An Investigation into the Risk Factors of Forest Fires and the Efficacy of Machine Learning Techniques for Early Detection

Asma Cherif[1]*, Sara Chaudhry[2], Sabina Akhtar[3]
Department of Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[1]
Center of Excellent in Smart Environment Research, King Abdulaziz University, Jeddah, Saudi Arabia[1]
Department of Computer Science, Bahria University, Shangrilla Road, Islamabad, Pakistan[2,3]

***Abstract*—Forest fires are a major environmental hazard that can have significant impacts on human lives. Early detection and swift action are crucial for controlling such situations and minimizing damage. However, the automatic tools based on local sensors in meteorological stations are often insufficient for detecting fires immediately. Machine learning offers a promising solution to forecast forest fires and reduce their rapid spread. In recent state-of-the-art solutions, only one or two techniques have been utilized for prediction. In this research, we investigate several methods for forest fire area prediction, including Long Short Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA), and Support Vector Regression (SVR). Our aim is to identify the most effective and optimal method for predicting forest fires. After comparing our results with other artificial intelligence and machine learning techniques applied to the same dataset, we found that the LSTM approach outperforms the ARIMA and SVR predictors by more than 92%. Our findings also indicate that the LSTM algorithm has a lower estimation error when compared to other predictors, thus providing more accurate forecasts.**

***Keywords*—Machine Learning; Forest Fire; LSTM; ARIMA; SVR**

I. Introduction

Forests play a crucial role in the earth's ecosystem and environmental sustainability. Wildland fires, commonly known as forest fires, are among the deadliest disasters that pose a threat to forest preservation, causing ecological devastation and casualties [1], [2]. In recent decades, human activities have caused the number of forest and land fires to significantly increase [3]. The 2019 wildfire in southeast Australia is an example of the devastating impact of forest fires, destroying over 11.2 million hectares of forest and leading to the extinction of numerous creatures [4], [5]. Detecting and controlling forest fires is becoming increasingly challenging. Rapid detection is essential to effective control. Detection techniques typically include smoke detection, satellite monitoring, and local perception (such as data analysis). While satellite monitoring is costly and subject to delays, smoke detection requires expensive equipment and maintenance. Data analysis, in contrast, is a less expensive and more accessible method for detecting and analysing forest fires [6], [7].

Human carelessness and natural factors like lightning are the two primary causes of forest fires. Recent studies have indicated that forest fires and climate change are related [8]. Inconsistencies in weather patterns, such as irregular rainfall, wind patterns, temperature swings, and precipitation, have contributed to an increase in forest fires in recent years [9]. The amount of money spent by the government to fight forest fires has also increased. Ecologists have employed several methods to understand how the forest landscape is changing, and statistical models have proven to be useful in examining how the patterns of forest fires are evolving for a particular location. However, current technologies cannot accurately predict the site of forest fires based on past data and environmental circumstances. Detecting forest fires early can help contain and reduce the scope of a disaster, and such solutions can aid firefighters in their efforts.

Artificial intelligence has emerged as a promising tool for predicting wildland fires with high accuracy [10], [11]. Neural networks (NN) are commonly used to forecast the occurrence of wildland fires and can reduce false alarms with the aid of infrared data [12]. Support vector machines (SVM) have also proven to be effective in increasing the accuracy and effectiveness of wildland fire predictions [13], [14]. Random Forest (RF) [15], [16] is considered to be a technique that performs quite well in these scenarios and its hybrid with SVM that is Random vector forest regression (RVFR) has also been studied to bring better results [17]. Additionally, data mining techniques such as logistic regression [18], [19], decision tree (DT) [15], and fuzzy logic [16] have been used to develop wildland fire prediction models. These techniques can help authorities detect forest fires early, allowing for more efficient and effective containment efforts.

Forest fires can cause significant losses to nature, the environment, and property. To prevent these losses, it is essential to have accurate and timely forest fire prediction systems in place. However, relying solely on automatic tools based on local sensors in meteorological stations may not be enough to instantly detect fires. Therefore, studying and selecting a suitable model for forest fire prediction can play a vital role in preventing fires from occurring or spreading. By accurately predicting forest fires, authorities can take proactive measures to ensure the safety of people, property, and the environment.

The main objective of this research is to identify the factors that contribute to the rapid spread of forest fires, including wind speed, humidity, temperature, precipitation, FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC

*Corresponding authors.

(Drought Code), and ISI (Initial Spread Index). The aim is to develop a predictive model that can assist authorities in assessing the long-term impacts of climate change. Furthermore, the study compares the outcomes with other machine learning and artificial intelligence techniques that were previously applied to the same dataset and reported in the literature. However, it is worth noting that only a limited number of techniques were utilized in the state-of-the-art solutions. Therefore, this research aims to fill this gap by applying various machine learning algorithms and selecting the best prediction model. By doing so, it can contribute to the development of a more accurate and effective forest fire prediction model that can help prevent losses due to forest fires.

This study is focused on addressing the following research questions:

- RQ1: What are the current state-of-the-art techniques and models used for predicting wild fires, as reported in the literature?
- RQ2: Which specific features or variables are most influential in accurately predicting the spread of forest fires?
- RQ3: Based on a thorough review of previous literature, what are the most suitable models that can be trained for our specific dataset?
- RQ4: What methods can be employed to improve the accuracy of the prediction model, such as reducing the root mean square error or mean absolute error?

The rest of the paper is structured in the following way: Section II provides a comprehensive literature review. Sections III and IV elaborate on the methodology and results, respectively. Finally, the paper concludes with a summary of findings and future research directions in the last section.

II. Literature Review

Numerous studies and research have been conducted to predict forest fires and minimize their damage.

Cortez et al. [14] proposed data mining algorithms, including support vector machine (SVM) and random forest, which utilized four different features (spatial, temporal, weather attributes, and FWI* component) to predict the burned area of a forest. However, this proposed solution based on SVM can only predict small fires and has a low accuracy rate for large fires.

Elshewy et al. [20] proposed three machine learning algorithms, which were applied to a dataset of 517 entries and 13 features per entry. The algorithms were tested in two scenarios: one with the entire dataset used for training and testing, and the other with 70% of the attributes used for training and 30% for testing. It was concluded that linear regression had the highest accuracy score of 0.99, surpassing the other two algorithms.

In their study, Nebot et al. [21] utilized fuzzy logic models to predict the burned area of forest fires. The authors employed two powerful fuzzy systems, the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Fuzzy Inductive Reasoning (FIR), to model burned sections of fires in Portugal. To validate their models, a 10-fold cross-validation method was used, involving model identification and validation repeated 10 times. The authors created 100 FIR and 100 ANFIS models to test the generalization performance of these hybrid fuzzy models. The FIR models exhibited the highest predictive power and the lowest MAE and RMSE errors when compared to all other models.

In their study, Al Janabi et al. [22] explored the use of soft computing algorithms for predicting forest fires. They collected over 500 entries for Montesinho Natural Park (MNP) in Portugal, which included 12 spatial and temporal parameters, a fire weather indicator, and burned area. To extract significant insights from the data, the researchers utilized Principle Component Analysis and Particle Swarm Optimization techniques. They employed five strategies simultaneously to compare and select the optimal solution. The SVM strategy with minimal estimating error was found to be the most successful approach.

Liang et al. [23] investigated the use of meteorological variables, including temperature, humidity, and precipitation, in combination with various predictive models for a fire-prone region in Alberta, Canada. The data was obtained from the National Fire Database of Canada. To estimate the size of forest fires, the researchers employed Back Propagation Neural Networks (BPNN), Recurrent Neural Networks (RNN), and Long Short Term Memory Models. The LSTM model exhibited the highest performance with an accuracy rate of 90.9%.

Zhang et al. [24] developed a Convolutional Neural Network (CNN) with a complex architecture for predicting forest fire susceptibility in Yunnan Province, China. The researchers manually assembled pictures to study the effects of various parameters. To optimize the model, they employed multicollinearity analysis and the Information Gain Ratio (IGR) approach. The model architecture was inspired by Google's AlexNet [25] and categorized photos into different classes. The CNN model exhibited a strong predictive potential with an AUC of 0.86.

Long Short-Term Memory (LSTM) algorithm has been extensively used to address time series prediction problems due to its high accuracy and speed [26]. The LSTM model is particularly advantageous in predicting trends, making it well-suited for predicting wildland fire burned areas. However, the complexity of factors and computations involved has limited the potential of the LSTM model in this domain. Therefore, further exploration and development of an enhanced LSTM model is necessary to improve the accuracy and applicability of predicting burned regions in wildland fires.

Li and Huang [27] proposed a modified model based on the Long Short-Term Memory (LSTM) network with multiple input layers and an attention mechanism module to predict the burned regions in wildland fires. The study used the Montesinho dataset and drew some interesting conclusions. First, to reduce computational complexity and interference, correlation analysis was used to identify potential related elements, and unnecessary factors were removed. This helped determine the various causes of wildfire speed. Second, a Multi-AM-LSTM

*Fire Weather Index (FWI) system was introduced in 1970s. It only required readings of four meteorological observations (i.e. temperature, relative humidity, rain and wind).

model was developed to learn effective features and predict burned areas. The proposed model achieved an impressive accuracy rate of over 96%. Comparison with other models showed that LSTM with an attention mechanism is particularly effective in this context.

George E. [11] proposed a novel mechanism for predicting forest fire risk using only meteorological data, independent of weather forecasting systems. The study utilized support vector machines to achieve a high accuracy rate of up to 96% for August in a two-class prediction of fire risk. The findings demonstrate the potential of this approach to accurately predict forest fire risk, which can aid in preventing and mitigating the damage caused by wildfires.

In their study, Richa.S et al [28] employed a variety of machine learning techniques to test eight classification models using the Cortez Morais dataset, which includes 517 instances and 13 attributes. The algorithms were evaluated using several metrics, including precision, recall, f-score, accuracy, and Area Under the Curve (AUC). The Boosted Decision Tree model achieved an accuracy of 72%, which outperformed the neural network model by 6%, the 2 Class Bayes machine by 14%, and the remaining five algorithms by 3%. These results suggest that the Boosted Decision Tree model may be a promising approach for predicting forest fires based on the Cortez Morais dataset.

Ahmed Al Janabi's study [20] utilized particle swarm optimization (PSO) to segment fire zones and principal component analysis (PCA) to identify key patterns or clusters. The study then employed five soft computing (SC) techniques based on neural networks in parallel to determine the best method for predicting forest fires with accuracy and optimality. The performance of these predictors was evaluated using five quality metrics, including information gain, relative absolute error, mean absolute error, and root mean squared error (RMSE). The results indicated that the SVM technique outperformed RBF, MPNN, PNN, and CCN predictors in terms of both effectiveness and efficiency. These findings suggest that the SVM technique may be a promising approach for predicting forest fires based on PSO and PCA segmentation.

In recent years, several projects have investigated the application of various fire detection methodologies and technologies. For instance, a study by ZQ et al. [6] presented an overview of the different forest fire detection methods that have been introduced, including wireless sensor networks, optical sensors, digital cameras, and satellite imaging. Similarly, Gibson et al. [5] explored the use of statistical methods in predicting forest fires while considering the various factors and challenges associated with it. Hong et al. [29] examined the use of different sensors, such as temperature, smoke, flames, and flammable materials, in fire detection systems.

David et al. [25] proposed an efficient machine learning method for predicting total burned areas for specific wildfire episodes with high accuracy. The transparent open box (TOB) algorithm does not rely on regression, correlation, or statistical distribution assumptions, nor does it use any hidden layers or complex calculations. This method can provide valuable insights to aid in mitigating specific burn events as they occur, resulting in various short- and long-term benefits. Furthermore, the proposed strategy can be adapted to anticipate and analyze data from other agricultural systems that rely on complex relationships between meteorological and environmental factors, in addition to regression and correlation-based approaches.

III. Methodology

This section provides a comprehensive overview of the dataset utilized in the study, along with the preprocessing steps taken to predict fire and details of the models used. First, we will discuss the dataset in detail, followed by a description of the preprocessing techniques employed to ensure the accuracy and reliability of the predictions. Finally, we will delve into the specific models utilized in this research, highlighting their strengths and limitations in predicting fire occurrences.

The workflow for our study is presented in Fig. 1. The diagram outlines the various steps taken to analyze the dataset, preprocess the data, and develop and evaluate the models used in the study. Each step in the workflow is described in detail in the subsequent sections.

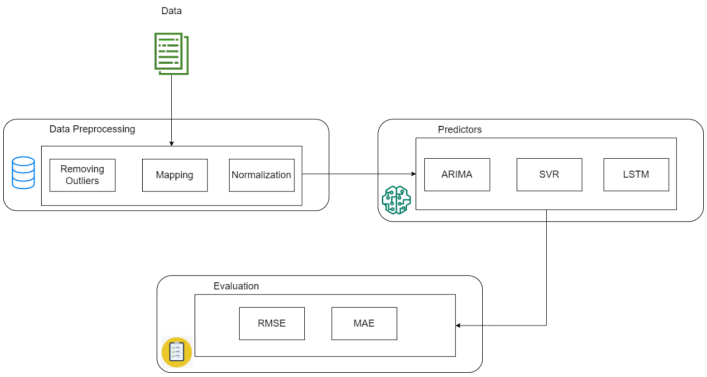


Fig. 1. Workflow process.

A. Dataset

The dataset used in this research was sourced from the UCI machine learning repository, consisting of 517 unique entries recorded at various periods between January 2003 and December 2020. The dataset comprises 12 attributes categorized into three groups: weather conditions, fire weather index (FWI), and geographical and temporal components, with the overall burned area serving as the output feature. The data was collected from two sources, as described in [7]. The first source recorded details such as date, time, location within a 9-by-9 grid, vegetation type, FWI elements, and burned area each time a forest fire occurred. The FWI, a Canadian system for categorizing fire hazard that was used to indicate the intensity of the fire. The second source consisted of weather measurements taken every 30 minutes by a meteorological station. The two databases were combined to create a single dataset. Table I provides a detailed explanation of each attribute.

B. Data Exploration

The dataset contains 517 rows of data. To gain insights into the distribution of the numerical variables, we created boxplots for each variable. Upon analysis, we identified the presence of outliers in some of the variables. The boxplots of these variables are presented in Fig. 2.

TABLE I. DESCRIPTION OF DATASET

Attributes	Description
Spatial Attribute	
X	Coordinates on x-axsis (1-9)
Y	Coordinates on y-axsis (1-9)
Temporal Attributes	
Month	January to December months start
Day	Day starts with Monday and ends on Sunday
Intensity of Weather Attributes	
FFMC	FFMS code in digits
DMC	DMC code in digits
DC	DC code in digits
ISI	ISI code in digits
Weather Attributes	
Temperature	Temperature in degree Celsius
RH	RH in percentage
Rain	Rain in mm
Wind	Relative humidity in percentage
Target	
Area	Total area burnt in hectare

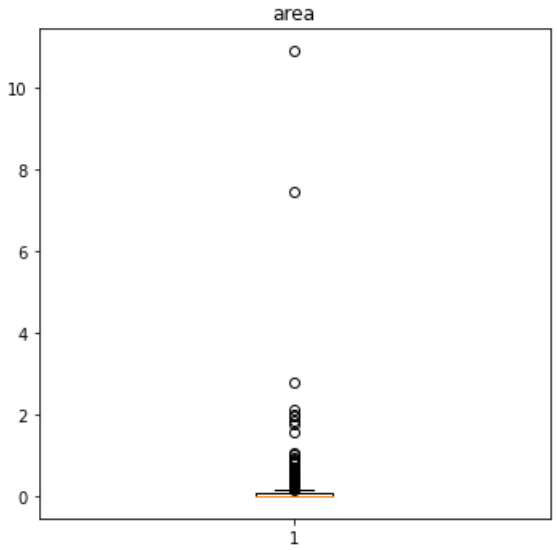


Fig. 2. BoxPlot of area.

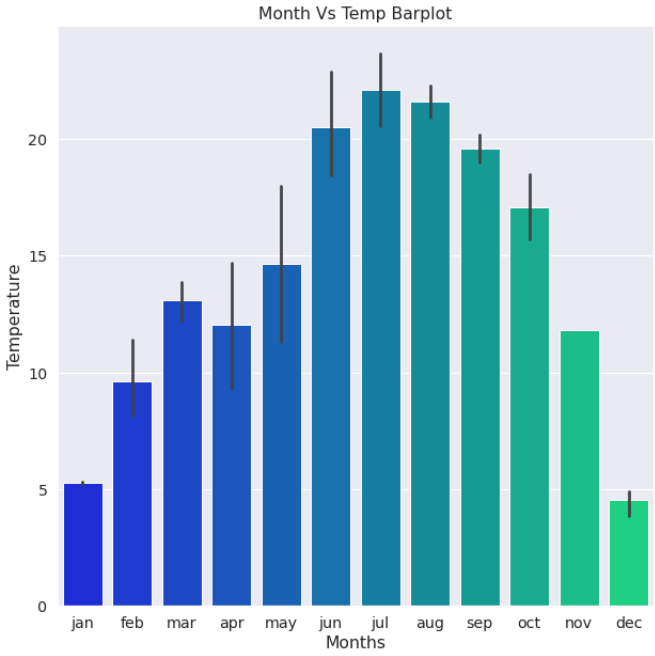


Fig. 3. Month vs Temperature.

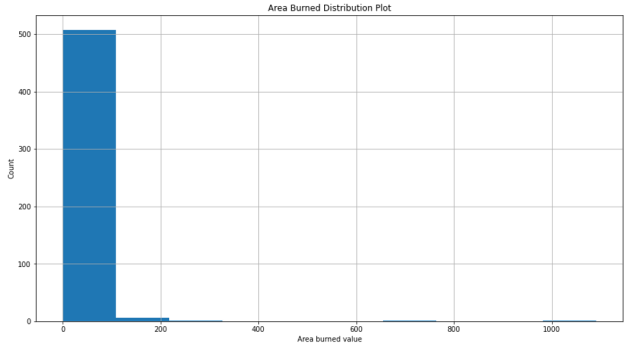


Fig. 4. Area distribution.

Our target variable is the "Area" variable which also has a significant number of outliers. However, we have decided to remove only the top 10% of outliers from the right side of the data. After removing these outliers, we were left with a total of 515 rows, down from the initial 517 rows.

The graph in Fig. 3 highlights that the months of June, July, August and September experience the highest temperatures. This phenomenon is primarily due to the summer season, resulting in high temperatures and increased humidity levels. As September concludes, the temperature starts to decline gradually.

1) Target transformation: The variable "area" has a highly skewed distribution that is positively skewed towards 0, as shown in Fig. 4. Modeling such a skewed distribution during training impacts the model performance. Ideally, the variable should have a normal distribution. To transform "area" into a normal distribution range, we performed a log transformation as shown in Fig. 5.

2) Mapping: To apply different models to our dataset, we needed to convert our categorical variables into numeric variables. Specifically, we mapped the "month" attribute to its corresponding month number and the "day" attribute to its corresponding day number, with Sunday being the first day. Additionally, we utilized one-hot encoding to apply LSTM to our dataset.

3) Normalization: The feature values in the dataset exhibited a wide range, potentially leading to increased computing complexity and inaccurate predictions. Thus, normalization is necessary to assign equal weight to each variable. This involves utilizing the Min-Max Scaling technique to quantify the variables, linearly transforming the original variable ranges to new ranges.

$$Z = \frac{x_i - min(x)_i}{max(x)_i - min(x)_i} \quad (1)$$

In Eq. 1, Z is the output value, x_i is the value of the

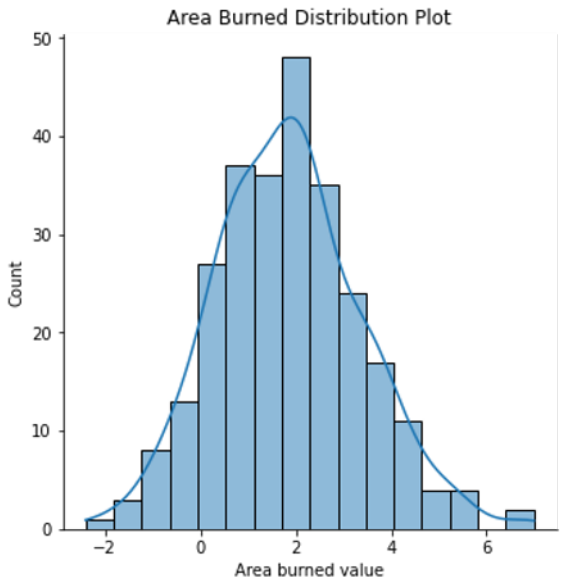


Fig. 5. Transformed area distribution.

variables, max and min denote the minimum and maximum values of each variable, respectively.

C. Feature Selection

Feature selection is a technique employed to improve accuracy in the machine learning process. Additionally, it enhances the predictive ability of algorithms by choosing the most essential variables while excluding redundant and irrelevant ones. In this research we applied two techniques: the Mutual Information (MI) and correlation analysis.

D. Mutual Information

MI is a measure of the amount of information that one random variable can provide about another random variable. It is often used for dimension reduction [30].

A formal way to express the mutual information between two random variables X and Y is as follows:

$$I(X,Y) = H(X) - H(X \mid Y) \tag{2}$$

Where $I(X,Y)$ is the mutual information for X and Y, $H(X)$ is the entropy for X and $H(X \mid Y)$ is the conditional entropy for X given Y.

Fig. 6 displays the results of applying the MI to our features and target. It indicates that the month has a great mutual information with the target feature "area". Besides, the temp attribute is showing some mutual information with the target. Thus, the events of fire have a relation with temperature.

E. Correlation Analysis

Correlation analysis is used to check which are the most important variables in the data set. The correlation between these characteristics is applied using the following equation:

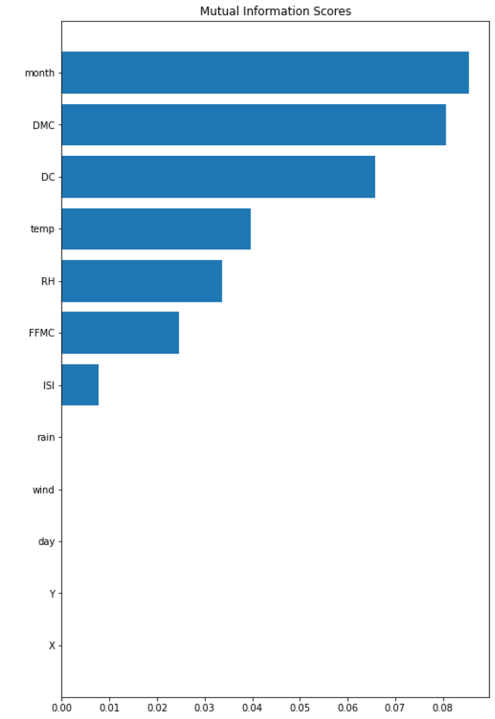


Fig. 6. The Mutual inclusion between the features and the target.

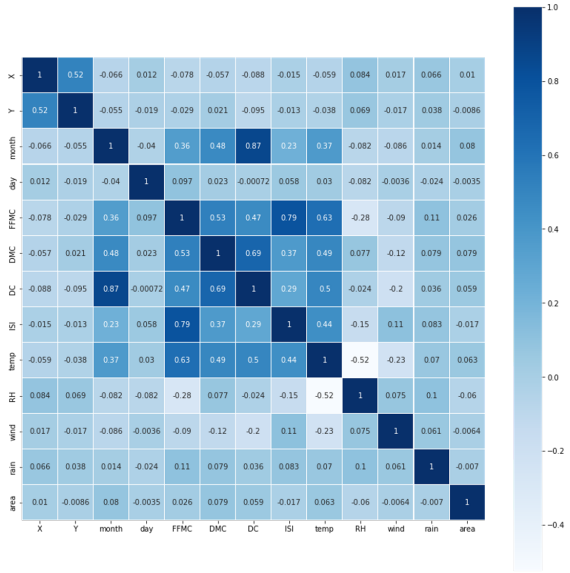


Fig. 7. Correlation analysis.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 (\sum x)^2][n\sum y^2 (\sum y)^2]}} \tag{3}$$

where n represents total observation and $-1 < r < 1$, the positive sign $(+)$ means positive correlation and negative sign $(-)$ means negative relationship between variables.

Fig. 7 indicates that 5 features have some correlation with the target which are: Month, DC, DMC, RH, Temp, ISI.

F. Predictors

In the following discussion, we will examine the predictors utilized for forecasting the area.

1) Long short term memory: Particular types of recurrent neural networks can be trained to identify long-term dependencies in data. As such, we utilized an LSTM network to address prediction problems using a dataset sourced from UCI Machine Learning respiratory. The LSTM network incorporates a cell state through which information is passed, and the addition or removal of information is regulated by input gates, forget gates, and output gates. To gain insight into the functioning of the LSTM network and how the gates acquire pertinent information, refer to Fig. 8[†].

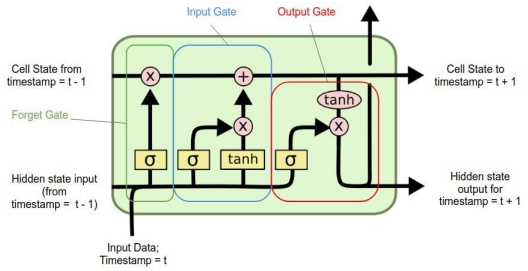


Fig. 8. LSTM network.

a) Forget gate: The forget gate plays a crucial role in determining which information should be kept or removed from the cell state. This gate processes both current input and relevant information using a sigmoid function, resulting in values that fall between 0 and 1. A value of 1 indicates that the information is to be retained, while a value of 0 indicates that it should be forgotten.

b) Input gate: The input gate is responsible for deciding which data to preserve in the cell state. This gate processes both input and hidden state information through a hyperbolic tangent (tanh) function. The resulting output is multiplied by a sigmoid function to determine whether to keep or discard the information. If the sigmoid output approaches 1, the information is retained, whereas a sigmoid output approaching 0 results in the information being discarded.

c) Output gate: The output gate plays a critical role in determining the final output of the cell state. This gate filters the cell state output to produce a refined version of the information that should be included in the final output.

2) ARIMA: ARIMA, short for AutoRegressive Integrated Moving Average, is a widely-used model in time series data analysis. It utilizes historical values of a time series to forecast future values. However, external factors can also have an impact on the time series and serve as accurate predictors of future values. This is where the SARIMAX model comes into play, by introducing an exogenous variable X. In statistics, predictors or input variables are referred to as exogenous, while the target variable is referred to as endogenous. The SARIMAX model enhances the $SARIMA(p,d,q)(P,D,Q)_m$ model by incorporating the effects of exogenous variables, producing more accurate forecasts [31]. The parameters p,d,q are trend elements which refer to autoregressive order, difference order and moving average order respectively. Similarly, the parameters P,D,Q are seasonal elements refering to autoregressive order, difference order and moving average order. The parameter m indicates seasonal length.

In Eq. 4, we may easily add any number of exogenous variables X_t to the $SARIMA(p,d,q)(P,D,Q)_m$ model to reflect the present value Y_t.

$$Y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i x_t^i \qquad (4)$$

Eq. 4 indicates that the SARIMA model is a linear combination of the historical values of a time series and the corresponding error terms. The SARIMAX model, an extension of SARIMA, incorporates exogenous variables, transforming it into a more comprehensive linear model. This extension allows to model the relationship between the time series and external factors in a more accurate and effective manner, thereby yielding improved forecasting results.

3) Support Vector Regression SVR: Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) algorithm, designed to handle regression problems in both linear and nonlinear domains. While SVM is primarily used for classification tasks with categorical variables, SVR is tailored for continuous variable prediction. This fundamental distinction enables SVR to model complex relationships between variables and generate accurate predictions from continuous datasets.

Despite a few minor differences, SVR operates on the same fundamental principles as SVM. Both algorithms aim to locate a curve that fits the given data points. However, as a regression procedure, SVR utilizes this curve to match the vector to the point of the curve, rather than as a decision boundary. This matching process is facilitated by the use of support vectors, which helps to identify the closest match between the data points and the function used to represent them.

IV. Results

This section presents the results of three models that were developed to predict forest fire areas based on historical data. By analyzing the performance of each model, we can gain insights into the effectiveness of their respective methodologies and techniques. Through these evaluations, we aim to identify the most accurate and reliable model that can be used to predict forest fire areas with the highest level of confidence.

A. LSTM

LSTM has been widely used to address time series prediction problems due to its high accuracy and speed. Here, we explain the development of the prediction model using LSTM.

[†]Figure from: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

1) Data Splitting: After preprocessing our data for use in LSTM, we were left with 514 rows and 30 features out of a total of 517 rows. To predict the fire area for the next day, we implemented a window length of 10. During the training phase, we utilized 80% of data, incorporating all features, and reserved 20% rows for testing. This model will be used to forecast the fire area for the next 38 days.

2) Model Training: Several models have been trained with different parameters as shown in Table II.

TABLE II. TRAINED LSTM MODELS

Model No.	Layers	1^{st} Layer Neurons	2^{nd} Layer Neurons	3^{rd}Layer Neurons
1	1	100		
2	1	50		
3	2	50	55	
4	3	100	100	55

All models utilize the ReLU activation function. Additionally, we have included a dense layer with eight neurons, as illustrated in Fig. 9. To optimize our model, we employed the Adam optimizer with default settings for the learning rate (lr=0.001) and weight decay (0.9). During training, we used a batch size of 25 and conducted 50 epochs to ensure that our models are able to learn from the data effectively.

Fig. 9 illustrates the second model summary, which consists of a total of 78,368 parameters, all of which have been trained.

```
Model: "Forest_fire_Prediction_Model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 LSTM_Hidden_Layer_1 (LSTM)  (None, 10, 100)           43600

 LSTM_Hidden_Layer_2 (LSTM)  (None, 55)                34320

 output_layer (Dense)        (None, 8)                 448

=================================================================
Total params: 78,368
Trainable params: 78,368
Non-trainable params: 0
_________________________________________________________________
```

Fig. 9. Model 2 summary.

To fine-tune the hyperparameters, we utilized the validation set. The hyperparameters for the proposed bidirectional LSTM models are presented in Table III.

TABLE III. HYPERPARAMETER OF LSTM

Hyperparameter	Values
Activation Function	ReLU
Number of LSTM layer	1,2,3,4
Dropout	Yes
Regularization	Yes
Optimizer	Adam
Learning rate	0.001
Batch size	25
Epochs	50

Loss indicates the performance of the model whether it is able to predict accurately or not. If loss is zero then it is accurate. Fig. 10 displays the loss with respect to the number of epochs for all the trained models. As shown in the figure, the loss decreases as the number of epochs increases. Besides, Models 3 and 4 demonstrate a better convergence between the training and validation losses.

Fig. 11 explains the ground truth or the actual area versus the predicted area for the models. We have predicted occurrence of forest fire area for thirty eight (38) days.

B. ARIMA

ARIMA is an extensively used model for analysis of time series data. It utilizes historical values of a time series to forecast future values. In this section, we explain how we prepared the data for creating a model using ARIMA and explain our findings.

1) Data Splitting: After preprocessing, we were left with 515 out of 517 rows and 13 features for ARIMA analysis. We trained the model on 477 days of data using all features, and reserved 38 rows for testing. The ARIMA model forecasted the area of fire for the next 38 days, as illustrated in Fig. 12.

2) SARIMAX summary: The SARIMAX model is utilized when an exogenous variable 'X' is included. In statistics, the term *exogenous* is used to describe predictors or input variables, while *endogenous* refers to the target variable- what we aim to predict. For ARIMA training, we determine the values of (P, D, Q). To determine these values for our dataset, we used the auto arima built-in function from the pdarima library, resulting in P=0, Q=1, and D=4. We also attempted to predict forest fire occurrences for the next 38 days, and Fig. 13 displays the SARIMAX model's results.

C. SVR

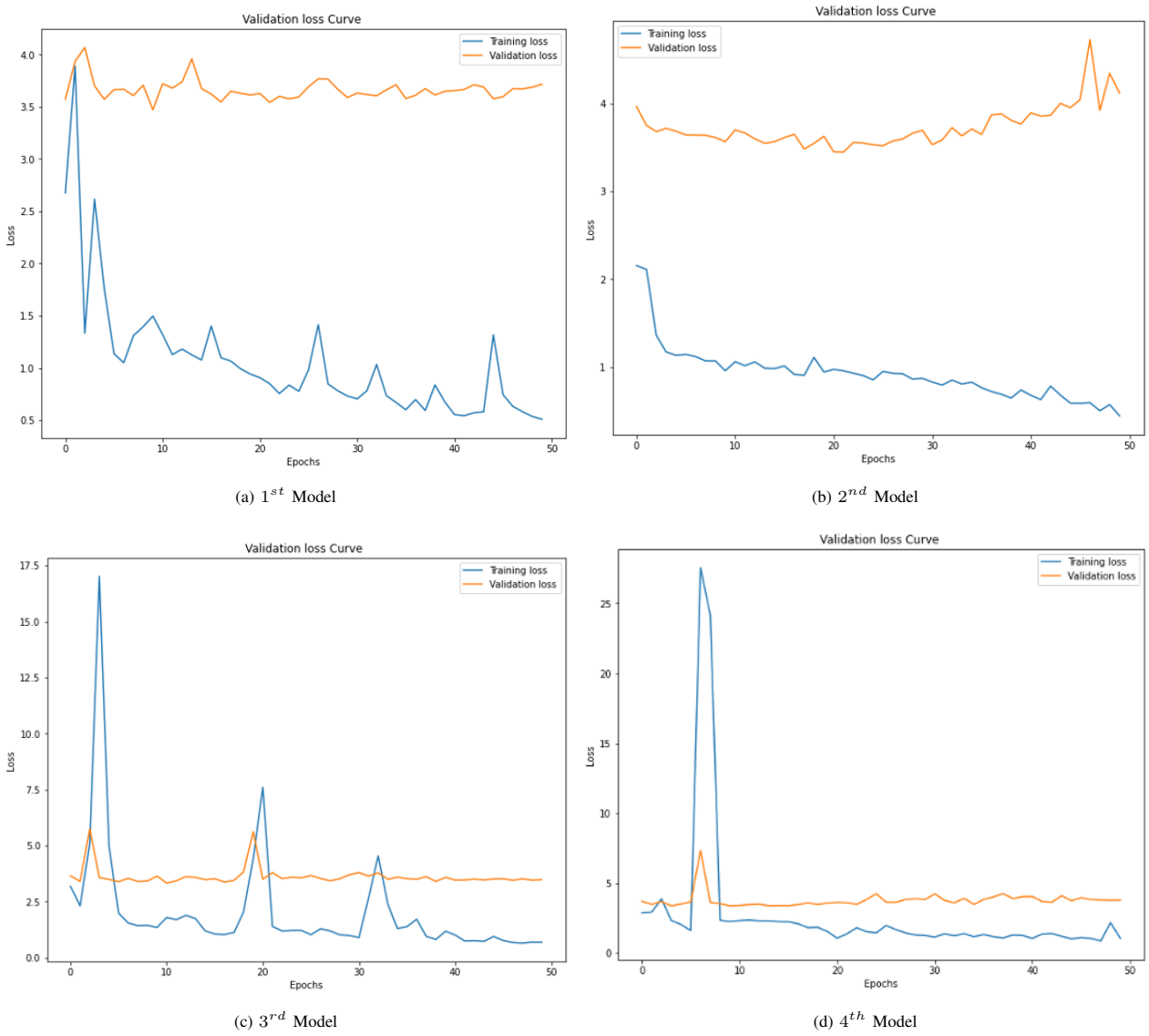
In SVR, we began with a total of 517 rows. After preprocessing, we were left with 515 rows and 13 features. For training, we used 477 days of data, including all features. The remaining 38 rows were used for testing. The model is designed to forecast the area of fire for the next 38 days. We utilized a linear kernel with C=1 and gamma of 2e-5 for the SVR. The SVR model forecasted the area of fire for the next 38 days, as illustrated in Fig. 14.

D. Overall Comparison of Models

To assess the model's effectiveness, we employ the RMSE (root mean square error) metric defined in Eq. 5. These error measurements are widely used in the field of data analysis and are valuable tools for determining the accuracy of a model's predictions. By analyzing the RMSE, we can gain insight into the model's strengths and weaknesses and identify areas for improvement.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2} \tag{5}$$

Based on the data presented in Table IV, it is evident that the ARIMA and SVR models have the highest RMSE error. Therefore, we can conclude with confidence that LSTM can be utilized for predicting fires in this dataset or other similar datasets. Indeed, it outperforms the two other method by more than 92%.



(a) 1^{st} Model

(b) 2^{nd} Model

(c) 3^{rd} Model

(d) 4^{th} Model

Fig. 10. Training loss for 4 LSTM models.

TABLE IV. COMPARING THE PERFORMANCE OF THE TESTED MODELS

Model	RMSE
ARIMA	16.59
LSTM 1	1.219
LSTM 2	1.233
LSTM 3	**1.15**
LSTM 4	1.16
SVR	18.82

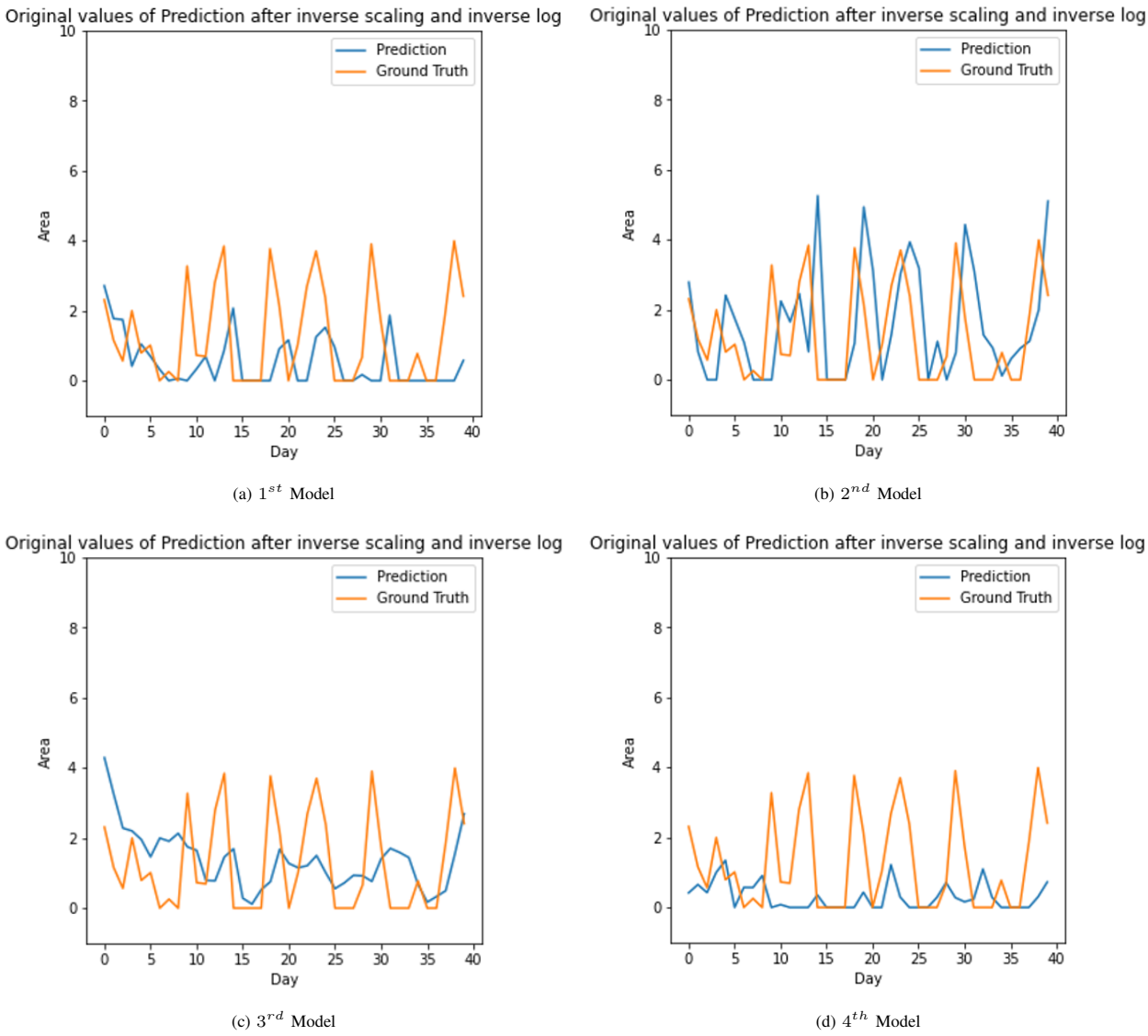
E. Discussion

To determine the predictive accuracy of artificial intelligence and machine-learning models, it is essential to compare their outcomes with those of other similar models. This is why this section showcases the results of previously published research that was conducted on the same datasets. For this comparison, we use the RMSE error, which is depicted in Eq. 5. By comparing these errors with those of other models, we can assess the effectiveness of our own model.

Several models have been applied and evaluated on the same datasets in the literature [7], [22], [21], [32]. Table V summarizes the assessment metric results of these models, along with the errors produced by LSTM, SVR, and ARIMA (shown at the bottom of the table). These datasets have shown promising results for both statistical models and other machine learning techniques. In this study, we have utilized comparable data from previous research to evaluate the performance of our models.

V. CONCLUSION AND FUTURE WORK

The aim of this study was to create an effective machine learning algorithm for predicting the spread of fire in hazardous incidents. This would enable workers to take immediate action, thereby reducing economic and natural losses. The study utilized data on forest fires from Portugal's Montesinho Natural



(a) 1^{st} Model

(b) 2^{nd} Model

(c) 3^{rd} Model

(d) 4^{th} Model

Fig. 11. Prediction of the 4 LSTM models.

Park, which is accessible through the UCI machine-learning repository. After thorough research, the study identified the most suitable machine learning and statistical models, including Long Short Term Memory (LSTM), Auto Regressive Integrated Moving Average (ARIMA), and Support Vector Regression (SVR). The study evaluated the models using the Root Mean Square Error (RMSE), and found that LSTM was the most effective model, producing the best results with an RMSE of 1.15.

The prediction results of LSTM, SVR, and ARIMA were compared to those published in the literature [7], [22], [21], [32]. The methodologies used in these papers included various types and structures of neural networks (NN), Support Vector Machines (SVMs), Decision Trees (DTs), Multiple Regression (MR), Random Forest (RF), Transparent Open-Box (TOB) Network, Radial Basis Function (RBF), Standard Genetic Programming Decision Trees (ST-GP), Linear Regression (LR), Adaptive Neuro-Fuzzy Inference System (ANFIS), Probabilistic Neural Networks (PNN), and Finite Impulse Response (FIR). Using the same data, the comparison aimed to evaluate the performance of the aforementioned methods and to identify the most accurate model for prediction.

The results indicate that the top-performing LSTM models exhibited the lowest RMSE error, as well as the highest predictive accuracy compared to all other models. To enhance accuracy in the future, we can make necessary parameter adjustments and augment data sources by incorporating various factors, such as forest vegetation, cover, types of trees, and Buildup Index. By integrating geographic information system (GIS) data and satellite imagery, we can further optimize this model and achieve even greater precision. This approach can significantly enhance our ability to analyze and understand forest dynamics and provide valuable insights for forest management and conservation efforts.

TABLE V. COMPARISON OF DIFFERENT METHODS

Method	Reference	Parameters	MSE
DT	Cortez et al. [7]	Reduction of the sum of squares	64.5
SVM	Al-Janabi et al. [22]	kernel RBF; C = 84.17; E = 0.001; G = 3800.3	54
RBF	Al-Janabi et al. [22]	Num. Neur. = 100; Min. Rad. = 0.01; Max. Rad. = 519.669; Min. Lambda = 0.01328; Max.Lambda = 9.953	54.2
TOB	David et al. [32]	Two stages; Wn = 0.54; Q = 10; Evol 2-6; Optimum	63.26
ANFIS	Angela Nebot et al. [21]	Hybrid algorithm constant funtion; 50 epochs	64.6
PNN	Al-Janabi et al. [22]	Gaussian kernel; p1 = 31.31; p2 = -27.31; p3 = 1.14; p4 = 6.31	63.2
FIR	Angela Nebot et al. [21]	EWP-EFP; 2-3 FS per variable	48.9
ARIMA	Our Research	P=0, q=1 ,d=4	16.59
SVR	Our Research	Kernal= linear, C=1, Gamma= 2e-5	18.82
LSTM	Our Research	Epoch=50, Batch Size= 25	1.15

SARIMAX Results

Dep. Variable:		y	No. Observations:		477	
Model:	SARIMAX(0, 1, 4)		Log Likelihood		-1.673	
Date:	Sun, 03 Jul 2022		AIC		31.346	
Time:	18:34:04		BIC		89.662	
Sample:	0		HQIC		54.277	
	- 477					
Covariance Type:	opg					
	coef	std err	z	P>|z|	[0.025	0.975]
X	0.0068	0.009	0.743	0.458	-0.011	0.025
Y	-0.0307	0.013	-2.424	0.015	-0.056	-0.006
month	0.0142	0.023	0.609	0.542	-0.031	0.060
day	-0.0031	0.006	-0.482	0.630	-0.016	0.009
DC	-0.0001	0.000	-0.638	0.524	-0.000	0.000
ISI	-0.0047	0.004	-1.074	0.283	-0.013	0.004
temp	0.0064	0.004	1.539	0.124	-0.002	0.015
wind	0.0108	0.011	0.999	0.318	-0.010	0.032
rain	-0.0967	0.143	-0.675	0.500	-0.378	0.184
ma.L1	-0.6290	0.019	-32.575	0.000	-0.667	-0.591
ma.L2	-0.0179	0.031	-0.580	0.562	-0.079	0.043
ma.L3	-0.0725	0.059	-1.230	0.219	-0.188	0.043
ma.L4	-0.0465	0.058	-0.800	0.424	-0.160	0.067
sigma2	0.0622	0.002	40.627	0.000	0.059	0.065

Ljung-Box (L1) (Q):	0.23	Jarque-Bera (JB):	41216.98
Prob(Q):	0.63	Prob(JB):	0.00
Heteroskedasticity (H):	44.06	Skew:	3.65
Prob(H) (two-sided):	0.00	Kurtosis:	48.00

Fig. 13. Sarimax summary.

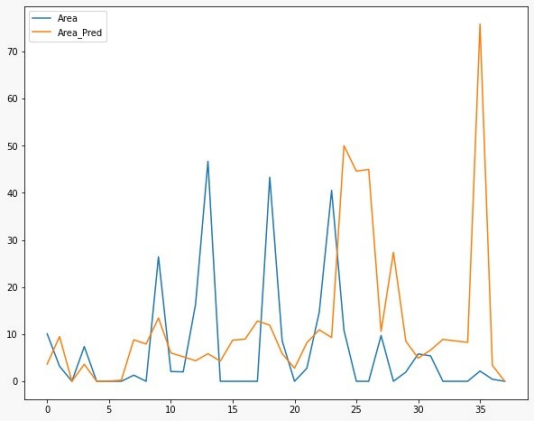


Fig. 14. Actual Area vs Predicted area.

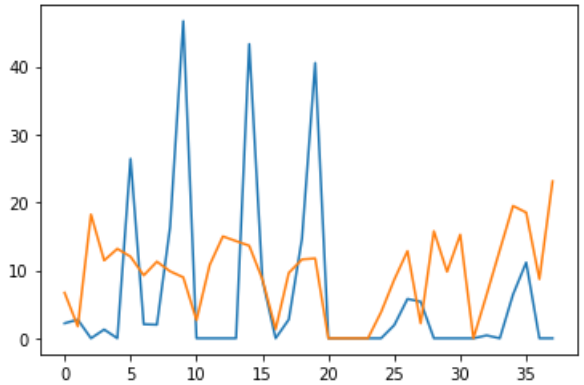


Fig. 12. Area vs Predicted area.

ACKNOWLEDGMENT

This research work was funded by Institutional Fund Projects under grant no. (FPIP:1564-612-1443). The authors gratefully acknowledge technical and financial support provided by the Ministry of Education and King Abdulaziz University, Deanship of Scientific Research, Jeddah, Saudi Arabia.

REFERENCES

[1] D. G. Neary, K. C. Ryan, and L. F. DeBano, "Wildland fire in ecosystems: effects of fire on soils and water," *U.S. Department of Agriculture, Forest Service*, vol. 4, p. 250, 2005.

[2] K. Nyongesa and H. Vacik, "Fire management in mount kenya: A case study of gathiuru forest station," 06 2018.

[3] "Mono-temporal and multi-temporal approaches for burnt area detection using sentinel-2 satellite imagery (a case study of rokan hilir regency, indonesia)," *Ecological Informatics*, vol. 69, p. 101677, 2022.

[4] M. Storey, O. Price, J. Sharples, and R. Bradstock, "Drivers of long-distance spotting during wildfires in south-eastern australia." *International Journal of Wildland Fire*, vol. 29, pp. 459–472, 2020.

[5] R. Gibson, T. Danaher, W. Hehir, and L. Collins, "A remote sensing approach to mapping fire severity in south-eastern australia using sentinel 2 and random forest," *Remote Sensing of Environment*, vol. 240, p. 13, 2020.

[6] Z. Xu, X. Su, and Y. Zhang, "Forest fire prediction based on support vector machine," *Chinese Agricultural Science Bulletin*, vol. 28, pp. 126–131, 2012.

[7] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data." pp. 512–523, 2007.

[8] J. E. Halofsky, D. L. Peterson, and B. J. Harvey, "Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the pacific northwest, usa," *Fire Ecology*, vol. 16, no. 1, p. 4, 2020.

[9] M. Flannigan, B. Stocks, and B. Wotton, "Climate change and forest fires," *Science of The Total Environment*, vol. 262, pp. 221–229, 2000.

[10] D. J. Arrue BC, Ollero A, "An intelligent system for false alarm reduction in infrared forest-fire detection." *IEEE Intell Syst Appl*, vol. 3, pp. 64–73, 2000.

[11] G. Sakr, I. Elhajj, G. Mitri, and U. Wejinya, "Artificial intelligence for forest fire prediction," *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, pp. 1311–1316, 07 2010.

[12] D.-T. Bui, B. Pradhan, H. Nampak, Q.-T. Bui, Q-A.Tran, and Q.-P. Nguyen, "Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using gis," *Journal of Hydrology*, vol. 540, pp. 317–330, 2016.

[13] R. S. Aakash, M. Nishanth, R. Rajageethan, R. Rao, and R. Ezhilarasie, "Data mining approach to predict forest fire using fog computing," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1582–1587, 2018.

[14] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data." pp. 512–523, 2007.

[15] "Predicting late-successional fire refugia pre-dating european settlement in the wenatchee mountains," *Forest Ecology and Management*, vol. 95, no. 1, pp. 63–77, 1997.

[16] H. reza Pourghasemi, M. Beheshtirad, and B. Pradhan, "A comparative assessment of prediction capabilities of modified analytical hierarchy process (m-ahp) and mamdani fuzzy logic models using netcad-gis for forest fire susceptibility mapping," *Geomatics, Natural Hazards and Risk*, vol. 7, no. 2, pp. 861–885, 2016.

[17] R. S. Bhadoria, M. K. Pandey, and P. Kundu, "Rvfr: Random vector forest regression model for integrated and enhanced approach in forest fires predictions," *Ecological Informatics*, vol. 66, p. 101471, 2021.

[18] R. S. Aakash, M. Nishanth, R. Rajageethan, R. Rao, and R. Ezhilarasie, "Data mining approach to predict forest fire using fog computing," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1582–1587, 2018.

[19] Z. Wu, H. HS, J. Yang, Z. Liu, and Y.Liang., "Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of northeastern china," *Sci Total Environ*, vol. 493, pp. 472–80, 2014.

[20] A. Elshewey and A. Elsonbaty, "Forest fires detection using machine learning techniques," *Xi'an Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture and Technology*, vol. XII, p. 2020, 10 2020.

[21] A. Nebot and F. Mugica, "Forest fire forecasting using fuzzy logic models," *Forests*, vol. 12.

[22] S. Al-Janabi, I. Al-Shourbaji, and M. A. Salman, "Assessing the suitability of soft computing approaches for forest fires prediction," *Applied Computing and Informatics*, vol. 14, no. 2, pp. 214–224, 2018.

[23] H. Liang, M. Zhang, and H. Wang, "A neural network model for wildfire scale prediction using meteorological factors," *IEEE Access*, vol. 7, pp. 1–1, 12 2019.

[24] Z. Guoli, W. Ming, and K. Liu, "Forest fire susceptibility modeling using a convolutional neural network for yunnan province of china," *International Journal of Disaster Risk Science*, vol. 10, pp. 386–406, 2019.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 25, 2012.

[26] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[27] Z. Li, Y. Huang, X. Li, and L. Xu, "Wildland fire burned areas prediction using long short-term memory neural network with attention mechanism." vol. 57.

[28] R. Sharma, S. Rani, , and Memon, "Intelligence approaches smart approach for fire prediction under uncertain conditions using machine learning," *Multimed Tools Appl*, vol. 79, 2020.

[29] H. Hong, M. Panahi, A. Shirzadi, T. Ma, J. Liu, A. Zhu, W. Chen, I. Kougias, and N. Kazakis, "Flood susceptibility assessment in hengfeng area coupling adaptive neurofuzzy inference system with genetic algorithm and differential evolution," *Science of The Total Environment*, vol. 621, pp. 1124–1141, 2018.

[30] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: a comprehensive study," *Journal of Cryptology*, vol. 24, no. 2, pp. 269–291, Apr 2011. [Online]. Available: https://doi.org/10.1007/s00145-010-9084-8

[31] N. Arunraj, D. Ahrens, and M. Fernandes, "Application of sarimax model to forecast daily sales in food retail industry," *International Journal of Operations Research and Information Systems*, vol. 7, pp. 1–21, 04 2016.

[32] D. A. Wood, "Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight," *Artificial Intelligence in Agriculture*, vol. 5, pp. 24–42, 2021.