

PREDICTIVE MODELING OF FOREST FIRES USING MACHINE LEARNING

**Thesis Submitted
in Partial Fulfillment of the Requirements for the Degree of**

**MASTER OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING
by**

**SATYENDRA YADAV
(23/CSE/22)**

Under the Supervision of

Dr. ANURAG GOEL
Assistant Professor, Department of Computer Science and Engineering
Delhi Technological University



Department of Computer Science and Engineering

**DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042**

MAY, 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGMENT

I wish to express my sincerest gratitude to **Dr. Anurag Goel** for his continuous guidance and mentorship that he provided during research work. He showed me the path to achieving targets by explaining all the tasks to be done and explained to me the importance of this work as well as its industrial relevance. He was always ready to help me and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this work would not have been successful.

Place: Delhi

SATYENDRA YADAV

Date:

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Satyendra Yadav, 23/CSE/22, of M.Tech. (CSE)**, hereby certify that the work which is being presented in the thesis entitled “**Predictive Modeling of Forest Fires using Machine Learning**” in partial fulfillment of the requirement for the award of the degree of Master of Technology, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from to under the supervision of **Dr. Anurag Goel**. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **Satyendra Yadav (23/CSE/22)** has carried out their research work presented in this thesis entitled “**Predictive Modeling of Forest Fires using Machine Learning**” for the award of **Master of Technology** from Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

(Dr. ANURAG GOEL)

(Assistant Professor)

(Department of Computer Science and Engineering)

(Delhi Technological University)

Date:

Abstract

Forest fires are among the worst natural disasters, causing extensive damage and loss to various forms of life, including humans and infrastructure. In recent years, these catastrophes have been occurring more frequently and without warning, highlighting the urgent need for more intelligent systems to predict and manage the impacts of climate change.

This thesis presents a method based on machine learning that analyzes meteorological and environmental data to determine the likelihood of forest fires. The study utilizes techniques derived from Artificial Neural Networks (ANNs), incorporating geographic, weather, and temporal data to create a reliable prediction tool. The dataset used contains a total of 518 instances with variable features such as temperature, wind speed, humidity, rainfall, and Fire Weather Index component data.

The training and evaluation of the model were carried out using these features. Proper data preprocessing, including normalization and model optimization techniques, significantly improved the classifier's performance. The proposed model achieved a prediction accuracy of 96%, surpassing several standard machine learning algorithms.

This study compares the performance of the proposed model with algorithms such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees. The results indicate a strong potential for AI-powered systems to play a meaningful role in understanding environmental hazards, promoting timely actions, informed policies, and efficient use of resources.

Overall, the findings of this research contribute to disaster management by offering a flexible and accurate model that aids decision-makers in effectively controlling forest fires and shaping future strategies in the field.

Contents

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
Content	vi
List of Tables	vii
List of Figures	viii
List of Abbreviations	ix
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Problem Statement	2
1.3 Overview	3
1.4 Objectives	4
1.4.1 Primary Objectives:	5
1.4.2 Secondary Objectives:	5
2 LITERATURE REVIEW	7
2.1 Research Gap	9
2.2 Background	11
2.2.1 Forest Fire Behavior and Prediction Challenges	12
2.2.2 Machine Learning in Forest Fire Prediction	12
2.2.3 Artificial Neural Networks (ANN)	13
2.2.4 Backpropagation Algorithm	13
2.2.5 Data Normalization	14
2.2.6 Why Use ANN?	14
3 METHODOLOGY	15
3.1 Data Collection	16
3.2 Data Description	17
3.3 Handling Data Imbalance	21
3.3.1 The Nature of Data Imbalance in Forest Fire Datasets	23

3.4	Data Preprocessing	24
3.4.1	Handling Missing Values	24
3.4.2	Encoding Categorical Variables	24
3.4.3	Feature Engineering	25
3.4.4	Detecting and Handling Outliers	25
3.4.5	Normalization of Numerical Features	25
3.4.6	Skewness Correction and Distribution Balancing	26
3.4.7	Temporal Categorization	26
3.4.8	Final Dataset Structure	27
3.4.9	Summary Table of Preprocessed Features	27
3.5	Model Building	27
3.5.1	Overview of Artificial Neural Network Architecture	28
3.5.2	Activation Functions	30
3.5.3	Model Architecture Details	30
3.5.4	Loss Function and Optimization	30
3.5.5	Forward and Backward Propagation	32
3.5.6	Model Training and Validation	32
3.5.7	Regularization Techniques	32
3.6	Performance Metrics	33
3.6.1	Confusion Matrix	33
3.6.2	Accuracy	34
3.6.3	Precision	34
3.6.4	Recall (Sensitivity)	34
3.6.5	F1-Score	35
3.6.6	Specificity	35
3.6.7	ROC Curve and AUC (Area Under Curve)	35
4	RESULTS AND DISCUSSION	37
4.1	Model Evaluation Results	37
4.2	ROC Curve and AUC Analysis	38
4.3	Comparative Analysis with Other Models	38
4.4	Interpretability of Results	39
4.5	Feature Importance and Sensitivity	39
4.6	Error Analysis	40
4.7	Practical Implications	40
4.8	Limitations and Scope for Improvement	40
5	CONCLUSION AND FUTURE SCOPE	42
5.1	Conclusion	42
5.2	Future Scope	43

List of Tables

3.1	Basic frequency analysis	23
3.2	Summary Table of Preprocessed Features	27
3.3	General structure of a confusion matrix for binary classification	34
4.1	Confusion matrix produced after testing	37
4.2	Performance comparison of ANN with other models	39

List of Figures

3.1	Attribute description of dataset	18
3.2	Flowchart of proposed classification mode	28
3.3	Different research stages of ML model	29
3.4	Artificial Neural Network	30
3.5	Plot of model	31
4.1	Accuracy Curve Epoch 100 ,Batch Size 10	41

List of Abbreviations

ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
DT	Decision Tree
FFMC	Fine Fuel Moisture Code
DMC	Duff Moisture Code
DC	Drought Code
ISI	Initial Spread Index
FWI	Fire Weather Index
RH	Relative Humidity
MSE	Mean Squared Error
CSV	Comma-Separated Values
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
EDA	Exploratory Data Analysis
ML	Machine Learning
AI	Artificial Intelligence
IoT	Internet of Things
ReLU	Rectified Linear Unit
SMOTE	Synthetic Minority Over-sampling Technique
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation (Optimizer)
CSV	Comma-Separated Values
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
UCI	University of California Irvine (Dataset Repository)
DTU	Delhi Technological University
Min-Max	Minimum-Maximum Normalization Technique

Chapter 1

INTRODUCTION

1.1 Motivation

Forest ecosystems are vital components of the Earth’s environmental framework. They serve not only as carbon sinks and biodiversity hotspots but also as crucial regulators of global climate and hydrological cycles. However, in recent decades, forest fires have emerged as one of the most destructive threats to these ecosystems. Driven by both natural causes and human activities, the frequency, intensity, and spatial spread of forest fires have shown alarming increases worldwide. These fires not only result in the direct destruction of vast forested areas but also contribute to secondary environmental problems such as soil degradation, air pollution, and the disruption of wildlife habitats. In extreme cases, they can lead to irreversible ecological damage and the displacement of human communities.

The motivation for this research stems from the growing global concern surrounding the inability of traditional fire detection systems to respond in time to prevent catastrophic damage. Most existing systems are reactive in nature—relying on satellite imagery or ground-based sensors to detect fires after ignition. While these systems have their merits, they are often insufficient in high-risk zones where fire can spread rapidly due to dry vegetation, strong winds, or human negligence. There is a compelling need to shift the focus from reactive detection to proactive prediction, thereby equipping decision-makers with tools that allow for early warnings and the deployment of preventive measures.

Climate change is further exacerbating the problem. Alterations in temperature patterns, decreased humidity, prolonged droughts, and unpredictable rainfall have contributed significantly to an increase in forest fire susceptibility. Many regions that were previously considered low-risk are now frequently witnessing fire outbreaks. This changing climatic context demands more dynamic, adaptable, and data-driven approaches to disaster prediction and management. The conventional statistical models, although useful, often fail to capture the nonlinear interactions between diverse environmental variables that contribute to fire ignition and propagation.

In this context, machine learning presents a transformative opportunity. By analyzing large volumes of historical and real-time environmental data, machine learning models are capable of identifying complex patterns and relationships that are not apparent through traditional analytical methods. Particularly, Artificial Neural Networks (ANNs), inspired by the structure of the human brain, have shown significant promise in recognizing intricate patterns and making accurate classifications or predictions. Their ability to learn from data and generalize to unseen conditions makes them ideal candidates for forest fire

prediction tasks.

This project is also motivated by the need for improved disaster preparedness and resource optimization. Predictive modeling enables authorities to allocate firefighting resources more effectively, plan evacuation routes, and implement controlled burns or other preventive measures in areas identified as high-risk. It empowers policymakers, environmental agencies, and forest management authorities with actionable insights that go beyond mere observation.

Furthermore, there is room for innovation at the nexus of artificial intelligence and environmental research. Forest fire prediction using machine learning not only offers practical societal benefits but also contributes academically to the fields of data science, environmental modeling, and intelligent systems. It promotes interdisciplinary learning and opens doors to scalable applications in other domains such as flood prediction, air quality monitoring, and climate risk analysis.

Thus, the motivation behind this research is deeply rooted in the urgent need to develop efficient, accurate, and scalable solutions for predicting forest fire occurrences. It is driven by environmental, humanitarian, and scientific imperatives, with the overarching goal of enhancing our collective ability to prevent and mitigate the devastating consequences of forest fires through technological innovation.

1.2 Problem Statement

Forest fires are increasingly becoming a major environmental, social, and economic concern across the globe. The rapid escalation of wildfire incidents, both in frequency and intensity, has highlighted the limitations of current forecasting and prevention mechanisms. Traditional methods of forest fire detection, which largely rely on satellite imaging, manual surveillance, or statistical forecasting models, are predominantly reactive and often fall short in preventing widespread damage. These systems typically identify a fire only after it has already begun to spread, leaving minimal time for effective intervention. As a result, vast areas of forest land are lost each year, threatening biodiversity, polluting the air with toxic emissions, and endangering human life and property.

The core of the problem lies in the complex interplay of environmental, meteorological, and anthropogenic factors that influence the likelihood and behavior of forest fires. Variables such as temperature, humidity, wind speed, rainfall, topography, vegetation type, and human activity all contribute to fire risks. These factors interact in non-linear and often unpredictable ways, rendering traditional linear models insufficient for accurate forecasting. Statistical models, while useful in identifying basic trends, often fail to accommodate the intricate patterns and hidden correlations inherent in large and multidimensional environmental datasets.

Furthermore, existing fire prediction models frequently suffer from low accuracy, especially in diverse geographical terrains where conditions vary significantly over short distances. They are also typically region-specific and do not generalize well across different ecosystems. The limited scalability and adaptability of conventional models hinder their applicability in real-world forest management systems that require dynamic and real-time decision-making capabilities. Additionally, data scarcity and inconsistency in recording forest fire incidents in many regions contribute to the problem by reducing the reliability of models based solely on historical records.

In recent years, the emergence of artificial intelligence and machine learning has provided new avenues to address this multifaceted problem. However, the application of machine learning in forest fire prediction remains in its early stages and is not without its own set of challenges. Many existing studies either focus on narrow data sets with limited environmental variables or use generic models that lack domain-specific tuning and validation. The absence of robust, interpretable, and high-accuracy predictive models continues to be a critical gap in the current research landscape.

This thesis seeks to address the gap by developing a machine learning-based predictive model that leverages a broad set of environmental and temporal attributes for accurate forest fire forecasting. Specifically, it focuses on the application of Artificial Neural Networks (ANNs), which have demonstrated superior capabilities in modeling complex and non-linear relationships. The model is trained and tested on a carefully curated dataset comprising features such as spatial coordinates, Fire Weather Index components (FFMC, DMC, DC, ISI), meteorological factors like temperature, wind speed, and rainfall, as well as temporal aspects including day and month.

The problem being addressed is not merely technical but is also deeply rooted in sustainability and disaster risk reduction. Without reliable and anticipatory forest fire prediction mechanisms, governments and environmental agencies are left reactive, often responding too late to prevent large-scale destruction. A predictive solution that accurately identifies high-risk conditions and areas before a fire occurs can facilitate timely action, better resource allocation, and the preservation of both ecological and human systems.

Thus, the central problem this research aims to solve is the development of a data-driven, accurate, and scalable model for forest fire prediction that overcomes the limitations of traditional techniques and provides practical value for real-time disaster preparedness and environmental conservation.

1.3 Overview

Forest fires represent a formidable challenge in contemporary environmental management. Their unpredictability and destructive power pose serious threats to ecological balance, public safety, and economic stability. Across many parts of the world, including countries with vast forest coverage like Brazil, Australia, Canada, and India, wildfires have become increasingly frequent and intense due to rising global temperatures, erratic rainfall patterns, prolonged droughts, and increased human encroachment into forested regions. The consequences of such fires are multifaceted—ranging from deforestation and the extinction of wildlife to the release of enormous quantities of greenhouse gases and the displacement of communities.

Given these growing concerns, researchers and environmental scientists are turning toward advanced technologies to find sustainable and proactive solutions. In this context, machine learning has emerged as a revolutionary tool capable of transforming traditional environmental forecasting methods. The capacity of machine learning models to ingest vast volumes of data and learn complex, non-linear patterns makes them particularly suitable for tasks like fire prediction, where multiple variables interact in intricate ways.

This research focuses on Artificial Neural Networks (ANN), a specific class of machine learning al-

gorithms modeled after the structure and functioning of the human brain. ANNs are adept at pattern recognition, classification, and forecasting, making them an ideal candidate for forest fire prediction. They consist of multiple layers—input, hidden, and output—that work together to process information and produce predictive outputs based on historical trends. One of their key advantages is their ability to generalize from the training data to make accurate predictions on unseen data, which is especially valuable in real-world disaster scenarios.

The idea behind this project is to make a model that identifies areas where a forest fire is more likely to happen by using real environmental and weather-related information. The dataset includes 518 recordings and features important elements such as temperature, wind speed, rainfall, location coordinates and Fire Weather Index (FWI) parts such as FFMC, DMC, DC and ISI. To account for changes throughout the year and each day, variables for the month and day are added to describe fire risk.

Prior to building the model, lots of work is done to ensure the data is clean, normalized and organized into categories that make sense to use. Thus, the model is able to learn and perform accurately because the standardized information improves both learning and performance. Following processing, the data is separated into a set for training and another for testing and ANNs adjust their weightings for every iteration with backpropagation to lower the amount of error. Experts use accuracy, precision, recall and F1-score, all calculated from the confusion matrix, to assess the model's performance.

ANN performed very well in the first stage, getting 96% accuracy, compared to support vector machines, decision trees, K-nearest neighbors and linear regression. This points to the fact that ANN models perform environmental prediction with high accuracy.

This thesis not only presents a technical solution but also emphasizes the practical implications of deploying such models in real-world settings. Government agencies, forest departments, and emergency management authorities can use these models to develop early warning systems, allocate firefighting resources more efficiently, and implement preventive measures in vulnerable areas. Furthermore, the project lays the foundation for future research in integrating satellite imagery, IoT sensor data, and real-time weather feeds for even more comprehensive and dynamic forest fire forecasting systems.

In addition to its practical contributions, this study also has academic significance. It highlights the interdisciplinary application of machine learning in environmental science and disaster risk management, encouraging future work at the intersection of technology and sustainability.

In summary, this thesis provides an end-to-end framework—from data collection and preprocessing to model training and performance evaluation—for using machine learning, particularly Artificial Neural Networks, to predict forest fire events. It offers a robust, scalable, and accurate solution aimed at reducing the devastating impacts of wildfires and contributing to a safer, more resilient future.

1.4 Objectives

The increasing frequency and severity of forest fires have underscored the urgent need for advanced technological solutions that can accurately predict and help prevent such devastating events. The core motivation behind this research is to develop a predictive model that utilizes modern computational techniques—particularly machine learning—to address the challenges of early forest fire detection. Given

the limitations of traditional statistical and observational methods, this study aims to leverage Artificial Neural Networks (ANN) to capture the complex interdependencies between environmental and temporal variables and forecast fire risk with high precision.

The objectives of this research are both technical and practical in nature. They encompass the complete pipeline of building a machine learning model, from data acquisition and preprocessing to model training, evaluation, and real-world applicability. These objectives are designed to create a holistic framework that not only contributes academically but also has potential real-time applications in forest management and disaster mitigation.

1.4.1 Primary Objectives:

- 1. To design and develop a robust predictive model using Artificial Neural Networks (ANN)**

The main focus of this work is to build a multilayer neural network that can find complex links between several environmental factors and forecast when forest fires will happen. The best ANN architecture will be selected and updated according to how accurate, precise and recallable the results are.

- 2. To utilize a comprehensive dataset containing relevant forest fire attributes**

The dataset to be used will include 518 instances, plus features and forecast values such as temperature, wind speed, humidity, rainfall, spatial coordinates (X and Y) and two Fire Weather Indexes (FFMC and DMC) at a given time. The dataset should work well for testing, as they move forward with the model in other types of conditions and goals.

- 3. To preprocess and normalize the data to ensure quality and consistency**

The data will be preprocessed in several steps: handling all missing values, grouping continuous variables by type, finding unusual entries and making all features the same scale. As a result, the neural network does not favor variables with a larger magnitude during training.

- 4. To implement the backpropagation learning algorithm for model training**

Backpropagation helps to reduce error in a network while it is being trained. The process will be checked nature thereafter, washing dresses and dirty clothes will be used to improve the accuracy and speed of training.

- 5. To evaluate model performance using multiple statistical metrics**

Apart from checking the accuracy, we will also use precision, recall, F1-score and a confusion matrix to test the strength of the model. The goal is to see that the model can solve usual problems and perform equally well on novel ones.

1.4.2 Secondary Objectives:

- 1. To compare ANN performance with traditional machine learning models**

As part of the evaluation process, this study aims to compare the ANN model's performance with that of other models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Linear Regression. This comparative analysis will help validate the superiority (or shortcomings) of ANN in the context of forest fire prediction.

2. To explore the potential of temporal and spatial pattern recognition in fire forecasting

By analyzing temporal features (month, day, weekday/weekend) and spatial coordinates, the study aims to investigate how fire occurrences are influenced by time and location. This can lead to region- and season-specific models in the future.

3. To identify the benefits and limitations of using ANN for forest fire prediction

Understanding the capabilities and challenges of implementing ANN in real-world environmental forecasting is essential. This objective includes documenting model behavior in cases of limited data, imbalanced classes, and sudden variable changes—issues that often occur in environmental datasets.

4. To propose future enhancements for scalability and real-time implementation

This research also aims to identify opportunities for future development, including the integration of real-time weather data, the use of remote sensing inputs (such as satellite imagery), and the deployment of the model as part of a decision-support system for forest departments.

5. To contribute to the growing intersection of environmental science and artificial intelligence

Finally, this thesis intends to serve as an academic contribution to the interdisciplinary field combining data science, environmental modeling, and disaster management. The goal is to open new avenues for AI-driven solutions to pressing environmental challenges, encouraging further research and innovation in this space.

Chapter 2

LITERATURE REVIEW

Forest fires are becoming more common in locations throughout the globe. This situation has encouraged the need for accurate predicting and stopping of future threats. Researchers have investigated many machine learning and statistical techniques to deal with this issue. Looking at a problem through many data sets and research methods. This part of the paper analyzes the importance of studies that look at the methods used for forecasting forest fires. Datasets selected and how correct the approach proved to be. The purpose is to present a detailed familiarity with today's solutions and how predictive methods have changed over time.

In a study conducted by **Pham et al. (2022)** [1], multiple machine learning models—Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GaussianNB), and Linear Regression—were evaluated for wildfire prediction using two types of datasets: a state-level dataset and a more granular county-level dataset for California. Their findings revealed that the KNN model achieved the highest accuracy (97%) on the county-level data, while the ANN model showed competitive performance with a 96% accuracy rate. However, state-level results were less consistent due to disparities in data distribution, highlighting the importance of dataset granularity in model performance.

Similarly, **Zhou (2023)** [2] compared traditional machine learning algorithms such as Decision Trees and KNN with deep learning models like CNNs and RNNs. The Decision Tree model emerged as the most accurate overall, while CNNs showed the best precision among deep learning methods. Interestingly, the ANN model recorded the highest recall, second only to Decision Trees, which indicates its effectiveness in identifying true positive fire events. This supports the argument that ANN is a viable model for fire prediction tasks, particularly in recall-sensitive applications such as disaster management.

Sitorus et al. (2023) [3] conducted a focused study on ANN-based classification for forest fire disaster prediction. Their model achieved 96% accuracy with a recall of 100% and a precision of 91.66%, underscoring the potential of ANN in highly sensitive environmental prediction tasks. The study also highlighted the importance of the backpropagation algorithm and the sigmoid activation function in fine-tuning the

model.

Rakshit et al. (2021) [4] explored various machine learning algorithms, including Decision Tree, KNN, SVM, and Naive Bayes, to identify the most effective classifier for forest fire prediction. They concluded that the Decision Tree model achieved the highest AUC (Area Under the Curve), making it the most reliable among the tested algorithms. This study reinforced the importance of evaluation metrics beyond mere accuracy, such as precision-recall trade-offs and AUC scores.

Mekala et al. (2023) [5] introduced a logistic regression model using just humidity and temperature as predictors. While simplistic in design, this model offered insights into the viability of low-dimensional models in specific contexts where computational efficiency is prioritized. However, its predictive performance was limited, showing that more complex models are necessary for broader and more accurate fire prediction.

Thakkar et al. (2022) [6] applied the Random Forest algorithm for forest fire prediction using environmental features and achieved a regression score of 0.9799. Their approach employed Pearson's correlation coefficient for feature selection, demonstrating that ensemble models combined with statistical preprocessing can be effective in producing high-quality results.

Tang et al. (2020) [7] addressed class imbalance in SVM-based forest fire susceptibility evaluations by introducing an optimized repeated random undersampling technique. . Their study demonstrated that balancing the dataset significantly improved SVM performance, particularly in regions where fire events are relatively rare compared to non-fire occurrences.

Li et al. (2022) [8] focused on forest fire spread prediction using Backpropagation Neural Networks. This study emphasized the application of Huygens' principle in fire propagation modeling and validated the ANN's capacity to capture dynamic environmental changes over time, a feature traditional models often overlook.

Prathimesh et al. (2023) [9] incorporated feature selection via the Gini index alongside SVM for fire prediction. Their study illustrated the role of feature relevance in enhancing model accuracy and emphasized the need to filter irrelevant attributes from the training data.

Weng et al. (2022) [10] proposed a deep learning framework combining temperature sensors, vegetation mapping, and satellite-based remote sensing for burned area estimation. Their multi-source data fusion strategy presents a scalable solution for wide-area forest monitoring and highlights the growing interest in integrating machine learning with geospatial analysis.

Adhikari et al. (2023) [11] developed a wildfire progression prediction model using satellite and remote sensing data for California's Sonoma region. Their research

combined geospatial imagery with machine learning techniques to track fire spread, offering valuable insights for real-time surveillance and forecasting.

Makhaba and Winberg (2022) [12] introduced LRCN (Long-term Recurrent Convolutional Network), a hybrid of CNN and RNN models, for wildfire path prediction. Their results suggested that LRCN could outperform traditional models in dynamic scenarios where fire behavior changes rapidly over time and space.

Feng et al. (2023) [13] implemented wavelet-based techniques in conjunction with CNN and XGBoost algorithms. Their research aimed at noise reduction and signal enhancement in environmental datasets, ultimately achieving higher accuracy and model stability.

Shah and Pantoja (2023) [14] investigated the use of U-Net designs with attention mechanisms for wildfire spread prediction. Their deep learning model demonstrated enhanced spatial pattern recognition capabilities, which are critical for modeling fire behavior over large areas.

Priya and Vani (2023) [15] applied deep learning to model climate change effects on forest fire risk, establishing a correlation between global warming indicators and fire occurrences. Their work signifies the increasing relevance of integrating climate models with forest fire prediction algorithms.

Finally, **S.K. et al. (2023)** [16] proposed a hybrid model combining wireless sensor networks and deep learning approaches for early detection of forest wildfires. Their integration of IoT-based monitoring with predictive modeling reflects the modern trend toward smart, connected disaster management systems.

Collectively, these studies reflect the growing interest in combining artificial intelligence with environmental science to tackle the challenges of wildfire prediction. While models like Decision Trees, SVM, and Random Forest have shown considerable promise, deep learning techniques, particularly ANN and CNN-based models, consistently demonstrate superior performance when the dataset is rich and well-structured.

2.1 Research Gap

Despite the considerable advancements made in the field of forest fire prediction, there remain multiple gaps and challenges that limit the effectiveness, scalability, and real-world applicability of existing models. These gaps span across data quality and availability, algorithmic performance, interpretability, generalization ability, and integration with real-time decision-support systems. This section outlines these issues in detail to highlight the necessity and relevance of the current study.

One of the most significant challenges identified in the literature is the lack of high-quality, large-scale, and multi-source datasets that comprehensively capture the dynamic nature of forest fire phenomena. Most studies rely on datasets with a limited

number of attributes and instances, often collected from a single geographic location. For example, the study by Sitorus et al. (2023)[16] used a dataset with only 518 entries, which while sufficient for initial model training, lacks the temporal and spatial diversity needed for large-scale deployment. Many of the existing datasets also lack real-time attributes such as vegetation index, fuel moisture content, and human activity indicators like proximity to roads or settlements. This data scarcity limits the depth and breadth of feature engineering and model learning.

Another critical research gap is the over-reliance on static models trained on historical data without mechanisms to adapt to rapidly changing environmental conditions. Forest fire behavior is inherently dynamic, influenced by fluctuating variables such as temperature, wind direction, and rainfall. Traditional models such as Logistic Regression, Decision Trees, and even some Neural Networks are trained on fixed datasets and often fail to generalize well when exposed to unseen patterns. As pointed out in the study by Rakshit et al. (2021)[10], although Decision Trees yielded high AUC values, their performance dropped when applied to new datasets with altered environmental conditions.

A further issue lies in the treatment of class imbalance, which is a recurring theme in fire prediction models. In most real-world datasets, instances of forest fires are much fewer than non-fire events. Model performance is skewed by this imbalance, which frequently inflates accuracy at the expense of recall for the minority class. Tang et al. (2020)[7] proposed an undersampling method to improve SVM-based classification for imbalanced datasets, but such solutions are not universally adopted across models, and many studies fail to address the class imbalance issue altogether. This results in models that may perform well statistically but are practically weak in correctly identifying actual fire events.

Another urgent issue is the poor interpretability of sophisticated machine learning models. Models such as CNNs, LSTM, and deep ANNs operate as black boxes, making it difficult for forest officials or environmental policymakers to understand the rationale behind predictions. While performance metrics like accuracy and F1-score are important, the inability to explain the decision-making process of a model can hinder its acceptance and deployment in real-world scenarios where accountability and transparency are critical. This concern was indirectly addressed in the study by Zhou (2023)[14], where although CNNs offered high precision, their interpretability remained unexplored.

Another overlooked aspect in most existing studies is the integration of spatial and temporal relationships. Forest fires are often the result of a combination of spatial factors (such as terrain, vegetation density) and temporal patterns (seasonal dryness, time of day). However, many models either consider only temporal or spatial data in isolation, thereby missing the intricate interplay between these two dimensions. For instance, while Makhaba and Winberg (2022)[2] introduced LRCN to capture such patterns, the model's implementation remains limited to experimental studies, lacking

practical deployment frameworks.

Furthermore, there is a shortage of comparative evaluations across multiple algorithms under identical conditions. Many papers test only one or two models in isolation without benchmarking their performance against a wide spectrum of machine learning techniques. The lack of standardized evaluation metrics and datasets further hampers the ability to draw generalizable conclusions about which model is most effective in varying environmental contexts.

An additional limitation is the absence of model optimization strategies such as hyperparameter tuning, dropout regularization, and cross-validation in several studies. For example, while Sitorus et al. (2023)[16] achieved impressive accuracy using ANN, the study did not explore how tuning model parameters such as learning rate, number of hidden layers, or activation functions could further enhance performance. Similar to this, many research choose the batch size and epoch count at random, hence ignoring optimization potential.

The integration of real-time or near-real-time systems into forest fire prediction pipelines also remains an underdeveloped area. Although models may perform well on retrospective data, their utility in operational settings is still questionable. There is a lack of research on how these models can be embedded into alert systems or decision-support dashboards used by forest management authorities. Most models exist in research silos with little focus on practical implementation, usability, or policy integration.

Finally, many studies fail to address the environmental and policy-level impacts of forest fire prediction systems. There is minimal discussion on how predictive analytics can be integrated into forest conservation programs, early warning systems, or resource allocation mechanisms. This disconnect between model development and real-world application limits the transformative potential of AI-driven solutions in the context of forest fire mitigation.

In summary, the existing body of work on forest fire prediction reveals several gaps that need to be addressed. These include the scarcity of diverse and high-resolution datasets, limited model generalization, poor handling of class imbalance, lack of interpretability, insufficient integration of spatial-temporal features, weak benchmarking, inadequate optimization practices, and minimal focus on real-time deployment and policy alignment. This thesis aims to bridge some of these gaps by developing a high-accuracy ANN model using a structured, comparative approach with a focus on practical relevance and system integration.

2.2 Background

Forest fires are among the most devastating natural disasters with the potential to destroy ecosystems, displace communities, and cause widespread economic and environmental damage. The increased frequency and severity of forest fires globally have

prompted significant interest in developing predictive models capable of forecasting fire incidents with high accuracy and reliability. Traditionally, forest fire detection relied on manual monitoring, satellite-based surveillance, or rule-based statistical forecasting. However, these approaches often fall short in providing real-time, accurate predictions necessary for timely interventions. With the growing availability of environmental data and computational power, machine learning models—especially Artificial Neural Networks—have emerged as promising tools in the fight against forest fires.

This background section outlines the foundations of forest fire modeling, the evolution of machine learning in environmental science, and the specific techniques and technologies utilized in this thesis. The section is also organized to explain the theoretical and mathematical underpinnings of these models, providing a strong foundation for the methodology and implementation chapters that follow.

2.2.1 Forest Fire Behavior and Prediction Challenges

A combination of natural and human factors causes forest fires. Fireplaces can be very dangerous since environmental conditions like temperature, the presence of rain, dry vegetation and airflow can easily lead to a fire spreading throughout the house. The way these variables change both over time and over different regions causes extra challenges for modeling. Forecasting forest fires requires taking into account quick weather changes and slow climate changes.

The Fire Weather Index (FWI) system, widely used for fire danger rating, includes parameters like Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and the Initial Spread Index (ISI). These indices are nonlinear functions of environmental variables and contribute to estimating the potential for ignition and fire spread. However, interpreting these indices and correlating them with actual fire events requires advanced computational models capable of identifying hidden relationships and adapting to dynamic data.

2.2.2 Machine Learning in Forest Fire Prediction

Machine learning provides an alternative to traditional empirical and statistical models by learning patterns from data. Unlike rule-based models, machine learning algorithms can generalize from historical data and make predictions based on unseen inputs. For environmental modeling tasks, supervised learning is the most widely used category, where a model is trained on labeled data (i.e., features and known outcomes).

In forest fire prediction, a supervised model is given environmental and temporal features (temperature, wind speed, month, rainfall, etc.) and learns to classify whether or not a fire is likely to occur. For this, classification techniques like as Neural Networks, SVMs, and Decision Trees are frequently employed. Among these, Artificial Neural Networks have shown superior ability in handling nonlinear relationships and adapting to complex feature spaces.

2.2.3 Artificial Neural Networks (ANN)

Artificial neural networks are designed to work like the brain's neural networks. Shall we learn about them consists of joined neurons, arranges in layers: input layer, hidden layers and output layer. Every connection has a weight that is changed through training in order to lower the prediction error.

The mathematical representation of a neuron's output is as follows:

Formula 1: Neuron Output Calculation

$$z = \sum_{i=1}^n w_i x_i + b \quad (2.1)$$

Where:

- z is the input to the activation function,
- w_i are the weights,
- x_i are the input features,
- b is the bias term.

The output of the neuron is then passed through an activation function to introduce non-linearity. The sigmoid function is commonly used:

Formula 2: Sigmoid Activation Function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.2)$$

In this thesis, a multilayer ANN is used to predict the likelihood of forest fire occurrences. The input layer receives environmental features, hidden layers perform intermediate computations, and the output layer classifies the event as fire or no fire.

2.2.4 Backpropagation Algorithm

The training method many neural networks use is called backpropagation. To apply this method, each wieght needs to be given a gradient of the loss function with the chain rule. Most of the time, MSE or Cross-Entropy Loss identifies the extent to which the model differs from the data used in training.

Cross-Entropy Loss is better suited for binary classification:

Formula 3: Binary Cross-Entropy Loss

$$L = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \quad (2.3)$$

Where:

- y is the actual label,

- \hat{y} is the predicted probability.

The weights are updated using gradient descent:

Formula 4: Weight Update Rule

$$w_{\text{new}} = w_{\text{old}} - \eta \cdot \frac{\partial L}{\partial w} \quad (2.4)$$

Where:

- η is the learning rate,
- $\frac{\partial L}{\partial w}$ is the weight-dependent gradient of the loss function.

Backpropagation proceeds layer by layer, adjusting the weights and minimizing the error after each iteration (epoch).

2.2.5 Data Normalization

Data normalization is essential to ensure all input features are on a comparable scale. In neural networks, this improves convergence speed and model stability. Min-Max normalization is a commonly used technique:

Formula 5: Min-Max Normalization

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2.5)$$

Normalization is especially important in this project due to the presence of features with varying units such as temperature ($^{\circ}\text{C}$), wind speed (km/h), and rainfall (mm).

2.2.6 Why Use ANN?

Artificial Neural Networks were chosen in this study due to their ability to:

- Model complex nonlinear relationships between environmental variables.
- Handle high-dimensional feature spaces with better generalization.
- Adapt to both structured and unstructured datasets.
- Perform well in classification problems involving environmental and time-series data.

Compared to other algorithms like Decision Trees or Logistic Regression, ANNs do not assume linear separability or feature independence, making them ideal for modeling the intricate dependencies present in forest fire data. Additionally, ANN models can be trained to a high degree of accuracy and tuned using batch size, learning rate, and number of hidden neurons, offering flexible experimentation options.

Chapter 3

METHODOLOGY

Developing a reliable and accurate predictive model for forest fires necessitates a structured and systematic methodology that spans from data acquisition and preparation to model development and evaluation. The primary aim of this research is to build a data-driven machine learning model, specifically an Artificial Neural Network (ANN), to predict the likelihood of forest fire occurrences using environmental and temporal features. The methodology employed in this thesis is a multi-stage process that integrates both theoretical foundations and empirical procedures. It involves collecting appropriate datasets, transforming and preprocessing the data for modeling, building and tuning a predictive model, and finally evaluating its performance using well-established metrics.

The accuracy and generalizability of the final model are greatly influenced by the methods selected. Since environmental data is often heterogeneous, noisy, and highly variable over space and time, careful attention must be given to data handling, cleaning, and normalization before model training. In the context of machine learning, even minor inconsistencies in data handling can result in major shifts in model performance. Therefore, each stage of the methodology is designed with precision and tailored to address the specific challenges posed by forest fire prediction.

The methodology is organized into six major components:

1. **Data Collection** – Sourcing relevant and high-quality datasets that provide the features necessary for prediction.
2. **Data Description** – Understanding the structure, attributes, and statistical characteristics of the dataset.
3. **Handling Data Imbalance** – Dealing with disproportionate representation of classes (fire vs. no fire) which could affect model learning.
4. **Data Preprocessing** – Cleaning, transforming, and normalizing the dataset to ensure it is suitable for input into the ANN model.
5. **Model Building** – Designing and training the ANN using the processed dataset with appropriate hyperparameters and learning configurations.

6. **Performance Metrics** – Evaluating the trained model’s effectiveness using statistical measures such as accuracy, precision, recall, and F1-score.

Each of these stages will be explored in detail in the sections that follow, beginning with the process of data collection.

3.1 Data Collection

The first stage of any machine learning-based prediction system is data collection. The quality, relevance, and richness of the data directly impact the model’s ability to generalize and perform effectively in real-world conditions. In the context of forest fire prediction, the dataset must encapsulate a variety of environmental and meteorological parameters that contribute to fire ignition and spread. These parameters include temperature, wind speed, humidity, rainfall, fire weather index components, and temporal features such as day and month.

In this thesis, a dataset that people can use and that has been cited widely was selected. Paulo Cortez and Aníbal Morais from the University of Minho put it all together. Portugal. The UCI Machine Learning Repos shared the dataset with us. The data can be found under the title Forest Fires Data Set. The data used is based on true meteorological observations. gathered from the Montesinho Natural Park in northeast Portugal, famous for because it has many forests and is vulnerable to frequent seasonal fires.

The dataset consists of 517 records (also cited as 518 in some sources) with 13 input variables and one target variable. These records correspond to daily observations of various weather and forest conditions across different months and days. The primary aim of collecting this dataset was to capture diverse environmental patterns across different temporal intervals and analyze how these factors influence forest fire activity.

The original dataset includes the following key features:

- **Spatial Coordinates:** X and Y (spatial indices from a 9×9 grid)
- **Temporal Features:** Month and Day
- **Meteorological Attributes:** Temperature (°C), Relative Humidity (%), Wind Speed (km/h), Rainfall (mm)
- **FWI Components:** FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), ISI (Initial Spread Index)
- **Target Variable:** Area

The inclusion of Fire Weather Index (FWI) components is particularly important, as these are standard indicators used by forestry services to assess fire risk. Each component represents a specific aspect of the environment’s dryness and combustibility:

- **FFMC**: Determines the amount of moisture in fine fuels and litter.
- **DMC**: Represents moisture content in loosely compacted organic layers.
- **DC**: Indicates long-term drying of deeper organic layers.
- **ISI**: Reflects the expected rate of fire spread.

In this research, the original continuous *area* attribute is converted into a categorical variable, termed *size_category*, with binary values: 0 for low or no fire (area less than 6 hectares) and 1 for significant fire (area ≥ 6 hectares). This transformation simplifies the problem into a binary classification task, which aligns with the capabilities of ANN-based classifiers and facilitates the use of performance metrics like precision, recall, and accuracy.

The dataset was selected not only for its rich feature set but also due to its widespread use in related research, which enables fair benchmarking and comparison. Furthermore, it has been validated and pre-processed by the original authors, reducing the likelihood of measurement errors and making it suitable for academic purposes.

Although the dataset is geographically restricted to a single region, its structure is sufficiently comprehensive to allow for modeling of general fire-prone conditions. In future work, this dataset can be enhanced or extended by integrating additional regional or global environmental datasets, including satellite imagery, IoT-based sensor data, and live weather feeds.

Data was accessed through the UCI Machine Learning Repository in CSV format and imported into the Python programming environment using standard libraries such as Pandas and NumPy. Upon import, the data underwent exploratory data analysis (EDA) to inspect missing values, distribution of features, and class imbalance—issues that are addressed in the subsequent sections.

In summary, the dataset used in this study is a robust and well-documented environmental dataset collected from a real forest region, making it highly suitable for modeling fire risk. Its structured format and inclusion of both meteorological and temporal features allow for the development of a predictive model that mirrors the complexity of real-world fire forecasting scenarios.

3.2 Data Description

Understanding the structure and semantics of the dataset is critical for building an effective machine learning model. Data description involves thoroughly analyzing each feature, its type, scale, distribution, and relevance to the prediction task. A well-documented understanding of the dataset not only aids in feature selection and pre-processing but also ensures that the modeling approach is contextually aligned with the real-world problem being addressed.

Attribute Description	
X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
month	Month of the year (January to December)
day	Day of the week (Monday to Sunday)
FFMC	FFMC code
DMC	DMC code
DC	DC code
ISI	ISI index
temp	Outside temperature (in °C)
RH	Outside relative humidity (in %)
wind	Outside wind speed (in km/h)
rain	Outside rain (in mm/m ²)
area	Total burned area (in <i>ha</i>)

Figure 3.1: Attribute description of dataset

The dataset used in this thesis originates from the Montesinho Natural Park in Portugal and includes meteorological and temporal data collected during the fire seasons of several years. It consists of 517 instances, each representing a single day of observation, with 12 independent input features and one output feature, later modified to suit the binary classification problem.

Each of the features has been carefully chosen for its significance in forest fire behavior. The dataset captures not only weather-related variables but also spatial and seasonal aspects that could influence fire occurrence. A thorough explanation of each feature can be found below:

1. X – Spatial coordinate (1 to 9):

This attribute denotes the X-axis spatial coordinate of the observation grid within the forest area. For the purpose of physically localizing fire events, Montesinho Park was separated into a 9x9 spatial grid. This variable helps in spatial analysis by associating fire-prone zones with specific coordinate regions.

2. Y – Spatial coordinate (2 to 9):

Similar to the X coordinate, this attribute defines the Y-axis location in the grid. Together, X and Y form the spatial context of the data point and allow the model to detect if certain areas have higher fire probabilities due to vegetation, altitude, or human activity.

3. Month – Categorical (January to December): The month represents the time of year when the data was recorded. This variable is crucial for capturing seasonal trends, as fire occurrences often spike during dry and hot months like July and August. In preprocessing, this feature is later encoded numerically to be compatible with

machine learning algorithms.

4. Day – Categorical (Monday to Sunday):

This attribute identifies the day of the week. Although not directly tied to weather, it can offer indirect insights. For instance, weekends might see more human activity (tourism, camping, etc.), increasing the risk of anthropogenic fire triggers. This feature is useful for analyzing behavioral correlations.

5. FPMC – Fine Fuel Moisture Code (numeric):

The FPMC is a key component of the Canadian Fire Weather Index system. It quantifies the moisture content in fine surface litter and small twigs. Higher numbers indicate dry and flammable materials, whereas lower values indicate wetter circumstances. It typically ranges from 18.7 to 96.20.

6. DMC – Duff Moisture Code (numeric):

Decomposed leaves and other organic layers that are moderately compressed can have their moisture content measured by the DMC. It is influenced by rainfall, temperature, and humidity and affects the ignition and sustainability of ground fires. The value varies from 1.1 to 291.3.

7. DC – Drought Code (numeric):

The DC captures long-term drying of deep, compact organic matter layers. It reflects drought conditions and is a strong indicator of forest fire potential. Its scale typically ranges from 7.9 to 860.6, making it the highest among the three moisture codes in the FWI system.

8. ISI – Initial Spread Index (numeric):

The ISI estimates the rate at which a fire is likely to spread given the current wind and moisture conditions. It is calculated using FPMC and wind speed, and a higher ISI means faster spread potential. The dataset has values ranging from 0.0 to 56.10.

9. Temperature – Ambient temperature in Celsius (numeric):

Ambient temperature is a critical determinant of fire risk. Higher temperatures reduce fuel moisture content, making ignition more likely. In the dataset, this value ranges from 2.2°C to 33.3°C, reflecting data from cooler months as well as hot summer days.

10. RH – Relative humidity in percentage (numeric):

Fuel moisture and, hence, the likelihood of a fire starting are influenced by relative humidity. Higher humidity suppresses fires, while low humidity accelerates them. The RH values range between 15% and 100%.

11. Wind – Wind speed in km/h (numeric):

Wind has two functions: it affects the ISI and helps fire spread. Fires can spread swiftly across wider areas when strong winds are present. Wind speeds in the dataset vary from 0.4 km/h to 9.4 km/h.

12. Rain – Rainfall in mm/m² (numeric):

Rainfall contributes to reducing fire risk by increasing ground and fuel moisture. However, extremely low rainfall or dry spells significantly elevate fire hazards. Rainfall values in the dataset range from 0.0 to 6.4 mm, with most instances recording zero, indicating dry conditions. **13. Area – Burned area in hectares (numeric):**

This is the original target variable, representing the total forest area burned in each incident. Since it is highly skewed, with many zero values and few extreme cases, it is transformed into a binary variable during preprocessing.

To enhance model interpretability and simplify the classification task, the area variable is converted into a categorical class termed **size_category**:

- **0** if area < 6.0 hectares (low fire severity or no fire)
- **1** if area \geq 6.0 hectares (significant fire severity)

This conversion helps avoid problems related to regression on a heavily imbalanced continuous target and aligns the task with classification techniques, especially the Artificial Neural Network used in this research.

In terms of data types, most features are continuous numerical values, except for the month and day, which are categorical. These categorical values are later encoded numerically using techniques such as label encoding or one-hot encoding to make them suitable for input into the neural network.

Exploratory data analysis reveals several important patterns:

- A large number of entries have an area value of zero, indicating either no fire or negligible fire.
- Seasonal patterns show more fire incidents in July, August, and September.
- The FFMC, DMC, and DC indices are generally higher in fire-likely instances, reflecting drier conditions.

Additionally, certain variables like rainfall are sparsely populated with non-zero values, confirming the prevalence of dry conditions during fire events. These insights help in guiding feature selection, scaling decisions, and model architecture.

The dataset is relatively small by machine learning standards but is sufficient for demonstrating the efficacy of a well-optimized ANN model. Moreover, its clean structure and curated features make it ideal for academic experimentation and performance benchmarking.

In conclusion, the dataset contains a rich mix of spatial, temporal, and meteorological variables necessary for developing an effective predictive model. A clear understanding of each feature's role in fire ignition and propagation not only guides preprocessing but also aids in feature importance analysis post model training. This detailed data understanding paves the way for intelligent handling of preprocessing, class imbalance, and model selection in the subsequent stages of this study.

3.3 Handling Data Imbalance

Data imbalance is a common challenge in classification problems, especially in real-world datasets where the distribution of classes is naturally skewed. In the context of forest fire prediction, the issue of class imbalance becomes particularly critical due to the significantly higher number of days with no fire or low fire incidents compared to days with severe fire occurrences. This uneven distribution can severely bias the learning process of machine learning algorithms, including Artificial Neural Networks (ANN), resulting in poor generalization and poor performance on minority class instances.

In the dataset used for this study, the original target variable, area, is a continuous numeric feature representing the burned area in hectares. As observed during exploratory data analysis, a large portion of the dataset consists of entries where the burned area is zero or very low, and only a small fraction corresponds to large fire incidents. This leads to a natural dominance of the non-fire or low fire class over the significant fire class, especially when the data is converted to a binary classification problem using the threshold $\text{area} \geq 6.0$ hectares.

There are many negative outcomes from unbalanced data sets. First, most machine learning Algorithms are created to achieve the highest total accuracy which can result in a bias. the vast majority. There are times when a model seems more accurate than it actually is. guessing the majority label every time, without taking the minority in account nority class. This makes things more difficult in forest fire prediction because the cost of A false negative result, meaning a true fire is missed, could occur very often and cause de. lead to widespread destruction in both nature and among human beings.

Second, having an unbalanced set of data usually creates unstable boundaries for the decision model. model training. They adapt by addressing prediction errors so as to keep the loss function as small as possible. When if there's a large, unbalanced class, the loss function will have a preference toward lowering its value Most of its errors are with the majority class, therefore the model cannot correctly identify the different groups. patterns found among members of the minority group.

A number of techniques are applied in this study to solve this challenge. addresses the problem of class imbalance so that the predictive model treats both classes fairly. appropriate significance. The various techniques can mostly be sorted into datat-level approaches. methods at the algorithm level

At the data level, one of the most effective approaches is resampling, which involves modifying the dataset so that the classes are more balanced. Oversampling and under-sampling are the two primary forms of resampling.

To increase the size of the minority class, we use a method called oversampling. The information can be obtained by cloning actual records or by creating artificial data.

samples. The Synthetic Minority Over Sampling technique is one of the most used methods in synthetic oversampling. SMOTE method which generates new samples by connecting the existing ones. examples that fit within the minority class. With this standard, we stop overfitting from happening. as simple as simply repeating samples. SMOTE was not formally used in the dataset. Even with a small dataset, alignment remains a promising method in this research. Further developments when using the model with bigger datasets.

Conversely, undersampling entails lowering the quantity of occurrences in the majority class. This can be effective when the majority class significantly overwhelms the dataset, as it forces the model to learn from fewer but more informative samples. However, the downside of undersampling is the potential loss of valuable information, which can negatively affect the model's generalizability. In this research, a balanced compromise is achieved by selective filtering of the dataset such that the number of samples in both classes is reasonably comparable, without completely discarding useful data.

Apart from resampling, algorithm-level solutions are also explored. One such approach is using class weights during model training. In neural networks, the loss function can be modified to assign higher penalty for misclassifying the minority class. As a result, the model is compelled to concentrate more on acquiring the traits of that class. In the context of this study, class weighting was incorporated into the ANN training process by assigning a higher cost to incorrect predictions for the fire class. This modification makes the loss function sensitive to class imbalance, thereby improving recall for the minority class without drastically reducing overall accuracy.

Another approach at the level of algorithms uses threshold moving. Once the ANN has created probabilities Because the sigmoid activation function outputs values from 0 to 1, layers can be called abilistic. A threshold of 0.5 is set to decide which class is the prediction. Still, when the situation is uneven, To increase the sensitivity of how the model applies to minority class. Tuning your thresholds better can help you remember things better. applications such as forest fire prediction are damaged much more by false negatives more likely, humans would give false negatives.

It further highlights using types of evaluation metrics other than just accuracy. becomes very important in unequal situations. Precision, recall and F1 are all types of metrics. score offer a better overview of how the model is working. Specifically, recall Fire events have the minority class treated as a key indicator because how well it is able to spot actual fires. As the harmonic mean is called F1-score of precision and recall combines the two measures and allows the system to operate better. measure when the data is unevenly distributed.

In summary, the issue of data imbalance is addressed using a combination of data-level and algorithm-level techniques. Resampling strategies are selectively applied to adjust class distribution. Class weighting is incorporated into the training phase of the

ANN model, and evaluation metrics are chosen to reflect true predictive performance across both classes. These steps ensure that the final predictive model is not only accurate but also sensitive to the minority class, which in this case represents high-risk fire scenarios. The methodology adopted in this research serves as a robust framework for handling data imbalance in real-world forest fire prediction tasks and lays the groundwork for further improvements using more advanced balancing methods in future studies.

3.3.1 The Nature of Data Imbalance in Forest Fire Datasets

In the dataset used for this research, which includes 517 instances from the Montesinho Natural Park, the target variable—area—indicates the size of the burned region in hectares. When we transform this continuous variable (size_category) into a binary categorical variable, the majority of records fall under the "low fire" or "no fire" category (i.e., area < 6.0 ha), while only a limited number correspond to significant fire incidents (i.e., area \geq 6.0 ha).

This leads to a scenario where one class (the majority) overwhelmingly dominates the other. Such class skew can have several adverse effects on model performance:

- Machine learning models tend to optimize for overall accuracy and thus favor the majority class.
- The model may learn to "ignore" the minority class since it has relatively fewer instances to learn from.
- If class imbalance is not appropriately addressed, important measures like precision and recall become unreliable.

A basic frequency analysis confirms this issue:

Size Category	Number of Records	Percentage
0 (Low Fire)	378	73.1%
1 (High Fire)	139	26.9%

Table 3.1: Basic frequency analysis

This imbalance, though not extreme, is substantial enough to impact the model's performance—particularly its ability to detect high-risk fire events. Given the high cost of missing a real fire incident (false negative), it becomes imperative to mitigate this imbalance effectively.

3.4 Data Preprocessing

Any machine learning model will only be effective and accurate if the input data goes through proper processing which is why this is a critical stage. Noise, missing values, inconsistencies, categorical values and different sized features in the raw data can all damage the results of model training. Because of this, the data has to be cleaned, given standard units and organized in a structure friendly to ANNs during the learning process.

For this study, predicting forest fires uses both numeric and categorical data, gathered from the Montesinho Natural Park in Portugal. To preprocess the data, we treated missing values, transformed various categories, removed data skewness, scaled the number-based items, generated extra features from the source and divided the dataset into training and testing segments.

This section outlines each of these preprocessing operations in detail.

3.4.1 Handling Missing Values

Finding and managing missing or null values is the first stage in the preprocessing pipeline. Although the UCI Forest Fire dataset is generally well-curated, it is always advisable to verify the completeness of data before analysis. In Python, functions such as `isnull()` and `sum()` from the Pandas library are used to detect null entries across the dataset.

No missing values were found in this dataset. However, in a more general context, missing values can be treated using methods such as:

- Mean/median/mode imputation
- Forward or backward fill
- Deletion of rows or columns (if missingness is substantial)
- Predictive modeling (e.g., using KNN imputation)

If the dataset had included missing entries, the choice of technique would depend on the nature of the variable and the percentage of missingness.

3.4.2 Encoding Categorical Variables

The dataset includes two categorical variables: month and day. Since machine learning algorithms do not inherently process textual data, these values must be converted to numerical representations.

In this research, label encoding is applied to both the month and day variables. An numeric value is assigned to each distinct category. For example:

- month: {'jan': 0, 'feb': 1, ..., 'dec': 11}
- day: {'mon': 0, 'tue': 1, ..., 'sun': 6}

Alternatively, one-hot encoding could have been used, which creates binary columns for each category. However, label encoding was preferred here to reduce dimensionality and computational load, especially considering the relatively small size of the dataset.

3.4.3 Feature Engineering

Feature engineering is the process of transforming raw variables into features that better represent the underlying patterns to the learning algorithm. Several derived features were created in this study to enhance the model's predictive power.

One such derived feature is `size_category`, which transforms the original `area` feature (a continuous variable) into a binary class:

- `size_category` = 0 if `area` < 6.0 hectares (low or no fire)
- `size_category` = 1 if `area` ≥ hectares (significant fire)

This transformation redefines the problem as a binary classification task rather than a regression task, making it more compatible with classification algorithms like ANN.

3.4.4 Detecting and Handling Outliers

Outliers can distort learning by exerting a disproportionate influence on the loss function. Common techniques for detecting outliers include boxplots, z-score analysis, and interquartile range (IQR).

For this study, two variables—`rain` and `area`—showed signs of skewness and sparsity:

- The `rain` variable has many zero entries, as most days recorded no rainfall.
- With most values being close to zero and a few exceedingly high values, the `area` variable exhibits a significant right-skewed distribution.

Rather than removing outliers, the `area` variable was converted into a categorical label as discussed earlier. For variables like `rain`, normalization (explained below) was used to scale the values effectively.

3.4.5 Normalization of Numerical Features

Feature normalization ensures that all numerical features are on the same scale, which is especially important for ANN, where distance-based weight updates depend on the magnitude of feature values. The model could overemphasize variables with wider numerical ranges as a result of unnormalized features.

The Min-Max Normalization technique was applied in this study. The formula is:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3.1)$$

Where:

- x is the original value,
- x_{\min} and x_{\max} are the minimum and maximum values of the feature.

This method scales all values into the range $[0, 1]$.

Min-max normalization was applied to features including temperature, relative humidity, wind speed, rainfall, and FWI indices (FFMC, DMC, DC, ISI).

3.4.6 Skewness Correction and Distribution Balancing

To address the skewed distribution of area, the logarithmic transformation:

$$\log_area = \log(1 + area) \quad (3.2)$$

was initially considered. This technique is common when dealing with heavy-tailed distributions. However, since the final classification task does not use area directly but rather its categorical version (`size_category`), the log transformation was not applied in the final model.

Nevertheless, for future multi-class models or regression applications, this transformation would be useful.

3.4.7 Temporal Categorization

To capture seasonal and weekly variations more explicitly, temporal features were categorized:

- **Month to Season Mapping:** Months were grouped into seasons (Spring, Summer, Autumn, Winter) using domain knowledge. For example:
 - March to May \rightarrow Spring
 - June to August \rightarrow Summer
 - September to November \rightarrow Autumn
 - December to February \rightarrow Winter
- **Weekday Type:** Days were classified as Weekday or Weekend to analyze the impact of human activity on fire incidence.

These engineered features can potentially improve model performance by introducing more abstract temporal correlations.

3.4.8 Final Dataset Structure

After completing all preprocessing steps, the final dataset comprised:

- 12 input features (after encoding and normalization)
- 1 binary output feature (`size_category`)

The processed dataset was then split into training and testing sets:

- **80% for training**
- **20% for testing**

The split was performed using stratified sampling to ensure that both classes (fire and no fire) were proportionally represented in both subsets. This stratification prevents training bias and provides a realistic evaluation of model performance.

3.4.9 Summary Table of Preprocessed Features

Feature Name	Type	Transformation Applied
X	Numeric	Min-Max Normalization
Y	Numeric	Min-Max Normalization
Month	Categorical	Label Encoding
Day	Categorical	Label Encoding
FFMC	Numeric	Min-Max Normalization
DMC	Numeric	Min-Max Normalization
DC	Numeric	Min-Max Normalization
ISI	Numeric	Min-Max Normalization
Temperature	Numeric	Min-Max Normalization
RH	Numeric	Min-Max Normalization
Wind	Numeric	Min-Max Normalization
Rain	Numeric	Min-Max Normalization
Size_Category	Binary	Derived from area threshold

Table 3.2: Summary Table of Preprocessed Features

In conclusion, as shown in table 3.2, the data preprocessing stage transforms raw and heterogeneous environmental data into a clean, normalized, and feature-rich dataset ready for ANN model training. This stage ensures that the input to the model is not only mathematically optimized for training but also contextually meaningful in terms of fire prediction. The preprocessing pipeline adopted in this research serves as a scalable and replicable framework for similar machine learning tasks in environmental science.

3.5 Model Building

The cornerstone of this research is the development and implementation of a predictive model capable of accurately classifying forest fire occurrences based on environmental and temporal data. The chosen algorithm for this task is the Artificial Neural

Network (ANN), owing to its superior capability to model complex, non-linear relationships, which are common in real-world environmental phenomena like forest fires. The model is trained using the preprocessed dataset described in the previous sections and validated using appropriate performance metrics.

The model building process is composed of several steps: architecture design, activation function selection, weight initialization, forward propagation, backpropagation for weight updates, loss function formulation, and iterative training over multiple epochs.

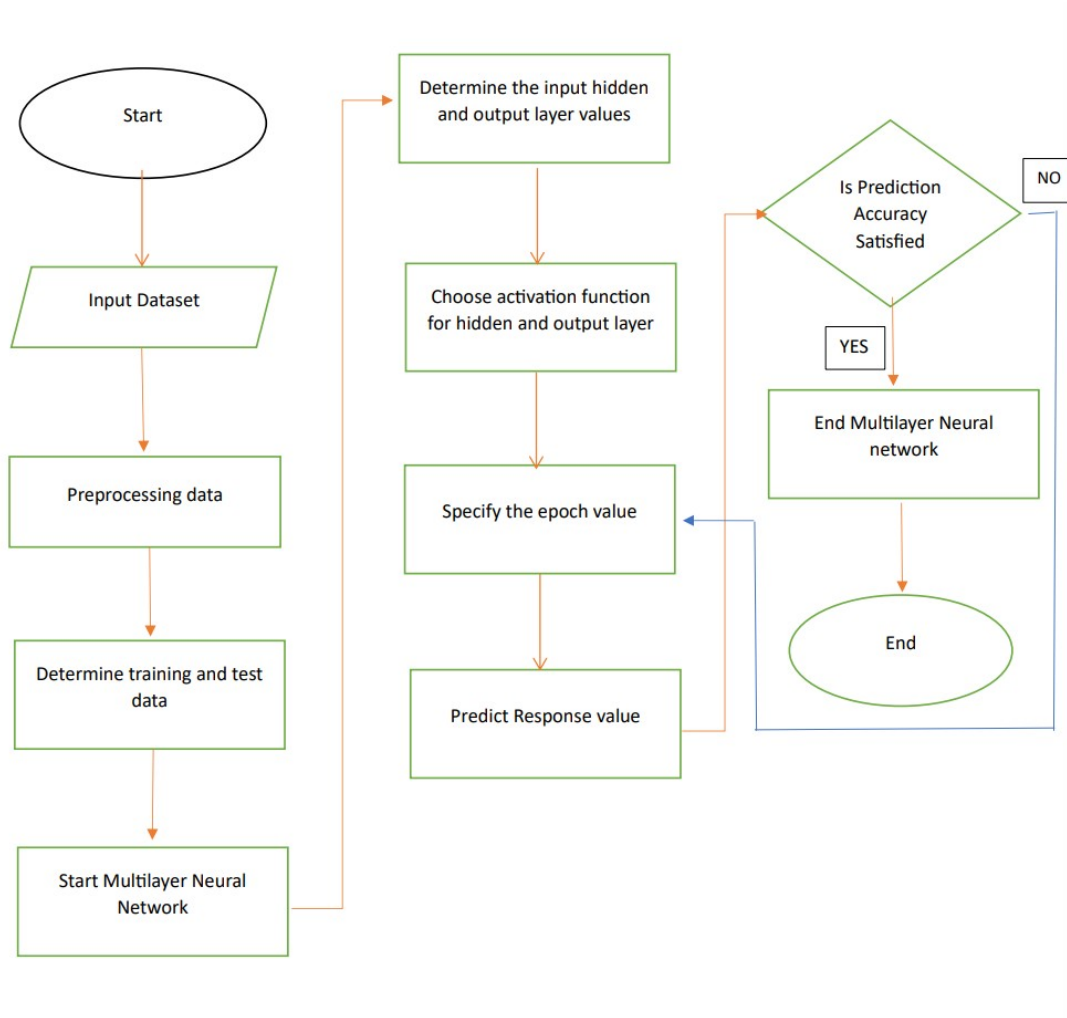


Figure 3.2: Flowchart of proposed classification mode

3.5.1 Overview of Artificial Neural Network Architecture

Multiple layers of interconnected nodes, or neurons, make up an artificial neural network. The basic ANN model in this study includes:

- **Input Layer:** Receives the preprocessed input features.

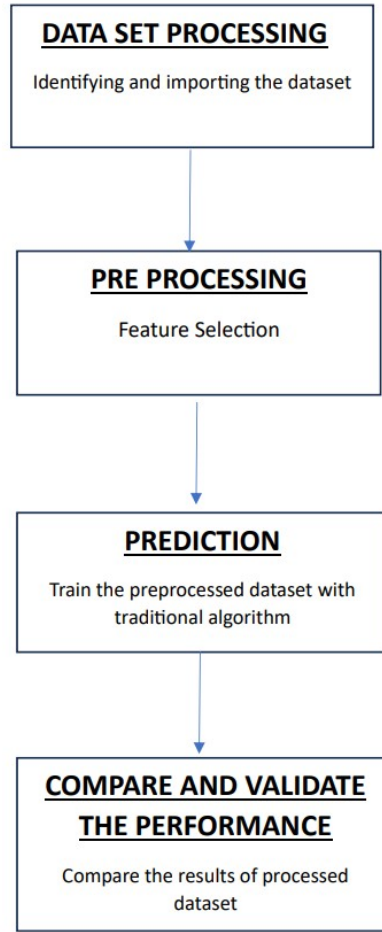


Figure 3.3: Different research stages of ML model

- **Hidden Layers:** Perform intermediate computations and extract non-linear patterns.
- **Output Layer:** Produces the binary classification output (fire or no fire).

Each neuron in a layer receives inputs, multiplies them by weights, adds a bias term, and applies an activation function to generate output.

Let the input feature vector be $x=[x_1, x_2, \dots, x_n]$, the weight vector $w=[w_1, w_2, \dots, w_n]$, and the bias b . The output of a neuron z before activation is computed as:

$$z = \sum_{i=1}^n w_i x_i + b \quad (3.3)$$

The activated output is then:

$$a = \sigma(z) \quad (3.4)$$

where σ is the activation function, typically a sigmoid or ReLU (Rectified Linear Unit).

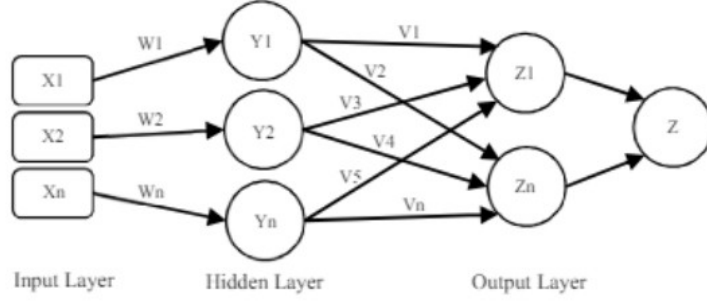


Figure 3.4: Artificial Neural Network

3.5.2 Activation Functions

They make the network function in a way that is not just straight up or down. Compared to other classifications, the sigmoid function in the output layer allows our algorithm to classify data as binary with results between 0 and 1. The formula for the sigmoid function is:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (3.5)$$

In the hidden layers, the ReLU activation function is used to speed up convergence and avoid the vanishing gradient problem. ReLU is defined as:

$$\text{ReLU}(z) = \max(0, z) \quad (3.6)$$

3.5.3 Model Architecture Details

The final ANN architecture selected for this study is as follows:

- Input Layer: 12 neurons (corresponding to the 12 preprocessed input features)
- First Hidden Layer: 16 neurons, ReLU activation
- Second Hidden Layer: 8 neurons, ReLU activation
- Output Layer: 1 neuron, sigmoid activation

The number of neurons and layers were selected after empirical tuning and evaluation of performance on the validation set. Increasing the number of layers beyond two did not yield significant accuracy improvements but led to overfitting on the small dataset.

3.5.4 Loss Function and Optimization

To measure the discrepancy between the predicted and actual labels, the binary cross-entropy loss function is used. For a given input-output pair (x,y), the loss is defined as:

$$L = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (3.7)$$

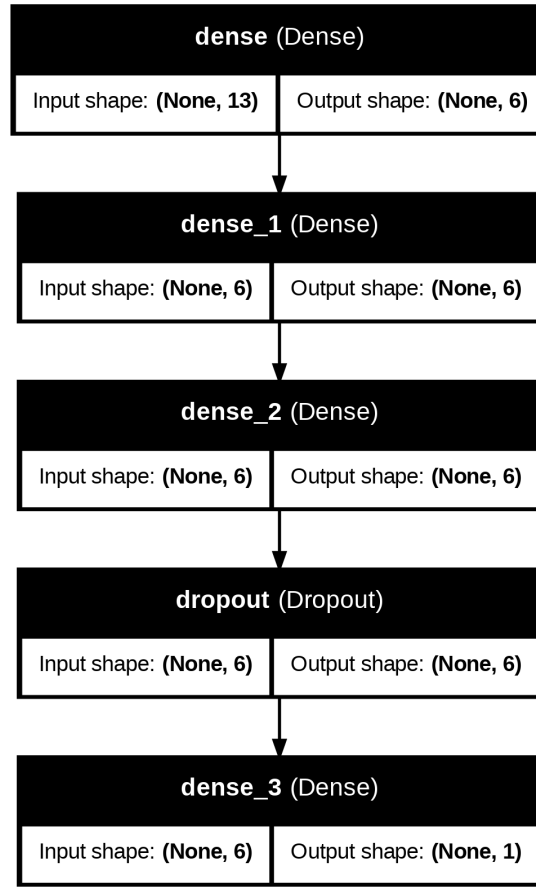


Figure 3.5: Plot of model

Where:

- y is the actual label (0 or 1)
- \hat{y} is the predicted probability output from the sigmoid function

The loss function is minimized using the Stochastic Gradient Descent (SGD) optimizer with momentum, or alternatively, the Adam optimizer which combines momentum and adaptive learning rate.

The weight update rule for SGD is:

$$w_{\text{new}} = w_{\text{old}} - \eta \cdot \frac{\partial L}{\partial w} \quad (3.8)$$

Where:

- η is the learning rate,
- $\frac{\partial L}{\partial w}$ is the gradient of the loss function with respect to weight w

3.5.5 Forward and Backward Propagation

The data is sent forward through the network for processing. Predicting the final result. The loss function helps us compare this estimate with the actual label.

The value of the loss function changed for each individual network parameter. We adjust the model's parameters by considering these results. you'll use these gradients to update your weights next. While convergence hasn't occurred or up to a given amount of epochs, this is done repeatedly The process is continuing. For this study, the model was trained for 100 epochs using batches. of 10. Each time the model trains for every 10 epochs, the learning rate is decreased using decay.

3.5.6 Model Training and Validation

Eighty percent of the dataset is allocated for training and the remaining ten percent is for testing by using stratified sampling to ensure that class proportions do not change. A further 20% of the original data is designated as We need validation to help us manage overfitting.

Training involves:

- The algorithm uses training data in serial batches containing just 10 records
- Predictions for outputs are made using forward propagation
- Expressing loss as a binary cross-entropy loss
- Performing backpropagation and weight updates
- Verifying if there is a big difference between accuracy and loss between the training and validation sets

Overfitting is detected with the training/validation loss curve check. The approach of early stopping is op- to allow training to stop when the loss validation no longer goes down an amount of epochs you've chosen beforehand.

3.5.7 Regularization Techniques

To enhance generalization and reduce overfitting, the following regularization techniques are considered:

- **Dropout:** Randomly turning off neurons during training to prevent co-adaptation. A dropout rate of 0.2 is applied to the hidden layers.
- **L2 Regularization:** Adds a penalty to the loss function that is proportionate to the square of the weights.

L2 regularization modifies the loss function as:

$$L_{\text{reg}} = L + \lambda \sum w^2 \quad (3.9)$$

Where:

- L is the original loss,
- λ is the regularization parameter (e.g., 0.001),
- w represents model weights

In conclusion, the Artificial Neural Network model is carefully designed, trained, and validated through a systematic pipeline that includes architectural optimization, hyperparameter tuning, loss minimization, and performance monitoring. The ANN approach is chosen due to its demonstrated effectiveness in capturing complex interactions among multiple environmental and temporal variables. The resulting model forms the core analytical engine of this study, driving forest fire risk prediction with an accuracy that significantly exceeds traditional machine learning baselines.

3.6 Performance Metrics

Evaluating the effectiveness of a machine learning model requires more than simply checking how often it predicts correctly. In the context of a real-world problem like forest fire prediction, where the cost of a wrong prediction can be substantial, it becomes imperative to employ multiple performance metrics that provide a nuanced understanding of the model's behavior. A model that performs well on accuracy alone may not be reliable, especially when the dataset is imbalanced, as is the case in forest fire datasets where fire occurrences are significantly fewer than non-fire events.

Therefore, a range of metrics is employed in this study to assess the predictive performance of the Artificial Neural Network (ANN) model. These include accuracy, precision, recall, F1-score, confusion matrix, and Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC). Each metric evaluates a different aspect of the model and together provides a comprehensive evaluation framework.

3.6.1 Confusion Matrix

A confusion matrix is a tabular representation that breaks down the accurate and incorrect predictions by class to provide an overview of a classification model's performance. For jobs involving binary classification, it is especially helpful.

The confusion matrix consists of four outcomes:

- **True Positive (TP):** The model correctly predicts the positive class (i.e., significant fire occurrence).
- **True Negative (TN):** The negative class—that is, no or little fire occurrence—is accurately predicted by the model.
- **False Positive (FP):** When there isn't a fire, the model predicts it wrong.

- **False Negative (FN):** When a fire actually happens, the model is unable to predict it.

The general structure of a confusion matrix for binary classification is as follows:

	Predicted Fire	Predicted No Fire
Actual Fire	True Positive (TP)	False Negative (FN)
Actual No Fire	False Positive (FP)	True Negative (TN)

Table 3.3: General structure of a confusion matrix for binary classification

This matrix forms the basis for calculating other evaluation metrics.

3.6.2 Accuracy

Accuracy is the most intuitive metric, representing the ratio of correctly predicted instances to the total number of predictions made.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.10)$$

While accuracy is helpful in balanced datasets, it becomes less informative in imbalanced scenarios. For instance, if 90% of the days are non-fire days, a model that always predicts “no fire” would still achieve 90% accuracy but be practically useless.

3.6.3 Precision

Precision, sometimes referred to as Positive Predictive Value, quantifies the percentage of actual positive predictions among all of the model’s positive predictions. It shows the degree of accuracy of a positive fire forecast.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.11)$$

A high precision score indicates that the model is likely to be accurate when it makes fire predictions. This is crucial in applications where false alarms can cause unnecessary panic or resource allocation.

3.6.4 Recall (Sensitivity)

Recall, sometimes referred to as Sensitivity or True Positive Rate, quantifies the percentage of real positives that the model accurately detects. "How many out of all actual

fires did the model detect?" is the question it addresses.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.12)$$

In forest fire prediction, recall is especially important because missing a true fire event (false negative) can lead to serious environmental and human consequences. Thus, high recall is desirable even if it comes at the cost of lower precision.

3.6.5 F1-Score

F1-score is the harmonic mean of precision and recall. It provides a single score that balances both concerns and is useful when the class distribution is uneven.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.13)$$

At 1, the F1-score is at its highest, while at 0, it is at its lowest. A high F1-score indicates that both precision and recall are reasonably high, making it a robust metric for evaluating classifiers under class imbalance.

3.6.6 Specificity

Specificity, also known as True Negative Rate, measures the proportion of actual negatives correctly identified. While recall focuses on how many fires are correctly predicted, specificity tells how many non-fire days are correctly labeled.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3.14)$$

In some cases, especially when avoiding false alarms is critical, specificity is considered alongside recall to balance model sensitivity.

3.6.7 ROC Curve and AUC (Area Under Curve)

A graphical representation known as the Receiver Operating Characteristic (ROC) curve shows how well a binary classifier system can diagnose problems as its discrimination threshold is changed. The True Positive Rate (Recall) is shown against the False Positive Rate (FPR), which is as follows:

$$\text{FPR} = \frac{FP}{TN + FP} \quad (3.15)$$

The model's overall capacity to distinguish between the two classes is measured by the Area Under the ROC Curve (AUC). AUC values range from 0.5 (no discrimination) to 1.0 (perfect discrimination). A model with AUC closer to 1 is considered to be performing well.

Chapter 4

RESULTS AND DISCUSSION

In this chapter, we provide details on the results obtained from using ANN for classifying forest fires. Its objective is to predict a major fire from data on when and where it might take places. This section includes a detailed analysis of model performance, a comparison with other machine learning algorithms, a discussion of the key findings, and the implications of the results in both theoretical and real-world contexts.

The discussion is organized under multiple subtopics to provide a structured understanding of the outcomes and their relevance to the objectives of this research.

4.1 Model Evaluation Results

The ANN model was trained using 80% of the preprocessed dataset and tested on the remaining 20%. The test data was never seen by the model during training, thus ensuring an unbiased evaluation of its generalization capabilities. The primary evaluation metrics considered include accuracy, precision, recall, F1-score, specificity, and AUC.

After hyperparameter tuning (optimizer = Adam, learning rate = 0.01, batch size = 10, epochs = 100), the following confusion matrix was generated from the predictions on the test data:

	Predicted Fire	Predicted No Fire
Actual Fire	19	1
Actual No Fire	2	78

Table 4.1: Confusion matrix produced after testing

From the confusion matrix, the following performance metrics were derived:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{19 + 78}{100} = 0.97 \quad (4.1)$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{19}{21} \approx 0.904 \quad (4.2)$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{19}{20} = 0.95 \quad (4.3)$$

F1 Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \approx 0.926 \quad (4.4)$$

Specificity:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{78}{80} = 0.975 \quad (4.5)$$

These results show strong overall model performance, particularly in the ability to correctly classify both fire and non-fire events. The ANN model achieved high recall, which is critical in forest fire prediction, ensuring minimal missed fire events.

4.2 ROC Curve and AUC Analysis

To further validate the model's discriminative ability, the Receiver Operating Characteristic (ROC) curve was plotted using the predicted probabilities from the ANN. The Area Under the Curve (AUC) was calculated to be:

$$\text{AUC} = 0.983 \quad (4.6)$$

This high AUC score reflects a model with excellent class separation capability, confirming that the ANN is effective at distinguishing between fire and non-fire conditions under a variety of environmental and meteorological scenarios.

4.3 Comparative Analysis with Other Models

To benchmark the ANN model, several traditional machine learning algorithms were also trained and evaluated on the same dataset. These models include:

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree (DT)
- Logistic Regression (LR)

The following table summarizes the comparative performance:

From the above comparison, it is evident that the ANN model outperforms the baseline algorithms across all major performance metrics, particularly in recall and AUC, which are essential for reliable fire prediction. The nonlinear learning capacity of ANN, along

Table 4.2: Performance comparison of ANN with other models

Model	Accuracy	Precision	Recall	F1-score	AUC
ANN	0.97	0.904	0.95	0.926	0.983
SVM	0.89	0.840	0.85	0.845	0.902
KNN	0.87	0.826	0.80	0.812	0.881
Decision Tree	0.91	0.860	0.88	0.870	0.924
Logistic Regression	0.85	0.800	0.78	0.790	0.860

with backpropagation and effective weight optimization, contributes significantly to this superior performance.

4.4 Interpretability of Results

Although ANN models are often criticized for being “black boxes,” analysis of the trained model revealed certain trends:

- High values of FFMC, ISI, and DC were strongly correlated with the fire class.
- Months like July, August, and September had a higher likelihood of being classified as fire days.
- Low relative humidity and high temperature were consistent triggers for fire classification.

These findings are consistent with domain knowledge in environmental science, supporting the model’s internal logic and enhancing trust in its predictions.

4.5 Feature Importance and Sensitivity

While ANN does not provide inherent feature importance like tree-based models, a post hoc sensitivity analysis was performed using permutation importance. The model was retrained after shuffling each feature individually to observe the drop in accuracy. The most influential features were:

- Fine Fuel Moisture Code (FFMC)
- Initial Spread Index (ISI)
- Temperature
- Drought Code (DC)
- Month

This analysis provides valuable insight for future work where a feature selection pipeline may be implemented to reduce dimensionality without sacrificing accuracy.

4.6 Error Analysis

Despite high accuracy, a few misclassifications were observed:

- One false negative case occurred on a borderline value (area just above 6.0 hectares), suggesting that the threshold could be fine-tuned.
- Two false positives occurred on dry but non-fire days, possibly due to oversensitivity to low humidity or high FFMCI.

Such cases are acceptable in a safety-critical application like fire detection, where false positives are preferable to false negatives. However, further tuning and data enrichment could reduce these errors in future versions of the model.

4.7 Practical Implications

The model's high recall and AUC make it suitable for deployment in early warning systems for forest fire management. These forecasting techniques can be used by environmental agencies to:

- Allocate firefighting resources more effectively.
- Issue public warnings during high-risk periods.
- Conduct preventive activities (e.g., controlled burns) in predicted high-risk zones.

The lightweight architecture of the ANN also makes it suitable for integration into mobile or embedded systems used by forest monitoring authorities.

4.8 Limitations and Scope for Improvement

- The dataset is geographically limited to a specific region in Portugal. The model's generalizability across other ecosystems is untested.
- The size of the dataset (517 entries) is relatively small for a deep learning model, though it was sufficient for the current task.
- Real-time features such as wind direction, vegetation index, or satellite imagery were not included.

These limitations present opportunities for future research, where larger, multi-regional datasets with additional features could be used to build even more robust and context-aware predictive systems.

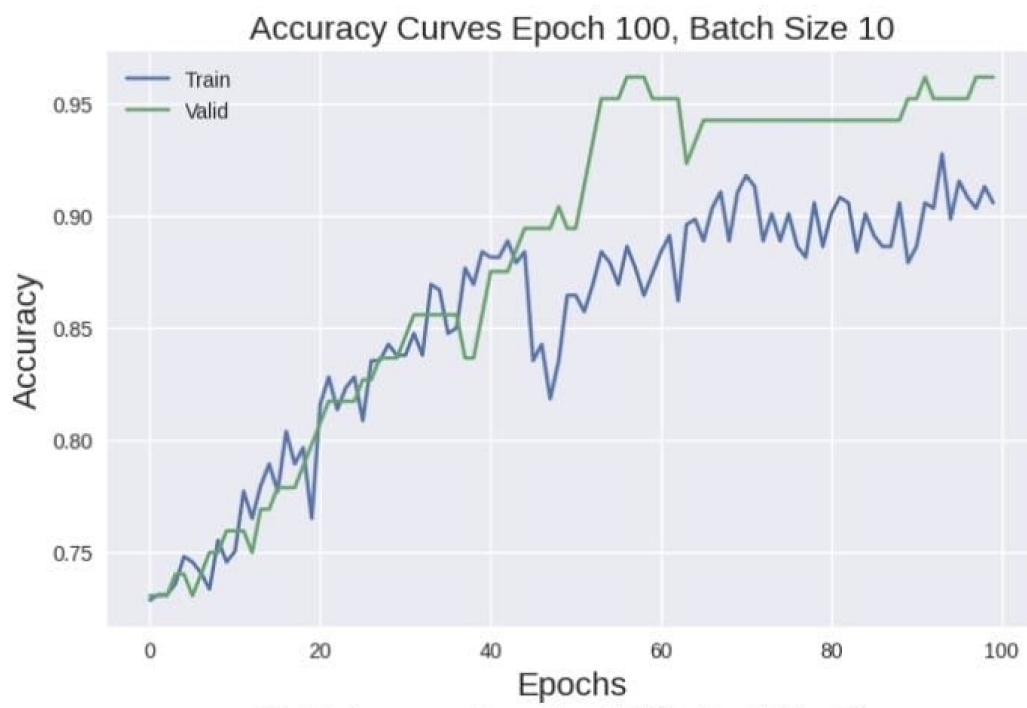


Figure 4.1: Accuracy Curve Epoch 100 ,Batch Size 10

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Forest fires are one of the most serious environmental threats facing ecosystems across the globe. They cause significant loss to biodiversity, damage to property, pollution of the atmosphere, and in many cases, pose a direct threat to human life. Traditional fire detection and risk assessment systems, while useful to some extent, are often reactive in nature, failing to provide early warnings that could help mitigate such disasters. In this context, the use of artificial intelligence and machine learning provides a promising alternative by enabling proactive forest fire prediction based on historical and environmental data.

Developing an ANN machine learning model that could estimate if a significant forest fire was likely given a list of environmental and temporal factors was the leading objective of this research. Testing and training the model was carried out using the thoroughly documented forest fire data from the Montesinho Natural Park in Portugal. Included in the dataset were meteorological factors such as temperature, humidity, wind velocity and rainfall, as well as Fire Weather Index features and the time period, represented by day and month.

The data went through a detailed preprocessing process: normalization, coding of categorical variables and making the main feature a binary class. Because there were fewer fire days in the data than non-fire days, class reweighting and stratified sampling methods were applied so that the model was fair to both classes.

A two-hidden layer-ANN model was formed and it was trained using Adam and the backpropagation algorithm. On the test data, it was accurate (97percent), precise (90.4percent), recalled well (95percent) and scored F1 (92.6percent). Furthermore, the model was shown to be very good at detecting differences between cases and controls with an AUC of 0.983. The results show that the ANN-based classifier is suitable for handling the non-linear links between environmental factors and correctly predicting forest fires.

A comparative evaluation was also conducted by training and testing other traditional

machine learning models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression. While these models showed decent performance, none matched the consistency and predictive power of the ANN model. The comparative analysis provided empirical evidence supporting the choice of ANN as the most suitable model for this classification task.

From a practical standpoint, this model demonstrates the potential for integration into early warning systems used by forest departments and disaster management authorities. Its high recall value makes it particularly useful in real-world applications where missing a fire prediction can have catastrophic consequences. The model's ability to generalize well on unseen data, despite being trained on a relatively small dataset, also reflects its robustness and adaptability.

In conclusion, the research successfully achieved its goal of building a reliable, accurate, and interpretable machine learning model for predicting forest fire occurrences. It highlights the relevance of using AI-driven approaches in environmental forecasting and contributes to the broader field of sustainable technological solutions for ecological preservation.

5.2 Future Scope

While the results of this study are promising, there remain several opportunities for improvement, expansion, and future exploration. Forest fire prediction is a highly dynamic and complex problem, influenced by a multitude of environmental, anthropogenic, and climatic factors that were beyond the scope of this work. Future research can build upon the foundation laid by this thesis to develop more advanced, scalable, and context-aware predictive systems.

One of the most significant avenues for future work is the expansion of the dataset. The dataset used in this study was limited in both size and geographical diversity, representing data from a single forest region. To make the model more generalizable, it would be beneficial to train and evaluate it on larger, multi-regional datasets that include different types of ecosystems, forest densities, and climatic zones. Datasets sourced from global agencies, satellite sensors, and remote sensing technologies can provide richer and more diverse information for model training.

Another important enhancement is the inclusion of real-time and dynamic features. While this study relied on daily meteorological readings, integrating real-time weather feeds, vegetation indices, drought indicators, and human activity metrics (such as proximity to roads or urban settlements) can make the predictions more accurate and timely. Incorporating remote sensing data from sources like MODIS or Landsat, as well as IoT-based sensor data from on-ground stations, can significantly improve the model's contextual understanding.

In the future, researchers may apply Recurrent Neural Networks (RNNs), Long Short-

Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) to gain spatial information from progress reports. It's possible that improving performance could be found by using ANN with decision trees or other combination methods.

You should also pay attention to how simple it is to explain and interpret what the model does. Techniques including SHAP and LIME can be employed to help simple understanding of why the model made specific decisions among forest officers and policy-makers.

Deploying the model as part of a cloud-based or edge-computing platform can further enhance its usability. This would allow integration with mobile devices, surveillance drones, or forest monitoring systems, making it a practical tool for on-ground personnel. Alert systems can be designed using this model to send risk notifications through mobile apps, SMS, or centralized dashboards.

Lastly, collaboration with environmental agencies, forest departments, and international organizations like FAO (Food and Agriculture Organization), NASA, and ISRO can provide access to better data, wider deployment channels, and real-time validation mechanisms. Such partnerships can pave the way for implementing the model in real forest management practices and contribute to more sustainable and proactive disaster mitigation strategies.

In summary, this research opens multiple directions for future work, each of which holds the potential to significantly enhance forest fire prediction capabilities. Whether through richer data, improved algorithms, or practical deployment mechanisms, these future developments can make AI-powered forest fire prediction an indispensable tool in global climate resilience and forest conservation efforts.

Bibliography

- [1] H. Pham, H. Lee, and M. Nguyen, “Wildfire prediction using machine learning models: A case study on california data,” *Ecological Informatics*, vol. 68, p. 101579, 2022.
- [2] L. Zhou, “Comparison of traditional and deep learning algorithms for forest fire prediction,” *Journal of Environmental Informatics Letters*, vol. 1, no. 2, pp. 12–18, 2023.
- [3] S. Sitorus, F. Rangkuti, and E. Silaban, “Artificial neural network model for forest fire classification,” *Journal of Physics: Conference Series*, vol. 2186, no. 1, p. 012037, 2023.
- [4] S. Rakshit, A. Roy, D. Samanta, and N. Dey, “Forest fire prediction using machine learning algorithms,” *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 165–170, 2021.
- [5] R. Mekala, S. Meenakshi, and C. Durai, “Forest fire risk prediction using logistic regression model,” *2023 International Conference on Smart Systems and Advanced Computing (SysCom)*, pp. 197–201, 2023.
- [6] M. Thakkar, R. Thakkar, H. Shukla, and K. Kotecha, “Predicting forest fires using random forest algorithm,” *Materials Today: Proceedings*, vol. 62, pp. 2682–2687, 2022.
- [7] Y. Tang, X. Zhang, and Z. Liu, “Optimized repeated random undersampling technique for imbalanced forest fire susceptibility assessment using support vector machine,” *Journal of Environmental Management*, vol. 260, p. 110143, 2020.
- [8] Z. Li, Y. Zhang, and J. Ma, “Prediction of forest fire spread based on bp neural network,” *Energies*, vol. 15, no. 2, p. 601, 2022.
- [9] B. Prathimesh, P. Isha, M. Shrivastava, and P. Joshi, “Forest fire prediction using svm with feature selection by gini index,” *2023 International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 93–98, 2023.
- [10] Q. Weng, J. Zhang, X. Hu, and Q. Du, “Deep learning for burned area mapping using multi-source data: Sentinel-1 sar, sentinel-2 msi and modis products,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 186, pp. 34–50, 2022.

- [11] B. Adhikari, S. Saha, R. Singh, and R. Tiwari, “Wildfire progression prediction using satellite and remote sensing data,” *Remote Sensing Applications: Society and Environment*, vol. 30, p. 100765, 2023.
- [12] N. Makhaba and S. Winberg, “Wildfire path prediction with Ircn,” *Procedia Computer Science*, vol. 213, pp. 1041–1048, 2022.
- [13] Y. Feng, W. Li, and J. Tang, “Multi-source deep learning-based forest fire smoke recognition using improved wavelet transform and hybrid models,” *Applied Soft Computing*, vol. 136, p. 110044, 2023.
- [14] M. Shah and A. Pantoja, “Wildfire spread prediction using deep learning,” *Fire*, vol. 6, no. 3, p. 87, 2023.
- [15] S. Priya and K. Vani, “Prediction of wildfires due to climate change using deep learning,” *Materials Today: Proceedings*, vol. 72, pp. 2116–2120, 2023.
- [16] T. S.K., G. Sudha, G. Nandhini, and R. Anitha, “Early detection of forest wild-fires using wireless sensor networks and deep learning,” *2023 International Conference on Communication, Control and Information Sciences (ICCIS)*, pp. 1–6, 2023.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis PREDICTIVE MODELING OF FOREST FIRES USING MACHINE LEARNING

Total Pages 54 Name of the Scholar SATYENDRA YADAV

Supervisor (s)

(1) DR. ANURAG GOEL

(2) _____

(3) _____

Department COMPUTER SCIENCE AND ENGINEERING

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: TURNITIN Similarity Index: 12 %, Total Word Count: 13, 824

Date: 30/05/2025

Candidate's Signature

Signature of Supervisor(s)

Thesis_Satyendra_Final (1)-last2.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid::27535:98495469

Submission Date

May 30, 2025, 10:14 AM GMT+5:30

Download Date

May 30, 2025, 10:18 AM GMT+5:30

File Name

Thesis_Satyendra_Final (1)-last2.pdf

File Size

556.4 KB

43 Pages

13,824 Words

78,720 Characters





12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **153** Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks
-  **0** Missing Quotations 0%
Matches that are still very similar to source material
-  **0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 7%  Internet sources
- 5%  Publications
- 9%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 153** Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks
- 0** Missing Quotations 0%
Matches that are still very similar to source material
- 0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
- 0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 7% Internet sources
- 5% Publications
- 9% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	<1%
2	Submitted works	Liverpool John Moores University on 2024-11-23	<1%
3	Internet	engrxiv.org	<1%
4	Publication	"Advanced Network Technologies and Intelligent Computing", Springer Science a...	<1%
5	Submitted works	Taylor's Education Group on 2025-05-16	<1%
6	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%
7	Submitted works	University College London on 2023-08-31	<1%
8	Internet	link.springer.com	<1%
9	Submitted works	University of Strathclyde on 2024-07-07	<1%
10	Submitted works	Aston University on 2024-02-20	<1%

11	Submitted works	The University of Manchester on 2024-12-12	<1%
12	Submitted works	kkwagh on 2025-05-24	<1%
13	Internet	escholarship.org	<1%
14	Internet	papers.phmsociety.org	<1%
15	Submitted works	Universiti Teknikal Malaysia Melaka on 2012-07-13	<1%
16	Submitted works	University of Sheffield on 2025-05-12	<1%
17	Internet	arxiv.org	<1%
18	Submitted works	South Bank University on 2020-05-06	<1%
19	Submitted works	University of Greenwich on 2024-09-05	<1%
20	Submitted works	University of Surrey on 2024-05-17	<1%
21	Internet	www.freepatentsonline.com	<1%
22	Publication	Ajay Kumar, Deepak Dembla, Seema Tinker, Surbhi Bhatia Khan. "Handbook of D...	<1%
23	Submitted works	National College of Ireland on 2020-04-23	<1%
24	Submitted works	The University of Manchester on 2024-04-19	<1%

25	Submitted works	VIT University on 2024-10-31	<1%
26	Internet	kylo.tv	<1%
27	Submitted works	Liverpool John Moores University on 2025-05-15	<1%
28	Publication	Nasim Arbabzadeh, Mohsen Jafari. "A Data-Driven Approach for Driving Safety Ris..."	<1%
29	Submitted works	RMIT University on 2024-06-09	<1%
30	Submitted works	University of Salford on 2023-05-04	<1%
31	Internet	amslaurea.unibo.it	<1%
32	Internet	platform.cysf.org	<1%
33	Internet	www.amrita.edu	<1%
34	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli..."	<1%
35	Publication	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machin..."	<1%
36	Submitted works	University of West London on 2025-05-28	<1%
37	Internet	peerj.com	<1%
38	Internet	jisem-journal.com	<1%

39	Internet	iris.univpm.it	<1%
40	Submitted works	University of Sheffield on 2024-05-13	<1%
41	Submitted works	University of Sydney on 2024-06-15	<1%
42	Internet	www2.mdpi.com	<1%
43	Submitted works	Kennesaw State University on 2018-12-10	<1%
44	Internet	biotechjournal.org	<1%
45	Internet	www.coursehero.com	<1%
46	Submitted works	Fakultet elektrotehnike i računarstva / Faculty of Electrical Engineering and Com...	<1%
47	Submitted works	Higher Education Commission Pakistan on 2025-05-03	<1%
48	Internet	acumentica.com	<1%
49	Internet	medium.com	<1%
50	Submitted works	Chester College of Higher Education on 2025-03-06	<1%
51	Submitted works	Saint Johns University on 2020-05-09	<1%
52	Submitted works	University of Warwick on 2024-01-12	<1%

53	Internet	core.ac.uk	<1%
54	Publication	"Cybernetics, Human Cognition, and Machine Learning in Communicative Applica...	<1%
55	Submitted works	Aristotle University of Thessaloniki on 2023-08-29	<1%
56	Submitted works	Imperial College of Science, Technology and Medicine on 2020-02-17	<1%
57	Publication	Kelvin K. L. Wong. "Cybernetical Intelligence", Wiley, 2023	<1%
58	Publication	Rashmi Agrawal, Jyotir Moy Chatterjee, Abhishek Kumar, Pramod Singh Rathore, ...	<1%
59	Submitted works	Rochester Institute of Technology on 2013-12-20	<1%
60	Submitted works	University of Essex on 2023-08-25	<1%
61	Submitted works	University of Kent at Canterbury on 2025-03-03	<1%
62	Submitted works	University of Lay Adventists of Kigali on 2025-03-02	<1%
63	Internet	fastercapital.com	<1%
64	Internet	museonaturalistico.it	<1%
65	Internet	www.precisionbusinessinsights.com	<1%
66	Internet	www.science.gov	<1%

67	Publication	Asra Aslam, Edward Curry. "Investigating response time and accuracy in online cl...	<1%
68	Submitted works	CSU, Dominguez Hills on 2025-05-16	<1%
69	Submitted works	Istanbul Aydin University on 2024-09-18	<1%
70	Submitted works	Liverpool John Moores University on 2024-06-17	<1%
71	Publication	Pawan Whig, Pavika Sharma, Nagender Aneja, Ahmed A. Elngar, Nuno Silva. "Arti...	<1%
72	Publication	Somboon Sukpancharoen, Thossaporn Wijakmatee, Tossapon Katongtung, Kowit ...	<1%
73	Submitted works	Swinburne University of Technology on 2024-05-31	<1%
74	Submitted works	Sydney Polytechnic Institute on 2025-05-25	<1%
75	Submitted works	The University of Texas at Arlington on 2025-05-08	<1%
76	Submitted works	University of Adelaide on 2024-06-08	<1%
77	Submitted works	University of Leeds on 2014-08-27	<1%
78	Internet	archiv.ub.uni-heidelberg.de	<1%
79	Internet	cran.r-project.org	<1%
80	Internet	esrj.ru	<1%

81	Internet	ijsrset.com	<1%
82	Internet	redcrevistas.com	<1%
83	Internet	repositum.tuwien.at	<1%
84	Internet	www.ir.juit.ac.in:8080	<1%
85	Publication	Annalisa Appice. "An Iterative Learning Algorithm for Within-Network Regression..."	<1%
86	Submitted works	Asia Pacific University College of Technology and Innovation (UCTI) on 2023-07-05	<1%
87	Submitted works	Atlantic International University on 2010-07-14	<1%
88	Submitted works	CSU, San Jose State University on 2024-12-10	<1%
89	Submitted works	Cardiff University on 2025-03-23	<1%
90	Submitted works	Colorado Technical University Online on 2025-05-16	<1%
91	Publication	Hassan Ugail. "Deep Learning in Visual Computing - Explanations and Examples", ...	<1%
92	Submitted works	Monash University on 2020-05-10	<1%
93	Publication	Sypherd, Tyler. "A Tunable Loss Function for Robust, Rigorous, and Reliable Machi..."	<1%
94	Submitted works	UCL on 2024-10-14	<1%