Preprints.org

Article

Not peer-reviewed version

Optimizing Forest Fire Prediction: A Comparative Analysis of Machine Learning Models through Feature Selection and Time-Stage Evaluation

Hamed Khosravi , Mohammad Reza Shafie , Ahmed Shoyeb Raihan * , Md Asif Bin Syed , Imtiaz Ahmed

Posted Date: 8 December 2023

doi: 10.20944/preprints202312.0577.v1

Keywords: Forest Fire Prediction; CRISP-DM Methodology; Feature Selection; Decision Tree

Article

Optimizing Forest Fire Prediction: A Comparative Analysis of Machine Learning Models through Feature Selection and Time-Stage Evaluation

Hamed Khosravi [1], Mohammad Reza Shafie [2], Ahmed Shoyeb Raihan [1,*], Md Asif bin Syed [1] and Imtiaz Ahmed [1]

[1] Department of Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV 26506, USA

[2] Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

* Correspondence: shoyebraihan85@gmail.com

Abstract: Despite being considered as natural components of many ecosystems, forest fires pose significant threats to the environment and human health. In order to ensure public safety and effective fire suppression planning, it is necessary to develop reliable prediction models to mitigate forest fire danger. These models should account for specific environmental conditions. The advent of big data in recent years has opened new avenues for improving forest fire predictions. Machine learning techniques, surpassing traditional forecasting methods, have shown significant promise in this area. By applying the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to a real-world dataset, this paper explores the application of machine learning approaches to understand forest fire patterns and predict fire danger. We consider six distinct time stages and incorporate feature selection to refine our predictions. It is important to note that forest fire behavior models are not universally effective due to geographical variations in data. Nevertheless, advanced decision-making techniques are vital in forest fire management. Our research presents a systematic exploration of the topic, comparing various machine learning models, thereby providing a comprehensive baseline for future investigations in this crucial environmental arena.

Keywords: Forest Fire Prediction, CRISP-DM Methodology, Feature Selection, Decision Tree

1. Introduction

Forests are essential for the preservation and sustainability of the natural environment, and they play a crucial role in supporting human habitat [1]. Forest fires are natural and essential components of several terrestrial ecosystems, including but not limited to boreal forests, temperate forests, Mediterranean ecosystems, savannas, and grasslands [2]. In most cases, forest fires can be extinguished quickly with human intervention, but in a few cases they become uncontrollable, causing great damage to the environment and posing a serious danger to human safety [3,4]. It is important to note that forest fires do not occur randomly, but rather in certain locations and under specific conditions [5]. Forest fire prevention has become a crucial research focus. The development of reliable forest fire danger prediction models holds significant importance for various applications, including public safety, forest management, and fire suppression planning [6–8]. Forest fire prediction models assist in taking necessary steps to prevent forest fires and their detrimental effects [9].

Recent studies have demonstrated that machine learning-based methods can discover more complex patterns in forest fire data than traditional methods [10]. Machine learning methods for forest fire prediction include but are not limited to, decision tree-based classifiers [11,12], artificial neural networks (ANN) [13,14], and support vector machines [15,16]. Currently, no universal model or approach is available that can effectively capture the behavior of forest fires across all regions. This is mainly due to variations in the training data originating from diverse geographical areas [17,18].

Effective management of forest fires encompasses a range of activities, including fuel management, fire detection and communication systems, fire weather forecasting, fire danger, and behavior indices, initial attack systems, identification of fire-sensitive resource areas, standby systems, training and pre-organization of co-operators, suppression resources and capability, and knowledge of fire and its ecological implications [19]. Given the multifaceted nature of forest fire management and the numerous factors involved, more systematic and advanced decision-making techniques are required, such as those based on operational research and decision information systems [20]. This study aims to provide the following primary contributions:

- We adopt CRISP-DM Methodology, analyzing six distinct time stages, enhancing the temporal accuracy of forest fire predictions.
- We Implement feature selection techniques to refine and improve the prediction quality and reduce potential noise from irrelevant data.
- We systematically compare a variety of machine learning models to determine the most effective one for forest fire prediction.
- To simulate the effectiveness of our procedure, we apply the models on a real-world dataset which resulted in superior predictive outcomes, showcasing an enhanced level of accuracy in forecasting when compared to results reported in previous literature.

Our systematic approach provides a robust framework that can be invaluable for researchers aiming to apply machine learning in environmental studies or related fields. In this study, a real-world dataset has been utilized to forecast the occurrence of forest fires with the aid of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. Six distinct time stages were considered and compared, and feature selection was implemented to predict with reduced data. The remaining sections of this paper are organized as follows: Section 2 presents a comprehensive literature review and bibliometric analysis of publications in this field. In Section 3, the methodology and its steps are delineated. Section 4 provides an analysis and comparison of the results. Finally, Section 5 summarizes the research outcomes and identifies potential areas for future research.

2. Literature Review

In this section, first we have conducted a bibliometric analysis of the papers in this research area. Then, 25 related papers were thoroughly chosen systematically and analyzed based on the data used, novelty, methods, and objectives.

2.1. Bibliometric Analysis

This section provides a detailed account of the bibliometric analyses carried out on 2088 research papers obtained from Scopus in RIS and BibTeX formats. The papers were selected based on relevant keywords of the field of study (forest fire prediction). To conduct these analyses, we employed two software tools: VOSviewer [21] and Bibliometrix [22]. These analyses aimed to gain insights into the patterns and trends in the research output of this field. As depicted in Figure 1, the three most commonly used keywords are forest fire, wildfire, and fire, with frequencies of 287, 174, and 100, respectively. The high frequency of machine learning, used 88 times, indicates its significance in the field.

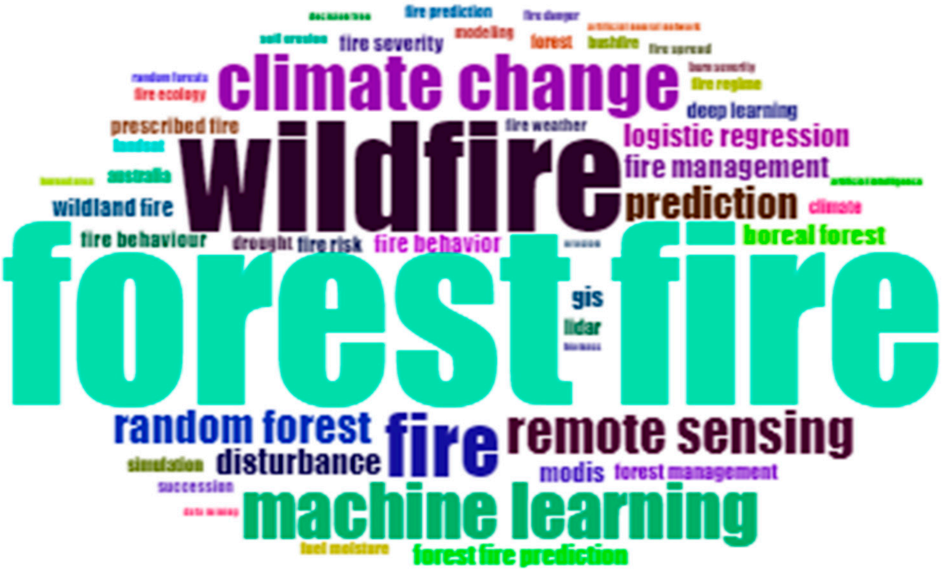


Figure 1. The word cloud of keywords in the area of study.

Figure 2 presents an analysis of the abstracts of papers, highlighting the different clusters of words used in the literature. The field can be divided into four main groups, with this study belonging to the largest group represented by the red nodes. This group is primarily focused on algorithms, prediction accuracy, machine learning, artificial intelligence, and forest fire. The other groups primarily focus on other characteristics such as soil, species, trees, and wind.

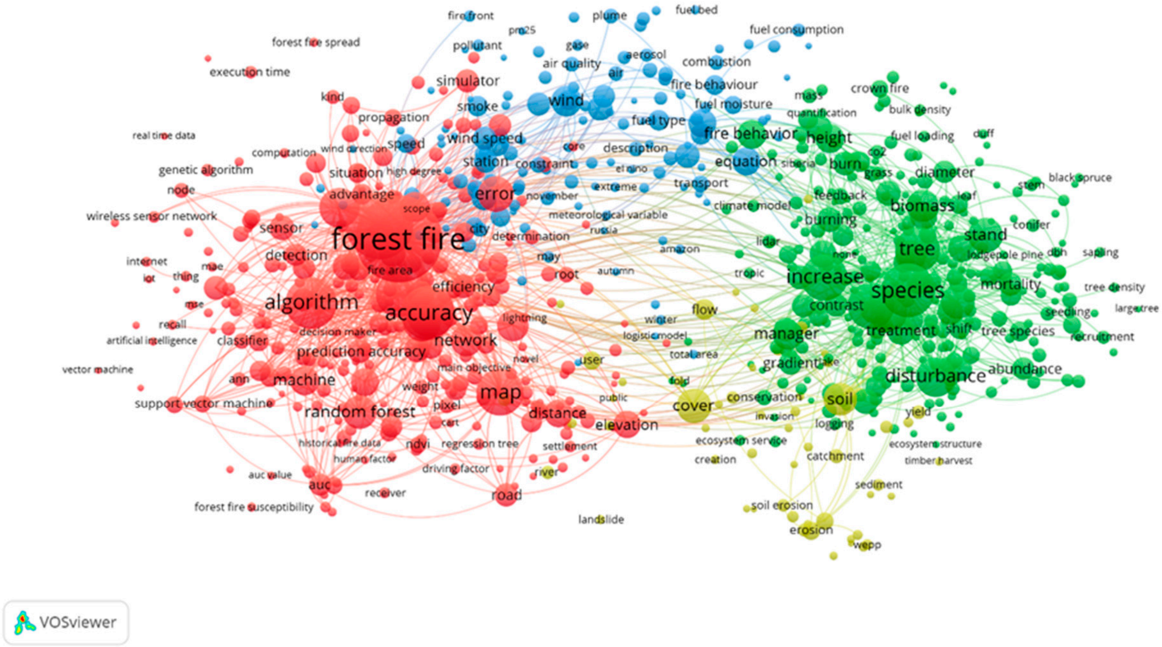


Figure 2. Analysis of words used in the abstracts of the research papers.

2.2. Systematic Review

In this section, 25 papers in the research area are analyzed in detail. The data used in the studies, methods, novelty, and their objectives have been discussed and shown in Table 1.

doi:10.20944/preprints202312.0577.v1

Table 1. A summary of 25 studies in the area of forest fire prediction.

Reference	Year	Data	Novelty/Objective	Method	Results
[23]	2005	Forest fire data from the central Forest Management Service in Athens.	Estimating the forest fire risky areas	Applying an inference mechanism based on fuzzy sets and fuzzy machine learning techniques using a decision support system	Providing successful estimates of areas at risk of forest fire
[24]	2009	NDVI values calculated from MODIS imagery	Predicting forest fire and detecting areas of high risk of forest fire in the Brazilian Amazon	Employing ANN and multitemporal imagery from the MODIS/Terra-Aqua sensors	Achieving an MSE value of around 0.07 in predicting forest fires in high-risk areas
[25]	2010	MESH database, which is formed of catastrophe related videos	Proposing computer vision-based fire detection method for identifying fire in videos	Exploiting Naïve Bayes for extraction of features and classification	average false-positive rate of 0.68% and a false-negative rate of 0.028%
[26]	2011	Weather data provided by the Lebanese Agricultural Research Institute (LARI)	Forest fire occurrence prediction by reducing the number of monitored features.	Artificial neural network (ANN) and support vector machine	outperformance of ANN over SVM by 0.17 on fires and SVM over ANN in the binary classification of fire/no fire scenario
[27]	2012	7,920 forest fire records from 2000 and 2009 provided by the Department of Forestry in Turkey	Estimating the burned areas using historical forest fire records and prediction of the lost area and the corresponding fire size	Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN), Support Vector Machines (SVM), and fuzzy logic	Indicating performance of above 60% in the estimation process. Demonstrating MLP model as the best one using only two inputs (humidity and wind speed) with a more than %65 success rate.
[28]	2013	Extracting 10,000 and 40,000 of a fire and non-fire samples from video images, of tunnels, downtown, and mountain area.	Presenting a fire alarm system based on image processing	Random forest (RF) and Markov chain	Detecting fires precisely, robustly, and with high reliability in public places.
[29]	2013	Forest fire data from Portugal	Exploring the impact of interactions between physical and political systems in forest fire management	System dynamics model	Presenting the unintended consequences of management decision-making when it focuses on fixing rather than

					preventing problems
[30]	2014	Meteorological data of the year 2012 for North Lebanon	Forest fires prediction	Using decision trees and backpropagation forward neural networks	Achieving 98.9% precision using a 4-inputs feed-forward neural network in prediction
[31]	2016	Collecting and creating a dataset with 237 fire images from various online resources	Forest fire occurrence detection using fire images	Exploiting SVM and CNN as classifiers	Achieving accuracy of 90% on global image-level testing using Deep CNN. Demonstrating the accuracy of 92.2% by SVM and 93.1% using CNN
[32]	2017	Using data from temperature, humidity, CO, and smoke sensors	Early fire detection and home monitoring based on fuzzy logic and wireless sensor network	Fuzzy logic in wireless sensor network	Error ratio: 6.67% (test for 30 sample data)
[33]	2018	CCTV surveillance cameras and 68,457 images collected from different fire datasets	Proposing an early fire detection framework using for CCTV surveillance cameras	Convolutional neural networks (CNNs)	Accuracy: 94.39% Precision: 0.82 Recall: 0.98 F-Measure: 0.89
[34]	2019	Forest fire dataset collected from the northeastern region of Portugal	Predicted the burned area of forest fires and the occurrence of large-scale forest fires	Ensemble learning, including random forests (RFs) and extreme gradient boosting (EGB)	Reaching a prediction accuracy of 72.3% by EGB
[35]	2020	57 historical fires and a set of nine spatially explicit explanatory variables	Performance evaluation of machine learning methods for forest fire modeling and prediction	Bayes Network (BN), Naïve Bayes (NB), Decision Tree (DT), and Multivariate Logistic Regression (MLP) machine learning methods	BN model (AUC = 0.96), DT model (AUC = 0.94), NB model (AUC = 0.939), and MLR model (AUC = 0.937)
[36]	2020	Topographical and meteorological data from South Kalimantan Province	Evaluation of machine learning methods for predicting forest fire occurrence in peatland areas.	Support vector machine (SVM), k-Nearest Neighborhood (kNN), Logistic Regression (logreg), Decision Tree (DT), Naïve Bayes (NB), and AdaBoost (DT based)	Accuracy: SVM = 91.8% KNN = 91.8% Logistic Regression = 83.6% Decision Tree (DT) = 90% Naïve Bayes = 86.9% Adaboost (DT Based) = 91.8%
[37]	2021	Publicly available raster data	Decision support for the selection of optimal tower site locations for early-	Single-site solution framework and System-site solution framework	Obtaining layouts by the optimization framework significantly outperforms the

			warning wildfire detection systems		initial layout concerning both covering objectives.
[38]	2021	Various resources	Providing a comprehensive review of the usage of different machine learning algorithms in a forest fire or wildfire management	Summarizing recent trends in the forest fire events prediction, detection, spread rate, and mapping of the burned areas	Identifying some potential areas where new technologies and data can help better fire management decision-making.
[39]	2021	Temporal data of wildfires collected in India	Investigation of wildfire prediction strategies dependent on computerized reasoning.	LSTM network, a time series forecasting Recurrent Neural Network (RNN)	Presenting the ability to predict forest fires with 94.77% accuracy
[40]	2022	Simulated fire spread raster data obtained by FlamMap	Proposing a method suitable for edge computing to use neural networks to predict the spread of forest fires	Backpropagation (BP) neural network	Achieving a computing speed of 5 seconds which is appropriate for edge computing
[41]	2022	Forest Fire Dataset	Predicting forest fire through environment parameters	Logistic regression (LR), support vector machine (SVM), and multiple linear regression	Accuracy: Logistic Regression = 80% SVM = 78% Multiple Linear Regression = 75%
[42]	2022	Forest fire activities and climate data over South Korea	Evaluating the effects of climatic conditions and drought phase on occurrence frequency (OF) of forest fire	Deep Belief Network	Model using only relative humidity (RH): R2 = 0.819 Model using a combination of RH and wind speed (WS): NSE = 0.828
[43]	2022	Forest fire management policy, forest fire control practices, and their constraints, and also FMU's capacity	Analyzing the performance of forest fire-related policy implementation based on Forest Management Units (FMUs)	Measuring the performance of the FMUs by the achievement of the policy objectives and effectiveness of policy implementation	Showing clarity of the policies, standards, and objectives to manage fire for FMUs, and challenges in their implementation, such as limited capacity and resources
[44]	2023	Dataset for Forest Fire Detection from Mendeley Data	Proposing a forest fire detection method based on a Convolutional Neural Network (CNN) architecture	Convolutional Neural Network (CNN) with separable convolution layers	Identifying forest fires within images with a 97.63% accuracy, 98.00% F1 Score, and 80% Kappa

[45]	2023	Nine years of data were gathered across Australia	Predicting the wildfire event probability based on a set of environmental predictors and forest vulnerability	Bayesian multiple logistic regression	Predicting a low probability for wildfire events during winter and autumn (< 6%), 31.5% during an average summer and 64.6% during extreme summer conditions
[46]	2023	Scopus, EBSCO, and SCIELO databases	Analyzing the interaction between both terms to identify what is known about the topic, the existence of previous studies	Searching protocol model with three phases: planning, execution, and results	Governance is inherent to forest fire management
[47]	2023	Using the temperature (TOA and Ground) and intensity values	Presenting a method for the quantitative evaluation of the efficiency of fire safety management in universities	Data envelopment analysis (DEA)	Proposing accurate method in detecting the forest fire

There have been several studies regarding forest fire forecasting in various aspects. Decision support systems [23] have been used to forecast forest fire in an earlier study. The decision support system applies an inference mechanism based on various aspects of fuzzy sets and fuzzy machine learning techniques to forest fire data to estimate forest fire risky areas. The author of this article used forest fire data from the Central Forest Management Service in Athens. The system successfully estimated forest fire risk areas. Several studies have also been conducted in this context using machine learning techniques. For example, in 2009, Maeda et al. [24] presented a method capable of detecting areas with high forest fire risk in the Brazilian portion of the Amazon. This method achieved a mean squared error of around 0.07. They used NDVI values calculated from MODIS imagery acquired during five periods preceding the 2005 fire season. These values were used to train and validate an artificial neural network (ANN) as a prediction model. In 2010, Borges et al. [25], proposed a computer vision-based fire detection method for identifying fire in videos by analyzing the frame-to-frame changes of specific low-level features describing potential fire regions. The behavioral change of each feature was evaluated by employing a Bayes classifier for robust fire recognition. The experiments in this study illustrated the method's functionality, with an average false-positive rate of 0.68% and a false-negative rate of 0.028%. In another study [26] in 2011, Sakr et al. tried to predict forest fire occurrence by reducing the number of monitored features and eliminating the need for weather prediction mechanisms. They used and evaluated artificial neural networks (ANN) and support vector machine (SVM) as fire danger prediction algorithms. These algorithms utilized relative humidity and cumulative precipitation to output a risk estimate. The weather data in this work were provided by the Lebanese Agricultural Research Institute (LARI). Özbayoğlu and Bozer conducted a study [27] in 2012 where the burned areas resulting from possible forest fires were estimated by Multilayer Perceptron (MLP), Radial Basis Function Networks (RBFN), Support Vector Machines (SVM), and fuzzy logic methods. In their study, they considered parameters such as geographical conditions of the existing environment, date and time when the fire broke out, meteorological data such as temperature, humidity and wind speed as well as the type and number of trees in one unit area. The data was collected from the Department of Forestry in Turkey and contained 7,920 forest fire records from 2000 and 2009. The results indicate that in some models, the estimation performances were above 60%, and the most accurate model turned out to be an MLP model involving only two inputs (humidity and wind speed) with more than 65% success rate.

Studies in forest fire prediction employing machine learning methods are still ongoing. In 2013, Onecue et al. conducted a study [28] regarding developing a fire alarm system based on image processing. This system employs an algorithm that classifies patches into fire or non-fire areas by using the random forest technique and introducing additive Markov chain to reduce false alarms. The algorithm detected fires precisely with a high reliability score in public places. Furthermore, it successfully minimized casualties and property damages through early fire detection. In 2013, one of the limited works in forest fire management was done by Collins et al. [29]. They explored how interactions between physical and political systems in forest fire management impact the effectiveness of different allocations using a System Dynamics (SD) model through a case study of Portugal. The results of this study showed that a balanced approach to suppression and prevention efforts could mitigate the self-reinforcing consequences of exacerbating fuel loads which leads to greater fires. Karouni et al. [30] evaluated the performance of decision trees and backpropagation forward neural networks in forest fire forecasting utilizing four meteorological attributes: temperature, relative humidity, wind speed, and daily precipitation. The results showed the effectiveness of the two techniques by using only the two most significant attributes (temperature and relative humidity). Another study related to the application of machine learning was conducted by Zhang et al. [31] where they implemented a convolutional neural network (CNN) and support vector machine (SVM) for forest fire occurrence detection based on fire images. They introduced and used a dataset containing 237 images gathered from online resources. Their proposed approach achieved an accuracy of 90% on global image-level testing through Deep CNN and demonstrated 92.2% and 93.1% accuracy through SVM and CNN, respectively, in determining the precise location of fire patches. In yet another study [32], Saputra et al. outlined an early fire detection and home monitoring system based on fuzzy logic and a wireless sensor network. This system used four sensors: temperature, air humidity, carbon monoxide, and smoke sensors. A sample size of 30 data points was used to test the performance of the proposed method, and the results showed an accuracy of 6.67% for test samples.

In recent years, numerous efforts have been made to increase the performance of forest fire forecasting systems. For instance, Muhammad et al. [33] presented an early fire detection framework using fine-tuned convolutional neural networks for CCTV surveillance cameras, which can detect fire in varying indoor and outdoor environments. They used 68,457 images collected from different fire datasets, such as Foggia's video and Chino's dataset, to train and validate their model. The proposed work reported further improvement of results from these datasets by increasing accuracy from 93.55% to 94.39%. Xie et al. [34] proposed two ensemble learning methods (Extreme Gradient Boosting and Random Forests) to predict the burned area of forest fires and the occurrence of large-scale forest fires using the forest fire dataset from the northeastern region of Portugal. In terms of the accuracy of predicting the burned area, the tuned random forest approach performed better than other regression models. Extreme gradient boosting outperformed other classification models in large-scale fire prediction. As discussed in another article, Pham et al. [35] evaluated machine learning methods for modeling and predicting forest fires. This study evaluates the abilities of Bayes Network (BN), Naïve Bayes (NB), Decision Tree (DT), and Multivariate Logistic Regression (MLP) machine learning methods for the prediction and mapping of fire susceptibility across the Pu Mat National Park, Vietnam. They utilized information from 57 historical fires and a set of 9 spatially explicit explanatory variables. In this study, the BN model with an AUC value of 0.96 showed to be more accurate than the other models in predicting future fires. A similar study [36] examined the application of various machine learning approaches to predict forest fire occurrence in peatlands. The variables collected from topographical and meteorological data from South Kalimantan Province were the time (of data collected), the district area, land surface temperature (LST), wind speed, humidity, height, and NDVI (normalized vegetation index). In comparison to other techniques, the KNN algorithm performed better in this study. One of the few studies conducted in the area of forest fire prediction through the use of decision support systems was conducted by Heyns et al. [37]. In this work, a decision support system was introduced for selecting multiple tower sites from a large number of potential site locations to maximize system visibility of smoke above a prescribed region which can be effective in early warning wildfire detection systems. Toward this, the authors presented two frameworks: a single-site solution and a multi-site solution. The results showed that the layouts obtained by the optimization framework were found to significantly outperform the initial layout concerning both the covering objectives.

Forest fire management systems have progressed in recent years. To introduce some limited efforts, Arif et al. [38] conducted a study that summarized recent trends in forest fire prediction, detection, spread rate, and mapping. This paper provides a comprehensive review of the usage of different machine learning algorithms in forest fire or wildfire management. In 2022, Natekar et al. [39] investigated wildfire prediction strategies dependent on computerized reasoning. They fed temporal data of wildfires collected in India to an LSTM network to predict fire propagation. Their proposed model could predict forest fires with 94.77% accuracy. Another article published by Li et al. [40] presented a method that used the BP neural network to predict the spread index of the raster map, solve the plane geometry, and obtain a better fire line fit. They trained the BP neural network model on the simulated fire spread raster data obtained by FlamMap and predicted fire spread direction and speed. The presented results in this paper show that the proposed model is faster than the algorithms provided by FlamMap, which is suitable for edge computing. Similarly, another study [41] regarding forest fire prediction is focused on predicting forest fire through environmental parameters such as oxygen, humidity, and temperature. The authors of this paper have used logistic regression (LR), support vector machine (SVM), and multiple linear regression models in their work. With 80% accuracy, the logistic regression model outperformed other models in this study. Furthermore, Sung et al. [42] evaluated the effects of climatic conditions and drought phases on forest fire occurrence frequency (OF) in South Korea. The authors proposed a model that uses deep learning to estimate the occurrence frequency of fires. Their RH-WS-AMOF model demonstrated the most impressive performance with an R-Squared value of 0.838 and an MSE score of 0.828.

Budiningsih et al. [43] investigated the performance of forest fire-related policy implementation based on five Forest Management Units (FMUs) in the fire-prone regions of Central Kalimantan and South Sumatra, Indonesia. Analysis of this work showed that the policies, standards, and objectives to manage fire are clear for FMUs. However, implementation challenges for this still exist, and as a result fire control activities have not been fully undertaken. Most FMUs have limited capacity and resources, complicated budget mechanisms, and low community participation. Strengthening FMU capacity will significantly improve forest fire control performance. In a recent study [44], Rahman et al. proposed a method for detecting forest fires using a deep convolutional neural network based on the dataset for Forest Fire Detection from Mendeley Data. Experimental results of this research showed that the method could identify forest fires within images with an accuracy of 97.63% and a F1 score of 98%. Charizanos et al. [45] developed a Bayesian model to predict wildfire event probability based on a set of environmental predictors and forest vulnerability represented by the normalized difference in vegetation index. According to the modeling results, forest vulnerability is the most significant predictor of wildfire probability. Bayesian hierarchical logistic regression modeling predicted a low probability for wildfire events during winter and autumn (< 6%), 31.5% during an average summer, and 64.6% during extreme summer conditions. In a recently published work by Holgado-Vargas [46], wildfire risk management and territorial governance were analyzed through a systematic review to evaluate the interaction between both terms to identify known information about the topic, the existence of previous studies and the unknown aspects that exist to date. The conclusion was that governance is inherent to forest fire management since the most common cause is anthropogenic. Hence the importance of all stakeholders, especially communities participating in the decision-making process regarding forest fire solutions. Anandaram et al., in a study [47], utilized a data envelopment analysis (DEA) technique to present a method for the quantitative evaluation of fire safety management efficiency in universities. This proposed technique was used to detect the forest fire based on the temperature (TOA and Ground) and intensity values.

In this study, it was shown that the proposed method of detecting forest fires is accurate and applicable.

3. Methodology

The choice of the right method is one of the most important steps in scientific research and plays a crucial role in the success of the research. The current research employed the CRISP-DM methodology, which stands for Cross-Industry Standard Process for Data Mining. This widely-used process model consists of six iterative phases that are not specific to any particular industry and provide guidance for the data mining process. The phases begin with understanding the business context and end with deploying the results [48,49]. In the following section, each phase will be comprehensively described, along with its application in the present study.

3.1. CRISP-DM Methodology

Research conducted by the Data Mining Research Association in 2007 suggests this method is the most widely used method for conducting data mining projects. It was introduced in 1997 and is a non-proprietary, documented, and free standard. This method does not rely on purely automatic theories but rather on the scientific experiences of data mining users, manufacturers, and service providers of data mining tools. Starting with an understanding of the core needs of the business, and ending with providing solutions to meet those needs, is the basis of this six-step process. The steps of this process follow one another in theory, but in practice, there are many reciprocal steps between them [48,49].

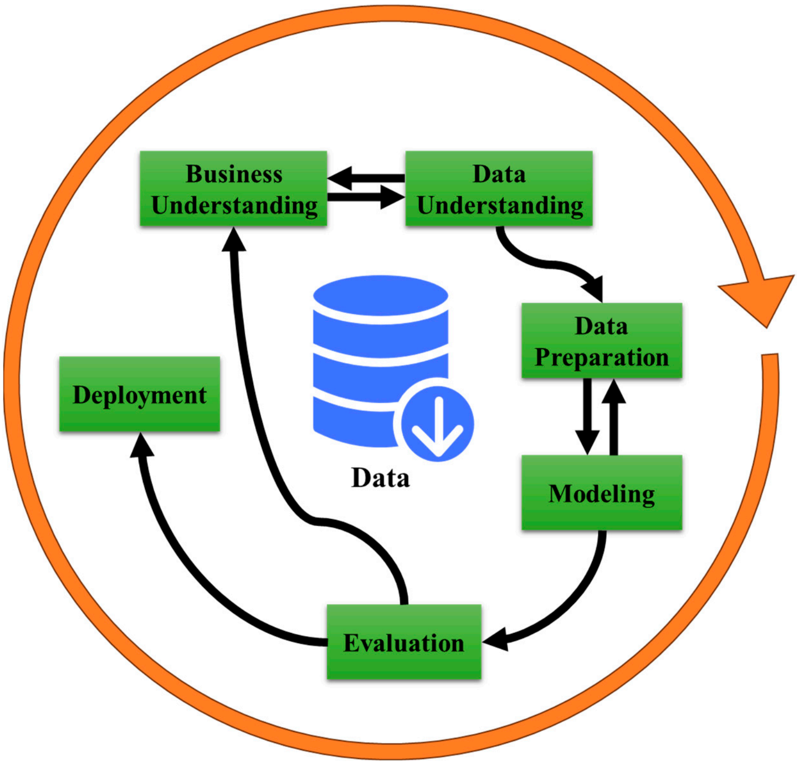


Figure 3. The connection between the phases of CRISP-DM.

3.2. Data Understanding

According to the needs of the business, a set of data that can be used to achieve the project goal is identified. The present study utilized a dataset consisting of data from two distinct regions in Algeria, namely Sidi Belabbas [50] in the northwest and Bejaia in the northeast. The dataset contains 244 instances, with each region contributing 122 instances. Meteorological observations from 2012 (June to September) were used as this is the period with the highest occurrence of forest fires, and

2012 had the highest recorded fire occurrence between 2007 and 2018. The dataset includes crucial weather elements that influence the occurrence of wildfires, including temperature, relative humidity, and wind speed. The attribute of rain was not used as it was deemed irrelevant to the decision tree model. Numerical attribute values were utilized to predict two possibilities: "fire" and "not fire." The forest fires dataset instances were first classified into "fire" and "not fire" classes using the two components of the FWI system, namely the FWI and the FFMC. The FFMC is an indicator of ignition probability and is calculated using meteorological elements such as temperature, relative humidity, wind speed, and rain through the Fire Weather Index Calculator [51]. The dataset comprises 138 instances of "fire" and 106 instances of "not fire," and its detailed description is provided in Table 2.

Table 2. Data description.

Attribute	Description
Region	The region where the data was collected: Sidi Belabbas or Bejaia
Date	The date of the meteorological observation
Temperature	The temperature in Celsius
RH	The relative humidity as a percentage
Wind	The wind speed in km/h
Rain	The amount of rainfall in mm
FFMC	The Fine Fuel Moisture Code, an indicator of the ignition probability
DMC	The Duff Moisture Code, an indicator of the fuel consumption potential
DC	The Drought Code, an indicator of the fuel moisture content
ISI	The Initial Spread Index, an indicator of the potential rate of spread of a fire
BUI	The Buildup Index, an indicator of the total amount of fuel available for combustion
FWI	The Fire Weather Index, a numerical rating of the potential fire intensity

3..3. Data Preparation

In this step, data mining methods are used to prepare the data for analysis. Most of the analysis time is spent during this step. It is possible for the data preparation phase of a project to consume as much as 80% of the total project time. In this phase, the analyst performs actions such as mixing, clearing, converting, and reducing data [48]. In this study, missing data was not encountered. The target variable consisted of "fire" and "not fire" labels, which were converted to 0 and 1, respectively. From ‘Date’ column, the month has been extracted. Additionally, exploratory data analysis was performed to assist with the modeling process. Figure 4 provides an overview of the interrelationships between various features. The analysis reveals that the target feature has a significant positive correlation with FFMC (0.77), while it exhibits the weakest correlation with the month feature (0.024). Furthermore, the target feature shows a negative correlation with RH, wind speed, and precipitation. These findings hold valuable implications for further research and modeling efforts.

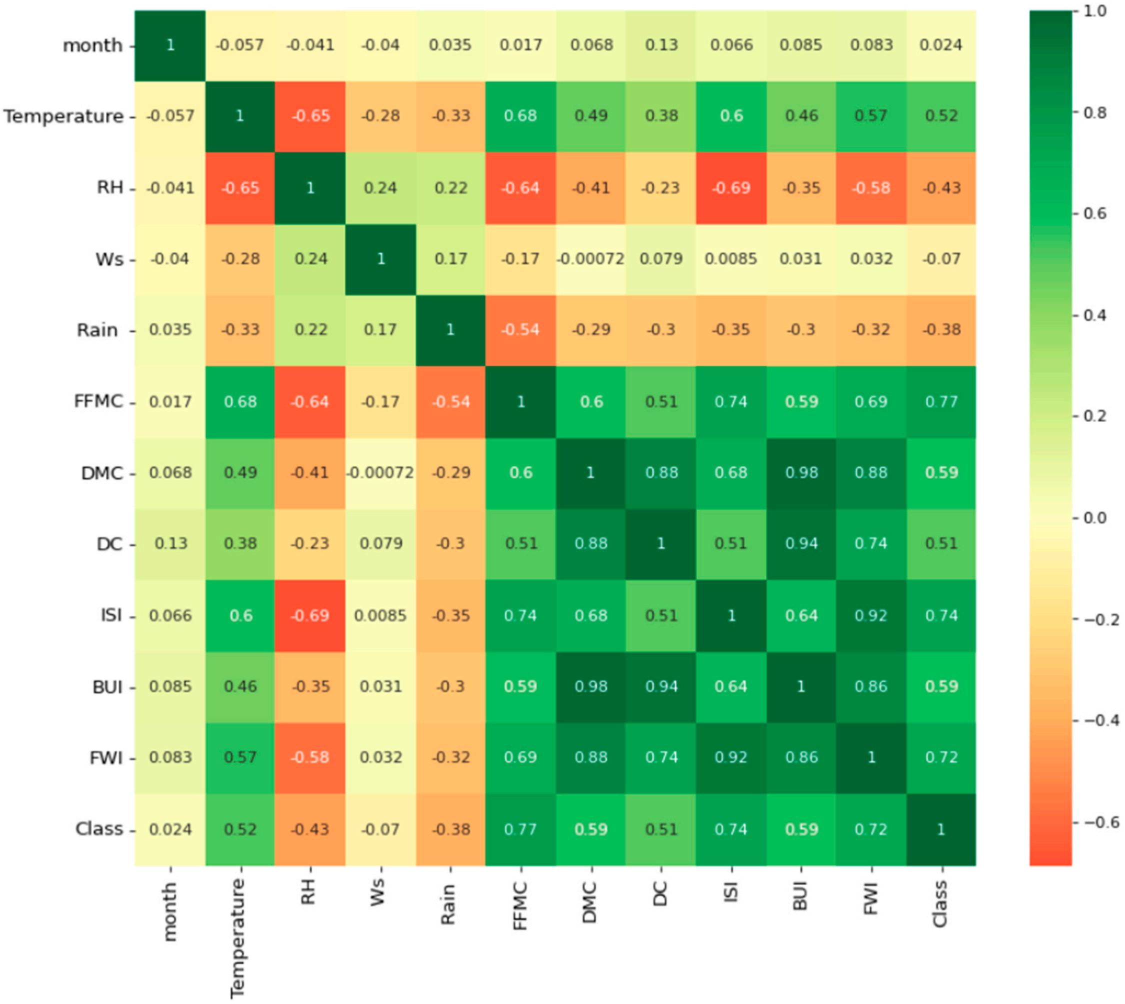


Figure 4. The correlation between features.

As illustrated in Figure 5, it is reasonable to observe that higher rainfall is associated with a lower fire incidence. However, the data also indicates that certain cases of fire have occurred despite rainfall levels of approximately 5 mm. These observations suggest that additional factors beyond rainfall may contribute to fire occurrence in certain scenarios, which warrants further investigation.

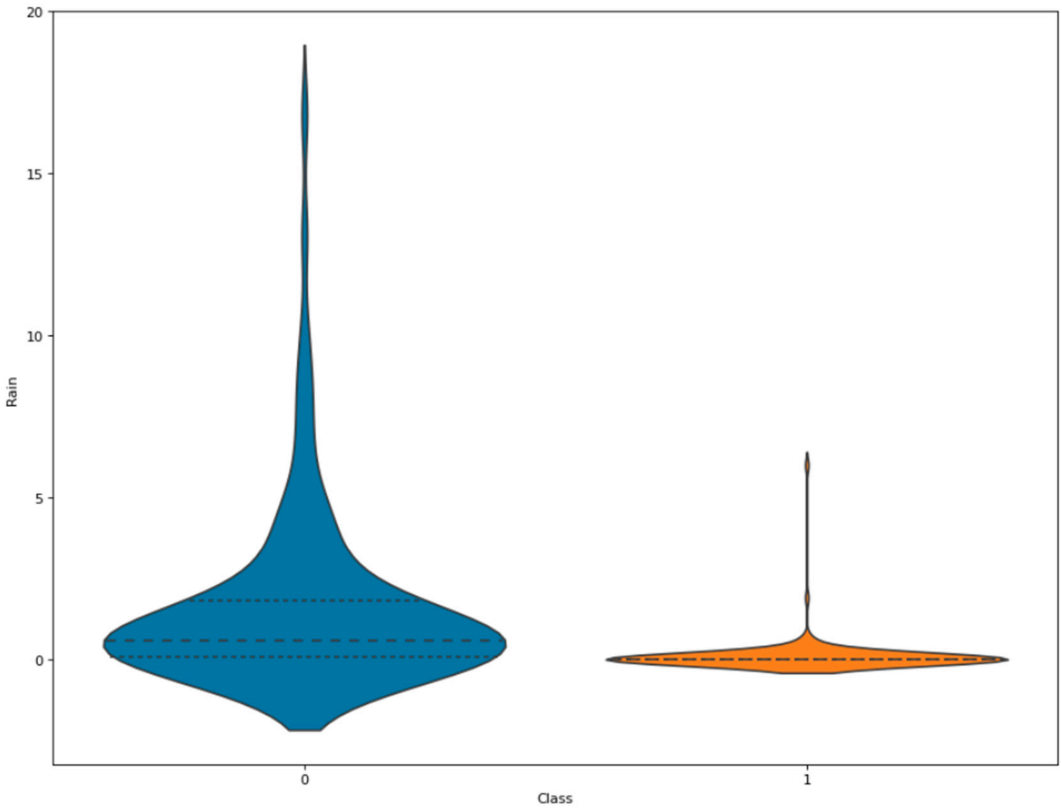


Figure 5. The amount of rainfall for the two classes.

Figure 6 displays the Fine Fuel Moisture Code (FWI) distribution plot for the two distinct classes. The graph indicates that when there is no fire, the FWI index is centered around 0. Conversely, the FWI index is distributed over a broader range when a fire occurs. Notably, the maximum value of the FWI index during a fire event is below 10, but it can reach up to approximately four times higher than the maximum value when there is no fire. These results provide valuable insights into the relationship between the FWI index and fire occurrence, which may inform fire prevention and mitigation efforts. To gain a deeper understanding of the FWI feature, Figure 7 depicts the box plot of the index for various months. The results indicate that, for both classes, the FWI index is substantially higher in August compared to the other months. Conversely, the FWI index is lowest in September. These findings suggest that August may be a critical period for fire prevention and management, while September may present a lower fire incidence risk.

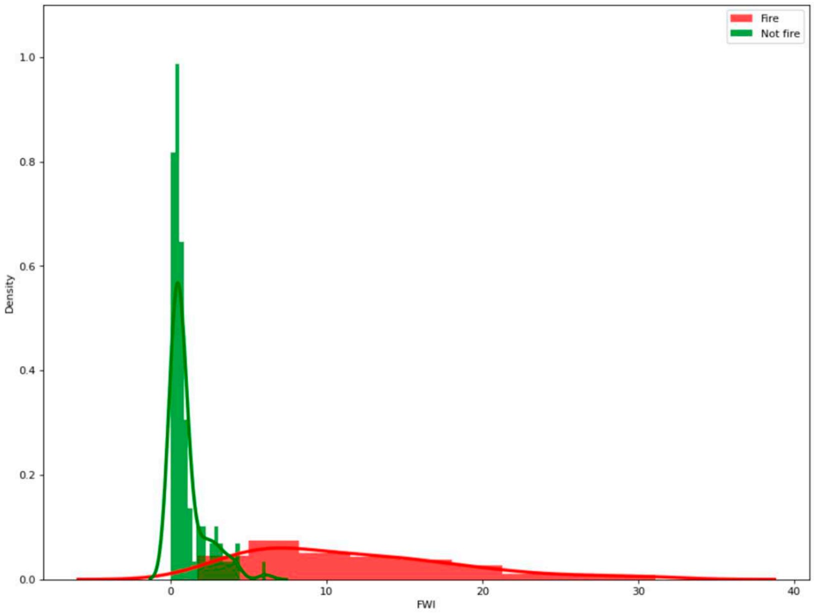


Figure 6. The distribution plot of FWI for the two classes.

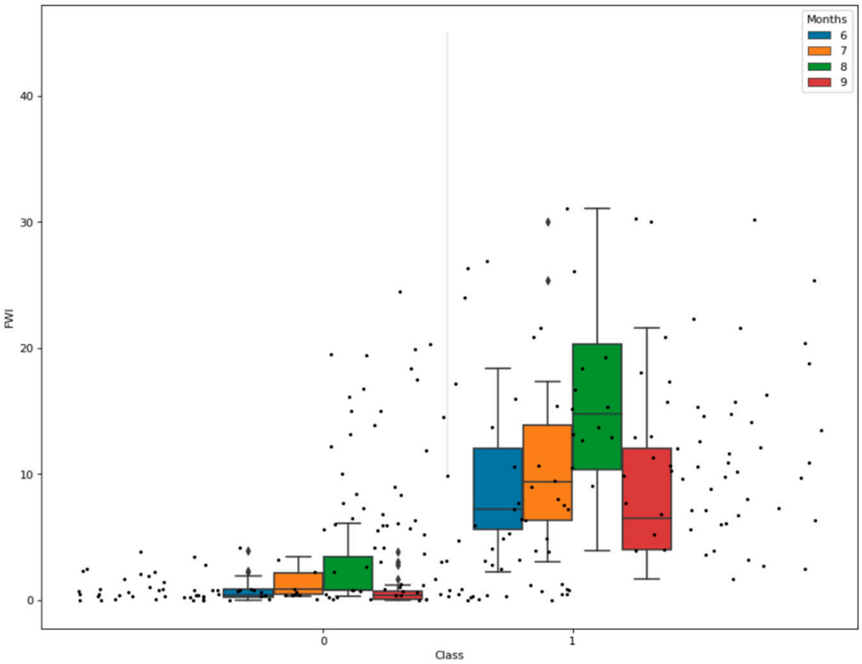


Figure 7. The box plot for the FWI index.

3.4. Modeling

In this stage of the Crisp method, different data mining methods can be applied to the prepared data to achieve the project's main goal and its intended result. The modeling process is not linear. It is a roundtrip trial-and-error process because there is no optimal model in data mining, and depending on the problem, different methods should be tested and their outputs compared, and sometimes it is necessary to go back to the previous step and prepare some data algorithms in a different way in order to reach the desired outcome [48]. Regarding the modeling process, we designated September data as the test dataset, while defining six distinct time stages based on the remaining months. Figure 8 illustrates the distribution of data across the months, indicating that the data is nearly evenly distributed among the months. This balanced distribution of data is crucial for ensuring the reliability and generalizability of the modeling outcomes.

doi:10.20944/preprints202312.0577.v1

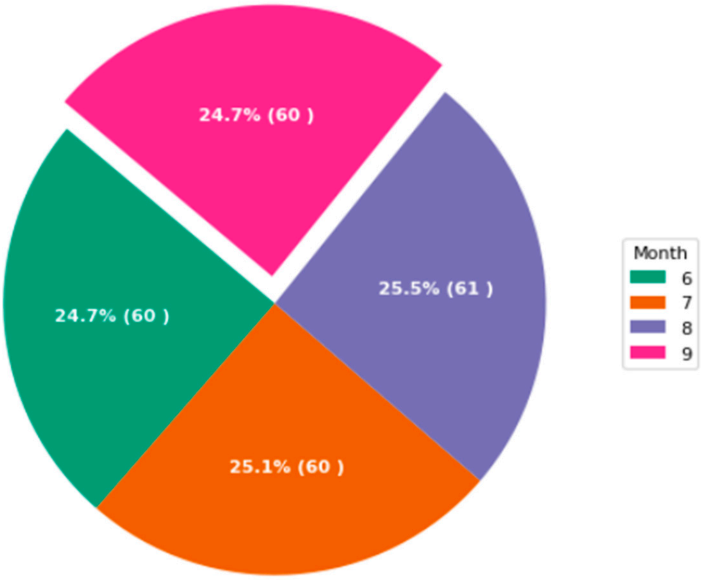


Figure 8. The distribution of the data across the months.

4. Result

In this section, at first, feature selection has been conducted to get the best features to use as inputs in the modeling. Then, different time stages are compared and the best one is identified. Finally, models in the best time stage are analyzed.

4.1. Feature Selection

The amount of high-dimensional data that exists and is publically available on the internet has greatly increased in the past few years. Therefore, machine learning methods have difficulty in dealing with the large number of input features, which is posing an interesting challenge for researchers. In order to use machine learning methods effectively, preprocessing of the data is essential. Feature selection is one of the most frequent and important techniques in data preprocessing and has become an indispensable component of the machine learning process [52,53]. The focus of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results. Feature selection methods provides us a way of reducing computation time, improving prediction performance, and a better understanding of the data in machine learning or pattern recognition applications [54].

We employed the correlation-based feature selection (CFS) method, which is a filter approach and, therefore, independent of the final classification model. The central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other [55]. It evaluates feature subsets only based on data intrinsic properties.

As discussed, we calculate the correlation value between each feature and class in each time stage. The correlation heatmap of each time stage is illustrated in Figure 1. Then, the features with more than a threshold value will be selected as a desired feature subset. We set the threshold value to 0.35 in this study. The selected features for each time stage are specified in Table 1.

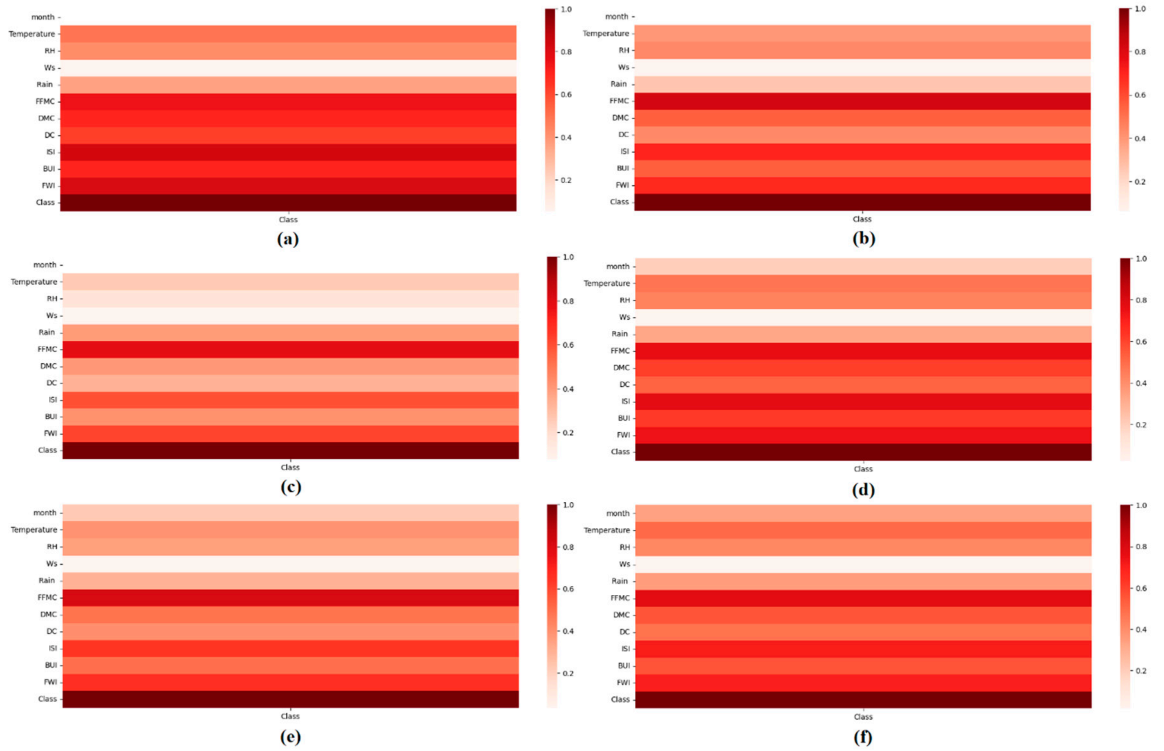


Figure 9. Correlation heatmap of each time stage: (a) time stage one, (b) time stage two, (c) time stage three, (d) time stage four, (e) time stage five, and (f) time stage six.

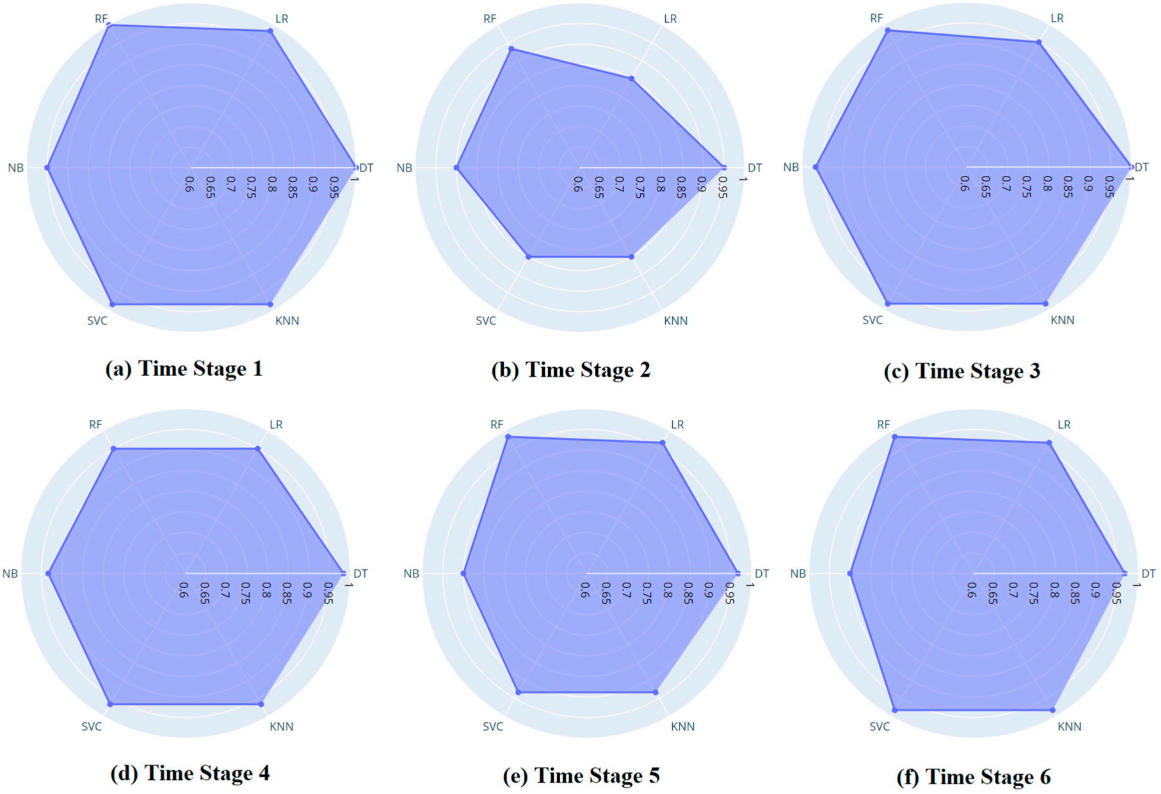
According to Table 3, a considerable number of features like FFMC, DMC, ISI, BUI, and FWI were selected for each time stage which can be representative of the high correlation of these features. Some features such as temperature, RH, and DC have been selected for all time stages except stage three. Among the features, Rain only was selected for time stages one and three. As illustrated in Table 1, WS (Wind Speed) has not been chosen for any time stage since this feature had a high correlation with other features, and including this feature will lead to redundancy.

Table 3. Selected features for each time stage.

	Time Stage one	Time Stage two	Time Stage three	Time Stage four	Time Stage five	Time Stage six
Month						
Temperature	☒	☒		☒	☒	☒
RH	☒	☒		☒	☒	☒
WS						
Rain	☒		☒			
FFMC	☒	☒	☒	☒	☒	☒
DMC	☒	☒	☒	☒	☒	☒
DC	☒	☒		☒	☒	☒
ISI	☒	☒	☒	☒	☒	☒
BUI	☒	☒	☒	☒	☒	☒
FWI	☒	☒	☒	☒	☒	☒

4.2. Best Time stage

We trained six machine learning models, Decision Tree (DT), Logistic Regression, Random Forest (RF), Gaussian Naïve Bays (Gaussian NB), Support Vector Classification (SVC), and K-Nearest Neighbors (KNN) employing selected features for six time stages. Accuracy of all models in each time stage has been depicted in Figure 10 through a spider chart. Based on Figure 10, time stage one demonstrates the best result, with an accuracy of (1) for RF and DT models, 0.98 in SVC, LR, and KNN, and 0.95 in NB.



(a) Time Stage 1

(b) Time Stage 2

(c) Time Stage 3

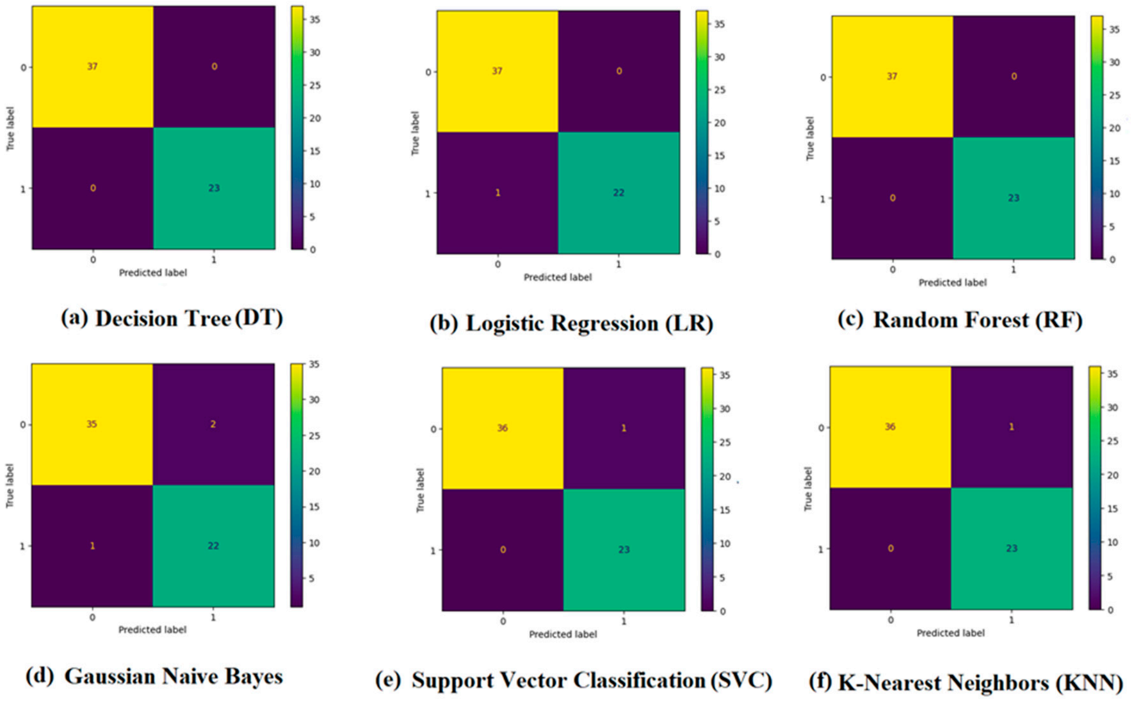
(d) Time Stage 4

(e) Time Stage 5

(f) Time Stage 6

Figure 10. Accuracy of all machine learning models in six-time stages, where time stage one is the best one.

As discussed in the previous section, time stage one was the best time stage according to Figure 2 and the accuracy of all six models in this time stage. In this section, we will evaluate the performance of each model in terms of accuracy, precision, recall and F2-score. The confusion matrix of each model is illustrated in Figure 11.



(a) Decision Tree (DT)

(b) Logistic Regression (LR)

(c) Random Forest (RF)

(d) Gaussian Naive Bayes

(e) Support Vector Classification (SVC)

(f) K-Nearest Neighbors (KNN)

Figure 11. Confusion matrix of all machine learning models.

Based on Figure 11, Decision Tree and Random Forest models have the best performance compared to other models regarding mentioned criteria. We have compared our result with Faroudja ABID et al [50] who worked with the same dataset. The result shows an improvement of 0.05 in recall, around 0.16 in accuracy, and around 0.20 in precision compared to their best model.

5. Conclusion

We conducted an in-depth study of forest fire prediction using machine learning models by applying a rigorous method to understand the nuances of the data and their implications. Using the the correlation-based feature selection (CFS) method, we identified features that were highly relevant for prediction and eliminated redundant or less informative ones. The Fine Fuel Moisture Code, Duff Moisture Code, Initial Spread Index, Buildup Index, and Fire Weather Index were found to be highly correlated across various time stages, indicating the importance of these features in predicting the future fire hazards. The study of six machine learning models across six distinct time stages indicated that time stage one was the most promising. A perfect accuracy, precision, recall, and F2-score were achieved by the Decision Tree and Random Forest models during this time period. As a result of the dominance of these two models over others, such as SVC, LR, and KNN, their suitability for this particular type of prediction task is evident. Furthermore, our approach and results demonstrated significant improvements over previous work on the same dataset, emphasizing the effectiveness of our method. Forest fire prediction is highlighted by this study as a critical issue that requires systematic data preprocessing and model evaluation. This research provides valuable insights and tools for future efforts to analyze and predict forest fires, which remain a critical concern for both environmental and human safety.

References

1. Cheng, T.; Wang, J. Integrated spatio-temporal data mining for forest fire prediction. Trans. GIS 2008, 12(5), 591-611.
2. Mitri, G.; Gitas, I. A semi-automated object-oriented model for burned area mapping in the Mediterranean region using Landsat-TM imagery. Int. J. Wildland Fire 2004, 13, 367–376.
3. Amil, M. Forest fires in Galicia (Spain): Threats and challenges for the future. J. For. Econ. 2007, 13, 1–5.

4. Bianchinia, G.; Denhama, M.; Cortésa, A.; Margalefa, T.; Luquea, E. Improving forest-fire prediction by applying a statistical approach. For. Ecol. Manag. 2006, 234(S1), 210.
5. Chen, F.; Du, Y.; Niu, S.; Zhao, J. Modeling Forest lightning fire occurrence in the Daxinganling Mountains of Northeastern China with MAXENT. Forests 2015, 6(5), 1422-1438.
6. Bhusal, S.; Mandal, R. Forest fire occurrence, distribution and future risks in Arghakhanchi district, Nepal. J. Geogr. 2020, 2, 10–20. Available online: https://www.researchgate.net/publication/341701669 (accessed on 13 January 2021).
7. Avilaflores, D.Y.; Pompagarcia, M.; Antonionemiga, X.; Rodrigueztrejo, D.A.; Vargasperez, E.; Santillan-Perez, J. Driving factors for forest fire occurrence in Durango State of Mexico: A geospatial perspective. Chin. Geogr. Sci. 2010, 20, 491–497.
8. 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 Sens. 2022, 14, 5546. Available online: https://doi.org/10.3390/rs14215546 (accessed on Day Month Year).
9. Singh, K.R.; Neethu, K.P.; Madhurekaa, K.; Harita, A.; Mohan, P. Parallel SVM model for forest fire prediction. Soft Comput. Lett. 2021, 3, 100014.
10. Ge, X.; Yang, Y.; Peng, L.; Chen, L.; Li, W.; Zhang, W.; Chen, J. Spatio-Temporal Knowledge Graph Based Forest Fire Prediction with Multi Source Heterogeneous Data. Remote Sens. 2022, 14, 3496. Available online: https://doi.org/10.3390/rs14143496
11. Jaafari, A.; Zenner, E.K.; Pham, B.T. Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree-based classifiers. Ecol. Inform. 2018, 43, 200–211.
12. Gholamnia, K.; Gudiyangada Nachappa, T.; Ghorbanzadeh, O.; Blaschke, T. Comparisons of Diverse Machine Learning Approaches for Wildfire Susceptibility Mapping. Symmetry 2020, 12, 604.
13. Thach, N.N.; Ngo, D.B.-T.; Xuan-Canh, P.; Hong-Thi, N.; Thi, B.H.; Nhat-Duc, H.; Dieu, T.B. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. Ecol. Inform. 2018, 46, 74–85.
14. Goldarag, Y.J.; Mohammadzadeh, A.; Ardakani, A. Fire risk assessment using neural network and logistic regression. J. Indian Soc. Remote Sens. 2016, 44, 885–894.
15. Jaafari, A.; Razavi Termeh, S.V.; Bui, D.T. Genetic and firefly metaheuristic algorithms for an optimized neuro-fuzzy prediction modeling of wildfire probability. J. Environ. Manag. 2019, 243, 358–369.
16. Moayedi, H.; Mehrabi, M.; Bui, D.T.; Pradhan, B.; Foong, L.K. Fuzzy-metaheuristic ensembles for spatial assessment of forest fire susceptibility. J. Environ. Manag. 2020, 260, 109867.
17. Tehrany, M.S.; Jones, S.; Shabani, F.; Martínez-Álvarez, F.; Tien Bui, D. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. Theor. Appl. Climatol. 2019, 137, 637–653.
18. Pham, B.T.; Jaafari, A.; Avand, M.; Al-Ansari, N.; Dinh Du, T.; Yen, H.P.H.; Phong, T.V.; Nguyen, D.H.; Le, H.V.; Mafi-Gholami, D.; Prakash, I.; Thi Thuy, H.; Tuyen, T.T. Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction. Symmetry 2020, 12, 1022. Available online: https://doi.org/10.3390/sym12061022.
19. Bonazountas, M.; Kallidromitou, D.; Kassomenos, P.; Passas, N. A decision support system for managing forest fire casualties. J. Environ. Manag. 2007, 84(4), 412-418.
20. Vakalis, D.; Sarimveis, H.; Kiranoudis, C.; Alexandridis, A.; Bafas, G. A GIS based operational system for wildland fire crisis management I. Mathematical modelling and simulation. Appl. Math. Model. 2004, 28(4), 389-410.
21. Van Eck, N. J.; Waltman, L. Software survey: VOSviewer, a computer program for Bibliometric mapping. Scientometrics 2009, 84(2), 523–538. Available online: https://doi.org/10.1007/s11192-009-0146-3.
22. Aria, M.; Cuccurullo, C. Bibliometrix: An R-tool for comprehensive science mapping analysis. J. Informetrics 2017, 11(4), 959–975. Available online: https://doi.org/10.1016/j.joi.2017.08.007 (accessed on Day Month Year).
23. Iliadis, L.S. A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation. Environ. Model. Softw. 2005, 20(5), 613-621.
24. Maeda, E.E.; Formaggio, A.R.; Shimabukuro, Y.E.; Arcoverde, G.F.B.; Hansen, M.C. Predicting Forest fire in the Brazilian Amazon using MODIS imagery and artificial neural networks. Int. J. Appl. Earth Obs. Geoinform. 2009, 11(4), 265-272.
25. Borges, P.V.K.; Izquierdo, E. A probabilistic approach for vision-based fire detection in videos. IEEE Trans. Circuits Syst. Video Technol. 2010, 20(5), 721-731.
26. Sakr, G.E.; Elhajj, I.H.; Mitri, G. Efficient Forest fire occurrence prediction for developing countries using two weather parameters. Eng. Appl. Artif. Intell. 2011, 24(5), 888-894.
27. Özbayoğlu, A.M.; Bozer, R. Estimation of the burned area in forest fires using computational intelligence techniques. Procedia Comput. Sci. 2012, 12, 282-287.
28. Kim, O.; Kang, D.J. Fire detection system using random forest classification for image sequences of complex background. Opt. Eng. 2013, 52(6), 067202.

29. Collins, R.D.; de Neufville, R.; Claro, J.; Oliveira, T.; Pacheco, A.P. Forest fire management to avoid unintended consequences: A case study of Portugal using system dynamics. J. Environ. Manag. 2013, 130, 1-9.
30. Karouni, A.; Daya, B.; Chauvet, P. Applying decision tree algorithm and neural networks to predict forest fires in Lebanon. J. Theor. Appl. Inf. Technol. 2014, 63, 282-291.
31. Zhang, Q.; Xu, J.; Xu, L.; Guo, H. Deep convolutional neural networks for forest fire detection. In Proceedings of the 2016 International Forum on Management, Education and Information Technology Application, Atlantis Press, 2016; pp. 568-575.
32. Saputra, F.A.; Al Rasyid, M.U.H.; Abiantoro, B.A. Prototype of early fire detection system for home monitoring based on Wireless Sensor Network. In 2017 International Electronics Symposium on Engineering Technology and Applications, IEEE, 2017; pp. 39-44.
33. Muhammad, K.; Ahmad, J.; Baik, S.W. Early fire detection using convolutional neural networks during surveillance for effective disaster management. Neurocomputing 2018, 288, 30-42.
34. Xie, Y.; Peng, M. Forest fire forecasting using ensemble learning approaches. Neural Comput. Appl. 2019, 31, 4541-4550.
35. Pham, B.T.; Jaafari, A.; Avand, M.; Al-Ansari, N.; Dinh Du, T.; Yen, H.P.H.; Phong, T.V.; Nguyen, D.H.; Le, H.V.; Mafi-Gholami, D.; Prakash, I. Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry 2020, 12(6), 1022.
36. Rosadi, D.; Andriyani, W.; Arisanty, D.; Agustina, D. Prediction of forest fire occurrence in peatlands using machine learning approaches. In Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, IEEE, 2020; pp. 48-51.
37. Heyns, A.M.; du Plessis, W.; Curtin, K.M.; Kosch, M.; Hough, G. Decision support for the selection of optimal tower site locations for early-warning wildfire detection systems in South Africa. Int. Trans. Oper. Res. 2021, 28(5), 2299-2333.
38. Arif, M.; Alghamdi, K.K.; Sahel, S.A.; Alosaimi, S.O.; Alsahaft, M.E.; Alharthi, M.A.; Arif, M. Role of machine learning algorithms in forest fire management: A literature review. J. Robot. Autom. 2021, 5, 212-226.
39. Natekar, S.; Patil, S.; Nair, A.; Roychowdhury, S. Forest fire prediction using LSTM. In Proceedings of the 2021 2nd International Conference for Emerging Technology, IEEE, 2021; pp. 1-5.
40. Li, B.; Zhong, J.; Shi, G.; Fang, J. Forest Fire Spread Prediction Method based on BP Neural Network. In 2022 9th International Conference on Dependable Systems and Their Applications, IEEE, 2022; pp. 954-959.
41. Pawar, S.; Pandit, K.; Prabhu, R.; Samaga, R. A Machine Learning Approach to Forest Fire Prediction Through Environment Parameters. In Proceedings of the 2022 International Conference on Artificial Intelligence and Data Engineering, IEEE, 2022; pp. 1-7.
42. Sung, J.H.; Ryu, Y.; Seong, K.W. Deep Learning-Based Prediction of Fire Occurrence with Hydroclimatic Condition and Drought Phase over South Korea. KSCE J. Civ. Eng. 2022, 26(4), 2002-2012.
43. Budiningsih, K.; Nurfatriani, F.; Salminah, M.; Ulya, N.A.; Nurlia, A.; Setiabudi, I.M.; Mendham, D.S. Forest Management Units' Performance in Forest Fire Management Implementation in Central Kalimantan and South Sumatra. Forests 2022, 13(6), 894.
44. Rahman, A.K.Z.; Sakif, S.M.; Sikder, N.; Masud, M.; Aljuaid, H.; Bairagi, A.K. Unmanned Aerial Vehicle Assisted Forest Fire Detection Using Deep Convolutional Neural Network. Intell. Autom. Soft Comput. 2023, 35(3).
45. Charizanos, G.; Demirhan, H. Bayesian prediction of wildfire event probability using normalized difference vegetation index data from an Australian forest. Ecol. Inform. 2023, 73, 101899.
46. HOLGADO-VARGAS, M.R. Forest Fire Management and Territorial Governance. J. Surv. Fish. Sci. 2023, 10(3S), 2391-2414.
47. Anandaram, H.; Nagalakshmi, M.; Borda, R.F.C.; Kiruthika, K.; Yogadinesh, S. Forest fire management using machine learning techniques. Meas. Sens. 2023, 100659.
48. Wirth, R.; Hipp, J. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000; pp. 29–39.
49. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. CRISP-DM 1.0. Step-by-step data mining guide, 2000.
50. Abid, F. et al. Predicting Forest Fire in Algeria using Data Mining Techniques: Case Study of the Decision Tree Algorithm. In Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD 2019), Marrakech, Morocco, 08-11 July 2019.
51. Zhang, Q.; Xu, J.; Xu, L.; Guo, H. Deep convolutional neural networks for forest fire detection. In 2016 International Forum on Management, Education and Information Technology Application, Atlantis Press, 2016; pp. 568–575.
52. Kumar, V. and Minz, S., 2014. Feature selection: a literature review. SmartCR, 4(3), pp.211-229.

53. Saadabadi, M S E: Malakshan, S R; Kashiani, H; Nasrabadi, N M. CCFace: Classification Consistency for Low-Resolution Face Recognition, 2023. Available online: https://doi.org/10.48550/arXiv.2308.09230
54. Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.
55. Hall, M.A., 1999. Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.