algorithms

MDPI

Article

Nuthur: An Intelligent Wildfire Prediction and Early Warning System for the Al-Soudah Region

Huda Abdulrahman Almuzaini *, Renad Abdullah Altoum, Aryaf Fayez Alotaibi, Sarah Mohammed Alowjan and Manar Mohammed Abutheeb

Department of Computer Science, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia; renadabdullah0987@gmail.com (R.A.A.); sarahalowjan1@gmail.com (S.M.A.); mabutheeb0@gmail.com (M.M.A.)

* Correspondence: hamozeani@imamu.edu.sa

Abstract

Forest fires remain a major threat to biodiversity, human settlements, and the climate. This study presents Nuthur, an intelligent wildfire prediction and early-warning system for Al-Soudah, Saudi Arabia, integrating near-real-time environmental data with artificial intelligence models. The system used the Algerian Forest Fire dataset and a newly created local Saudi Arabian dataset. L1 regularization and Recursive Feature Elimination with Cross-Validation (RFECV) were used to examine relevant environmental variables, while oversampling, undersampling, and (Conditional Tabular Generative Adversarial Network) CTGAN-based synthetic augmentation were evaluated to address class imbalance. Multiple ML and DL models were evaluated, including Random Forest (RF), SVM, XGBoost, CatBoost, ensemble models, MLP, TabNet, and exploratory LSTM and CNN models, which were not interpreted as temporal or spatial models. Under random five-fold cross-validation, ML models achieved accuracy values from 0.89 to 0.97, with XGBoost with oversampling achieving the highest accuracy of 0.97. Deep-learning models achieved accuracy values from 0.75 to 0.94, with TabNet using RFECV achieving the best deep-learning result. A separate spatial cross-validation analysis of the Saudi dataset showed lower geographic generalization performance. CatBoost without oversampling achieved the highest mean spatial accuracy (0.8317) and weighted F1-score (0.7782), while logistic regression with oversampling achieved the highest fire-class recall (0.5616). In contrast, XGBoost with oversampling had a fire-class recall of 0.1096. These results highlight the need for further geographically and temporally diverse Saudi data before operational deployment.

Keywords: wildfire prediction; early fire detection; machine learning; deep learning; XGBoost; environmental monitoring; Al-Soudah region; CTGAN

Check for updates

Academic Editor: Roberto Montemanni

Received: 2 August 2026
Revised: 31 August 2026
Accepted: 1 September 2026
Published: 3 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Forests account for more than 31% of the world's land surface [1] and play a vital role in maintaining environmental balance. Nevertheless, forest fires pose a severe threat to biodiversity, human settlements, and the climate, while also resulting in substantial economic losses. Their unpredictable nature and the difficulty of containment, particularly when detection is delayed, make them among the most destructive natural disasters. In recent years, the frequency and intensity of forest fires have increased significantly, driven by climate change, prolonged drought, and rising temperatures. Consequently, understanding

the environmental factors that influence fire occurrence is essential for developing effective prediction systems capable of reducing associated risks and damages.

The Al-Soudah region, located in the Asir Province of southwestern Saudi Arabia, is characterized by mountainous terrain, dense juniper forests, and high ecological value. At approximately 3015 m above sea level, its moderate climate supports biodiversity that is uncommon in much of the Arabian Peninsula. However, seasonal dry periods and dense vegetation increase the risk of wildfire ignition and spread. The Asir region recorded 1371 VIIRS fire alerts between July 2020 and February 2025 [2], highlighting a growing threat to ecosystems and nearby communities. Despite this risk, wildfire prediction research focused on Al-Soudah remains limited. This study aligns with Saudi Vision 2030 and the Saudi Green Initiative, which aim to grow 10 billion trees, rehabilitate more than 74 million hectares of land, and protect 30% of the Kingdom's terrestrial and marine areas by 2030 [3,4]. Nuthur therefore aims to support earlier wildfire-risk assessment for Al-Soudah's vulnerable natural environment.

Several studies have presented different methods for predicting forest fires. Some studies rely on satellite imagery and deep learning techniques, while others depend on remote sensing data to improve Forest Fire Detection and Forecasting Systems. In addition, several researchers have applied artificial intelligence and machine learning approaches, including ensemble learning, fuzzy logic models, and Bayesian networks, to predict wildfire occurrence [5–7].

The proposed solution is to develop a smart wildfire prediction system that relies on environmental data and artificial intelligence to predict wildfires before they occur. The system will provide warnings and alerts to users, including authorities and volunteers, enabling relevant organizations to reduce potential losses. This will involve analyzing critical environmental factors to monitor high-risk areas. Furthermore, a web-based platform will be developed to assist responsible authorities in managing fire incidents efficiently. The platform will also support volunteer participation by allowing users to report fire-triggering conditions observed near affected areas. The system is named Nuthur (Arabic: نُذر), a term rooted in the Arabic language meaning warnings or omens of danger, reflecting the system's core purpose of delivering early alerts before wildfire threats escalate.

The principal contributions of this study are summarized as follows:

- Dataset Construction: A new local dataset was collected for the Al-Soudah region by integrating NASA FIRMS fire hotspot data with meteorological data from WeatherAPI, and was unified with the Algerian Forest Fire dataset to form a combined training corpus.
- Data Balancing and Augmentation: Oversampling, undersampling, and (Conditional Tabular Generative Adversarial Network) CTGAN-based synthetic data generation were evaluated and compared to investigate their effects on class imbalance and model performance.
- Feature Selection: Two complementary feature selection methods, L1-based regularization and Recursive Feature Elimination with Cross-Validation (RFECV), were employed to identify the most influential environmental variables.
- Comprehensive Model Evaluation: A broad range of ML models (Random Forest (RF), SVM, XGBoost, CatBoost, and ensemble methods) and DL models (LSTM, CNN, MLP, and TabNet) were systematically evaluated across multiple data balancing strategies, with XGBoost achieving the best overall accuracy of 0.97.
- System Deployment: The best-performing model was integrated into the Nuthur web-based platform, providing real-time fire risk prediction, automated scheduled monitoring, and structured incident reporting for both authorities and volunteers.

The following section reviews previous studies and related work on wildfire prediction and detection systems, focusing on the techniques used, the datasets applied, and the performance of the proposed models and approaches.

2. Related Work

The review is organized around two main topics, as shown in Figure 1. The first is Wildfire Occurrence Prediction Approaches, which includes data-driven methods using satellite imagery and historical environmental data, as well as sensor-based technologies for real-time monitoring. The second category is wildfire spread-based approaches, which aim not only to predict fire occurrence but also to estimate how a fire is likely to spread across geographical areas over time.

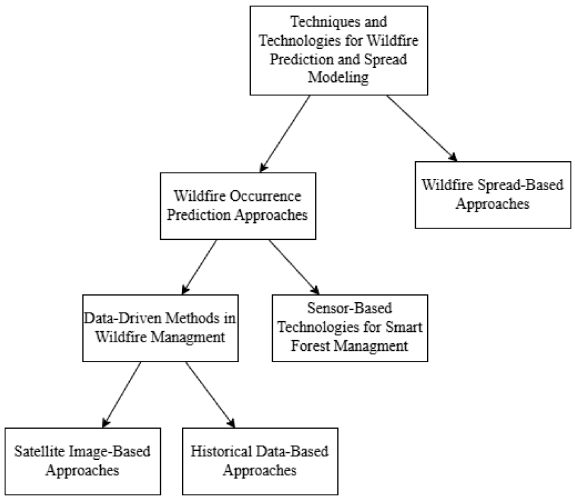


Figure 1. Techniques for wildfire prediction and spread modeling.

2.1. Wildfire Occurrence Prediction Approach

2.1.1. Data-Driven Methods in Wildfire Management

Data-driven methods for wildfire prediction have varied in many recent studies. The data used can be classified into two main types: image data extracted from satellites, drones, and other sources, and historical data extracted from records of previous wildfires or weather data at the time of the fire. In many cases, image data is used to monitor high-risk areas for early detection, while historical data is often used to train models to develop highly effective early warning systems. The following subsections explain the role of each type of data in previous studies.

- **Satellite Image-Based Approaches**

Satellite imagery has become essential in the detection, prediction, and monitoring of wildfires. The ability to handle wildfire issues has been significantly improved by the combination of cutting-edge remote sensing technology with AI and ML approaches [8].

One of the significant systems is presented in [9], where the authors introduce a method to enhance DL models for wildfire detection. The study utilizes the California Wildfire GeoImaging Dataset (CWGID), which is based on bi-temporal Sentinel-2 imagery from Google Earth Engine. The CWGID contains 100,000 high-resolution labeled image pairs that capture pre- and post-wildfire conditions across California; each pair includes labels indicating burned or unburned areas. The study evaluates the performance of three CNN models, showing that the EfficientNet-B0 model achieved the highest accuracy, exceeding 92% in the wildfire detection task.

Ref. [10] introduced a mission using a small satellite known as the KITSUNE satellite, a 6U CubeSat equipped with a camera capable of capturing high-quality images at a 5-m resolution. Image classification is performed onboard instead of transmitting all images to Earth for analysis. The CNN model was trained and implemented on the satellite computer, which runs the model when needed and attaches the results to the image data. The system achieved 98% accuracy using the MiniVGGNet model, while two other tested models (LeNet, ShallowNet) also performed well with 95% accuracy. The KITSUNE satellite was launched from the International Space Station (ISS) in March 2022, demonstrating that a small satellite can successfully apply DL in space. Ref. [11] presented a novel approach for detecting wildfires using small satellites called CubeSats. The authors used two ML models, MobileNetV2 and ResNet152V2, to detect fires from satellite images. Since CubeSats have limited camera quality, a super-resolution technique called Real-ESRGAN is applied to enhance image clarity. The dataset was created from Landsat-8 satellite images, with 3818 images for training, 954 for validation, and 1194 for testing. The results show that super-resolution improves forest fire detection. However, the authors noted that the dataset remains relatively small.

Ref. [12] compared traditional ML models with CNN-based approaches for wildfire prediction. They evaluated three ML models: SVM, AdaBoost, and XGBoost, along with a DL approach using CNN. The dataset is compiled by merging several smaller datasets and adding extra testing images from Google. It is organized into merged datasets of "fire" and "non-fire" images. The authors conclude that the CNN model is the most accurate, achieving an accuracy of 89.9%, making it highly effective for predicting fire-prone areas. However, they also note a challenge: the need for real-time data processing. Ref. [13] proposed a new deep CNN called MultiScale-Net to detect active fires using Landsat-8 satellite images. The model uses different filter sizes and special layers to capture details at various scales, improving fire detection accuracy. It also introduced a new fire index called Active Fire Index (AFI), created by combining three specific satellite image bands to enhance detection. They tested 27 versions of the model with various settings by changing filter sizes, input combinations, and layer configurations. The results show that using multiple filter sizes and the AFI achieved a high F1-score of 91.62% and an IoU of 84.54%, outperforming traditional methods.

Ref. [14] reviewed various DL models used for detecting and predicting wildfires from satellite images. It highlights models such as U-Net, VGG16, InceptionV3, and other CNN-based architectures applied to imagery from satellites including Sentinel-2, MODIS, Landsat-8, and Himawari-8. The datasets include both real forest-fire and non-fire satellite images. The review concludes that many models achieved high accuracy, with some reaching up to 99% in detecting fires and burned areas. The Fire-Net model using Landsat-8 achieved 99.95% accuracy. The paper also notes that model performance varies depending on the satellite and the region.

The semi-arid oak forests of Lorestan Province in western Iran are also exposed to increasing wildfire risk. In this study [15], the wildfire susceptibility of the region was investigated using environmental and human-related data. The study area covered approximately 28,300 km^2, and the dataset was prepared from composite Landsat imagery. Forest cover was classified into five density classes, from forest-edge areas to dense forests. An ensemble machine-learning model provided by the Biomod2 package was developed to generate a wildfire-susceptibility map. The results showed that the Normalized Difference Vegetation Index (NDVI) was an important factor in wildfire susceptibility. Human activities near road networks and urban centers also strongly influenced wildfire patterns. The final susceptibility map identified approximately 2774 km^2, representing 25.6% of the study area, as highly susceptible to wildfire.

- **Historical Data-Based Approaches**

In many studies, historical data from past wildfire records or previous weather conditions have been used to develop ML models for predicting wildfires. This data has helped researchers understand the causes and influencing factors of wildfires and is also useful in environments lacking satellite imagery and sensor data.

The state of California is one of the regions with high wildfire risk. In this study [16], the increase in wildfires in California was discussed, driven by climatic factors such as drought and extreme heat. The dataset extracted from California's environmental data consists of approximately 128,125 instances with 18 features. A model was developed based on environmental data to predict the risk of wildfire outbreaks using ML techniques, including SVM, KNN, DT, LR, and the random forest algorithm. A final model accuracy of up to 92% was achieved using the RF model. In Ref. [17], a model based on climatic data was developed using Algeria's 2012 wildfire dataset, which consists of 244 instances from two regions in Algeria. The focus was on five features: month, temperature, humidity, wind speed, and rainfall. Several ML models were applied, including DT, LR, RF, and ANN. The RF model achieved the highest accuracy, reaching 86.486% in wildfire prediction.

Ref. [18] aimed to evaluate the performance of ML methods in wildfire prediction. The study relied on wildfire data from Bu Mat National Park in Vietnam between 2014 and 2016, covering 57 fire incidents. The environmental factors considered included elevation, slope, temperature, aspect, and drought index. Four models were evaluated: BNN, NB, DT, and LR. BNN outperformed the others with an AUC score of 0.96. The study is considered an important reference for improving wildfire prediction strategies.

2.1.2. Sensor-Based Technologies for Smart Forest Management

In study [19], the authors discussed the concept of "smart forests", a modern approach that integrates AI with sensor technologies and robotic systems with the aim of developing more efficient and sustainable forest management. The study confirms that AI cannot play an effective role without high-quality data, making data collection using sensors a crucial stage in this digital transformation. The paper highlights the importance of introducing humans into the "human-in-the-loop" to ensure system flexibility and accuracy. It also showcases the researchers' experience with a comprehensive sensor platform capable of integrating multiple types of sensors and analyzing data in the field using autonomous robots. The system proposed in [20] for early detection of forest fires relies mainly on wireless sensor networks (WSNs) to collect environmental data from forests. These sensors can measure a range of parameters such as temperature, humidity, smoke, and other indicators indicating the risk of fire. Using the Internet of Things (IoT), these sensors are connected to an integrated network that allows for efficient and continuous data transmission over a low-energy network (LPWAN), ensuring wide coverage of remote or hard-to-reach areas. This data is analyzed using AI techniques, specifically DL, to predict fires before they escalate. The system provides high accuracy in the early detection of environmental hazards, enabling faster and more efficient preventive action.

The proposed paper [21] deals with the topic of "early detection of forest fires" using an automatic system based on smoke and fire sensors. The system proposed in this paper relies on the use of sensors to measure physical changes such as smoke and temperature, which helps in monitoring the onset of fire quickly and accurately. One of the most prominent features of the system is the "remote alert" using the "GSM" module, which allows instant notifications to be sent to users via mobile phone, allowing them to respond quickly and efficiently. The system is also characterized by its low cost and ease of maintenance compared to traditional systems, such as human monitoring or remote video. Ref. [22] focused on the use of multiple sensors in wildfire detection using unmanned aerial vehi-

cles (UAS). The proposed system relies on the synchronous integration of visual cameras (VIS) and thermal cameras (IR) to improve fire detection accuracy. Thermal cameras capture the hot spots that appear initially when a fire breaks out, while visual cameras provide additional information on the surrounding environmental situation. These images are combined to give a uniform image, helping to identify smoke or hot spots more accurately. The data collected by these sensors is sent to a control center via a 5G network for automated analysis. This technology improves the effectiveness of the system in terms of "rapid response" to fires and reduces potential losses. The use of multiple sensing technologies enhances the ability to monitor forests in real time and contributes to the effective protection of the environment and natural resources.

2.2. Wildfire Spread-Based Approaches

Wildfire spread prediction is an important field of wildfire management where scientists study environmental measures to estimate fire direction and speed. The prediction is often made using simulation models, ML techniques, or even remote sensing data to capture the temporal and spatial fire behavior. The following studies illustrate how different techniques can be used to model wildfire spread while enriching conclusions about the relationship between environment and fire behavior.

In study [23], the authors evaluated the use of genetic algorithms (GAs) in calibrating wildfire spread models, particularly the widely used Rothermel model. Wildfire prediction models rely on physical factors such as fuel moisture, wind speed, bulk density, surface area-to-volume ratio, and slope. Small changes in these parameters can lead to large differences between predicted and actual fire spread, making calibration essential. The study classified GA approaches into two main categories: methods that directly compare simulation results with real fire data, and surrogate-based methods. Using 37 prescribed-burn datasets, the results showed that GA techniques significantly improved model accuracy. The study also examined different objective functions and highlighted the need for standardized evaluation frameworks to ensure fair comparisons between calibration methods.

In study [24], the authors introduced FireCast, a DL framework for predicting wildfire spread by learning spatiotemporal patterns from past fires. Unlike physical models such as FARSITE, FireCast is data-driven and does not require manual parameter tuning. The framework uses CNNs to extract features from satellite images and combines topography, meteorology, and land cover data such as wind speed, temperature, vegetation, and elevation from sources including MODIS and land cover databases. The system divides landscapes into grids and predicts whether each cell will burn or not as a binary classification task. By analyzing sequences of images over time, FireCast learns the relationship between fire spread and environmental conditions and automatically applies fire behavior rules from training data. It generates probability maps that help emergency services identify high-risk zones and improve decision-making during wildfire events. The model was validated using real wildfire data and compared with random prediction methods and traditional wildfire-spread models such as FARSITE, outperforming them in accuracy, recall, and F-score, demonstrating the feasibility of DL-based wildfire spread modeling at scale.

In study [25], the authors developed an ML framework for predicting the next-day spatial spread of wildfires using remote-sensing data. The study introduced the "Next Day Wildfire Spread" database, containing 18,545 wildfire events across the United States from 2012 to 2020. For each event, wildfire masks for consecutive days and 11 environmental and human-related variables such as temperature, humidity, wind speed and direction, elevation, vegetation index, drought index, and population density were included. The problem was modeled as a semantic segmentation task using a convolutional autoen-

coder to capture spatial fire propagation patterns. The DL model outperformed traditional methods such as logistic regression and random forest, achieving higher AUC-PR, precision, and recall while preserving the spatial structure of fire regions. Ablation experiments showed that the previous day's fire mask had the greatest impact on prediction performance, followed by vegetation and elevation. The study also found that coarser output resolutions improved overall AUC but reduced spatial precision, highlighting a trade-off between localization and predictive accuracy. Overall, the study presents a scalable and reproducible approach for wildfire forecasting and provides publicly available tools and datasets for future research.

Although existing approaches have improved wildfire prediction, several limitations remain. Satellite-image and geospatial approaches are useful for identifying long-term wildfire-susceptibility areas, but they may not provide immediate risk assessments at a specific time and location. Models based on historical environmental data can support prediction of fire occurrence; however, their performance may be affected by limited fire records, class imbalance, uncertain non-fire labels, and differences in environmental conditions between regions. Sensor-based systems provide real-time monitoring but may require extensive infrastructure and continuous data availability. These limitations motivate the development of Nuthur as an on-demand wildfire-risk assessment and alerting system for the Al-Soudah region.

3. Datasets

In this research, two datasets were utilized to build a reliable wildfire prediction model: the existing Algerian Forest Fire dataset and a locally collected Saudi Arabian dataset. A systematic pipeline of steps was followed for data collection, preprocessing, and unification of these heterogeneous datasets. A summary of both datasets is presented in Table 1.

We began by collecting open-access datasets from different regions, which included the Portugal forest fire dataset [26] and the Algerian forest fire dataset [27]. These datasets contain environmental variables relevant to wildfire behavior. Initially, we considered using the Portugal dataset [26]; however, after preliminary analysis and experiments, we found that it was not well-suited to our research goal. The dataset was originally designed for regression tasks. Although we tried to adapt it for classification by labeling instances where the area was 0 as 0 (no fire) and any area greater than 0 as 1 (fire), this transformation did not yield good results for our project; it reduced the accuracy of the model and made the results less reliable.

Although the Algerian Forest Fire dataset was collected in a different geographical setting, it contains core meteorological variables, including temperature, relative humidity, wind speed, and rainfall, which are widely recognized as influential predictors of wildfire occurrence. However, the vegetation composition, climatic regime, topography, and fire drivers of northern Algeria differ from those of Al-Soudah. In particular, Al-Soudah is a high-elevation region in the Asir Mountains characterized by juniper woodland and a semi-arid climate with locally distinct rainfall patterns. Therefore, the Algerian dataset was not used as a direct representation of Al-Soudah conditions. Instead, a local Saudi Arabian dataset was constructed using NASA FIRMS fire hotspot records and location-specific meteorological data to incorporate the environmental characteristics of the study area. The combined analysis was intended to evaluate model robustness while maintaining a local dataset for regionally relevant prediction.

The Algerian dataset [27] was designed for binary classification to predict whether a given day's condition will lead to a forest fire. It contains meteorological data from two regions of Algeria (Bejaia located in the northeast, and Sidi Bel-abbes, located in the north-

west) recorded during the months of June and September 2012. These are the peak months for forest fires in the region. It contains 244 instances (122 for each region) and 12 attributes: Date, Temp, RH, Ws, Rain, FFMC, DMC, DC, ISI, BUI, FWI, and Region. A description of each attribute we used is provided in Table 2. It presents the list of attributes selected for building the wildfire prediction model. Each attribute is described briefly to explain its role and relevance to fire risk assessment. These variables were unified across all datasets to ensure consistency in training and testing.

Table 1. Summary of datasets.

Dataset	Description	Access
Portugal Forest Fires	The Portugal Forest Fires dataset contains 517 observations. It aims to predict the burned area of forest fires using variables like temperature, wind, humidity, and fire weather indices. This dataset is used for regression tasks and environmental modeling.	Open Access
Algerian forest fires dataset	The Algerian Forest Fires dataset contains daily meteorological data and Fire Weather Index (FWI) components related to fire occurrences in Algeria. It supports research on forest fire prediction using ML.	Open Access
South Saudi Arabia wildfire integrated dataset (2019–2025)	A custom dataset was developed by the research team by integrating NASA FIRMS [28] fire hotspot data with meteorological data sourced from WeatherAPI. The dataset underwent thorough preprocessing and cleaning, resulting in a curated collection of 613 records.	collected

Table 2. Description of the attributes used in the wildfire prediction model.

Attribute	Meaning
Year	Year: 2012 to 2025
Month	Month of the year: “jan” to “dec”
Day	Day of the week: “mon” to “sun”
FFMC	Fine Fuel Moisture Code (FFMC) index from the Forest Fire Weather Index (FWI) system: 18.7 to 96.20
DMC	Duff Moisture Code (DMC) index from the FWI system: 1.1 to 291.3
DC	Drought Code (DC) index from the FWI system: 7.9 to 860.6
ISI	Initial Spread Index (ISI) index from the FWI system: 0.0 to 56.10
Temperature	temperature in Celsius: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
Rain	outside rain in mm/m^2: 0.0 to 6.4
Ws	Wind speeds in kilometers per hour.
Classes	The label: whether the day had a fire or no fire

3.1. Creating a Local Dataset for Saudi Arabia

In addition to the Algerian dataset, a new dataset was created for the forested area in southern Saudi Arabia to provide local data for this study. The following steps were used to construct the dataset:

3.1.1. Getting Fire Data from NASA FIRMS

We requested and used fire incident data from NASA FIRMS [28] for the years 2019 to 2025. FIRMS is a satellite-based platform that provides near-real-time global fire detection data derived from MODIS and VIIRS sensors. The data includes fire locations (latitude and longitude), time of detection, fire intensity measured by Fire Radiative Power (FRP), and confidence level assigned to each active-fire detection. At first, we obtained 7543 records. To ensure data reliability and quality, we implemented a filtering process for selecting the fire and non-fire instances.

- *Fire instance (Label = 1)*: If the satellite confidence was high and the fire intensity FRP was greater than 5.
- *Non-fire instance (Label = 0)*: We selected dates when no fires were recorded in the NASA FIRMS database. For these dates, we picked geographical coordinates within fire-prone regions of southern Saudi Arabia. This was done intentionally so that the model could learn the weather conditions associated with fire occurrence within the target ecosystem, rather than simply separating different geographic areas. After applying these filters, we obtained a total of 613 records (73 fire and 540 non-fire instances). This refined dataset allowed us to use more recent and accurate fire incident data, which helped improve the quality and performance of our model.

3.1.2. Adding Weather Data from WeatherAPI

Since the dataset from NASA FIRMS does not include weather data. We used WeatherAPI [29]. This API provides global, location-based historical weather data derived from meteorological reanalysis models. For each fire and non-fire record, meteorological variables were associated with the corresponding geographic coordinates and date. The retrieved variables include air temperature, relative humidity, wind speed, and precipitation. Integrating these parameters allowed us to enrich the fire dataset and unify it with the Algerian dataset by associating relevant weather conditions with every data point. Using the collected weather data, the FWI-derived indicators were calculated, comprising the FFMC, DMC, DC, and ISI. These fire risk indicators were calculated using published equations provided by Natural Resources Canada [30]. The implementation follows the published Canadian FWI equations without modifying the underlying mathematical formulations. However, the input weather variables in our framework are obtained from the API for the corresponding event date and location rather than being restricted to the standard daily noon observation schedule. While the standard Canadian FWI System uses daily noon observations and 24-h accumulated precipitation, our framework operates in near-real-time via on-demand API queries. Accordingly, the resulting FWI-derived indicators should be interpreted as event-time estimates used consistently with the operational prediction pipeline, rather than as strict standard daily Canadian FWI observations. These indicators effectively capture critical ignition and spread dynamics, providing consistent, regionally relevant features for wildfire classification.

3.2. Dataset Balancing and Evaluation Protocol

The final Saudi Arabian dataset contains 613 records, including 73 fire records (11.9%) and 540 records with no FIRMS-detected fire (88.1%). This distribution indicates a substan-

tial class imbalance. Therefore, overall accuracy alone may be influenced by the majority class and should not be interpreted as a sufficient measure of fire-detection performance.

To examine the effect of class imbalance on model performance, the original dataset was retained as a baseline and compared with a random oversampling configuration. Oversampling was applied only to the training data within each cross-validation iteration to increase the representation of the minority fire class without modifying the evaluation fold. For the combined Algerian–Saudi Arabian dataset, oversampling, undersampling, and CT-GAN -based synthetic augmentation were evaluated as separate balancing strategies. The comparative results are presented in Section 5.

Model performance was evaluated using stratified five-fold cross-validation with random shuffling. The dataset was divided into five folds while preserving the class distribution. In each iteration, four folds were used for training, and the remaining fold was used for evaluation. This process was repeated five times, and the reported accuracy and F1-score values represent the average across all folds. As the folds were randomly generated, this evaluation assesses generalization across the available records rather than strict chronological forecasting. Future work will evaluate the system using a time-based train–test split when more temporal data become available.

In addition to random stratified five-fold cross-validation, a separate spatial five-fold cross-validation analysis was conducted for the Saudi Arabian dataset to assess generalization across geographically separate locations.

4. Materials and Methods

The Nuthur system operates as an intelligent platform designed to track and forecast wildfires through a comprehensive monitoring framework. The system architecture integrates environmental monitoring, fire risk analysis, report management, and real-time notification services. This multi-layered approach ensures that critical data is processed swiftly, facilitating early detection and rapid response. The system supports two primary user groups: volunteers and authorities. Volunteers are responsible for submitting fire reports to authorities, who subsequently manage report assessment, track associated risks, and execute necessary safety measures to prevent or mitigate incidents.

The architecture of the proposed Nuthur system is structured to ensure efficient reporting, data validation, and an effective response to wildfire incidents. By compartmentalizing functionalities into distinct layers, the system achieves high scalability and maintainability. Figure 2 illustrates the high-level workflow and the interactions between the various subsystem components.

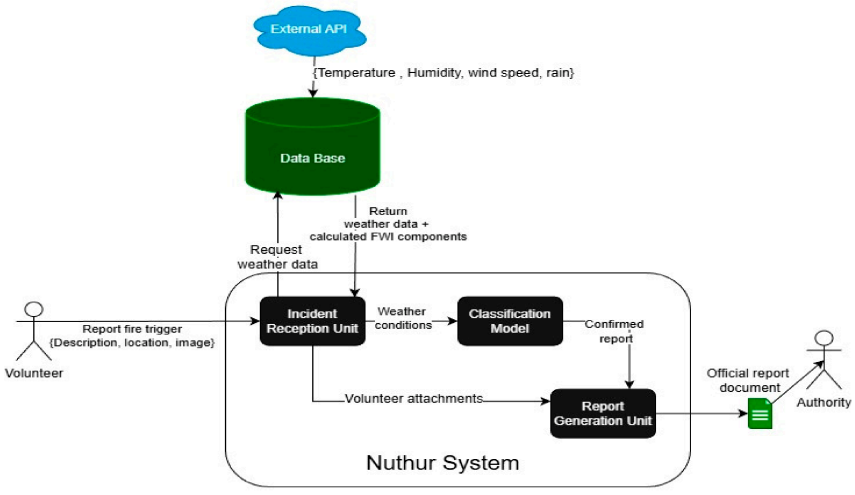


Figure 2. System architecture.

4.1. System Architecture and Subsystems

The system follows a layered client-server architecture comprising several interconnected subsystems. The Client Layer (Volunteer Website) allows users to submit incident reports, which include essential details such as the fire's location, a descriptive text, and photographic evidence. The Website Layer (Incident Reception Unit) receives these reports and validates them to ensure the completeness of the submitted data. Once validated, this unit requests real-time meteorological parameters—specifically temperature, humidity, rainfall, and wind speed for the reported location from the External Service (Weather API).

The Data Layer (Database) serves as the central repository, securely storing the validated volunteer reports, the retrieved weather data, and all system logs. The Processing Layer (Classification Model) then analyzes the combined dataset, evaluating current weather conditions against historical patterns to determine the actual risk of a wildfire occurring. Following the risk assessment, the Output Layer (Report Generation Unit) synthesizes the findings into an official report summarizing all relevant information concerning the incident. Finally, this generated report is transmitted to the End Users (Authorized Agencies), enabling them to take appropriate and timely action based on the verified intelligence.

Upon detecting a potential fire threat, the system triggers an immediate alert, automatically dispatching notifications to registered users and official agencies. Environmental conditions are assessed at fixed intervals of every eight hours through a scheduled monitoring process, ensuring continuous and proactive surveillance. Authorized personnel can additionally export PDF summaries of each scheduled check, providing a structured record of the system's periodic fire risk assessments. Alerts are generated based on both automated environmental evaluations and volunteer-submitted reports, creating a dual-source intelligence mechanism. This continuous alerting capability reduces emergency response times and strengthens overall wildfire prevention efforts.

4.2. Fire Prediction Model

To study how class distribution affects the performance in predicting fire occurrences, machine learning (ML) and deep learning (DL) models were evaluated under three data balancing strategies: oversampling, undersampling, and no sampling.

4.2.1. Machine Learning Models

Several traditional ML models were applied to predict fire occurrences based on environmental and weather features. The RF model was utilized as an ensemble of decision trees to improve prediction accuracy and mitigate overfitting. A logistic regression (LR) model served as a simple linear baseline, predicting the probability of fire occurrence using a sigmoid function. The SVM model was selected for its capability to identify the optimal separating boundary between classes, particularly in high-dimensional data spaces. Furthermore, XGBoost was employed due to its computational efficiency and robust performance on structured data, while CatBoost was chosen for its effectiveness in handling categorical features and preventing overfitting through ordered boosting.

In addition to individual models, two ensemble strategies were implemented to enhance overall performance and stability. The first approach was a Voting Ensemble model, which combined CatBoost, SVM, and XGBoost using soft voting, aggregating predictions based on the averaged probabilities from each constituent model. This approach aims to reduce individual model bias and improve overall prediction reliability. The second strategy involved a stacking ensemble model, which integrated RF, XGBoost, and CatBoost as base learners, with logistic regression serving as the meta-model to synthesize their out-

puts and further improve generalization. Together, these ensemble methods leveraged the distinct strengths of multiple algorithms to achieve more balanced and accurate fire occurrence predictions.

For all ML models, hyperparameters were tuned using Grid Search, and five-fold cross-validation was applied to ensure consistent and reliable results. StandardScaler was applied only to scale-sensitive models, such as logistic regression and SVM, to normalize feature values before training. Tree-based models, including random forest, XGBoost, and CatBoost, were trained without feature scaling because their threshold-based learning process is not affected by feature magnitude. To further optimize model efficiency, feature selection was subsequently applied to identify the most influential variables for the top-performing models; detailed results of this analysis are presented in the subsequent section.

4.2.2. Deep Learning Models

Several DL models were developed as exploratory baselines to examine their ability to capture nonlinear relationships in the tabular data. The LSTM model received the tabular features reshaped into a 3D representation (N, "features", 1), treating the feature dimension as the sequence axis rather than as a true temporal sequence. For CNN2D, the features were reshaped into a two-dimensional representation (N, "features", 1, 1) with a single column and processed using convolutional filters. These transformations were used to provide the input structures required by the respective architectures and do not imply inherent temporal or spatial relationships in the original tabular data. As the available data consist of independent tabular observations, LSTM and CNN were used only as exploratory benchmark architectures.

Additionally, a multilayer perceptron (MLP) was utilized as a fully connected network to model nonlinear relationships between input variables and the target class. The study also incorporated TabNet, a DL architecture specifically designed for tabular data, which applies sequential attention to focus on the most relevant features during training [31]. To prevent overfitting and ensure model stability, all DL models were trained using Dropout, L2 Regularization, and Early Stopping mechanisms. Consistent with the approach taken for the ML models, feature selection was conducted on the best-performing DL models to highlight the variables most critical to prediction accuracy.

4.3. Feature Selection

Feature selection is the process of identifying and retaining only the most significant variables in a dataset while discarding less relevant ones. The primary objectives of this process are to simplify the model architecture, reduce computational training time, and enhance interpretability by focusing exclusively on features that contribute most substantially to accurate predictions.

In this study, L1-based feature selection [32] was applied using the 'SelectFromModel' meta-transformer, which selects features with non-zero coefficients from a linear model penalized with the L1 norm. This approach produces sparse solutions where many feature coefficients are reduced to zero. The regularization parameter 'C' in the LinearSVC model controls this sparsity: smaller values of 'C' enforce stronger regularization, resulting in fewer retained features. As illustrated in Figure 3, varying the 'C' parameter produced different feature subsets.

A value of C = 0.01 was selected, which effectively reduced the original dataset from 13 features to 5 key features: ['year', 'Ws', 'FFMC', 'BUI', 'FWI']. This L1-based feature selection method was applied to the top three ML models (XGBoost, CatBoost, and Ensemble Learning) as well as the top three DL models. The L1-based feature-selection experiment

was conducted to examine the effect of a compact feature subset on the performance of the best ML and DL models. It was used as a comparative feature-reduction analysis rather than as an independent external validation procedure. This systematic reduction not only streamlined the computational process but also allowed for the identification of the most influential variables, facilitating an examination of which features remained consistently important across different modeling paradigms.

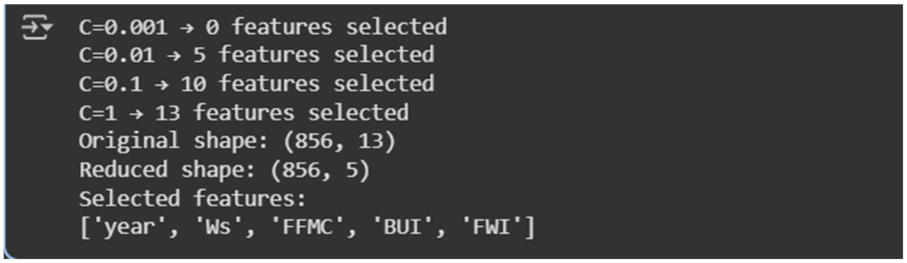


Figure 3. Selected features using L1-based feature selection.

Furthermore, RFECV was applied to identify the most critical features within the combined dataset. RFECV is an iterative process that repeatedly trains a model, ranks features by their importance, and systematically removes the least significant ones. During each iteration, cross-validation evaluates model performance to determine the optimal number of features to retain.

An XGBoost classifier was selected as the base estimator for the RFECV process due to its proficiency in handling nonlinear feature interactions and providing reliable importance scores. Following the elimination process, the algorithm identified 10 optimal features from the original 13: [‘day’, ‘month’, ‘year’, ‘Temperature’, ‘Ws’, ‘Rain’, ‘FFMC’, ‘DC’, ‘ISI’, ‘BUI’]. These selected features were subsequently utilized as inputs for both the ML and DL models. The evaluation and performance comparison based on these refined feature sets are discussed in the subsequent section. Although feature selection reduced the original feature set from 13 to 10 variables, several selected FWI components remain mathematically related, including FFMC, DC, ISI, and BUI. This may affect coefficient stability and feature interpretability in logistic regression and linear SVM. Therefore, these models were used for comparative prediction rather than coefficient-based interpretation, while tree-based models were also evaluated because they are less sensitive to multicollinearity.

4.4. Software and Implementation Environment

Data preprocessing, feature selection, model training, and evaluation were performed in Google Colab (Google LLC, Mountain View, CA, USA; https://colab.research.google.com/) using Python version 3.13.15. The following libraries were used: NumPy 2.1.3, pandas 2.2.3, scikit-learn 1.6.1, imbalanced-learn 0.14.2, Matplotlib 3.10.0, SDV 1.38.2, XGBoost 3.4.1, CatBoost 1.2.10, TensorFlow 2.20.0, and Keras 3.13.2. RF, LR, and SVM were implemented using scikit-learn; XGBoost and CatBoost used their respective libraries; CTGAN was implemented using SDV; and the deep-learning models were developed using TensorFlow and Keras.

The Nuthur web platform was developed using JavaScript, Node.js, React, Next.js, and MongoDB. The official websites of these software tools are https://www.javascript.com/, https://nodejs.org/, https://react.dev/, https://nextjs.org/, and https://www.mongodb.com/.

5. Results and Discussion

This section presents the experimental results of fire prediction models under different data scenarios. It also describes our systematic approach to training and evaluating both ML and DL models on individual regional datasets as well as their combination. For each dataset configuration, we conducted experiments with and without oversampling, with undersampling techniques, and with the use of CTGAN to generate synthetic samples for class balancing. All reported results are based on 5-fold cross-validation.

The experiments were conducted under three distinct settings: the Algerian dataset alone, the Saudi Arabian dataset alone, and the combined Algerian–Saudi Arabian dataset. The individual-dataset experiments evaluate model performance within each dataset, whereas the combined-dataset experiments provide an exploratory assessment of model robustness under heterogeneous data conditions.

5.1. Algerian Dataset

The performance of several ML algorithms and DL architectures was evaluated on the Algerian dataset using accuracy and F1-score as metrics, under two conditions: with oversampling and without oversampling. This comparison highlights how data balancing affects classification performance across both model types. The results for the ML models are summarized in Table 3, while the corresponding DL results are presented in Table 4.

Table 3. Performance of ML models using the Algerian dataset. Bold values indicate the highest performance for the evaluation metric.

Model	Without Oversampling		Oversampling	
	Accuracy	**F1**	**Accuracy**	**F1**
RF	0.98	**0.98**	0.98	**0.98**
LR	0.96	0.96	0.95	0.96
XGBoost	0.98	**0.98**	0.98	**0.98**
SVM	0.95	0.95	0.95	0.95
CatBoost	0.98	**0.98**	0.98	**0.98**
Stacking	0.98	**0.98**	0.98	**0.98**

Table 4. Performance of DL models using the Algerian dataset. Bold values indicate the highest performance for the evaluation metric.

Model	Without Oversampling		Oversampling	
	Accuracy	**F1**	**Accuracy**	**F1**
LSTM	0.93	**0.93**	0.93	**0.93**
CNN1D	0.85	0.85	0.83	0.83
MLP	0.81	0.81	0.87	0.87
CNN2D	0.91	0.91	0.91	0.91

The original Algerian dataset exhibited only a slight class imbalance, comprising 137 fire samples and 107 non-fire samples. The machine learning models achieved stronger overall performance under five-fold cross-validation evaluation. Random forest and XGBoost achieved the highest accuracy and weighted F1-score of 0.9836, both with and without oversampling. CatBoost and the stacking model also performed strongly, each achieving an accuracy and weighted F1-score of 0.9795. In contrast, logistic regression and SVM obtained relatively lower results, with accuracy scores ranging from 0.9548 to 0.9632. Over-

all, oversampling did not improve performance on the Algerian dataset, likely because the original class distribution was already relatively balanced.

Among the deep learning models, LSTM and CNN2D achieved accuracy scores of 0.93 and 0.91, respectively. After class balancing through the method of oversampling to 109 per class, DL models showed small changes after oversampling. LSTM stayed the same at 0.93 accuracy, CNN2D also kept its accuracy at 0.91, and CNN1D and MLP showed slight drops or increases but remained close to their original performance. This shows that oversampling can reduce class imbalance but not necessarily improve performance in DL models trained on tabular data. Traditional ML models became more robust and consistent in both imbalanced and balanced settings.

To contextualize the Algerian-dataset findings, Table 5 compares the results of this study with directly relevant published wildfire prediction studies that used the same dataset. The comparison presents datasets, evaluated models, and best reported performance.

Table 5. Comparison with published wildfire prediction studies. Bold values indicate the highest performance for the evaluation metric.

Study	Dataset and Input Features	Models Compared	Best Reported Result
Mohanty and Lee [17] (2024)	Algerian Forest Fire dataset; reduced to five meteorological features.	RF, DT, LR, and ANN.	RF: accuracy = 0.8649.
Zaidi [33] (2023)	Algerian Forest Fire dataset; reduced to six principal components.	ANN, LR, KNN, SVM, and RF.	ANN: accuracy = 0.967; F1-score = 0.971 $\pm$ 0.023.
This study (Nuthur)	Algerian Forest Fire dataset; all features	RF, LR, XGBoost, SVM, CatBoost, and stacking.	**RF and XGBoost: accuracy = 0.9836; weighted F1-score = 0.9836.**

As shown in Table 5, the Nuthur evaluation achieved the strongest reported performance among the compared studies; however, direct numerical comparison should be interpreted carefully because the studies use different feature sets and preprocessing procedures. Therefore, this comparison is included to contextualize the results rather than to claim universal state-of-the-art superiority.

5.2. Saudi Arabia Dataset

Both machine learning (ML) and deep learning (DL) algorithms were applied to the Saudi Arabian dataset to evaluate their classification performance under two data distribution conditions: with oversampling and without oversampling. This comparison allows for a systematic assessment of how class balancing affects model performance. The performance of the ML models is summarized in Table 6, while the corresponding results for the DL architectures are presented in Table 7.

The Saudi dataset is imbalanced, containing 540 instances of the "not fire" class and 73 instances of the "fire" class. Even with this imbalance, most ML models performed very well without oversampling. XGBoost gave the best results with 0.92 F1 and 0.91 accuracy, followed by CatBoost, random forest, and the stacking model, which also showed strong performance. After applying oversampling, the performance of ML models stayed almost the same or changed only slightly. However, logistic regression was the only model that dropped clearly in performance from 0.87 to 0.67 accuracy. These findings show that the effect of oversampling varied across models; it improved the performance of the stacking

and CatBoost models but did not improve XGBoost or logistic regression. The separate spatial cross-validation analysis in Section 5.6 provides a stricter assessment of generalization to geographically separate locations.

Table 6. Performance of ML models using the Saudi dataset. Bold values indicate the highest performance for the evaluation metric.

Model	Without Oversampling		Oversampling	
	Accuracy	F1	Accuracy	F1
RF	0.90	0.89	0.85	0.86
LR	0.87	0.82	0.67	0.73
XGBoost	0.92	**0.91**	0.90	0.89
SVM	0.88	0.85	0.87	0.88
CatBoost	0.90	0.89	0.91	0.90
Stacking	0.91	0.90	0.93	**0.93**

Table 7. Performance of DL models using the Saudi dataset. Bold values indicate the highest performance for the evaluation metric.

Model	Without Oversampling		Oversampling	
	Accuracy	F1	Accuracy	F1
LSTM	0.88	**0.83**	0.82	0.81
CNN1D	0.88	**0.83**	0.88	0.83
MLP	0.88	**0.83**	0.86	**0.84**
CNN2D	0.86	0.82	0.79	0.82

For DL models such as LSTM, CNN, and MLP, the performance was good without oversampling, up to 0.88 accuracy. However, oversampling did not help these models. In some cases, the performance even dropped, like LSTM, CNN2D, and MLP. Overall, the results show that ML models, especially ensemble models like XGBoost and CatBoost, work better and more consistently on this type of tabular data. In contrast, DL models were more sensitive to oversampling and did not gain any clear improvement from it.

5.3. Combined Dataset

The combined Algerian and Saudi Arabian dataset was used to evaluate both ML algorithms and DL models under three data balancing conditions: without oversampling, with oversampling, and with undersampling. The detailed results for the ML models are presented in Table 8, providing a comparison of accuracy scores across the three balancing strategies. The corresponding DL results are summarized in Table 9, illustrating the accuracy scores across the same conditions.

The combined dataset consists of 856 samples with 14 features. On the original data (imbalanced), most ML models achieved high performance: XGBoost reached the highest accuracy of 0.96, while Stacking, Ensemble Classifier, and CatBoost followed with 0.95. Other models like RF and LR also gave strong results, with accuracy scores between 0.90 and 0.91. By comparison, DL models performed slightly lower, with CNN2D and LSTM scoring around 0.89–0.90. The TabNet model also did well, with an accuracy of 0.93.

After applying oversampling, the performance of ML models stayed mostly the same or improved slightly, with XGBoost and Stacking still leading at 0.97 and 0.96. DL models like CNN1D and MLP also improved noticeably. However, LSTM's accuracy slightly

decreased with oversampling, likely because LSTM is designed for sequence data rather than tabular data, which makes it less suitable in this scenario.

Table 8. Performance of ML models using the Algerian and Saudi data. Bold values indicate the highest performance for the evaluation metric.

Model	Without		Oversampling		Undersampling	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
RF	0.91	0.91	0.95	0.95	0.93	**0.93**
LR	0.90	0.90	0.86	0.86	0.87	0.87
XGBoost	0.96	**0.96**	0.97	**0.97**	0.91	0.92
SVM	0.89	0.88	0.93	0.93	0.90	0.89
CatBoost	0.95	0.95	0.95	0.95	0.91	0.92
Ensemble Classifier	0.95	0.95	0.95	0.95	0.93	**0.93**
Stacking	0.95	0.95	0.96	0.96	0.92	0.92

Table 9. Performance of DL models using the Algerian and Saudi data. Bold values indicate the highest performance for the evaluation metric.

Model	Without		Oversampling		Undersampling	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
LSTM	0.89	0.88	0.85	0.85	0.85	0.85
CNN1D	0.82	0.82	0.87	0.86	0.88	0.87
MLP	0.75	0.65	0.90	0.90	0.90	**0.89**
CNN2D	0.90	0.89	0.88	0.88	0.87	0.87
TabNet	0.93	**0.93**	0.92	**0.92**	0.88	0.88

With undersampling, some DL models like CNN1D performed even better than with oversampling, reaching 0.88 accuracy. However, LSTM's accuracy and the accuracy of some other models slightly decreased with undersampling. ML models dropped slightly in performance, but Stacking and CatBoost stayed strong with scores above 0.90. TabNet also remained stable, only dropping a little with undersampling.

Overall, ML models, especially XGBoost and CatBoost, were stable and performed well with all data balancing methods. DL models gained more from oversampling, except for LSTM, which did not benefit as much. Ensemble methods like Voting and Stacking showed consistent and reliable performance across all sampling strategies.

5.4. Conditional Tabular GAN

To enhance data diversity and address class imbalance in the combined dataset, a CTGAN was used to generate synthetic tabular samples. CTGAN is designed specifically for structured data and can generate new data points while preserving the relationships between features and class labels. In this study, CTGAN was trained on the original combined dataset and used to generate 500 additional synthetic samples, resulting in a larger and more balanced dataset of 992 "not fire" and 364 "fire" samples. The performance of several ML and DL models trained on this augmented dataset was evaluated using accuracy and F1-score. The results are presented in Table 10.

Table 10. Performance of ML and DL models using the Algerian and Saudi data with conditional tabular GAN. Bold values indicate the highest performance for the evaluation metric.

Model	Accuracy	F1
RF	0.81	0.79
XGBoost	0.85	**0.85**
CNN2D	0.81	0.78
CatBoost	0.86	**0.85**
Stacking	0.84	0.82

The models trained on the CTGAN-augmented dataset showed slightly lower performance compared to those trained on the original imbalanced data. CatBoost achieved the highest accuracy of 0.86 and an F1-score of 0.85, followed closely by XGBoost with an accuracy of 0.85 and an F1-score of 0.85. Stacking also performed well, with an accuracy of 0.84 and an F1-score of 0.82. Traditional models like RF and DL models like CNN2D reached accuracies of 0.81 and 0.81, and F1-scores of 0.79 and 0.78, respectively.

This performance reduction was mainly due to the limited number of real fire instances in the combined dataset, which contained only 210 real fire instances. This limited number of samples was not enough for CTGAN to learn the data patterns accurately. As a result, the synthetic samples added noise to the training data and affected the relationships between weather variables and FWI-derived indicators. In contrast, random oversampling kept the original data patterns and achieved more reliable performance on this dataset.

5.5. Feature Selection Results for Combined Dataset

Feature selection techniques were applied to the combined dataset to identify the most significant attributes influencing model performance. Two approaches, L1 regularization and RFECV, were employed to extract optimal feature subsets. The selected features were then applied to the best ML and DL models to evaluate their performance. The outcomes, measured using accuracy and F1-score, are summarized in Table 11.

Table 11. Performance of the ML and DL models using 2 feature-selection models, L1 and RFECV. Bold values indicate the highest performance for the evaluation metric.

Model	L1		RFECV	
	Accuracy	F1	Accuracy	F1
Machine Learning				
XGBoost With Oversampling	0.93	0.93	0.94	0.94
CatBoost	0.95	**0.95**	0.95	**0.95**
Ensemble Classifier	0.95	**0.95**	0.95	**0.95**
Deep Learning				
LSTM	0.88	0.88	0.90	0.89
CNN2D	0.91	**0.90**	0.90	0.89
TabNet	0.87	0.87	0.94	**0.94**

As shown in Table 11, the effect of feature selection differs between the models. For the ML models, XGBoost with oversampling improves slightly when using RFECV, rising from 0.93 to 0.94 in both accuracy and F1. This means that RFECV, which uses XGBoost as the estimator, was able to identify a more informative subset of features that enhanced the model's performance. CatBoost and the ensemble classifier kept the same results under

both L1 and RFECV, with 0.95 accuracy and F1, which means that these models are stable and not strongly affected by reducing the number of features.

For the DL models, LSTM improved when using RFECV, moving from 0.88 to 0.90. This indicates that selecting a smaller and more useful set of features helped the model learn more effectively. The CNN2D model changes only slightly from 0.91 to 0.90, showing that removing a small number of features does not strongly affect its performance. TabNet, however, shows a clear improvement under RFECV, moving from 0.87 to 0.94 in both metrics, showing that it benefits from selecting only the most important features.

Overall, while both L1 and RFECV feature selection methods contribute to dimensionality reduction and faster training, their impact varies by model type. RFECV shows particular advantages for sequential models such as LSTM and ensemble-based methods like XGBoost, whereas CNN-based models perform best with all features available.

5.6. Spatial Cross-Validation Analysis for the Saudi Arabian Dataset

To complement the random cross-validation experiments reported in Sections 5.1–5.5, a separate spatial cross-validation analysis was conducted for the Saudi Arabian dataset. The dataset was divided into five geographically separated groups using latitude and longitude, and each group was held out once for evaluation while the models were trained on the remaining groups. Latitude and longitude were used only to define the spatial groups and were not included as model predictors. Hyperparameter tuning and oversampling were performed only on the corresponding training data. The results are reported as mean ± standard deviation across the five spatial folds in Table 12.

Table 12. Spatial cross-validation performance on the Saudi Arabian dataset. Bold values indicate the highest performance for the evaluation metric.

Model	Without Oversampling		Oversampling	
	Accuracy	**F1**	**Accuracy**	**F1**
RF	0.8153 ± 0.1847	0.7558 ± 0.2383	0.7991 ± 0.1582	0.7604 ± 0.2290
LR	0.8262 ± 0.1773	0.7584 ± 0.2350	0.5304 ± 0.1146	0.6010 ± 0.1070
XGBoost	0.8056 ± 0.1634	0.7505 ± 0.2279	**0.7902 ± 0.1373**	**0.7611 ± 0.2001**
SVM	0.8212 ± 0.1762	0.7559 ± 0.2343	0.6479 ± 0.0627	0.7014 ± 0.0865
CatBoost	**0.8317 ± 0.1684**	**0.7782 ± 0.2173**	0.7886 ± 0.1367	**0.7653 ± 0.1860**
Stacking	0.8217 ± 0.1743	0.7613 ± 0.2339	**0.8148 ± 0.1703**	0.7526 ± 0.2314

The spatial cross-validation results were lower and more variable than the random cross-validation results, indicating that model performance is sensitive to geographic separation within the Saudi dataset. For XGBoost with oversampling, the mean spatial-fold accuracy and weighted F1-score were 0.7902 ± 0.1373 and 0.7611 ± 0.2001, respectively. The aggregated out-of-fold predictions showed a fire-class recall of 0.1096 and a false-negative rate of 0.8904, meaning that 65 of the 73 fire observations were classified as not fire (Figure 4). These findings indicate that, although the models performed well under random cross-validation, predicting fire occurrences at geographically separate locations remains challenging with the current Saudi dataset.

5.7. Critical Comparison with Existing Wildfire Prediction Approaches

Table 5 provides a numerical comparison between the Algerian-dataset results of this study and directly relevant published wildfire prediction studies. Table 13 further provides a critical comparison of Nuthur with existing wildfire prediction approaches in terms

of their objectives, data inputs, outputs, advantages, and limitations. However, numerical performance values alone do not fully describe the differences between wildfire prediction approaches because the studies use different objectives, data sources, predictors, and evaluation protocols.

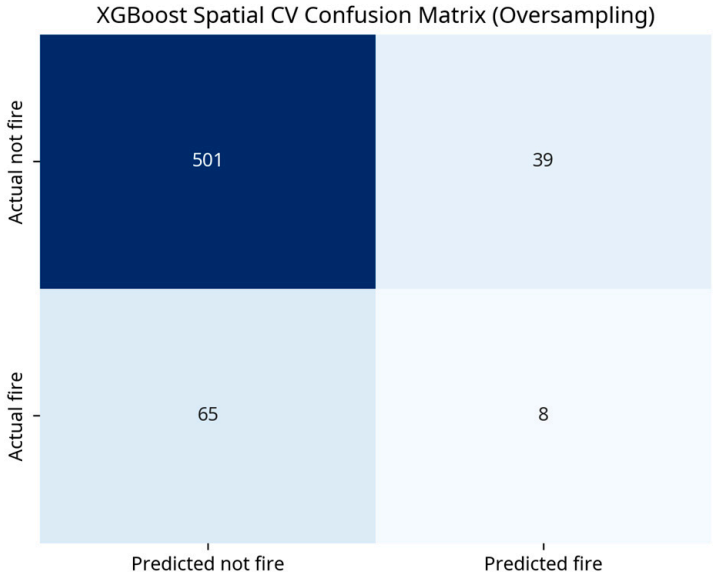


Figure 4. Confusion matrix of XGBoost under spatial five-fold cross-validation for the Saudi Arabian dataset.

Table 13. Critical comparison of Nuthur with existing wildfire prediction approaches.

Approach	Typical Data and Output	Advantage	Limitation	Nuthur Comparison
Geospatial wildfire susceptibility mapping	Satellite imagery, terrain, vegetation, land use, and human-related spatial factors; produces a long-term risk map.	Supports regional planning and identification of high-susceptibility areas.	May not provide immediate risk assessment for a specific time and location.	Nuthur supports on-demand assessment and scheduled alerts using current location-specific environmental conditions; however, it does not produce continuous high-resolution susceptibility maps.
Historical environmental-data ML classification	Historical weather and fire data; predicts the fire/no-fire class for individual records.	Can learn nonlinear relationships from tabular environmental predictors.	Can be sensitive to sample size, class imbalance, pseudo-absence labels, and geographic or temporal dataset shift.	Nuthur evaluates several ML models and integrates the workflow into an alerting platform. Its Saudi spatial results show that further local data and validation are needed.
Wildfire spread modeling	Fuel, wind, terrain, and fire-behavior variables; estimates how an existing fire may evolve over space and time.	Supports operational understanding of active-fire movement.	Requires detailed, time-resolved fuel, weather, and spatial data and is computationally demanding.	Nuthur predicts risk occurrence rather than validated fire spread; it should not be interpreted as a fire-spread simulation system.
Nuthur: current study	NASA FIRMS active-fire records, local meteorological data, Algerian benchmark data, and volunteer reports; supports risk assessment and alerts.	Combines model evaluation with an on-demand and web-based early-warning workflow for Al-Soudah.	Limited local fire records, pseudo-absence labels, non-standard event-time FWI-derived indicators, and lower spatial generalization constrain deployment readiness.	The approach is intended as a region-specific prototype. Its advantages are operational integration and local focus; its limitations are transparently reported in the revised manuscript.

Geospatial wildfire-susceptibility studies, such as Sadeghi et al. [15], use satellite imagery and spatial factors to identify high-risk areas over a longer period. These approaches support regional planning but do not provide an immediate assessment for a specific location and time. Historical environmental-data models, such as Zaidi [33], use weather and fire-related variables to classify wildfire occurrence. Nuthur differs by combining location-specific environmental assessment with scheduled monitoring, automated alerts, and volunteer reporting. However, as shown by the Saudi spatial cross-validation results in Section 5.6, further geographically and temporally diverse local data are needed before operational deployment.

6. Model Selection and Graphical User Interface Design

Model selection considered both overall classification performance and the ability to identify fire cases. Although XGBoost showed strong performance under random cross-validation, the spatial cross-validation analysis showed a low fire-class recall of 0.1096 and a false-negative rate of 0.8904. In contrast, logistic regression with oversampling achieved the highest fire-class recall of 0.5616, reducing the number of missed fire cases from 65 to 32. However, this improvement was accompanied by a higher number of false-positive alerts. As Nuthur is designed as an early-warning system, reducing missed fire events was considered an important priority. Therefore, logistic regression with oversampling was selected as a preliminary risk-screening model for the current prototype. Further geographically and temporally diverse Saudi wildfire data, together with threshold optimization, are required before confirming a final operational model.

The Nuthur interface was designed to be simple and intuitive, supporting both Arabic and English, and catering to two distinct user roles: volunteers and employees. Employees can navigate the home screen to view and manage volunteer reports, run manual weather checks, and access scheduled monitoring results, as shown in Figure 5. Volunteers, upon logging in, can submit new fire reports by describing the incident, attaching photographs, and sharing location coordinates, after which the data is automatically analyzed by the model to assess fire risk (Figures 6 and 7). Both user types can also manage their profiles and track report statuses through their respective dashboards.

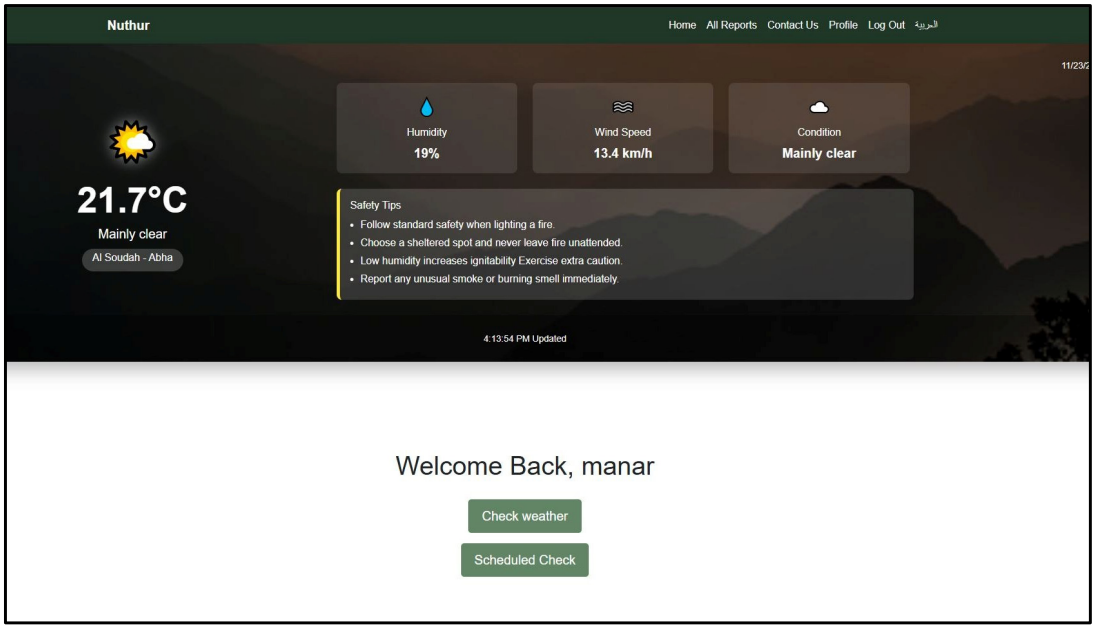


Figure 5. Employee Home page.

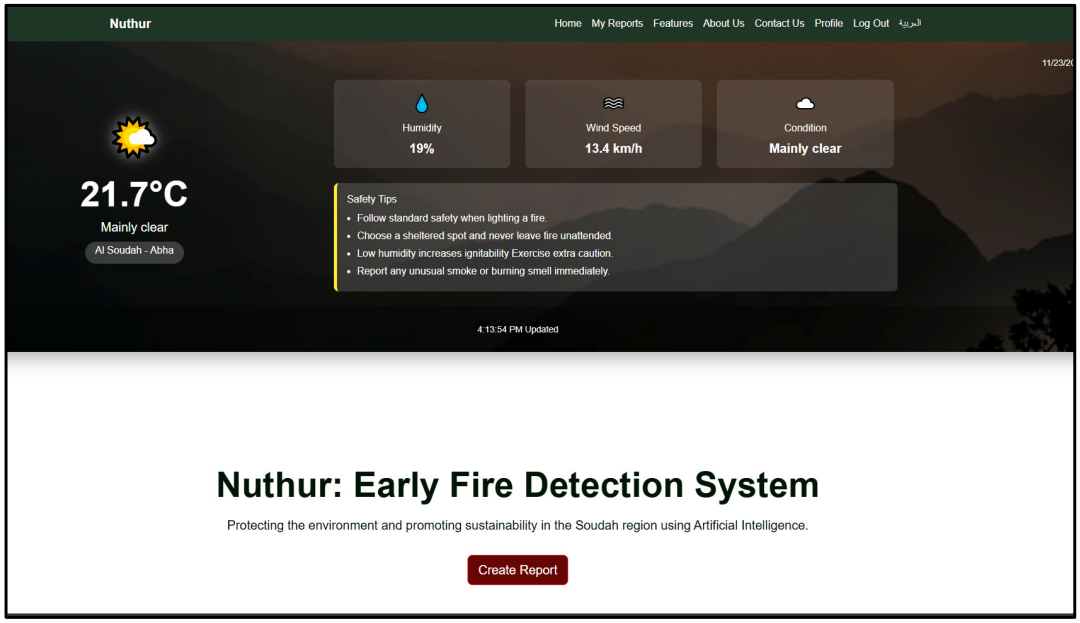


Figure 6. Volunteer Home page.

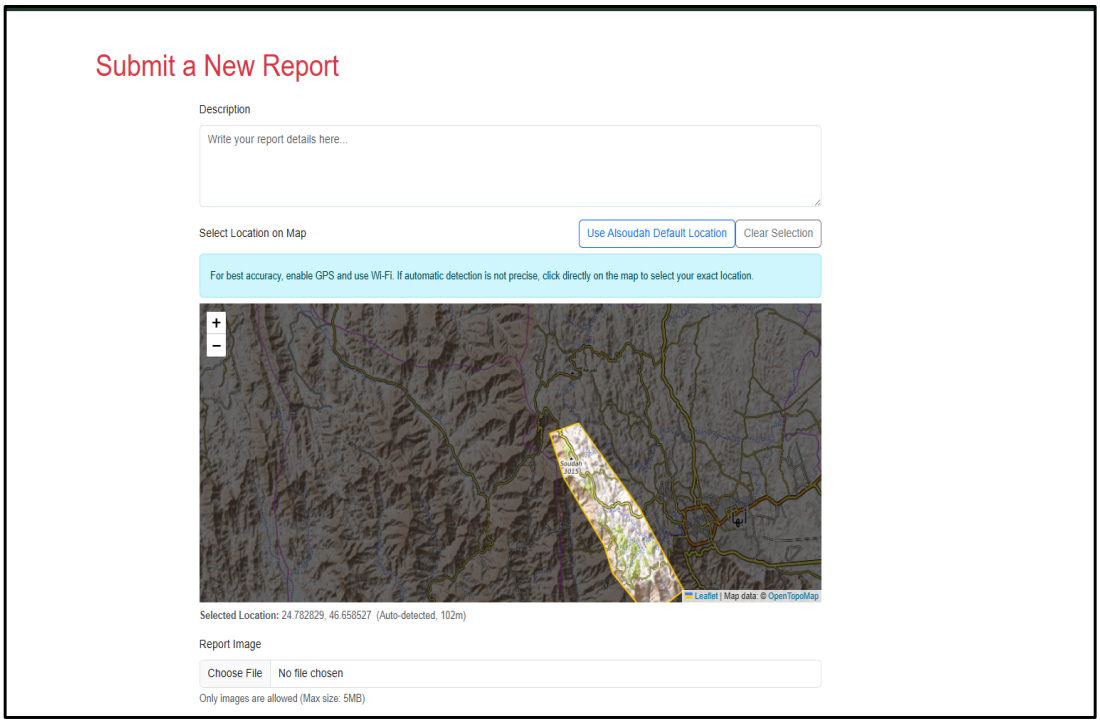


Figure 7. Create Report page.

7. Conclusions

This study presented Nuthur, an intelligent wildfire prediction and early-warning system for the Al-Soudah region of Saudi Arabia, an ecologically sensitive area characterized by dense juniper forests and rich biodiversity. The system integrates near-real-time meteorological data with historical NASA FIRMS active-fire records and the Algerian Forest Fire dataset. Feature-reduction methods and data-balancing strategies were evaluated to examine their effects on predictive performance, while CTGAN-based synthetic augmentation was assessed as an exploratory approach. Multiple ML and DL models were evaluated;

LSTM and CNN architectures were included only as exploratory benchmarks because the available data are tabular and do not represent temporal sequences or spatial grids.

Under random five-fold cross-validation, XGBoost with oversampling achieved the highest accuracy of 0.97. However, the separate Saudi spatial cross-validation analysis produced lower and more variable results, demonstrating the difficulty of generalizing to geographically separate locations. CatBoost without oversampling achieved the highest mean spatial accuracy (0.8317) and weighted F1-score (0.7782), while logistic regression with oversampling achieved the highest fire-class recall (0.5616). In contrast, XGBoost with oversampling achieved a fire-class recall of 0.1096 and missed 65 of the 73 fire records in the pooled spatial evaluation. Therefore, the 0.97 random cross-validation accuracy alone should not be interpreted as evidence of reliable operational fire-detection performance. The Nuthur web-based platform provides on-demand risk assessment, scheduled monitoring, automated alerts, and structured incident reporting for volunteers and authorities.

Several limitations should be considered. The FWI-derived indicators were calculated using meteorological observations retrieved at the satellite fire-detection time to support Nuthur's on-demand, near-real-time workflow. This differs from the standard Canadian FWI procedure, which requires daily local-noon observations, 24-h precipitation, and antecedent weather conditions. In addition, the samples with no FIRMS-detected fire are pseudo-absences rather than confirmed no-fire observations, and their sampling may introduce spatial bias. The spatial cross-validation findings further show that the current Saudi dataset does not yet support confirmation of a final operational model.

Future work will incorporate daily weather data and a rolling temporal window to calculate standard FWI components, collect geographically and temporally diverse local fire and control observations, and perform further external validation. The system can also be extended through integration with localized National Center for Meteorology APIs, a mobile application for time-critical reporting, and multimodal analysis of volunteer photographs and text reports.

Author Contributions: Conceptualization, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; methodology, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; software, R.A.A., A.F.A., S.M.A. and M.M.A.; validation, H.A.A., R.A.A., A.F.A. and M.M.A.; formal analysis, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; investigation, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; resources, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; data curation, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; writing—original draft preparation, H.A.A., R.A.A., A.F.A., S.M.A. and M.M.A.; writing—review and editing, H.A.A.; visualization, H.A.A.; supervision, H.A.A.; project administration, H.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. FAO. *Global Forest Resources Assessment 2020*; FAO: Rome, Italy, 2020. [CrossRef]
2. Al-Qthanin, R.; Aseeri, R. Assessment of Forest Fire Impact and Vegetation Recovery in the Ghalahmah Mountains, Saudi Arabia. *Fire* **2025**, *8*, 172. [CrossRef]
3. Saudi Green Initiative. Greening Saudi: Grow 10 Billion Trees Across Saudi Arabia. Available online: https://www.sgi.gov.sa/about-sgi/sgi-targets/greening-saudi (accessed on 31 August 2026).
4. Protecting Land & Sea | Saudi Green Initiative Target. Available online: https://www.sgi.gov.sa/about-sgi/sgi-targets/protecting-land-and-sea/ (accessed on 31 August 2026).
5. Nebot, À.; Mugica, F. Forest Fire Forecasting Using Fuzzy Logic Models. *Forests* **2021**, *12*, 1005. [CrossRef]

6. Sevinc, V.; Kucuk, O.; Goltas, M. A Bayesian network model for prediction and analysis of possible forest fire causes. *For. Ecol. Manag.* **2020**, *457*, 117723. [CrossRef]
7. Xie, Y.; Peng, M. Forest fire forecasting using ensemble learning approaches. *Neural Comput. Appl.* **2019**, *31*, 4541–4550. [CrossRef]
8. Honary, R.; Shelton, J.; Kavehpour, P. A Review of Technologies for the Early Detection of Wildfires. *ASME Open J. Eng.* **2025**, *4*, 040803. [CrossRef]
9. Martin, V.; Venable, K.B.; Morgan, D. Development and application of a Sentinel-2 satellite imagery dataset for deep-learning driven forest wildfire detection. *SN Comput. Sci.* **2025**, *6*, 921. [CrossRef]
10. Azami, M.; Orger, N.; Schulz, V.; Oshiro, T.; Cho, M. Earth Observation Mission of a 6U CubeSat with a 5-Meter Resolution for Wildfire Image Classification Using Convolution Neural Network Approach. *Remote Sens.* **2022**, *14*, 1874. [CrossRef]
11. Choi, J.; Ahn, J. Early wildfire detection with CubeSat images using single image super-resolution. In Proceedings of the 37th Annual Small Satellite Conference, Logan, Utah, 5–10 August 2023.
12. Devi, K.L.S.R.; Kumar, G.N.; Narayana, P.A.; Ramana, K.V.; Amarendra, K.; Gullipalli, T.R. Forest Fire Prediction and Management using AI (Artificial Intelligence), ML (Machine Learning) and Deep Learning Techniques. In *2024 8th International Conference on Inventive Systems and Control (ICISC)*; IEEE: New York, NY, USA, 2024; pp. 324–327. [CrossRef]
13. Rostami, A.; Shah-Hosseini, R.; Asgari, S.; Zarei, A.; Aghdami-Nia, M.; Homayouni, S. Active Fire Detection from Landsat-8 Imagery Using Deep Multiple Kernel Learning. *Remote Sens.* **2022**, *14*, 992. [CrossRef]
14. Aarich, M.; Rouijel, A.; Amine, A. Deep Learning Approaches for Forest Fires Detection and Prediction using satellite Images. *Procedia Comput. Sci.* **2024**, *251*, 758–763. [CrossRef]
15. Sadeghi, A.; Nadoushan, M.A.; Sani, N.A. Segment-level modeling of wildfire susceptibility in Iranian semi-arid oak forests: Unveiling the pivotal impact of human activities. *Trees For. People* **2024**, *15*, 100496. [CrossRef]
16. Hahs, B.; Sood, K.; Gomez, D. A Data-Driven Model for Wildfire Prediction in California. In *2024 International Conference on Smart Applications, Communications and Networking (SmartNets)*; IEEE: New York, NY, USA, 2024; pp. 1–6. [CrossRef]
17. Mohanty, Y.; Lee, J. A Machine Learning Approach to Predict the Occurrence of Forest Fires with Meteorological Parameters. *J. Stud. Res.* **2024**, *13*. [CrossRef]
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.; et al. Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction. *Symmetry* **2020**, *12*, 1022. [CrossRef]
19. Bengio, Y.; Lecun, Y.; Hinton, G. Deep learning for AI. *Commun. ACM* **2021**, *64*, 58–65. [CrossRef]
20. Benzekri, W.; El, A.; Moussaoui, O.; Berrajaa, M. Early Forest Fire Detection System using Wireless Sensor Network and Deep Learning. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 64. [CrossRef]
21. Priya, I.R.; George, R.; Naveen, N.; Banerjee, D.; Srija, A. Forest fire detection using wireless sensor network. *Int. Res. J. Eng. Technol.* **2021**, *8*, 3626–3630.
22. Kanand, T.; Kemper, G.; König, R.; Kemper, H. Wildfire Detection and Disaster Monitoring System Using Uas and Sensor Fusion Technologies. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *XLIII-B3-2020*, 1671–1675. [CrossRef]
23. Pereira, J.; Mendes, J.; Júnior, J.S.S.; Viegas, C.; Paulo, J.R. A Review of Genetic Algorithm Approaches for Wildfire Spread Prediction Calibration. *Mathematics* **2022**, *10*, 300. [CrossRef]
24. Radke, D.; Hessler, A.; Ellsworth, D. FireCast: Leveraging Deep Learning to Predict Wildfire Spread. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization: Macao, China, 2019; pp. 4575–4581. [CrossRef]
25. Huot, F.; Hu, R.L.; Goyal, N.; Sankar, T.; Ihme, M.; Chen, Y.-F. Next Day Wildfire Spread: A Machine Learning Dataset to Predict Wildfire Spreading From Remote-Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
26. Cortez, P.; Morais, A. Forest Fires Dataset. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/dataset/162/forest+fires (accessed on 21 September 2025).
27. Faroudja ABID. Algerian Forest Fires Datase. IEEE. Available online: https://ieee-dataport.org/documents/algerian-forest-fires-dataset (accessed on 21 September 2025).
28. NASA. NASA | LANCE | FIRMS. Available online: https://firms.modaps.eosdis.nasa.gov/ (accessed on 6 September 2025).
29. WeatherAPI. Weather API Documentation. WeatherAPI.com. Available online: https://www.weatherapi.com/docs/ (accessed on 20 December 2025).
30. Wang, Y.; Anderson, K.R.; Suddaby, R.M. Updated Source Code for Calculating Fire Danger Indices in the Canadian Forest Fire Weather Index System. 2015. Available online: https://ostrnrcan-dostrncan.canada.ca/handle/1845/243650 (accessed on 9 September 2025).
31. Arik, S.Ö.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 6679–6687. [CrossRef]

32. Developers, S.-L. Feature Selection. Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/feature_selection.html (accessed on 27 October 2025).
33. Zaidi, A. Predicting wildfires in Algerian forests using machine learning models. *Heliyon* **2023**, *9*, e18064. [CrossRef]

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.