, A. Beelagi, S. Malagi, and A. Sudi, "Forest fire prediction using machine learning techniques," 2021 International Conference on Intelligent Technologies (CONIT), pp. 1–6, 2021.
[12] M. R. Spoorthy and H. Kumar, "Detection of forest fire areas using machine learning," Int. J. Adv. Res. Sci. Comput. Technol., vol. 2, no. 2, 2022.
[13] P. Mimboro, B. Yanuargi, R. Surimbac, K. Kusrini, and K. Khusnawi, "Forest fire prediction using K-mean clustering and random forest classifier," CSRID Journal, vol. 14, no. 2, pp. 157–165, 2022.
[14] P. Rakshit, S. Sarkar, S. Khan, P. Saha, S. Bhattacharyya, N. Dey, S. Islam, and S. Pal, "Prediction of forest fire using machine learning algorithms: The search for the better algorithm," 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), pp. 1–6, 2021.
[15] Y. Xie and M. Peng, "Forest fire forecasting using ensemble learning approaches," Neural Comput. Appl., vol. 31, pp. 4541–4550, 2018.
[16] M. Castelli, L. Vanneschi, and A. Popović, "Predicting burned areas of forest fires: An artificial intelligence approach," Fire Ecol., vol. 11, no. 1, pp. 106–118, 2015.
[17] T. Niranjan, D. Swetha, V. Charitha, and A. J. Stephen, "Predicting burned area of forest fires," Int. Res. J. Comput. Sci., vol. 6, pp. 132–136, 2019.
[18] R. Coughlan, F. Di Giuseppe, C. Vitolo, C. Barnard, P. Lopez, and M. Drusch, "Using machine learning to predict fire-ignition occurrences from lightning forecasts," Meteorol. Appl., vol. 28, no. 1, p. e1973, 2020.
[19] Y. Pang, Y. Li, Z. Feng, Z. Feng, Z. Zhao, S. Chen, and H. Zhang, "Forest fire occurrence prediction in China based on machine learning methods," Remote Sens., vol. 14, no. 21, p. 5546, 2022.
[20] J. Chen, X. Wang, Y. Yu, X. Yuan, X. Quan, and H. Huang, "Improved prediction of forest fire risk in central and northern China by a time-decaying precipitation model," Forests, vol. 13, no. 3, p. 480, 2022.
[21] C. Lai, S. Zeng, W. Guo, X. Liu, Y. Li, and B. Liao, "Forest fire prediction with imbalanced data using a deep neural network method," Forests, vol. 13, no. 7, p. 1129, 2022.
[22] X. J. Walker, J. L. Baltzer, L. L. Bourgeau-Chavez, N. J. Day, W. J. De Groot, C. Dieleman, E. E. Hoy, J. F. Johnstone, E. S. Kane, M. A. Parisien, S. Potter, B. M. Rogers, M. R. Turetsky, S. Veraverbeke, E. Whitman, and M. C. Mack, "ABoVE: Synthesis of burned and unburned forest site data, AK and Canada, 1983-2016," ORNL DAAC, Oak Ridge, Tennessee, USA, 2022.
[23] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, "On evaluating data preprocessing methods for machine learning models for flight delays," 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2018.
[24] Ramagundam, S. (2022). Ai-Driven Real-Time Scheduling For Linear Tv Broadcasting: A Data-Driven Approach. *International Neurourology Journal*, *26*(3), 20-25.
[25] H. Lin and H. Ding, "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition," J. Theor. Biol., vol. 269, no. 1, pp. 64–69, 2011.
[26] J. Qiu, H. Wang, J. Lu, B. Zhang, and K. L. Du, "Neural network implementations for PCA and its extensions," Int. Scholarly Res. Notices, vol. 2012, 2012.
[27] Ramagundam, S. (2023). Predicting broadband network performance with ai-driven analysis. *Journal of Research Administration*, *5*(2), 11287-11299.
[28] M. M. Uddin, A. A. Bakar, and S. Chien, "Machine learning for prediction and optimization of forest fire outbreaks," Int. J. Comput. Appl., vol. 178, no. 6, pp. 15–23, 2019.
[29] T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.
[30] D. Singh, A. Prakash, and P. Gupta, "A study on K-NN algorithm for classification and its applications," Int. J. Comput. Sci. Inf. Technol., vol. 8, no. 4, pp. 1–8, 2016.

[31] S. Islam, S. U. Rehman, and W. Ali, "A comparative study of K-NN algorithm with different distance metrics," J. Artif. Intell., vol. 2, no. 1, pp. 7–12, 2007.
[32] Ramagundam, S. (2021). Next Gen Linear Tv: Content Generation And Enhancement With Artificial Intelligence. *International Neurourology Journal*, *25*(4), 22-28.
[33] S. Ray, "Decision tree classifier: A survey," Int. J. Comput. Sci. Inf. Technol., vol. 10, no. 1, pp. 45–52, 2019.
[34] A. C. Lorena, A. C. de Carvalho, and G. Zaverucha, "Random forest for classification: A survey," Comput. Intell., vol. 27, no. 3, pp. 184–214, 2011.
[35] P. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," Pattern Recognit. Lett., vol. 27, no. 4, pp. 294–300, 2006.
[36] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[37] L. Xie, X. Li, and Y. Han, "Gradient boosting based machine learning for prediction of forest fires," Comput. Stat. Data Anal., vol. 128, pp. 45–57, 2019.
[38] I. Rish, "An empirical study of the naive Bayes classifier," Proc. IJCAI 2001 Workshop Empirical Methods in Artif. Intell., pp. 41–46, 2001.