Distributed Random Forest for Predicting Forest Wildfires Based on Weather Data

Robertas Damaševičius[1] and Rytis Maskeliūnas[2]

[1] Department of Applied Informatics, Vytautas Magnus University, Kaunas Lithuania
robertas.damasevicius@vdu.lt
[2] Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania
rytis.maskeliunas@ktu.lt

Abstract. Forest fires pose a significant threat to ecosystems, economies, and human settlements. Accurate prediction of forest fires can aid in timely interventions, resource allocation, and effective management strategies. This study aimed to develop a machine learning model to predict forest fire occurrences based on various environmental and meteorological variables. Using a dataset comprising variables such as temperature, humidity, wind speed, and moisture codes (FFMC, DMC, DC, and ISI), we employed Distributed Random Forest (DRF) and a 5-fold cross-validation approach on training data to assess the model's performance. The model demonstrated high discriminatory power with an AUC of 0.989 and a low Mean Squared Error (MSE) of 0.041. The results underscored the critical role of weather conditions and fuel moisture content in influencing fire occurrences. The study's findings have implications for forest management, emphasizing the potential of machine learning in shaping fire prevention strategies and safeguarding forest ecosystems.

Keywords: Forest fires · Predictive modeling · Weather data · Wildfire prediction.

1 Introduction

Forest fires, also known as wildfires, have been a natural occurrence shaping ecosystems for millions of years. These fires play a crucial role in the ecological processes, aiding in nutrient recycling, habitat creation, and the natural succession of species. However, in recent decades, the frequency, intensity, and scale of forest fires have escalated dramatically, causing widespread ecological destruction, economic losses, and even human casualties. Several factors contribute to the onset and spread of forest fires. While some are anthropogenic, such as land clearance and arson, natural factors, particularly weather conditions, play a significant role. Weather variables like temperature, humidity, wind speed, and precipitation have been observed to influence the likelihood and intensity of forest fires. For instance, prolonged dry spells, coupled with high temperatures and

strong winds, create conducive environments for fires to ignite and spread. The changing global climate has further exacerbated the situation. With rising global temperatures, many regions are experiencing longer and more intense droughts, making them more susceptible to forest fires. This changing dynamic underscores the need for a deeper understanding of the relationship between weather patterns and forest fires. The devastating impacts of forest fires are felt globally, transcending geographical boundaries and affecting both developed and developing nations. Recent catastrophic events, such as the Australian bushfires of 2019-2020 and the wildfires in Greece [30] and Portugal [25], have brought the issue to the forefront of global attention. These events have not only resulted in the loss of millions of acres of forest but have also led to significant human displacement, loss of life, and billions in economic damages. Yet, despite the increasing frequency and intensity of these fires, our predictive capabilities remain limited. Current early warning systems and predictive models often rely on a combination of historical data and real-time observations, which, while valuable, may not always provide adequate lead time for preventive measures or resource allocation.

Weather patterns, being a primary natural driver for forest fires, offer a promising avenue for enhancing our predictive capabilities [9]. If we can harness the vast amounts of weather data available and develop robust models that accurately predict forest fires, we can significantly improve our preparedness and response strategies [7]. Furthermore, as climate change continues to alter global weather patterns, understanding the nuanced relationship between weather variables and forest fires becomes even more critical. This research is motivated by the urgent need to bridge the gap between our current understanding and the evolving challenges posed by forest fires in a changing climate. In this context, predicting forest fires based on weather data becomes not just an academic exercise but a pressing necessity. Accurate predictions can aid in early warning systems, better resource allocation, and more informed forest management strategies, potentially saving lives, preserving biodiversity, and reducing economic losses [3].

The primary objective of this study is to explore the potential of weather data as a predictive tool for forest fires. To achieve this, the research aims to:

1. Analyze and quantify the relationship between key weather variables, namely temperature, humidity, wind speed, and precipitation, and the occurrence and intensity of forest fires.
2. Evaluate the performance of these models in terms of their accuracy, reliability, and applicability in real-world scenarios.

2 Literature Review

2.1 Overview of Forest Fire Prediction Methods

The prediction of forest fires has been a topic of interest for several decades, with methodologies evolving alongside advancements in technology and data

availability. Early prediction methods were primarily deterministic, relying on field observations and expert judgment. These methods often utilized the Fire Danger Rating System (FDRS), which considered factors like fuel moisture content, wind speed, and temperature to assess fire risk [4].

With the advent of computational capabilities, statistical models became prominent. Logistic regression has been employed to predict the probability of fire occurrence based on meteorological variable [21, 29, 31]. Time-series analysis, especially autoregressive integrated moving average (ARIMA) models, have been used to forecast fire occurrences based on historical data [28].

The recent two decades have witnessed a surge in the application of machine learning techniques for forest fire prediction [23, 5, 24, 13]. Decision trees and random forests have been popular due to their ability to handle non-linear relationships and provide insights into sniprtance [1]. Neural networks, with their capacity to model complex patterns, have also been explored, especially with the rise of deep learning architectures [19]. Despite these advancements, challenges persist. Many models, while accurate in controlled settings, struggle with real-world applicability due to the dynamic and multifaceted nature of forest fires. Additionally, the integration of diverse data sources, from satellite imagery to on-ground sensors, remains a complex task. While significant strides have been made in forest fire prediction methods, there remains a pressing need for models that are both accurate and applicable in diverse real-world scenarios.

2.2 Weather Variables and Their Impact on Forest Fires

Weather variables play a pivotal role in influencing the occurrence, spread, and intensity of forest fires. Their impact on forest fires has been extensively studied, revealing intricate relationships that vary across different geographical and temporal scales [12, 17].

Temperature is one of the most influential factors in forest fire dynamics. Elevated temperatures lead to increased evaporation rates, drying out vegetation and making it more susceptible to ignition. Moreover, high temperatures can increase the intensity and spread rate of fires once ignited. Several studies have shown a direct correlation between prolonged heatwaves and an increase in the number and intensity of forest fires.

Humidity, or the amount of moisture in the air, inversely affects the likelihood of forest fires. Low humidity levels result in drier conditions, reducing the moisture content in vegetation and making it more flammable. Conversely, high humidity levels can act as a mitigating factor, reducing the fire's intensity and spread. The interplay between temperature and humidity is especially crucial, with their combined effects often determining the overall fire risk.

Wind plays a dual role in forest fire dynamics. On one hand, strong winds provide the necessary oxygen to fuel the fire, increasing its intensity and spread rate. On the other hand, winds can carry embers and firebrands over considerable distances, leading to spot fires and rapid fire propagation. Wind direction, in conjunction with speed, can influence the direction of fire spread, making it a critical factor in fire management and containment strategies.

Precipitation, both in terms of its amount and distribution, has a direct impact on forest fire risk. Regular rainfall can maintain moisture levels in vegetation, reducing its flammability. Irregular precipitation patterns, such as prolonged droughts followed by short intense rainfall, can create conditions conducive to fires. Rain can lead to rapid vegetation growth, which, when followed by dry periods, results in an abundance of dry fuel, elevating the fire risk.

While each weather variable has its distinct impact on forest fires, it's their combined and interactive effects that determine the overall fire risk in a region. Understanding these intricate relationships is crucial for developing accurate predictive models.

2.3 Machine Learning Models and Their Limitations

Machine learning has emerged as a powerful tool in the realm of forest fire prediction and detection. Over the years, various models have been proposed, each with its strengths and weaknesses [1]. This section delves into some of the most prominent machine learning models used for fire detection and highlights their associated limitations.

Decision Trees are a popular choice for fire prediction due to their interpretability and ability to handle non-linear relationships. They work by recursively splitting the data based on feature thresholds, leading to a tree-like model of decisions. Random Forests, an ensemble of decision trees, further enhance the model's accuracy by aggregating predictions from multiple trees [27, 26, 8]. While Random Forests mitigate this to some extent, individual decision trees are prone to overfitting, especially with noisy data. In regions where fires are rare, the model might be biased towards non-fire predictions. Decision trees are inherently local models and might not generalize well to conditions outside the training data.

Support Vector Machines (SVM) have been employed for fire detection, especially in scenarios where the dataset is not vast [21, 26, 8]. They work by finding the hyperplane that best separates the classes in a high-dimensional space. SVMs can be computationally intensive, especially with large datasets. The performance of SVMs can vary significantly based on the choice of kernel and regularization parameters. Unlike decision trees, SVMs do not provide an intuitive understanding of feature importance.

With the rise of deep learning, Artificial Neural Networks (ANN) have been explored for fire detection [27, 10, 32]. These models consist of interconnected layers of neurons that can learn complex patterns from data. Deep neural networks, with their vast number of parameters, are prone to overfitting, especially with limited data. Training deep neural networks requires significant computational resources. Neural networks, especially deep ones, lack the interpretability of models like decision trees, making it challenging to understand their predictions.

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm. When used For fire prediction [31, 11], it would classify a new instance based on the majority class of its 'k' nearest training instances. KNN's performance

can degrade if the dataset has irrelevant or redundant features. The algorithm can be slow for large datasets as it requires computing distances to all training instances for each prediction. KNN assumes that data points in the same class are homogeneous, which might not always hold true for fire prediction.

While machine learning offers promising avenues for fire detection, it's evident that no single model is universally optimal. The choice of model depends on the specific context, data availability, and desired outcomes. Moreover, the inherent limitations of these models underscore the importance of continuous research, model refinement, and the potential benefits of hybrid models that combine the strengths of multiple approaches.

2.4 Deep Learning Models and Their Limitations

Deep learning, a subset of machine learning, has revolutionized numerous domains, including the realm of forest fire detection. Leveraging neural networks with many layers, deep learning models can automatically learn hierarchical representations from data. This section delves into some of the most prominent deep learning models used for fire detection and highlights their associated limitations.

Convolutional Neural Networks (CNNs) are primarily designed for image processing and have been employed for detecting fires using satellite and aerial imagery [14, 16, 2]. They consist of convolutional layers that automatically and adaptively learn spatial hierarchies from the data. CNNs require vast amounts of labeled data to train effectively, which can be a challenge given the rarity of fire events. Training CNNs, especially deeper architectures, demands significant computational resources and time. While transfer learning using pre-trained CNNs can mitigate the data demand, these models might not always be optimal for specific fire detection tasks.

Recurrent Neural Networks (RNNs), and their advanced variant Long Short-Term Memory (LSTM), are designed to handle sequential data. They've been explored for predicting forest fires based on time-series weather data [20, 18, 15]. Traditional RNNs suffer from the vanishing gradient problem, making them challenging to train. While LSTMs mitigate this issue, they introduce additional complexity. LSTMs, with their gating mechanisms, can be memory-intensive. While they handle sequences, LSTMs might struggle with very long-term dependencies without additional architectural modifications.

Deep learning models, with their ability to learn intricate patterns from data, offer promising avenues for fire detection. However, they come with their set of challenges, primarily related to data demand, computational intensity, and interpretability. As with traditional machine learning, the choice of a deep learning model for fire detection should be context-specific, considering the nature of the data, available resources, and desired outcomes. Continuous research in this domain is essential to harness the full potential of deep learning for forest fire detection and mitigation.

3 Methodology

3.1 Dataset

We use the dataset [6]. The dataset captures forest fires in the northeast region of Portugal and integrates meteorological data with data on forest fires. The primary aim of this dataset is to predict the burned area of forest fires using these attributes. The primary objective when using this dataset is to predict the "Area" attribute (i.e., the burned area) based on the other attributes. This can be approached as a regression problem, where the goal is to predict a continuous value, or as a classification problem (e.g., by categorizing the burned area into 'low', 'medium', 'high'). Here we predict if $Area > 0$, i.e., whether a fire has occured. The dataset captures a diverse range of conditions, from different times of the year to varying weather scenarios.

The attributes of the dataset are summarized in table 1. The Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI) are components of the Fire Weather Index (FWI) system, which is widely used for assessing wildfire potential. This system is used to estimate the risk of wildfire in various regions based on meteorological data. Each of these components provides insights into different aspects of fire potential.

FFMC is an index that represents the moisture content in the top litter and other fine fuels present in the forest floor. It gives an indication of the relative ease of ignition and the flammability of fine fuel. The FFMC value increases as the moisture content decreases (i.e., as the fine fuels become drier). A high FFMC value indicates that surface fuels are dry and can easily ignite. It's particularly sensitive to changes in relative humidity and temperature.

DMC represents the moisture content in the organic layers beneath the surface, specifically in the upper duff layers. This index gives an indication of fuel consumption in moderate-depth duff layers and medium-sized woody material. A higher DMC value indicates that the organic materials in the subsurface are dry. This means that if a fire were to start, it could burn more deeply and intensely, consuming the organic material in the forest floor.

DC is an index that represents the moisture content in deeper, compact organic layers. It's an indicator of seasonal drought effects and the flammability of the deeper organic layers. A high DC value indicates that the deeper organic layers are dry, suggesting a long-term drying trend. This can be particularly concerning as it means fires can burn deeply into the ground, making them harder to extinguish and causing them to smolder for longer periods.

ISI is an index that represents the rate at which a fire will spread. It combines the effects of wind and the FFMC on fire spread. A high ISI value indicates that if a fire were to start, it would spread rapidly. This is particularly influenced by wind speed; strong winds can quickly spread embers and flames, leading to a faster-moving fire.

These indices provide a comprehensive view of the potential fire behavior, from ignition to spread, based on the moisture content in various layers of the forest floor and the effects of weather conditions. They are crucial tools for forest

management and fire prevention, helping authorities make informed decisions about fire risk and resource allocation.

The spatial coordinates (X and Y) provide a sense of the location of the fire within the Montesinho park, which can be crucial for understanding patterns or specific regions more prone to fires.

Table 1: Description of the Forest Fires Dataset Attributes

Attribute	**Description**
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
Month	month of the year: "jan" to "dec"
Day	day of the week: "mon" to "sun"
FFMC	Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
DMC	Duff Moisture Code index from the FWI system: 1.1 to 291.3
DC	Drought Code index from the FWI system: 7.9 to 860.6
ISI	Initial Spread Index from the FWI system: 0.0 to 56.10
Temp	temperature in Celsius degrees: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
Wind	wind speed in km/h: 0.40 to 9.40
Rain	outside rain in mm/m2: 0.0 to 6.4
Area	the burned area of the forest (in ha): 0.00 to 1090.84 (target variable)

3.2 Data Preprocessing

Data preprocessing is a critical step in the machine learning pipeline, ensuring that the dataset is well-suited for model training. This section discusses common preprocessing steps, including handling missing values and feature engineering.

Missing values in a dataset can arise due to various reasons, such as data entry errors, unrecorded observations, or sensor malfunctions. Handling them is crucial as most machine learning algorithms require complete datasets for training [22]. Several strategies can be employed:

- **Deletion:** Simply remove the rows with missing values. This method is mathematically represented as:

$$D' = \{d \in D | \text{value}(d, a) \neq \text{missing}, \forall a \in A\} \tag{1}$$

  where D' is the dataset after deletion, D is the original dataset, and A is the set of all attributes.
- **Mean Imputation:** Replace the missing values with the mean of the observed values for that feature. For a feature X:

$$x_{\text{missing}} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

  where x_i are the observed feature values and n is the number of observations.

– **Median or Mode Imputation:** For numerical and categorical features, respectively, replace missing values with the median or mode.

Data augmentation is a technique used to artificially increase the size of a dataset by creating modified versions of existing data. This is especially useful in domains like image processing, where deep learning models require large datasets to train effectively. Augmenting data can help improve the performance and generalization of models by providing them with a more diverse set of training examples. In this study, we used AugmenterR library in R. It employs a data enhancement method grounded in conditional entropy. This approach can generate new data points based on a specified value of a categorical feature, enhancing the dataset for classification purposes. Additionally, it demonstrates notable enhancements for machine learning models working with limited data.

3.3 Distributed Random Forest (DRF)

Distributed Random Forest (DRF) is an ensemble learning method that constructs multiple decision trees during training and outputs the mode (for classification) or mean (for regression) prediction of the individual trees for unseen data [33]. DRF is designed to be distributed and scalable, making it suitable for large datasets. Given a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the DRF algorithm proceeds as follows:

1. For each tree t in the forest:
   (a) A bootstrap sample (with replacement) of the data is taken.
   (b) A decision tree $h_t(x)$ is grown using the bootstrap sample. At each node:
      i. A random subset of features is selected.
      ii. The best split based on these features is used to split the node.
2. The final model $H(x)$ is the aggregation of the predictions of all trees:

$$H(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

for regression, or

$$H(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

for classification, where T is the number of trees.

3.4 Model Evaluation Metrics

Evaluating the performance of predictive models is crucial to understand their accuracy and reliability. Various metrics can be employed, each providing a different perspective on the model's performance. This section delves into three commonly used metrics for regression tasks.

The Mean Absolute Error (MAE) provides a measure of the average magnitude of errors between predicted and observed values, without considering their direction. It is given by the formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where y_i represents the observed values, $\hat{y}_i$ denotes the predicted values, and n is the number of observations.

The Root Mean Square Error (RMSE) is another metric that measures the average magnitude of errors. However, by squaring the differences before averaging, RMSE gives more weight to larger errors. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Similar to MAE, y_i and $\hat{y}_i$ are the observed and predicted values, respectively, and n is the number of observations.

The R-squared value, often termed the coefficient of determination, provides a measure of how well the observed outcomes are replicated by the model. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables. The R-squared value is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed values.

The Gini coefficient, often used in economics to measure income inequality, can also be applied in the context of binary classification as a performance measure. It quantifies the disparity between the distribution of the positive and negative classes in predictions. Mathematically, the Gini coefficient (G) can be defined in terms of the Area Under the ROC Curve (AUC) as:

$$G = 2 \times \text{AUC} - 1 \tag{6}$$

The Gini coefficient ranges between -1 (perfect inequality) and 1 (perfect equality). In the context of classification, a Gini coefficient close to 1 indicates that the model has good discriminatory power, while a value close to 0 suggests that the model is no better than random guessing. For a perfectly discriminating model, AUC = 1 and thus $G = 1$. For a model that discriminates no better than random guessing, AUC = 0.5 and $G = 0$.

LogLoss is a performance metric used to evaluate the accuracy of a classification model where the prediction output is a probability value between 0 and 1. It penalizes both the type I and type II errors in predictions. The closer the predicted probabilities are to the actual outcomes, the lower the LogLoss value,

making it a suitable metric for models that output probabilities. For a binary classification problem, the LogLoss is defined as:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{7}$$

Where: N is the number of samples or instances. y_i is the actual class label of the i^{th} instance (0 or 1). p_i is the predicted probability that the i^{th} instance belongs to class 1. A perfect model would have a LogLoss of 0. However, it's important to note that a smaller LogLoss is better, with 0 indicating a perfect log-likelihood. Conversely, a model with predictions that are off from the actual values will incur a larger LogLoss.

4 Results

4.1 Descriptive Statistics of the Dataset

The analysis of the variables' values is presented in Figure 1. The FFMC values are generally high, indicating that fine fuels are typically dry, which can be a fire risk. Most of the values are clustered around the 90s, indicating that fine fuels are generally dry, which can be a fire risk. The DMC and DC values show significant variability, suggesting diverse moisture content in both shallow and deep organic layers across regions. The DC values indicate significant variability in the moisture content of deeper organic layers. The median is higher than the mean, suggesting that most regions have a higher drought code, indicating drier conditions. The ISI has some regions with exceptionally high fire spread rates.The maximum value of ISI is notably higher than the 3rd quartile, suggesting some regions with exceptionally high fire spread rates. There's a wide range of temperatures, but most regions have little to no rain, which can contribute to dry conditions and increased fire risk. The distribution of wind speed is fairly even, but the maximum value suggests there might be occasional strong winds.

4.2 Model Performance

Table 2 summarizes the performance of the trained model. The model has a high mean accuracy of approximately 96.3%. The low standard deviation (0.00696) suggests that the accuracy is consistent across different runs or subsets of the data. The low error rate (0.03714) further confirms the model's high accuracy. The high Matthews Correlation Coefficient (mcc) value (0.92584) suggests that the model performs well across both positive and negative classes.

The model's performance is summarize in Table 3 in terms of Area Under Curve (AUC), accuracy (ACC), Precision (PRC), True Positive Rate (TPR) and True Negative Rate (TNR).

Performance Metrics from 5-fold cross-validation are reported in Table 4. The model appears to perform well based on the provided metrics. The high AUC and AUCPR values indicate excellent discriminatory power, and the low MSE

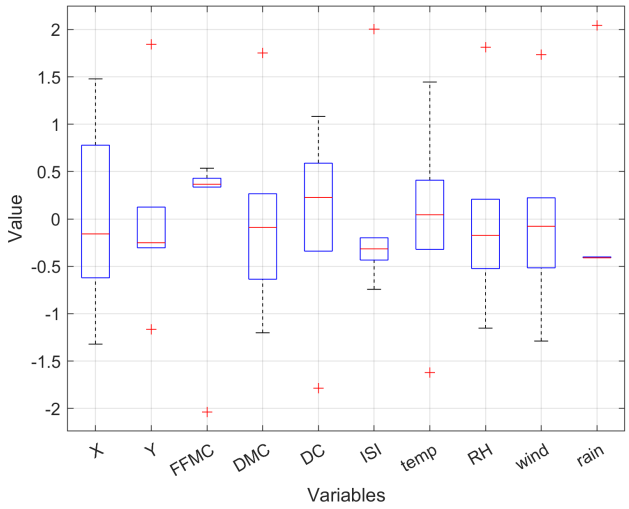


Fig. 1: Boxplot of the normalized values of dataset variables

Table 2: Descriptive Statistics of the Model Metrics

Metric	**Mean**	**Standard Deviation (SD)**
accuracy	0.962857	0.006962
auc	0.989500	0.004102
err	0.037143	0.006962
err_count	10.400000	1.949359
logloss	0.204777	0.066532
max_per_class_error	0.050543	0.013681
mcc	0.925840	0.013846
mean_per_class_accuracy	0.962847	0.007596
mean_per_class_error	0.037153	0.007596
mse	0.041276	0.004419
pr_auc	0.987651	0.010687
precision	0.968042	0.015978
r2	0.834264	0.017558
recall	0.956842	0.019023
rmse	0.202935	0.010779
specificity	0.968852	0.014449

Table 3: Classification Performance Metrics

Metric	**AUC**	**ACC**	**PRC**	**TPR**	**TNR**
Value	0.98942	0.95333	0.95608	0.94966	0.95695

and LogLoss values suggest accurate predictions. The R^2 value indicates that a high portion of the variance in the target variable is explained by the model.

Table 4: Performance Metrics from 5-fold Cross-validation

Metric	**Value**
MSE	0.0411854
RMSE	0.2029419
LogLoss	0.203288
Mean Per-Class Error	0.04288606
AUC	0.9891877
AUCPR	0.9883058
Gini	0.9783753
R^2	0.8352571

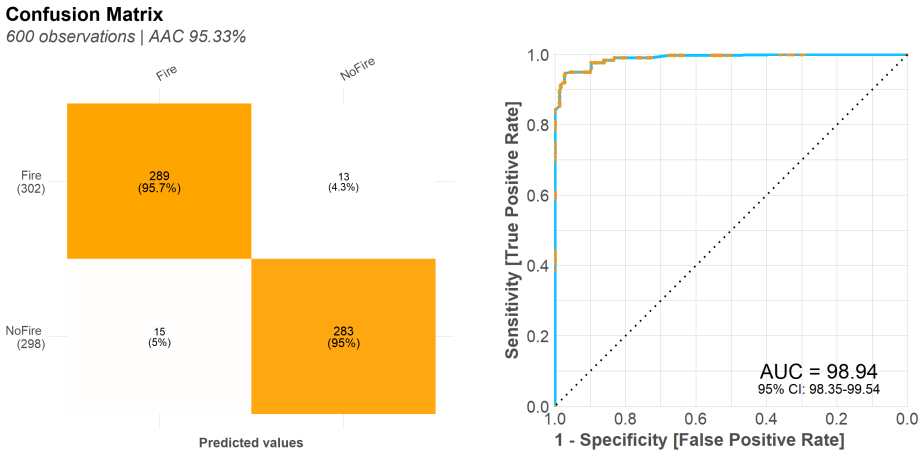
The confusion matrix of the classification results is presented in Figure 2a. The error rate for the Fire class is 0.022923, which means about 2.29% of the total Fire instances were misclassified. The error rate for the NoFire class is 0.032764, which means about 3.28% of the total NoFire instances were misclassified. The overall error rate for the model is 0.027857, meaning about 2.79% of all instances (both Fire and NoFire) were misclassified. The model seems to perform relatively well with an overall accuracy of about 97.21% (100% - 2.79%). However, there's a slightly higher misclassification rate for the NoFire class compared to the Fire class. The results of ROC (Receiver Operating Characteristic) analysis are presented in Figure 2b.

4.3 Feature Importance Analysis

Table 5 and Figure 3 present the results of feature importance analysis. The most important feature is temperature (*temp*) explaining 16.5% of variability followed by Duff Moisture Code (DMC) explaining 12.7% and and relative humidity (RH) explaining 12.6%.

Table 5: Variable Importance Metrics

Variable	**Relative Importance**	**Scaled Importance**	**Importance**
temp	1048.5476	1.0000000	0.1651155
DMC	811.0522	0.7735006	0.1277169
RH	802.3088	0.7651620	0.1263401
DC	695.1674	0.6629812	0.1094685
wind	678.9797	0.6475431	0.1069194
X	657.9071	0.6274461	0.1036011



(a) Confusion matrix (b) ROC analysis

Fig. 2: Performance of the model

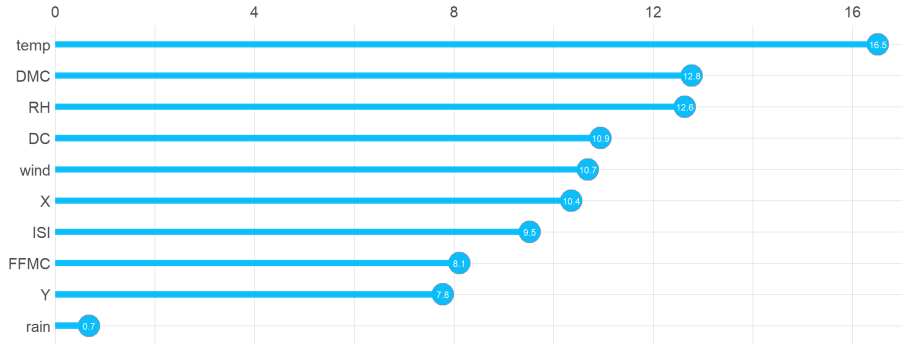


Fig. 3: Importance of features

5 Conclusion

In this study, we employed a comprehensive approach to predict forest fires based on various environmental and meteorological variables. The results from our 5-fold cross-validation on training data provide compelling evidence of the model's robustness and accuracy in predicting forest fires.

Several key findings emerged from our analysis: The model demonstrated high discriminatory power, as evidenced by the high AUC and AUCPR values. These metrics, being close to 1, indicate that the model can effectively distinguish between instances of fires and no fires. The low MSE and LogLoss values further attest to the model's accuracy in its predictions. Variables such as temperature, humidity, and the various moisture codes (FFMC, DMC, DC) played pivotal roles in influencing the model's predictions. Their relative importance in

the model underscores the critical role of weather conditions and fuel moisture content in forest fire occurrences. The R^2 value of 0.8352571 suggests that our model accounts for approximately 83.53% of the variance in the target variable. This high R^2 value, combined with other performance metrics, indicates that our model is not only accurate but also robust in its predictions across different scenarios and conditions.

The results of this study have significant implications for forest management and fire prevention strategies. By understanding the key variables that influence fire occurrences and their relative importance, forest management can devise more targeted and effective strategies to mitigate fire risks. For instance, during periods of high temperatures and low humidity, combined with high FFMC, DMC, or DC values, forest managers can increase surveillance, restrict certain activities, or deploy resources in anticipation of potential fires.

This study, focused on predicting forest fires based on weather data, has some notable limitations:

- The dataset covers the northeast region of Portugal. While valuable, the findings might not be directly applicable to other regions with different climatic conditions, vegetation types, or socio-economic factors.
- The study emphasizes weather data, but potentially influential variables, such as land use patterns, vegetation health, and human activities, were not included in the analysis, possibly limiting the model's predictive power.
- The dataset represents a specific time frame. Forest fire patterns and their relationship with weather might evolve over longer periods, especially in the context of global climate change.
- The machine learning models employed, though state-of-the-art, have their inherent limitations and assumptions. The real-world is often more complex than what can be captured by a single model or algorithm.
- Given the complexity of models and granularity of the dataset, there's a potential risk of overfitting, where the model might perform exceptionally well on the training data but fail to generalize to new, unseen data.
- Events like policy changes, significant infrastructural developments, or large-scale human migrations can influence forest fire patterns. Such external factors were not considered in this study.

Acknowledgement

This research paper has received funding from Horizon Europe Framework Programme (HORIZON), call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) - Creation of the centre of excellence in smart forestry "Forest 4.0" No. 101059985.

References

1. Abid, F.: A survey of machine learning algorithms based forest fires prediction and detection systems. Fire Technology **57**(2), 559 – 590 (2021)

2. Ahmad, K., Khan, M.S., Ahmed, F., Driss, M., Boulila, W., Alazeb, A., Alsulami, M., Alshehri, M.S., Ghadi, Y.Y., Ahmad, J.: Firexnet: an explainable ai-based tailored deep learning model for wildfire detection on resource-constrained devices. Fire Ecology **19**(1) (2023)
3. Aljumah, A.: Iot-inspired framework for real-time prediction of forest fire. International Journal of Computers, Communications and Control **17**(3) (2022)
4. Amelia, J.P., Dupe, Z.L., Prasasti, I.: Analysis and verification of fire danger rating system (fdrs) parameters in land and forest fire in west kalimantan in 2019 and its relationship with hotspots and rainfall. vol. 275, p. 247 – 264 (2022)
5. Bera, B., Shit, P.K., Sengupta, N., Saha, S., Bhattacharjee, S.: Forest fire susceptibility prediction using machine learning models with resampling algorithms, northern part of eastern ghat mountain range (india). Geocarto International **37**(26), 11756 – 11781 (2022)
6. Cortez, P., Morais, A.d.J.R.: A data mining approach to predict forest fires using meteorological data (2007)
7. Damaševičius, R., Bacanin, N., Misra, S.: From sensors to safety: Internet of emergency services (ioes) for emergency response and disaster management. Journal of Sensor and Actuator Networks **12**(3) (2023)
8. Dong, H., Wu, H., Sun, P., Ding, Y.: Wildfire prediction model based on spatial and temporal characteristics: A case study of a wildfire in portugal's montesinho natural park. Sustainability **14**(16) (2022)
9. Flannigan, M., Wotton, B.: Climate, Weather, and Area Burned. Elsevier (2007)
10. Gaikwad, A., Bhuta, N., Jadhav, T., Jangale, P., Shinde, S.: A review on forest fire prediction techniques (2022)
11. Ghate, S.N., Sapkale, P., Mukhedkar, M.: Forest wildfire detection and forecasting utilizing machine learning and image processing (2023)
12. Ivchenko, O., Tiutin, A., Kozachenko, M., Pankin, K.: A relationship between weather conditions and a number of forest fires. vol. 979 (2022)
13. Li, L., Sali, A., Noordin, N.K., Ismail, A., Hashim, F.: Prediction of peatlands forest fires in malaysia using machine learning. Forests **14**(7) (2023)
14. Li, X., Wang, X., Sun, S., Wang, Y., Li, S., Li, D.: Predicting the wildland fire spread using a mixed-input cnn model with both channel and spatial attention mechanisms. Fire Technology **59**(5), 2683 – 2717 (2023)
15. Liang, H., Zhang, M., Wang, H.: A neural network model for wildfire scale prediction using meteorological factors. IEEE Access **7**, 176746 – 176755 (2019)
16. Mittal, P., Sharma, A., Singh, R.: Deformable patch-based-multi-layer perceptron mixer model for forest fire aerial image classification. Journal of Applied Remote Sensing **17**(2) (2023)
17. Mohammadian Bishe, E., Norouzi, M., Afshin, H., Farhanieh, B.: A case study on the effects of weather conditions on forest fire propagation parameters in the malekroud forest in guilan, iran. Fire **6**(7) (2023)
18. Murali Mohan, K.V., Satish, A.R., Mallikharjuna Rao, K., Yarava, R.K., Babu, G.C.: Leveraging machine learning to predict wild fires. p. 1393 – 1400 (2021)
19. Mutakabbir, A., Lung, C.H., Ajila, S.A., Zaman, M., Naik, K., Purcell, R., Sampalli, S.: Spatio-temporal agnostic deep learning modeling of forest fire prediction using weather data. vol. 2023-June, p. 346 – 351 (2023)
20. Natekar, S., Patil, S., Nair, A., Roychowdhury, S.: Forest fire prediction using lstm (2021)
21. Pahuja, N.K., Rivero, M.H.: Predicting the impact of wildfire using machine learning techniques to assist effective deployment of resources. p. 201 – 205 (2022)

22. Palanivinayagam, A., Damaševičius, R.: Effective handling of missing values in datasets for classification using machine learning methods. Information **14**(2) (2023)
23. Pang, Y., Li, Y., Feng, Z., Feng, Z., Zhao, Z., Chen, S., Zhang, H.: Forest fire occurrence prediction in china based on machine learning methods. Remote Sensing **14**(21) (2022)
24. Pham, B.T., Jaafari, A., Avand, M., Al-Ansari, N., Du, T.D., Hai Yen, H.P., Phong, T.V., Nguyen, D.H., Van Le, H., Mafi-Gholami, D., Prakash, I., Thuy, H.T., Tuyen, T.T.: Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry **12**(6) (2020)
25. Pinto, M.M., et. al.: The extreme weather conditions behind the destructive fires of June and October 2017 in Portugal. Imprensa da Universidade de Coimbra (2018)
26. Rodrigues, M., De la Riva, J.: An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environmental Modelling and Software **57**, 192 – 201 (2014)
27. Rubí, J.N., de Carvalho, P.H., Gondim, P.R.: Application of machine learning models in the behavioral study of forest fires in the brazilian federal district region. Engineering Applications of Artificial Intelligence **118** (2023)
28. Slavia, A.P., Sutoyo, E., Witarsyah, D.: Hotspots forecasting using autoregressive integrated moving average (arima) for detecting forest fires. p. 92 – 97 (2019)
29. Wu, Z., Li, M., Wang, B., Quan, Y., Liu, J.: Using artificial intelligence to estimate the probability of forest fires in heilongjiang, northeast china. Remote Sensing **13**(9) (2021)
30. Xanthopoulos, G., Roussos, A., Giannakopoulos, C., Karali, A., Hatzaki, M.: Investigation of the weather conditions leading to large forest fires in the area around Athens, Greece. Imprensa da Universidade de Coimbra (2014)
31. Yue, W., Ren, C., Liang, Y., Liang, J., Lin, X., Yin, A., Wei, Z.: Assessment of wildfire susceptibility and wildfire threats to ecological environment and urban development based on gis and multi-source data: A case study of guilin, china. Remote Sensing **15**(10) (2023)
32. Zaidi, A.: Predicting wildfires in algerian forests using machine learning models. Heliyon **9**(7) (2023)
33. Zhou, G., Chen, F.: Drfmm: A map-matching algorithm based on distributed random forest multi-classification. vol. 189 (2018)