Forest Fire Prediction Using Machine Learning and Deep Learning Techniques

M. Shreya, Ruchika Rai, and Samiksha Shukla

Abstract Forests are considered synonyms for abundance on our planet. They uphold the lifecycle of a diversity of creatures, including mankind. Destruction of such forests due to environmental hazards like forest fires is disastrous and leads to loss of economy, wildlife, property, and people. It endangers everything in its vicinity. Sadly, the presence of flora and fauna only increase the fire spread capability and speed. Early detection of these forest fires can help control the spread and protect the nearby areas from the damage caused. This research paper aims at predicting the occurrence of forest fires using machine learning and deep learning techniques. The idea is to apply multiple algorithms to the data and perform comparative analysis to find the best-performing model. The best performance is obtained by the decision tree model for this work. It gave an accuracy of 79.6% and a recall score of 0.90. This model was then implemented on front-end WebUI using the flask and pickle modules in Python. The front-end Website returns the probability that a forest fire occurs for a set of inputs given by the user. This implementation is done using the PyCharm IDE.

Keywords Forests · Fire · Prediction · Machine learning · Deep learning · Flask implementation

1 Introduction

Forests have always been the synonym for abundance on our planet. As a source of food, vegetation, and wildlife, they have played a significant role in human lives ever

M. Shreya · R. Rai · S. Shukla (✉)
Department of Data Science, Christ University, Bangalore, India
e-mail: samiksha.shukla@christuniversity.in

M. Shreya
e-mail: m.shreya@science.christuniversity.in

R. Rai
e-mail: ruchika.rai@christuniversity.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 141,
https://doi.org/10.1007/978-981-19-3035-5_51

since humankind emerged. They control the soil and water cycles, thus impacting the agriculture levels in the nearby areas. Being home to more than half of the animal species, they keep the ecological diversity in check. These giant green land masses play a crucial role in our daily life, from the food we eat or the air we breathe to the climate we see around us.

Countries rely on their forests for the economy as well, as they provide a livelihood to millions of their people. The day-to-day production activities of goods like paper, wood, fruits, etc., even medicines and cosmetics, rely heavily on forests. Forests also improve the touristic value of a country. In places like Hawaii, Bali, Maldives, and most islands, the government relies heavily on tourism, reflected in their budget plans. Local forests provide a beautiful ecotourism experience to their customers.

Forest fires are many environmental hazards, controlling whom becomes complicated with every passing minute. The presence of vegetation acts as fuel to the fire, and every minute, the fire intensifies by multiple folds, thus increasing the water required to put it out. The previous century has marked numerous forest fire events. From the Black Friday Bushfire in Australia's Victoria State in January 1939, which burnt a massive 4.9 million acres, to the recent Californian wildfire in the USA, which burnt around 18,000 acres in 2020, wildfires have been causing havoc and resulting in human and wildlife fatalities. The increase in demand to detect and control these fires with a minimum delay has been growing ever since. Researchers and scientists have been trying to develop solutions to avoid or reduce the effects of these fires so that the damage can be minimized. While the fire control techniques and strategies are taken care of by the Forest Department of every country, there exists a lot of scope in coming up with novel ideas for faster forest fire detection.

Motivation: Machine learning (ML) and deep learning (DL) techniques have been used in every discipline over the past decade. It has provoked scholars to implement these computer techniques even in the forest fire domain. Research has been conducted to study the patterns in forest fires. Many papers have elaborated on the insights from data, used ML or DL models for various purposes, and proposed solutions for implementation. While a few papers suggested prototypes of new sensor devices, a few analysed the satellite images for a particular region, and others only used specific ML or DL techniques with already existing data.

Contribution: This paper aims to analyse and predict forest fires using machine learning and deep learning techniques. Two datasets containing meteorological and Fire Weather Index (FWI) parameters are collected. These datasets correspond to two different regions, namely Portugal and Algeria. To gain insight into the characteristics of the data, a thorough exploratory data analysis (EDA) is performed. The data is visualized using principal component analysis (PCA) and split into train and test sets. Nine machine learning models and one deep learning model are applied to the data, and results are obtained by comparing accuracy, precision, and recall metrics. The best model is obtained based on the metrics, and the same is implemented on a front-end web UI using flask. This WebUI takes the input of certain features from the user and generates an output that shows the probability of forest fire occurrence for the corresponding input conditions.

Organization: This paper is organized as follows. Section 2 provides a brief overview of related research work. The problem definition, research challenges, and dataset description are depicted in Sect. 3. Section 4 presents the methodology and module description. Module description explains pre-processing, model, and architecture of the model. The performance analysis and results are discussed in Sect. 5. Section 6 concludes the paper, and future work is discussed.

2 Related Work

One of the renowned works done in this area was a contribution of Cortez and Morais [1], who collected the dataset for forest fires found in the UCI machine learning repository. There were four different feature selection setups: spatial, temporal, FWI components, and weather attributes. Several data mining models like decision trees (DT), random forest (RF), multiple regression, and neural networks (NN) were used, among which SVM gave the best results with just four attributes such as temperature, wind speed, relative humidity, and rain. Their work was extended by Rishickesh et al. [2], who used bagging and boosting techniques with and without PCA. In their paper [2], logistic regression gave the best result with PCA, and without PCA, gradient boosting gave the best result.

Multiple papers have collected and used NASA's MODIS terra and aqua data for analysis. In contrast, Slobodan Milanovíc et al. [3] collected data for Serbia, and Ma et al. [4] collected data for six geographical regions in China. The random forest model gave the best accuracy for both papers. In Ref. [3], the model performance was assessed by ROC curves, while the other paper [4] concluded that climate and vegetation variables had a significant impact on forest fire occurrence. A new methodology for predicting forest fires in Greece based on the Canadian Fire Weather Index system components (FWI—fire weather index) was introduced by Varela et al. [5]. This paper used simple Geographic Information Systems (GIS) functionality. The methodology outcomes are provided as indicators for individual areas and as maps at a regional or national level. Identifying the advantages of both methods, Stojanova et al. [6] used a combination of NASA's MODIS, Aladin, and GIS data for three regions of Slovenia to predict forest fires. Various machine learning and ensemble methods from the WEKA data mining system were used, and it was seen that bagging of decision trees gave the best result.

Manual data collection was observed in a few research works like Pragati et al. [7] and Varela et al. [8]. The paper [7] used Wireless Sensor Networks (WSN), which took regular readings and sent them to the cloud for storage. In contrast, the paper [8] collected data for only two features, namely temperature and humidity. The data sent to the cloud was experimented with using decision tree and support vector machines. The decision tree gave the best results, whereas the other data used only the regression technique. If the result of the regression technique met some fire threshold conditions, then the ID, location, temperature, and humidity parameters were sent forward for fire control.

While the famous work by Cortez and Morais [1] used five different data mining techniques, Wijayanto et al. [9] used a data mining technique called adaptive neuro-fuzzy inference system (ANFIS). Data on wildfire hotspots in Central Kalimantan, Indonesia, were analysed, and this technique was used to classify fire alarms into true or false ones. The results showed low training and testing errors.

Some papers experimented with image analysis of the forest fire hotspots. Ananya et al. [10] implemented several ML models, among which random forest gave the best results. Then, the satellite image map space of the area was extracted for prediction and converted to an HSV model. If the number of hotspots in the image is more significant than a threshold limit, a fire was detected. However, Syarifudin et al. [11] used a 1D convolutional neural network to predict the number of hotspots for the early treatment of forest fires. They used variations in output, the learning rate, and the number of nodes applied on neural networks. The final results included the MAAPE for daily, monthly, and 12-month predictions, respectively.

A few papers and book chapters were studied to understand forests in general. A broad spectrum of information was gained around forests and their evolution through the paper written by Wodzicki [12], translated to English by Barbara Przybylska. The author elaborated on the evolution of forests on earth, starting from the formation of wood-like plant tissues to the current day forests. It was emphasized that forests ecosystems should be characterized by their entirety and not by specific features. Few other important points like the size of forest lands, effects on gases like nitrogen, soil fertilization, and impact on biodiversity were covered in the book chapter 'Tourism and Forest Ecosystem' by Gössling and Hickler [13]. They explained how tourism impacted the forest ecosystem, the various forest-based activities, and the interdependence of forests and climate changes.

The research work by Zong et al. [14] does not fit any models onto the data. Instead, it mainly focuses on estimating the forest fire seasons and patterns. The data were collected for five countries, namely Kazakhstan, Kyrgyzstan, Tajikistan, Uzbekistan, and Turkmenistan, and the fire seasons were analysed for different vegetation types like shrubs, grasslands, mountain forests, etc.

A literature review paper by Arif et al. [15] and a comparative study by Ahmad A. A. Alkhatib [16] were utilised to understand the current forest fire analysis and prediction trends. The paper [15] reviewed and summarized the recent trends in forest fire events prediction, detection, spread rate, and mapping of the burned areas. It stated the four significant research areas related to forest fires and emphasized that accurate prediction was a challenging task. On the other hand, the paper [16] summarized and elaborated the various methods and then pointed out the advantages and disadvantages of each.

While hybrid models were used by researchers like Shidik and Mustofa [17], an innovative approach was presented in the paper by Silvester et al. [18]. The hybrid model used a combination of clustering and classification approaches where several algorithms were combined and tested with fuzzy c-means (FCM). This model's best result was obtained using the combination of FCM with BPNN and classified into one of the three categories, namely no burn area, light burn, and heavy burn. However, in the paper [18], bird sounds were classified using deep learning (DL) algorithms like convolutional neural network (CNN) into two classes, namely under normal conditions and threatened or panic conditions. ReLU activation function was used, and the results obtained showed that this classification surprisingly gave 96.45% accuracy.

3 Problem Definition, Challenges, and Dataset Description:

3.1 Problem Definition

The problem is to predict the occurrence of forest fires using machine learning and deep learning techniques. This paper aims at a thorough understanding of the dataset. Multiple supervised and ensemble techniques are used to analyse the data. The best model obtained is implemented on a front-end Web UI using the flask module.

3.2 Challenges

The analysis and prediction were performed on a combination of two datasets. These datasets correspond to two different geographic locations, namely Portugal and Algeria. Also, the feature set of one dataset is a subset of the other dataset features. Building the remaining features from the official document is a time-consuming task.

3.3 Dataset Description

The first dataset is called 'Algerian forest fires'. The second dataset is called 'forest fires'. Both the datasets are taken from the UCI machine learning repository, and they correspond to the forest fire data of Algeria and Portugal, respectively.

The final dataset had a total of 14 features, which were Temp, RH, Windspeed, Rain, FFMC, DMC, DC, ISI, BUI, FWI, and Classes (same description as in Table 1) along with month (month of the year 1–12), year (year of forest fire), and location (two classes, Algeria or Portugal). Since the data for BUI and FWI columns were missing in the Portugal dataset, they were imputed using the formulae from the official document [19]. The area and day features were dropped since imputing their missing values would not be meaningful.

Table 1 Features of both the datasets

Algerian data features	Portugal data features
Date—date of forest fire (dd/mm/yy)	X—X-axis coordinate (0–9)
Temp—Temperature in Celsius	Y—Y-axis coordinate (0–9)
RH—Relative humidity in %	Month—Month of the year (Jan. to Dec.)
Ws—Wind speed in km/h	Day—Day of the week (Mon. to Sun.)
Rain—Rainfall in mm/m^2	FFMC—Fine fuel moisture code
FFMC—Fine fuel moisture code	DMC—Duff moisture code
DMC—Duff moisture code	DC—Drought Code
DC—Drought code	ISI—Initial Spread Index
ISI—Initial spread index	Temp—Temperature in Celcius
BUI—Buildup index	RH—Relative Humidity in %
FWI—Fire weather index	Wind—Wind speed in km/h
Classes—two classes (fire and not fire)	Rain—Outside rain in mm/m^2
	Area—Burned area of forest

4 Methodology

Initially, domain knowledge is acquired by reading multiple papers and understanding the kind of work done. The proposed methodology is applied to the combined dataset with 14 columns. A total of nine machine learning algorithms and one simple neural network under deep learning are applied. The coding is done in the Jupyter Notebook using the Python programming language (Fig. 1).

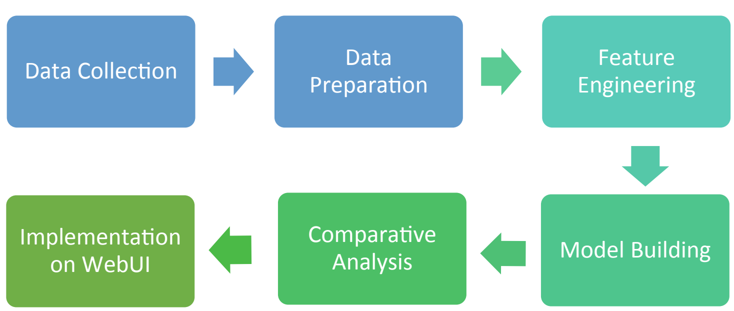

Fig. 1 Methodology

The following steps are followed in methodology:

Step 1: Data Aggregation: The data were collected from two different sources for this study. So, the first step was to aggregate it.

Step 2: Data Preparation: Since a few columns had no data in any one of the datasets, they were either imputed using the official document [19], or they were dropped if deemed not necessary, as mentioned in Sect. 3.3.

Step 3: Exploratory Data Analysis: This step was performed in three stages: univariate, bivariate, and multivariate. Unsupervised techniques like principal component analysis (PCA) and clustering were used under multivariate analysis. Each of them gave significant insights into the data at hand.

Step 4: Training and Test Set: In this step, the data are split into training and testing data in the ratio of 80–20. As a result, 80% of the data is used for training purposes, and the remaining 20% is used for testing the performance of the previously built model. The more training data is, the better it is to learn from it.

Step 5: Model Building: A total of ten models are trained and tested on the data. Among them, nine are machine learning models, which include ensemble techniques. They are support vector machine, decision tree, logistic regression, naïve Bayes, k-nearest neighbours, random forest, AdaBoost classifier, XGBoost classifier, and voting classifier. Under deep learning, a neural network is built, trained, and tested on the data, and the results are analysed.

Step 6: Model Comparison: All the models built in the previous step are compared based on various performance metrics like accuracy, precision, and recall. The best accuracy of 79.6% was obtained using the decision tree algorithm. It also gave a recall score of 0.90 for the 'fire' target class.

Step 7: Feature Identification: Since the decision tree model gave the best results, feature importance is drawn from the model. The top three critical features are selected for the front-end WebUI.

Step 8: WebUI implementation: PyCharm IDE is used for this implementation. Essentially, the WebUI takes input from the user and gives an output along with a safe or dangerous message (Fig. 2).

4.1 Module Description

4.1.1 Proposed Model

The dataset is split into features, targets, training, and testing sets. All the columns excluding the target are considered as features. The data are divided into an 80:20 ratio for training and testing purposes. It means that 80% of the data are used for training the model, and then, the other 20% is used to test the pre-trained model. All the ten algorithms mentioned in Step 5 of the methodology section are trained and tested on the data.

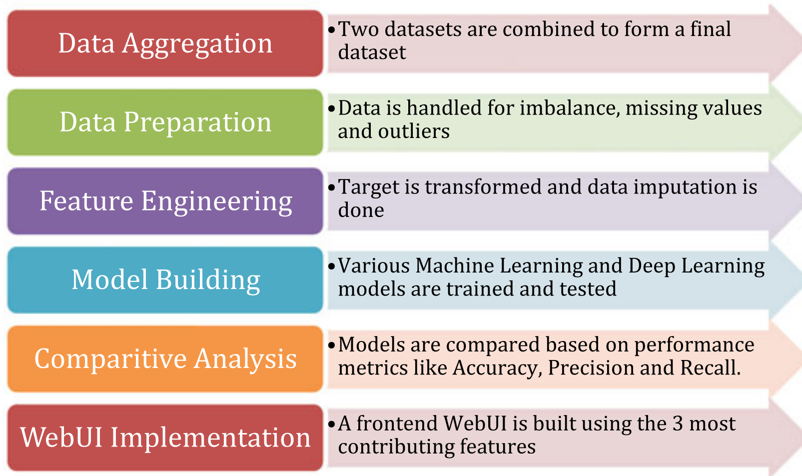

Fig. 2 Overview of workflow

It was seen that the decision tree gave the highest accuracy of 79.6%. Since accuracy can sometimes be biased, the model performances were also checked using the confusion matrix. A classification report was obtained for each model, clearly stating the precision and recall of the fire and non-fire classes. Since fire events are more important than non-fire ones, the precision and recall corresponding to the fire class are considered.

4.1.2 Feature Importance

Since the decision tree algorithm gives the best accuracy, this model is considered the best-performing model on the data. The feature importance is then obtained using Python's sklearn library. The top three features with the highest priority are noted.

Figure 3 clearly shows that features corresponding to 'FFMC', 'windspeed', and 'temperature' have the highest importance, respectively. They will be used for the front-end WebUI implementation.

4.1.3 WebUI Implementation Using Flask

A front-end web page is created where the user inputs three features: FFMC, wind speed, and temperature. These values are captured as a form and sent as a POST or GET request to the back-end. In the back-end, the pickle library is used for storing the models. The decision tree is trained once on the data, and the trained model is loaded into the pickle file. The pickle file is called in the back-end whenever a new input is received from the front-end. Since the model is already trained, the pickle file generates an output for the new data points, and this output is displayed back to

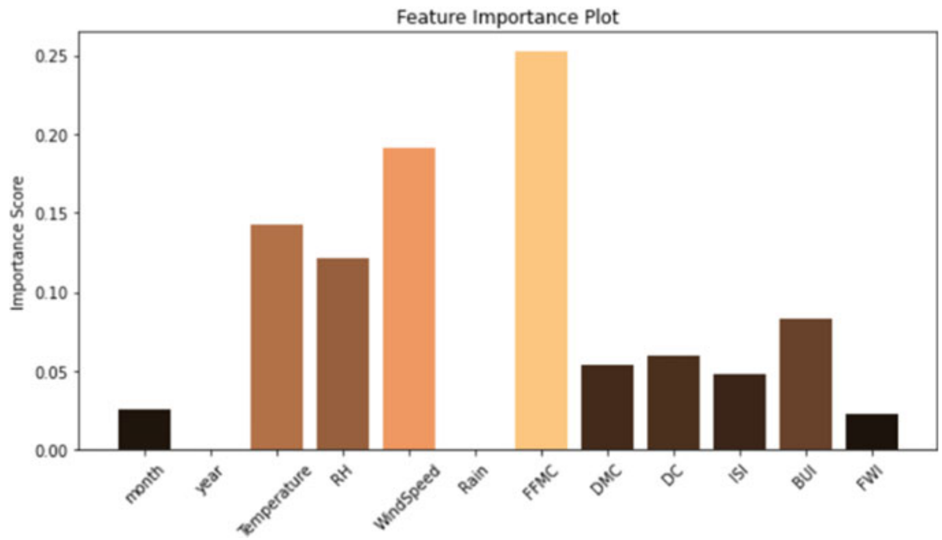

Fig. 3 Feature importance

the user on the front-end. The probability of a forest fire occurring or non-occurring is printed on the front-end along with danger or safe message.

5 Experimental Results

Table 2 shows the results obtained from all the ten models, including the accuracy, precision, recall, f1-score, and AUC-ROC score of each model. The decision tree

Table 2 Model performances

Model name	Accuracy	Precision	Recall	F1-score	AUC-ROC score
Support vector machine	73.6	0.68	0.90	0.80	0.71
Decision tree	79.6	0.73	0.90	0.80	0.75
Logistic regression	69.7	0.67	0.88	0.76	0.68
Naive Bayes	67.1	0.63	0.93	0.75	0.65
K-nearest neighbours	68.4	0.72	0.68	0.70	0.68
Random forest	73.6	0.73	0.80	0.77	0.73
AdaBoost classifier	69.7	0.70	0.78	0.74	0.69
XgBoost classifier	73.6	0.73	0.80	0.77	0.73
Voting classifier	73.6	0.74	0.85	0.80	0.75
Artificial neural network	61.8	–	–	–	–

Fig. 4 Confusion matrix of the decision tree model

```
print(confusion_matrix(y_test,y_preds_dt))

[[21 14]
 [ 4 37]]
```

gives the best accuracy. Also, its recall score is 0.90, which means that the model can correctly identify a fire occurrence 90% of the time.

The confusion matrix gives an insight into how many data points were classified correctly and how many of them were misclassified. The values along the diagonal are correct predictions, and every other value is where the model is misclassified (Fig. 4).

The rows in this confusion matrix indicate the actual classes, whereas the columns indicate the predicted classes. So, there are four values printed, which are true negatives (top left value), false negatives (top right value), false positives (bottom left value), and true positives (bottom right value).

Various metrics like true positive rate, false positive rate, accuracy, precision, recall, sensitivity, specificity, and many more are used to assess the model's performance. These are calculated from the values in the confusion matrix.

Figure 5 shows a sample WebUI input that gives an output probability of forest fire occurrence as 0.87 along with a message that the forest is in danger. If the likelihood of forest fire occurrence is too low, its probability, along with the message 'Your Forest is Safe', is printed instead.

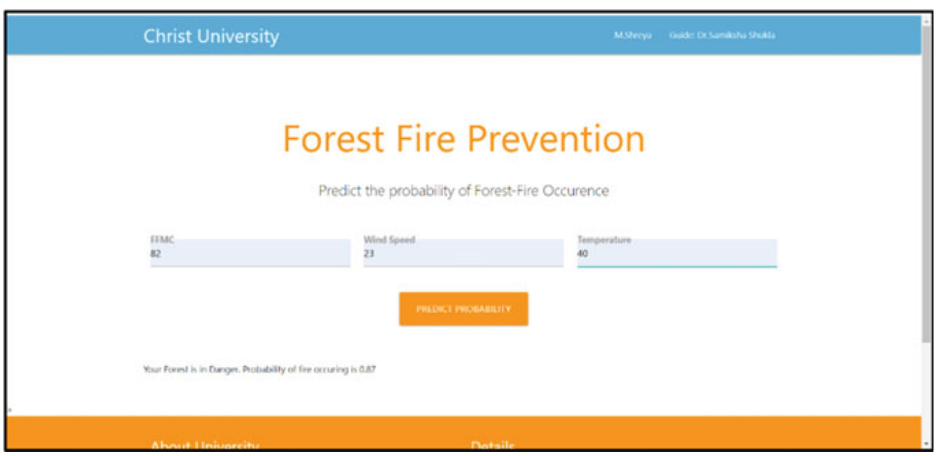

Fig. 5 Front-end WebUI

6 Conclusion and Future Work

This study presents the forest fire-related challenges and prediction model using machine learning and deep learning techniques. Multiple research papers have been examined to understand their objectives, analyse the technique utilised, and identify any gaps in them.

Based on the data collected from various resources, the data cleaning process is applied. After addressing the missing values, various visualization approaches are used to gain insights. Specifically, univariate, bivariate, and multivariate analysis techniques were applied to gain insights. After this, several ML and DL models were applied and tested for their performance. The decision tree algorithm obtained the best results with an accuracy of 79.6%, and a recall score of 0.90 for the fire class. The model was implemented on a front-end WebUI using the flask and pickle Python modules. This WebUI gives the probability of a forest fire for a given set of input parameters.

In the future, the work can be extended by collecting more data from various geographical locations around the globe for further analysis. Using more data would give more precise results for the machine learning model as it would have more data to learn from. Also, a few other deep learning techniques can be applied.

References

1. Cortez P, Morais A (2007) A data mining approach to predict forest fires using meteorological data
2. Rishickesh R, Shahina A, Nayeemulla Khan A (2019) Predicting forest fires using supervised and ensemble machine learning algorithms. Int J Recent Technol Eng (IJRTE) 08(02)
3. Milanovíc S, Markovíc N, Pamucar D, Gigovíc L, Kostíc P, Milanovíc SD (2021) Forest fire probability mapping in eastern serbia: logistic regression versus random forest method. In: MDPI Forests
4. Ma W, Feng Z, Cheng Z, Chen S, Wang F (2020) Identifying forest fire driving factors and related impacts in China using random forest algorithm. In: MDPI Forests
5. Varela V, Vlachogiannis D, Sfetsos A, Politi N, Karozis S (2020) Methodology for the study of near-future changes of fire weather patterns with emphasis on archaeological and protected touristic areas in Greece. In: MDPI Forests
6. Stojanova D, Panov P, Kobler A, Džeroski S, Taškova K (2020) Learning to predict forest fires with data mining techniques
7. Pragati, Shambhuwani S, Umbrajkar P (2019–2020) Forest fire detection using machine learning. Int J Adv Sci Res 04(12)
8. Varela N, Martinez D, Jorge L, Ospino A, Zelaya NAL (2020) Wireless sensor network for forest fire detection. In: The 15th international conference on future networks and communications (FNC), Aug 9–12, 2020
9. Wijayanto AK, Sani O, Kartika ND, Herdiyeni Y (2017) Classification model for forest fire hotspot occurrences prediction using ANFIS algorithm. In: IOP Conference Series: Earth and Environmental Science
10. Shahdeo A, Shahdeo A, Reddy PS, Chaitra K (2020) Wildfire prediction and detection using random forest and different colour models. Int Res J Eng Technol 07(06)

11. Syarifudin MA, Novitasari DCR, Marpaung F, Wahyudi N, Hapsari DP, Supriyati E, Farida Y, Amin FM, Nugraheni RD, Nariswari R, Setiawan F (2020) Hotspot prediction using 1D convolutional neural network. In: 5th International conference on computer science and computational intelligence, procedia computer science, pp 845–853
12. Wodzicki TJ (2020) Forest—the photosphere of life in the Earth's atmosphere (Translated by: Barbara Przybylska). Leśne Prace Badawcze, Forest Research Papers, vol 81, pp 133–138
13. Gössling S, Hickler T (2020) Tourism and global environmental change—chapter X: tourism and forest ecosystems. In: Gössling S, Hall MC (eds) Tourism and global environmental change
14. Zong X, Tian X, Yin Y (2020) Impacts of climate change on wildfires in central Asia. MDPI Forests
15. Arif M, Alghamdi KK, Sahel SA, Alosaimi SO, Alsahaft ME, Alharthi MA, Arif M (2021) Role of machine learning algorithms in forest fire management: a literature review. J Robot Autom 5(1):212–226
16. Alkhatib AAA (2014) A review on forest fire detection techniques. Int J Distrib Sens Netw 2014:12
17. Shidik GF, Mustofa K (2014) Predicting size of forest fire using hybrid model. In: International federation for information processing (IFIP), ICT-EurAsia, pp 316–327
18. Permana SDH, Saputra G, Arifitama B, Yaddarabullah, Caesarendra W, Rahim R (2021) Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm. J King Saud Univ Comput Inf Sci
19. Van Wagner CE (1985) Equations and FORTRAN program for the Canadian forest fire weather index system. In: Pickett TL (ed) Canadian forestry service, Petawawa National Forestry Institute, Chalk River, Ontario, Forestry Technical Report