**A COMPARATIVE STUDY OF FOREST FIRE PREDICTION USING MACHINE LEARNING MODELS**

by

Sachin Malego st125171@ait.asia  
 Sila N Mahoot st125127@ait.asia

A Proposal Submitted in Fulfillment of the Requirements for   
AT82.01 Computer Programming for Data Science and Artificial Intelligence Course   
by Dr. Chantri Polprasert

Asian Institute of Technology

School of Engineering and Technology

Department of Information and Communication Technologies

Thailand

October 2024

ABSTRACT

Forest fires are one of the most pressing incidents that have a significant impact on both the environment and human life, making early prediction is crucial for minimizing the damage. This proposal presents a comparative study of various machine learning models for forest fire prediction leveraging the historical data from various dataset. Techniques such as XGBoost, Decision Tree, Random Forest, and Linear and Logistic Regression are explored, with models evaluated based on accuracy and other performance metrics. The study highlights the strength and weakness of each approach with geographic maps of the country (Portugal and Algeria), demonstrating that machine learning can significantly enhance predictive capabilities, providing a valuable tool for fire prevention and disaster management. This project does not determine if a forest fire will take place or not in any place, however we are predicting the confidence probability of the forest fire based on some attributes.

CONTENTS

**Page**

[ABSTRACT 2](#_Toc179533190)

[CONTENTS 3](#_Toc179533191)

[LIST OF FIGURES 4](#_Toc179533192)

[CHAPTER 1 INTRODUCTION 5](#_Toc179533193)

[1.1 Background 5](#_Toc179533194)

[1.2 Objectives 7](#_Toc179533195)

[1.3 Business Understanding 8](#_Toc179533196)

[1.4 Impact 8](#_Toc179533197)

[CHAPTER 2 PROBLEM STATEMENT 9](#_Toc179533198)

[CHAPTER 3 RELATED WORKS 10](#_Toc179533199)

[CHAPTER 4 DATASETS 11](#_Toc179533200)

[4.1 Description 11](#_Toc179533201)

[4.2 Features 12](#_Toc179533202)

[CHAPTER 5 METHODOLOGY 13](#_Toc179533203)

[5.1 Data Acquisition 13](#_Toc179533204)

[5.2 Exploratory Data Analysis 13](#_Toc179533205)

[5.3 Pre-processing 18](#_Toc179533206)

[5.4 Modeling 19](#_Toc179533207)

[5.5 Training 19](#_Toc179533208)

[5.6 Evaluation 20](#_Toc179533209)

[5.7 Deployment 20](#_Toc179533210)

[CHAPTER 6 PRELIMINARY RESULTS 23](#_Toc179533211)

[REFERENCES 24](#_Toc179533212)

LIST OF FIGURES

**Figures** **Page**

[Figure 1 Understanding El Niño 6](#_Toc179533213)

[Figure 2 Area of Interest 11](#_Toc179533214)

[Figure 3 Density Plot of Algerian Fire Dataset 14](#_Toc179533215)

[Figure 4 El Niño Occurrence and Average Temperature 15](#_Toc179533216)

[Figure 5 Multicollinearity Check in Algerian Fire Dataset 15](#_Toc179533217)

[Figure 6 Monthly Fire Analysis of Algerian Fire Dataset 16](#_Toc179533218)

[Figure 7 Temperature Data Representation of Algerian Fire Data 16](#_Toc179533219)

[Figure 8 Rainfall in Millimeters 17](#_Toc179533220)

[Figure 9 Wind Speed in Kilometer Per Hour 17](#_Toc179533221)

[Figure 10 Percentage of Relative Humidity 18](#_Toc179533222)

[Figure 11 K-Fold Cross Validation 19](#_Toc179533223)

[Figure 12 Web Application Deployment 21](#_Toc179533224)

# INTRODUCTION

With the high incident rates of forest fires, its prediction holds a significant environmental and scientific importance (Sharma & Khanal, 2024). In recent years, forest fires have emerged as a critical concern, posing a recurring threat to vast forested areas worldwide. Observing the data from Nepal alone, it has experienced 965 forest fire incidents in the last one year (08-10-2023 to 07-10-2024) with an estimated economic loss of around USD 644,696. Such is not only the scenario of Nepal but at a global level. Here in this project, we review such dataset of forest fire from Forest fire dataset of Algeria, and Portugal and run different models to make the best possible prediction.

## Background

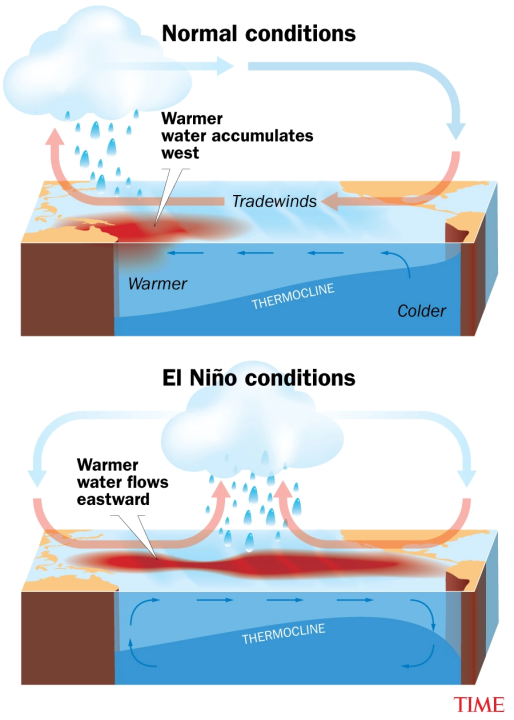
Forest fire forecasting is critical for mitigating the environmental, economic, and public health impacts of wildfires. According to existing literature, forest fire prediction methods can be broadly categorized into three main approaches: physics-based models, statistical models, and machine learning models (Zaidi, 2023). Physics-based models rely on simulating fire behavior based on physical laws, such as fluid dynamics and thermodynamics. Statistical models focus on identifying correlations between past fire events and contributing factors, often using regression techniques. However, machine learning (ML) models have recently gained prominence due to their ability to learn complex, non-linear relationships from large datasets, making them highly suitable for dynamic and uncertain phenomena like forest fires.

This project leverages data-driven algorithms in the machine learning domain to predict the likelihood of forest fire occurrences. Our goal is to train ML models on historical weather and environmental data and use these models to predict fire risks based on current conditions. The data, sourced from regions in Portugal and Algeria, contains key variables such as temperature, relative humidity, wind speed, and climate events like El Niño. These variables are integral to capturing the complex interplay of factors that contribute to fire ignition and spread.

Before delving into the technicalities of forest fire prediction, it's crucial to explore the climatic conditions that exacerbate wildfire risks. As depicted in Figure 1, El Niño is a particularly significant factor, characterized by the abnormal warming of sea surface temperatures in the central and eastern Pacific Ocean. Although El Niño is a Pacific-based event, it has far-reaching implications for global weather patterns, leading to drier or wetter conditions in various regions, including Europe and North Africa. In both Portugal and Algeria, El Niño can disrupt typical weather patterns, resulting in reduced rainfall, prolonged dry spells, and increased temperatures – conditions that are conducive to the ignition and spread of forest fires.

In addition to El Niño, several other environmental and climatic variables influence fire risk, including rainfall patterns, wind speed, relative humidity, fine fuel moisture, duff moisture, and drought conditions. These factors, when integrated into machine learning models, can provide a powerful predictive tool for early fire detection and risk assessment.

Figure   
El Niño Conditions



## Objectives

The objective of forest fire prediction is to improve early detection and risk assessment, a crucial need globally. Despite the experience of forest departments, human limitations in processing multiple variables hinder accurate predictions. Hence in such situations Machine learning can analyze numerous factors simultaneously, making fire prediction more efficient and effective. The major objectives of this project are:

1. **Evaluate Model Performance**

Compare the accuracy, precision, recall, F1 score and overall predictive performance of various machine learning models in predicting forest fires.

1. **Identify Key Features**

Analyze the importance of various environmental and weather-related factors (e.g., temperature, humidity, wind speed, El Niño) across different models to identify which features contribute most to fire prediction accuracy.

1. **Assess Generalization Capability**

Investigate how well each machine learning model generalizes to unseen data, particularly when predicting forest fires in different geographic regions or under varying climate conditions.

1. **Model Interpretability and Usability**

Evaluate the interpretability of each model, determining how easily stakeholders (e.g., forest management agencies) can understand and utilize the model's predictions for decision-making.

1. **Predictive Alerts for Health Precautions**

Evaluate how machine learning models can provide early fire risk alerts, enabling timely health interventions, such as air quality warnings or evacuation notices, to protect vulnerable populations (e.g., children, elderly, people with pre-existing conditions).

1. **Support Sustainable Forest Management**

Explore how machine learning predictions can aid in sustainable forest management by allowing for preemptive actions like controlled burns or forest thinning to mitigate the severity of future fires, reducing long-term environmental damage.

## Business Understanding

With climate change intensifying conditions such as droughts, and unpredictable weather patterns, forest fires are becoming more frequent and destructive. Effective prediction is crucial but challenging due to the complexity of interacting factors.

This project will focus on forest fire prediction in regions prone to fires, specifically Portugal and Algeria. It will utilize weather and environmental data from these regions and compare multiple machine learning models to assess their performance in predicting fire occurrences. The project will evaluate the impact of different variables such as temperature, humidity, wind, and the presence of El Niño on fire risk prediction.

## Impact

The proposed study on forest fire prediction using machine learning models has the potential to generate significant and far-reaching impacts across multiple sectors. These impacts can be categorized as environmental, social, economic, technological, policy-related, and scientific, all contributing to a more informed and effective approach to forest fire management.

The comparative study of forest fire prediction using machine learning models has the potential to significantly impact environmental conservation, public safety, economic stability, technological progress, policy development, and scientific advancement. By addressing a critical global challenge, this research will contribute to a safer and more resilient world in the face of increasing forest fire risks due to climate change.

# PROBLEM STATEMENT

The increasing frequency and intensity of forest fires due to climate change pose significant threats to ecosystems, human health, and economies worldwide. To address this pressing issue, there is a need to develop accurate predictive models that can forecast forest fire occurrences based on historical data and environmental conditions. This study aims to investigate the relationships between various factors contributing to forest fires, including:

* **Temporal Trends**: Analyzing seasonal and monthly trends in forest fire occurrences to identify patterns and peak risk periods.
* **Correlations Between Environmental Parameters**: Examining the correlations between key weather parameters (e.g., temperature, humidity, wind speed) and the incidence of forest fires to understand how these factors influence fire risk.
* **Impact of Climate Phenomena**: Assessing the effects of significant climate phenomena, such as El Niño, on forest fire occurrences, providing insights into how broader climatic trends affect fire behavior.
* **Comparative Analysis of Machine Learning Models**: Evaluating the performance of different machine learning algorithms in predicting forest fire risk, determining which model offers the best accuracy and reliability for real-time decision-making.
* **Visualization**: Visualize impact areas in maps for precision and effective decision making.

By comprehensively analyzing these elements, the project seeks to enhance forest fire prediction capabilities, enabling more effective prevention and response strategies to mitigate the environmental and societal impacts of forest fires. The main goal of this project is to create a reliable machine learning model that can predict the chances of forest fires based on weather and environmental data. By using data from Portugal and Algeria, this project aims to solve the problems with traditional prediction methods and provide a tool to help authorities make better decisions for preventing and managing fires. The challenge is to make a model that can handle different weather patterns and conditions, ultimately helping reduce the damage caused by forest fires.

# RELATED WORKS

The project goal is to train the model to predict forest fires. The focus is on using different factors along with the El Niño climate phenomenon to generalize and predict fire occurrence scenarios for better decision-making.

Jing et al., in the research paper “Toward a more resilient Thailand: Developing a machine learning-powered forest fire warning system” (Tang, et al., 2023), describes a machine learning-based forest fire warning system in Thailand. It uses satellite data and gas measurements to predict forest fires, with models like linear classifiers, gradient boosting classifiers, and neural networks. The XGBoost model had the best performance with an accuracy of 99.6%. Our project aims to improve this approach by adding climate factors like El Niño.

Abdelhamid Zaidi, in the paper “Predicting wildfires in Algerian forests using machine learning models” (Zaidi, 2023), focuses on developing a predictive system for forest fires in Algeria, a region that has seen an increase in fire occurrences over recent years. The study aims to create an accessible, low-budget system for predicting wildfires based on climatic data and machine learning algorithms. The author utilizes machine learning algorithms to analyze weather data and predict potential fire occurrences. The study concludes that a well-designed system could significantly benefit local authorities and the public by enabling early fire detection and facilitating timely interventions.

In another paper from IEEE, Viswa et al., “Comparison of Forest Fire Prediction System using Machine Learning Algorithms” (Bharathi & PeddaReddy, 2023), presents a comparative analysis of various machine learning algorithms for predicting forest fires. It emphasizes the importance of early detection in minimizing the devastating effects of wildfires on ecosystems and human settlements. The study evaluates the performance of different algorithms based on accuracy, precision, and recall. The conclusion suggests future research directions, including integrating additional data sources and advanced algorithms.

# DATASETS

## Description

In this project, we are using three main datasets from Kaggle and other public sources. These datasets represent key information about forest fire occurrences in Algeria and Portugal which both are bordering the Atlantic Ocean (see Figure 2). We are also adding weather and climate data to provide more context.

* Algerian Forest Fires Dataset
* Portugal Forest Fires Dataset
* Weather and Environmental Data
* El Niño Data (Sea Surface Temperature)

Figure   
Area of InterestA map of europe with different countries/regions

Description automatically generated

## Features

The dataset contains following features:

1. Algerian Forest Fires Dataset:
   1. Date: The date of the observation.
   2. Temperature: Daily average temperature in degrees Celsius.
   3. Humidity: Daily average relative humidity as a percentage.
   4. Wind Speed: Daily average wind speed in km/h.
   5. Rainfall: Daily rainfall in mm.
   6. Region: The specific region in Algeria where the data was collected.
   7. Fire Occurrence: Binary indicator (Yes/No) for whether a fire occurred.
2. Portugal Forest Fires Dataset:
   1. Date: The date of the observation.
   2. Temperature: Daily average temperature in degrees Celsius.
   3. Humidity: Daily average relative humidity as a percentage.
   4. Wind Speed: Daily average wind speed in km/h.
   5. Rainfall: Daily rainfall in mm.
   6. FFMC (Fine Fuel Moisture Code): Index representing the moisture content of surface litter.
   7. DMC (Duff Moisture Code): Index indicating the moisture content of loosely compacted organic layers.
   8. DC (Drought Code): Index indicating long-term drying conditions.
   9. ISI (Initial Spread Index): Index for predicting the rate of fire spread.
   10. Fire Occurrence: Binary indicator (Yes/No) for whether a fire occurred.
3. Weather and Environmental Data:
   1. Temperature: Daily average temperature in degrees Celsius.
   2. Humidity: Daily average relative humidity as a percentage.
   3. Wind Speed: Daily average wind speed in km/h.
   4. Precipitation: Daily rainfall in mm.
4. El Niño Data (Sea Surface Temperature):
   1. Niño 3.4 Index: Represents sea surface temperature anomalies in the central equatorial Pacific, which is a critical indicator of El Niño events.

# METHODOLOGY

## Data Acquisition

The data for this project was collected from publicly available sources on Kaggle and other weather data platforms. The main datasets are the Algerian Forest Fires Dataset and the Forest Fires Dataset - Portugal, which have valuable information on past forest fires and weather conditions. The data is provided in CSV format, making it easy to load and work with using Python libraries like Pandas.

We also got weather data for Algeria and Portugal from Meteostat and Weather Underground. Meteostat provides historical weather metrics, while Weather Underground allows us to gather more recent daily weather data using web scraping. This weather data is very important for understanding the factors that contribute to forest fires.

To understand the role of climate events like El Niño, we also used El Niño sea surface temperature data from the Climate Prediction Center (CPC). This helps us see how larger climate patterns might influence forest fire occurrences. All the data was downloaded or accessed through APIs, then integrated for further exploration and analysis.

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) helps us understand the patterns and relationships in our data. For this project, we are doing EDA on the Algerian and Portugal datasets, as well as weather and El Niño data, to find trends that might predict forest fires.

The following steps and best practices will be applied during the EDA process:

1. Data Cleaning

We will check for missing values, mistakes, and inconsistencies. Missing data will either be filled in or removed based on its importance.

1. Data Transformation

We will turn date features into more useful information, like year or month, to help find seasonal trends.

1. Data Visualization

We will use different types of plots to understand the data better:

* 1. Histograms or Density Plots show how individual features are distributed. The example is shown in Figure 3.
  2. Line Plots to show trends over time, like changes in temperature and humidity. The example is shown in Figure 4.
  3. Correlation Heat Maps to find which features are most important for predicting fires. The example is shown in Figure 5.
  4. Bar Plots show the frequency of fire occurrence. The example are shown in Figure 6, Figure 7, Figure 8, Figure 9, and Figure 10.
  5. Box Plots to find and deal with outliers in the data.

EDA will give us important insights that will help us prepare the data, choose features, and build better models.

Figure   
Density Plot of Algerian Fire Dataset

A group of blue and white graphs

Description automatically generated

Figure   
El Niño Occurrence and Average Temperature

A green line graph on a white background

Description automatically generated

Figure   
Multicollinearity Check in Algerian Fire Dataset

A screenshot of a chart

Description automatically generated

Figure   
Monthly Fire Analysis of Algerian Fire Dataset

A graph of a fire analysis

Description automatically generated

Figure   
Temperature Data Representation of Algerian Fire Data

A graph of different colored bars

Description automatically generated

Figure   
Rainfall in Millimeters

A screenshot of a graph

Description automatically generated

Figure   
Wind Speed in Kilometer Per Hour

A graph of different colored bars

Description automatically generated

Figure   
Percentage of Relative Humidity

A graph of different colored lines

Description automatically generated

## Pre-processing

Data preprocessing is important for making sure our datasets are clean and ready for training. Here are the steps we will take:

1. Handling Missing Values: Missing values will be filled in or removed, depending on their impact.
2. Encoding Categorical Variables: We will turn categorical features, like region, into numerical data using one-hot or label encoding.
3. Feature Scaling: We will normalize or standardize features like temperature and wind speed to make sure they're on a similar scale.
4. Date Feature Extraction: Dates will be transformed to get useful information like month or day of the week.
5. Handling Outliers: We will find and handle outliers, so they don't affect our model badly.
6. Data Integration: We will combine all datasets—fire data, weather data, and El Niño data—into one comprehensive dataset for modeling.
7. Feature Engineering: We will create new features that could improve the model's performance, like combining temperature, humidity, and wind speed.
8. Data Balancing: Since forest fires are rare, we will balance the dataset so the model can learn well from both fire and non-fire cases.
9. Splitting Data: The data will be split into training, validation, and test sets to make sure our model performs well on new data.

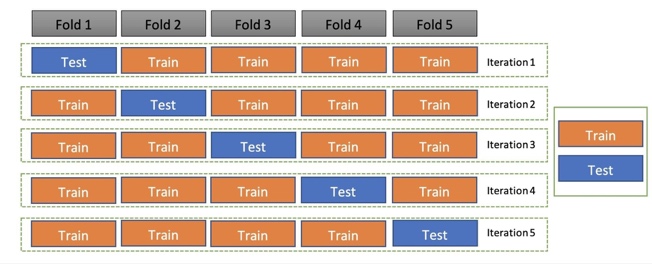
These preprocessing steps will help make sure our data is suitable for training a reliable machine learning model.

## Modeling

To find the best model for predicting forest fires, we will use cross-validation and grid search to compare different machine learning algorithms and tune their settings. Cross-validation will help make sure the models can work well on new data, and grid search will help find the best settings for each model. Figure 11 illustrates the iteration process of k-fold cross validation. We will then compare these models based on accuracy, precision, recall, and F1 score, then choose the best one for predicting forest fires.

1. Random Forest and XGBoost are good options because they handle many features and complex relationships well.
2. Gradient Boosting is also strong for modeling interactions between features.
3. Regression models will be used as a simple baseline model to compare with the others.

Figure   
K-Fold Cross Validation



## Training

Training the model means using the training data to teach the model how to predict forest fires. We will use the scikit-learn library to train our models. The model will learn by finding the relationships between input features, like temperature and wind speed, and the target feature, fire occurrence.

Steps in Training:

1. Model Initialization: Start the model with the best settings found in grid search.
2. Fitting the Model: Train the model with the .fit() method.
3. Monitoring Performance: Keep track of how well the model is learning by checking the loss function and metrics like accuracy.
4. Addressing Overfitting: Use methods like early stopping to prevent the model from overfitting.

After training, we will evaluate the model with the validation set. If needed, we will adjust the settings to improve performance, then test the final model with new data.

## Evaluation

We will evaluate the model to make sure it accurately predicts forest fires. We will use these metrics:

* Accuracy: How often the model makes the correct prediction.
* Precision: How often the model correctly predicts a fire when it says there is one.
* Recall: How many actual fires the model can predict.
* F1 Score: A balance between precision and recall.

These metrics will help us understand how well the model works and what areas need improvement.

## Deployment

The forest fire prediction model will be deployed using Azure Machine Learning for hosting and Dash for the web-based interface. Azure Machine Learning provides a safe and scalable way to host machine learning models. The deployment steps are illustrated in Figure 12.

Figure   
Web Application Deployment

A screenshot of a computer

Description automatically generated

Here are the steps for deployment:

1. Create Prediction Model and Web Application Locally
   1. Develop ML Model (model.py): Build the model in a Python script named model.py.
   2. Convert Model to Deployable Format: Save the trained model using pickle or joblib to prepare it for deployment.
   3. Create Dash Web Application (app.py): Make a Dash web app that lets users input data and see predictions, including maps showing fire risk.
2. Upload Code to GitHub: The relevant codes, including the model.py and app.py scripts, along with dependencies, will be uploaded to GitHub for version control and collaboration.
3. Azure Machine Learning Setup
   1. Create Azure Account: Use Azure to access Machine Learning services.
   2. Register Model on Azure: Register the model in Azure to manage it easily.
   3. Create an Inference Endpoint: Create an endpoint that the Dash app can use to get predictions from the model.
   4. Containerize the Model: Use Docker to containerize the model, making it easier to deploy.
4. Deploy Dash Application on Azure Web App
   1. Create Azure Web App Service: Use Azure App Service to host the Dash app.
   2. Configure Azure App Service: Connect the web app to the model endpoint.
   3. Clone Repository and Deploy: Clone the GitHub repository to Azure and deploy the app.
5. Testing and Monitoring
   1. Test the Deployment: Make sure the deployed app works well and interacts properly with the model.
   2. Monitor the Application: Use Azure's monitoring tools to track performance and fix any issues.

By using Azure Machine Learning and Dash, we will make the forest fire prediction system accessible online. This system will help users input data and see predictions in real time, including maps that show areas at risk of forest fires.

# PRELIMINARY RESULTS

This chapter presents preliminary results of the analysis conducted on various environmental factors that influence forest fire occurrence. The findings provide insights into the conditions under which forest fires are most likely to occur, helping to identify critical thresholds for different parameters that contribute to fire risk.

* Temperature: The highest fire counts occurred between 30-37 degrees Celsius.
* Rain: The highest fire counts occurred when there was no rain or very little rain, i.e., between 0.0 to 0.3.
* Wind Speed: The highest fire counts occurred when the wind speed was between 13 to 19 km/hr.
* Relative Humidity: The highest fire counts occurred when the relative humidity (RH) was between 50 to 80%.
* Fine Fuel Moisture Code (FFMC) index, which ranges between 28.6 to 92.5, shows that above 75 there is a higher chance of forest fires.
* Duff Moisture Code (DMC) index, which ranges between 1.1 to 65.9, indicates that values between 1.1-10 have a lower chance of forest fires, whereas above 10-30 DMC has very high evidence of forest fires in the past.
* Drought Code (DC) index, which ranges between 7 to 220.4, shows that values between 0-25 are safe and have a lower chance of forest fires, whereas values above 25 DC have a higher chance of forest fires.
* Initial Spread Index (ISI) index, which ranges between 0 to 18, shows that values between 0-3 have lower fire chances, and values above 3 ISI have a higher chance of forest fires.
* Buildup Index (BUI) index, which ranges between 1.1 to 68, shows that values between 1.1 to 10 have a lower fire chance, and values above 10 BUI have a higher chance of forest fires.
* Fire Weather Index (FWI) index, which ranges between 1 to 31.1, shows that values between 0-3 have a lower chance of forest fires, and values between 3-25 FWI have a higher chance of forest fires.

REFERENCES

Bharathi, V., & PeddaReddy, C. (2023). Comparison of Forest Fire Prediction System using Machine Learning Alogrithms. *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1539-1542). IEEE. doi:https://doi.org/10.1109/ICACITE57410.2023.10182818

Sharma, S., & Khanal, P. (2024). Forest Fire Prediction: A Spatial Machine Learning and Neural Network Approach. *Fire, 7*(6), 205. doi:https://doi.org/10.3390/fire7060205

Tang, J., Weeramongkolkul, M., Suwankesawong, S., Saengtabtim, K., Leelawat, N., & Wongwailikhit, K. (2023). Toward a More Resilient Thailand: Developing a Machine Learning-Powered Forest Fire Warning System. *Heliyon, 10*(13). doi:https://doi.org/10.1016/j.heliyon.2024.e34021

Zaidi, A. (2023). Predicting Wildfires in Algerian Forests using Machine Learning Models. *Heliyon, 9*(7). doi:https://doi.org/10.1016/j.heliyon.2023.e18064