Dataset Creation and Imbalance Mitigation in Big Data: Enhancing Machine Learning Models for Forest Fire Prediction

by

Fatemeh Tavakoli

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made available electronically to the public.

Abstract

Historically, forest fire prediction methods have leaned on heuristics, local insights, and basic statistical models, often neglecting the complex interplay of variables such as temperature, humidity, wind speed, and vegetation type. The lack of real-time prediction capabilities, paired with unpredictable weather patterns attributed to climate change, underscores the shortcomings of traditional methods, especially in geographically varied regions like Canada. In contrast, machine learning provides the adaptability needed for real-time responses, effectively harnessing updated data and addressing region-specific forest fire risks. The shift towards machine learning is both a timely and revolutionary approach.

This research addresses the urgent need for effective forest fire prediction and management strategies, specifically in the Canadian context, by harnessing machine learning methodologies. Using Copernicus's reanalysis data, this study establishes a comprehensive predictive framework employing four cutting-edge machine learning algorithms. Random Forest, XGBoost, LightGBM, and CatBoost. The study features a robust data preprocessing pipeline, class imbalance correction, and rigorous model evaluation measures. Key contributions include the creation of a feature-rich dataset, comprehensive methods for addressing the class imbalance in large scale datasets, and the development of a machine learning framework tailored for forest fire classification. The findings have significant implications for data-driven forest management strategies, with the aim of facilitating proactive fire prevention measures on a large scale.

One primary challenge encountered was the inherent class imbalance in fire classification datasets, with a striking 158:1 ratio between "non-fire" and "fire" events. To address this, the study utilized various re-sampling strategies, encompassing under-sampling, oversampling, and hybrid techniques. Specific methods employed included NearMiss, SMOTE, and SMOTE-ENN. The NearMiss method with a 0.09 sampling ratio was found to be particularly effective in addressing this imbalance. When combined with NearMiss version 3 at a 0.09 ratio, the XGBoost model outperformed its peers, showcasing an accuracy of 98.08%, a sensitivity of 86.06%, and a specificity of 93.03%. The findings indicate that while high recall from NearMiss Version 3 optimized sensitivity, there was sometimes a trade-off with precision.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Sagar Naik, for his unwavering support, insightful critiques, and patient mentorship throughout this research journey.

Special thanks go to my committee members, Dr. Ayman El-Hag and Dr. Ramadan El Shatshat, for their invaluable feedback and constructive suggestions.

I am grateful to my research group Dr. Marzia Zaman, Dr. Srini Sampalli, Dr. Chung-Horng Lung, Parveen Kaur, Richard Purcell, and Abdul Mutakabbir for their companionship and support.

Dedication

I dedicate this thesis to my parents, Aliakbar and Hooriyeh, who have always taught me to be courageous and to step outside my comfort zone. Just as they did when they journeyed from our hometown of Janah, Iran, to Kish Island, then to Tehran, and eventually to Dubai, seeking a brighter future. Soon, they will be embarking on a new chapter in Toronto. I am forever grateful for your love and support.

To Uncle Farhad and Aunt Maryam, and my cousins, Saraa and Ahmad: When Dubai was unfamiliar territory, you became my second family. You always reminded me of the values of kindness and generosity. Your unwavering support and encouragement have been invaluable in my journey of knowledge and personal growth.

To my sister, Samin, and my brother, Sharif: You have always been there for me in every way. You understand and know me in ways I sometimes can't even comprehend, seeing depths in me I often overlook.

To my partner, who taught me to simply be myself and find happiness in little things. Your love and understanding have been a constant source of strength throughout my journey.

To my late grandmother, Nana, whom we lost on 9 September 2023. Despite raising a large family of 10 children, she opened her arms to me when I relocated to Dubai.

Finally, to the new beginnings.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CanFIRE Canadian Fire Effects Model 12

CatBoost Categorical Gradient Boosting 3, 4, 7, 8, 18, 21, 27, 32, 47, 54, 55, 57

CFFWI Canadian Forest Fire Weather Index 12, 13, 40

CIFFC Canadian Interagency Forest Fire Center 1, 11

CNNs Convolutional Neural Networks 2

CSV Comma Separated Values 40, 41

CWFIS Canadian Wildland Fire Information System 6, 8, 11–13, 41

DTs Decision Trees 22, 23

ECMWF European Centre for Medium-Range Weather Forecasts 39

EFB Exclusive Feature Bundlin 25

ENN Edited Nearest Neighbors 31, 50, 51

ERA European Environment Agency 39, 40, 45

FWI Fire Weather Index 12, 13

GANs Generative Adversarial Networks 4, 6, 16, 18, 28, 30, 31, 49, 50, 57, 58

GBDT Gradient Boosted Decision Trees 24

GOSS Gradient-based One-Side Sampling 25

LightGBM Light Gradient Boosting Machine 3, 4, 7, 8, 18, 21, 25–27, 32, 47, 49, 52, 53, 55–57

LSTM Long Short-Term Memory 17

ML Machine Learning 2, 3, 5–8, 10, 14, 16–18, 20, 21, 23, 24, 28, 32, 35, 47, 55, 57, 58

NearMiss3 NearMiss Version 3 4, 6, 18, 28, 30, 49, 50, 52, 53, 55–57

NetCDF Network Common Data Form 40

QGIS Quantum Geographic Information System 41

SMOTE Synthetic Minority Oversampling Technique 4, 6, 16, 18, 28, 30, 31, 49–53

SMOTE-ENN Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) 4, 6, 16, 18, 28, 31, 49–53

SVMs Support Vector Machines 2, 17

XGBoost Extreme Gradient Boosting 3, 4, 7, 8, 16, 18, 21, 24, 27, 32, 47, 49, 52, 53, 55–57

Chapter 1

Introduction

1.1 Motivation

The growing occurrence of forest fires poses a significant threat not only to wildlife and the environment but also to human settlements. Factors like climate change, human activities, and limited resources for prevention and management intensify the challenges associated with forest fires. Forest fires can have devastating impacts, including the loss of biodiversity, destruction of habitats, air pollution, and economic setbacks. Consequently, there is an essential need to develop effective strategies for forest fire prevention, early detection, and rapid response to mitigate their adverse effects and protect both the environment and communities.

The alarming rise in global forest fire incidents serves as an urgent call to action for innovative solutions. A recent study from the University of Maryland starkly illustrates the severity of the situation: forest fires have resulted in nearly double the amount of tree cover loss today compared to two decades ago [44]. In 2021 alone, forest fires accounted for a shocking 9.3 million hectares of global tree cover loss, signaling that the situation has reached a critical tipping point.

Particularly in Canada, the recent surge in forest fires serves as an imperative for innovative intervention. According to the Canadian Interagency Forest Fire Center (CIFFC), an unprecedented 9.5 million hectares of land were burned between January and July 2023 alone, an area equivalent to the size of Portugal [7]. This dramatic increase in forest fire activity in Canada coincides with warmer than average temperatures and drought conditions, highlighting an urgent need for predictive models that can adapt to rapidly changing environmental variables.

ML presents a particularly promising avenue for addressing this issue. Traditional methods of prediction and management have proven inadequate given the scale and complexity of forest fires in Canada. ML algorithms have the ability to analyze vast and intricate datasets, incorporating variables such as temperature, humidity, and drought conditions, to provide more accurate and timely predictions. Considering the increasing rates of tree cover loss and Canada's unique climate challenges, such as rapidly warming high-latitude regions, there is a pressing need for a ML-based approach to forest fire prediction. This is crucial not only for safeguarding Canada's natural resources but also for protecting communities and mitigating the devastating economic and environmental impacts of increasingly frequent fires. Therefore, research that uses ML for forest fire prediction is not merely an academic exercise, but a social imperative, especially in the Canadian context.

ML models can be used in the forest fire domain to improve our ability to detect and predict forest fires, as well as to optimize the allocation of resources for prevention and management. There are some ways that ML models can be used in this domain:

- Forest fire detection: ML models can be trained on satellite imagery and other sensor data to detect the signs of a forest fire, such as smoke plumes or changes in temperature. These models can use various techniques such as Convolutional Neural Networks (CNNs) or Support Vector Machines (SVMs) to identify patterns that indicate a fire and alert authorities to its location.

- Forest fire prediction: ML models can be trained on historical data on weather conditions, vegetation, and other factors that contribute to forest fires to predict the likelihood of a fire occurring in a given area. These models can use techniques such as decision trees, random forests, or neural networks to identify the factors that are most predictive of forest fires and generate forecasts for future periods.

- Resource allocation: ML models can be used to optimize resource allocation for the prevention and management of forest fires. For example, models can be trained on data on the location, size, and intensity of past fires, as well as data on the availability of resources such as firefighters and equipment. These models can then be used to determine the optimal allocation of resources to different areas to minimize the damage caused by fires.

- Postfire Analysis: ML Models can be used to analyze the impact of forest fires on the environment and predict the recovery rate. For example, models can be trained on satellite imagery and other data to track changes in vegetation over time and predict the rate of regrowth after a fire.

1.2 Objective

The principal objective of this research is to establish a reliable and scalable predictive framework for forest fires utilizing ML methodologies. Given the focus on supervised classification, the study aims to precisely distinguish between "fire" and "non-fire" conditions based on Copernicus reanalysis data and four prominent ML algorithms: Random Forest, XGBoost, LightGBM, and CatBoost. The feature set comprises **T**emperature, **S**oil **W**ater, **E**vaporation, **R**unoff, **W**ind, **P**ressure, **P**recipitation, and **V**egetation. Figure 1.1 illustrates this research objective. Labeled boxes represent the contributions, which are explained in the following section in order.

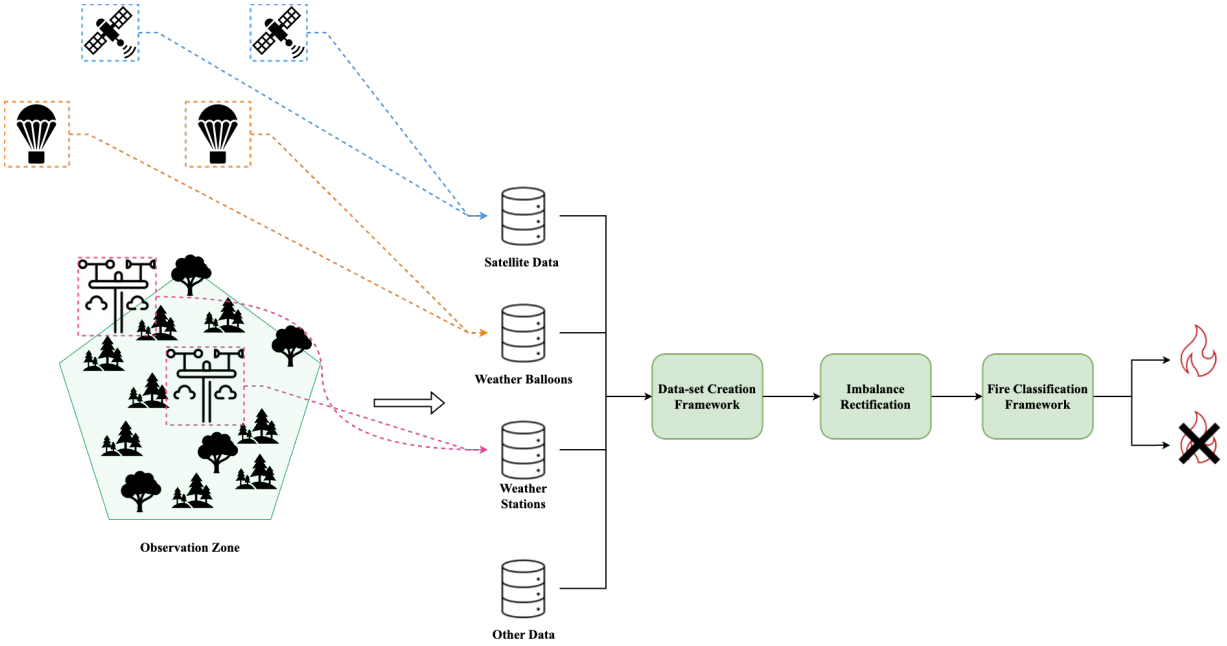


Figure 1.1: System Model

Rationale for Using a Single Data Source

A unique aspect of this study is the exclusive reliance on Copernicus reanalysis data, rather than combining multiple data sources. This decision is motivated by the inherent inconsistencies in spatial and temporal resolutions when fusing data from multiple sources.

Additionally, using satellite imagery for large areas is computationally intensive and time-consuming, making it less feasible for the scope of this study, which encompasses extensive geographic coverage over several years.

Specific Objectives

1. **Data Collection and Pre-processing:**
   - To collect and compile Copernicus reanalysis data focused on features $\mathcal{F} = \{T, SW, E, R, W, P, Pr, V\}$.
   - To perform robust pre-processing techniques for data normalization, outlier treatment, and handling of missing values.
2. **Handling Imbalanced Datasets:**
   - To identify and correct for class imbalance in the dataset, which is especially crucial given the large spatial and temporal extent of the study.
   - To employ techniques such as NearMiss3, SMOTE, SMOTE-ENN, or GANs links to achieve a balanced dataset.
3. **Model Development and Training:**
   - To implement Random Forest, XGBoost, LightGBM, and CatBoost algorithms ($M \in \{\text{Random Forest}, \text{XGBoost}, \text{LightGBM}, \text{CatBoost}\}$).
   - To train the models on a labeled dataset $\mathcal{D}$ tagged as "fire" and "non-fire." Mathematically, this is framed as finding a function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the feature space and $\mathcal{Y} = \{\text{"fire"}, \text{"non-fire"}\}$.
4. **Model Evaluation and Comparison:**
   - To use performance metrics like accuracy, recall (Sensitivity), Specificity, Weighted F1 score and ROC-AUC, formally defined as $\text{Metric}_M(\mathcal{D})$.
   - To conduct statistical tests to distinguish the most effective model from the set M.
5. **Interpretation and Analysis:**
   - To provide ecological and computational perspectives on the results.

- To identify the most influential features based on feature significance metrics.

6. **Validation and Generalization:**
   - To validate the model on an unseen dataset, thereby assessing its applicability in real-world scenarios.
   - To evaluate the model's generalization capabilities across varying geographic and climatic conditions.

By fulfilling these objectives, the research aims to significantly contribute to the domains of ML and environmental science. The findings are expected to influence data-driven forest management strategies and facilitate proactive fire-prevention measures on a large scale.

1.2.1 Research Question

1. How reliable is a single data source like Copernicus reanalysis data for predicting forest fires?
2. What are the limitations, if any, of exclusively relying on Copernicus reanalysis data for forest fire prediction?
3. What challenges and opportunities arise when working with large, imbalanced datasets in the context of forest fire prediction?
4. What types of environmental characteristics are the most predictive of forest fires?
5. How do different ML models perform in terms of forest fire prediction, and is there a "best" model?
6. How well do the predictive models generalize to new, unseen data or different geographical areas?
7. What are the practical implications of this research for forest management and fire prevention policies?

1.2.2 Contributions

This research endeavors to bridge existing gaps in the field of forest fire prediction through methodological, computational, and practical contributions. Spanning multiple facets from data collection to model application, the study pioneers several key advancements. Specifically, this research contributes in three main ways:

C1 Data-set Creation Framework
One of the pivotal contributions of this study is the assembly of a comprehensive dataset. Covering the area of interest, this data set uniquely offers 27 distinct features that encapsulate a broad spectrum of information, ranging from spatial and topographical elements to temporal aspects. We combined data from sources such as Copernicus Reanalysis Climate data [5], the CWFIS Datamart [37], Statistics Canada [6], and ArcGIS RESET [14]. By meticulously gathering and encoding historical "fire" data points as 1 and "non-fire" cells as 0, the dataset serves as a robust foundation for evaluating ML models. This feature-rich dataset is invaluable for its potential to contribute to future research and practical applications in the realm of forest fire management.

- *Use of Copernicus Reanalysis Data*
  The dataset is primarily sourced from Copernicus reanalysis data, chosen for its reliability and comprehensive coverage. This choice enables consistency in the spatial and temporal dimensions across the data.

C2 Handling Imbalance and Class Distribution in Large-Scale Datasets
This study addresses the pervasive issue of class imbalance commonly found in large datasets, a challenge that can compromise the performance and reliability of ML models. We introduce a method to handle the pronounced 158:1 class imbalance ratio between "non-fire" and "fire" events. To counteract this imbalance, sophisticated under-sampling, over-sampling, and hybrid re-sampling techniques are employed. Specifically, we utilize NearMiss3, SMOTE, and SMOTE-ENN to adjust the distribution of the minority and majority classes in the dataset.

Furthermore, this research ventures into exploratory analyses concerning the application of GANs for the purpose of oversampling in tabular data settings. By pioneering the use of GANs in this specific context, the study opens new avenues for enhancing model performance when tackling imbalanced datasets.

C3 ML Framework for Forest Fire Classification

A secondary, yet vital, contribution lies in the development of a ML framework tailored specifically for forest fire classification.Forest fire classification, in this context, refers to the process of categorizing a given geographical region or a dataset into "fire" or "non-fire" based on certain environmental and climatic features. Let $\mathcal{F}$ represent the feature set for forest fire classification, defined as:

$$\mathcal{F} = \{T, SW, E, R, W, P, Pr, V\}$$

Where:

$$\begin{aligned} T &: \text{Temperature} \\ SW &: \text{Soil Water} \\ E &: \text{Evaporation} \\ R &: \text{Runoff} \\ W &: \text{Wind} \\ P &: \text{Pressure} \\ Pr &: \text{Precipitation} \\ V &: \text{Vegetation} \end{aligned}$$

These features serve as predictors to determine the likelihood or risk of a forest fire occurrence in a particular area. Leveraging state-of-the-art algorithms like Random Forest, XGBoost, LightGBM, and CatBoost, the framework aims to provide robust and scalable solutions for predicting forest fires. This represents a significant step forward in employing advanced ML techniques for environmental and ecological applications.

Finally, the research goes beyond mere theoretical postulations by rigorously testing the best-performing model on unseen, out-of-sample data. This provides a critical evaluation of the model's real-world applicability and generalizability, thus affirming its utility for practical, actionable insights in forest fire prevention and management.

In conclusion, this research offers novel data collection methods and ML applications for forest fire prediction with proven real-world applicability. It stands as a pivotal reference for future academic, governmental, and industry efforts in forest fire risk mitigation.

To conduct experiments and demonstrate contribution $C1$, a dataset was collected for the province of Saskatchewan, Canada, spanning the years 2000–2018. The data sources include Copernicus Reanalysis Climate data, CWFIS Datamart, the provincial boundary shapefile provided by Statistics Canada, and provincial water body information provided by ArcGIS RESET. This effort yielded a total of 5,655,825 raw data points, which will be explained further in Chapter 4. For contributions $C2$ and $C3$, a joint methodology is presented in Chapter 3.

1.3 Thesis Outline

The structure of this thesis is organized to facilitate a comprehensive understanding of the methodologies, findings, and implications of the research. It is outlined as follows:

Chapter 2 begins with an exhaustive Literature Review, offering a scholarly context for the research by surveying critical advancements, methodologies, and gaps in the field of forest fire prediction.

Chapter 3 gives an overview of the modeling framework used in this research followed by a section on Notion, which serves to lay the conceptual groundwork for the study. This section aims to provide definitional clarity and theoretical underpinnings for the principles and terminologies employed throughout the research.

Chapter 4 delves deeply into the Dataset Creation Framework. This chapter provides detailed explanations of each data source used, exploring how they individually and collectively contribute to the robustness of the dataset. Special attention is paid to the methodology employed for the inclusion of the target column, "fire", which is determined based on historical data. This allows for the establishment of a reliable foundation upon which ML models are trained and evaluated.

Chapter 5 presents results and analysis, articulating the research outcomes derived from the methodologies used. It elaborates on the performance metrics of various ML models, such as Random Forest, XGBoost, LightGBM, and CatBoost. The chapter also interprets these results in the context of forest fire prediction and outlines their ecological and practical implications.

Finally, Chapter 6 serves as the Conclusion and Future Work section. It synthesizes the key findings of the study, discusses their relevance, and suggests directions for future research in the domain.

This structural organization ensures a logical flow of content, facilitating a coherent and in-depth exploration of the research topic.

Chapter 2

Literature Review

This chapter delves into literature of Canada forest fire in early days and leveraging ML techniques in forest fire prediction domain which is the building blocks of this research. Forest fires detection are a growing area of research that is to benefit the advancement of ML. The aim is to develop accurate and timely prediction models to support proactive fire management and prevention strategies. To achieve this goal, an in-depth understanding of the current literature on forest fire prediction and ML is crucial. This chapter offers a well-organized overview of relevant studies, spotlighting the latest trends and approaches in the area of forest fire prediction.

2.1 Canada Forest Fire in Early Days

Roughly a third of the boreal forest (Fig. 2.1, where the dashed box represents the Canadian portion) encircles the top half of the globe and belongs to Canada. These forests are primarily situated above the latitude of 50 degrees North [1].

In the last two decades, approximately 70% of the total loss of tree cover due to fires occurred in boreal zones. While fires serve as an integral component of the ecological balance in boreal forests, the loss of tree cover in these regions has increased significantly. Specifically, there has been an annual increase of approximately 110,000 hectares, or a 3% rise, over this two-decade period. This accounts for almost half of the total global escalation in fire-caused tree cover losses from 2001 to 2022 [26].

According to [26], both the eastern and western parts of Canada experienced unprecedented levels of wildfires in just the initial two months of the wildfire season in 2023. These

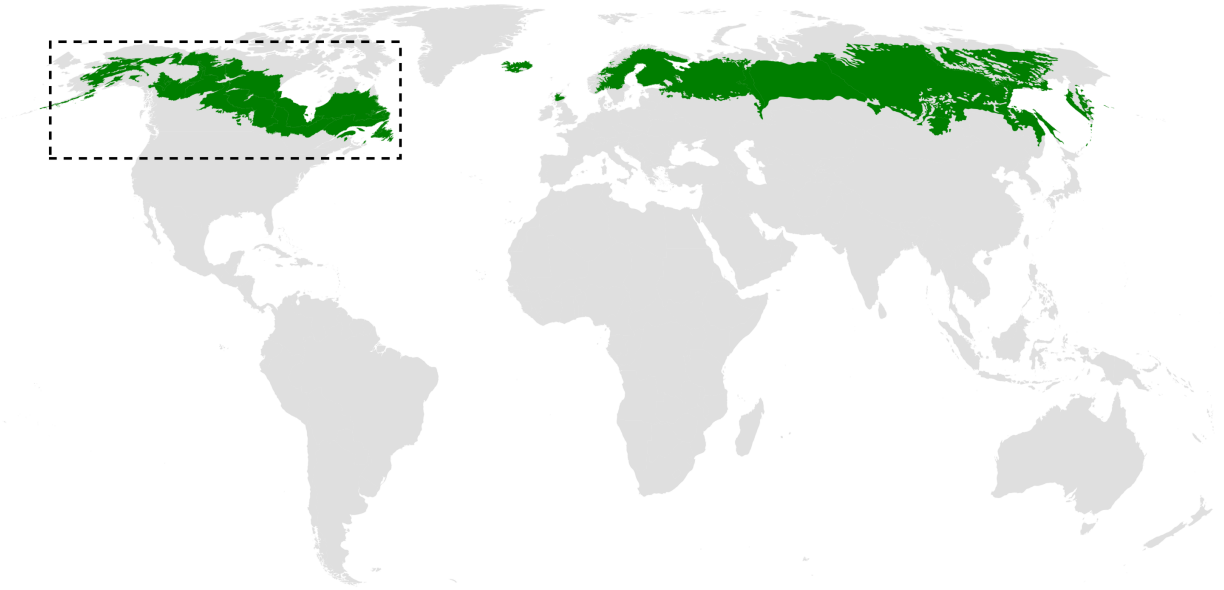


Figure 2.1: Geographical Distribution of Boreal Forests, Highlighting the Canadian Portion Represented by the Dashed Box.[3]

were driven by above-average temperatures and drought conditions. Data from the CIFFC indicate that an estimated 9.5 million hectares were scorched from January to July 2023 [7].

Since Canada contains the majority of the boreal forest, it is crucial for the country to have a system for collecting and managing information on wildland fires. In the early 1980s, Canada established the CWFIS. Managed by Natural Resources Canada, the CWFIS is designed to collect, monitor, and disseminate information essential for understanding fire risks, current fire activity, and the overall state of wildland fires across the country. It integrates data from various sources, including satellite imagery, weather stations, and ground reports, offering a multifaceted view of fire activity. The system serves a diverse audience, including government agencies, fire managers, researchers, and the general public, by offering tools, maps, and statistical data that aid in fire prediction, management, and research. Through its real-time updates and historical archives, CWFIS plays a critical role in supporting Canada's efforts in wildfire preparedness, response, and recovery.

2.1.1 Evolution and Efficacy of CWFIS

The CWFIS [45] has undergone substantial development since its early days. Initially established as a national fire management tool in 1975, it later evolved into the Canadian Fire Effects Model (CanFIRE) in 1994 before transforming into the CWFIS we know today. Operated with a modular architecture, it collects data from more than 1,200 weather stations across Canada, a significant increase from its original 250.

Data for the system are sourced both manually and through automated fire weather stations and then sent to a centralized network via multiple communication pathways, including telephone lines, digital broadband, and satellite telemetry. Although hourly weather data is collected, the system performs its risk assessment calculations on a daily basis.

One of the key analytical frameworks employed by the CWFIS is the CFFWI, an empirical model developed in 1987. This model is instrumental in providing detailed insights into multiple aspects of wildfire risk, such as ignition ease, potential fuel consumption, and fire spread rate. These metrics are consolidated into the Fire Weather Index (FWI), a comprehensive measure of wildfire danger. It has been suggested that the FWI could also serve as a foundational element in predictive models for fire occurrences.

The FWI System generates six key metrics: three are moisture codes and the other three are numerical scores indicating the likelihood of wildfires, as illustrated in 2.2. These metrics collectively serve as the FWI, offering a comprehensive gauge of fire risk. Figure 2.2 depicts the elements that make up the FWI System. These elements are calculated based on sequential daily measurements of factors such as temperature, relative humidity, wind speed, and precipitation over a 24-hour period.

Despite its broad capabilities, the CFFWI itself does not directly forecast when or where a fire might ignite. However, Lawson and Armitage [23] argue that the FWI could conceptually be used to predict the occurrence of fires when applied to specific weather data points.

Despite its capabilities, the CWFIS does have limitations. Its geographic coverage is vast, using a grid cell resolution of 1,000 meters for the most fire-prone regions, but the number of weather stations is considered insufficient to capture the intricacies of rapidly changing terrains or unusual geographic features. This shortfall impacts the accuracy of the FWI and limits the system's efficacy in certain regions.

The latest version of the CWFIS introduced a comprehensive guide for the placement and instrumentation of fire weather stations. This guide emphasizes the importance of accurate meteorological data for both the FWI and the Canadian Forest Fire Behavior

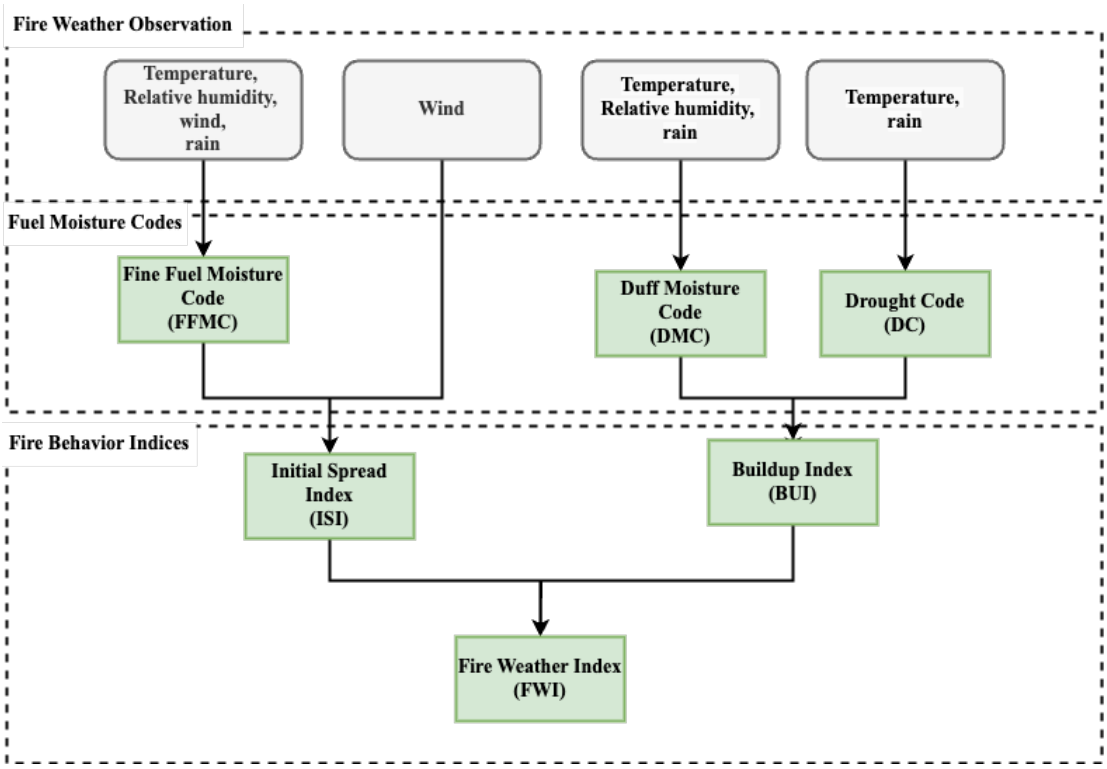


Figure 2.2: CFFWI Structure. (Adapted from [45][23][24])

Prediction system. It also identifies challenges in obtaining precise forecasts, especially in complex terrains, and calls for a specialized fire weather forecasting service supported by an expanded network of stations.

Thus, while CWFIS serves as a cornerstone in the realm of Canadian wildfire management, offering valuable real-time and predictive data, it also presents avenues for future research and enhancement, particularly in the granularity of its data and its geographical coverage.

2.2 Forest Fire and ML

2.2.1 Data Collection

Data collection is a critical aspect of developing effective ML models for detecting forest fires. Several studies have explored different approaches to collecting data in the forest fire domain. Variety of data collection methodologies are employed, each with its own set of advantages and applications. Satellite imagery, highlighted in studies by [31] and [13], offers broad geographical coverage and is instrumental for early fire detection. Remote sensing techniques, showcased in works by [2], [20], and [16], provide detailed environmental insights, including vegetation health and temperature. Real-time meteorological variables are captured through weather stations as indicated by [30], offering invaluable data for immediate risk assessment. IoT sensors, a focus of research by [41], [49], and [16], offer granularity and real-time responsiveness, enhancing the robustness of detection systems. UAVs and ground-based sensors, as explored by [40], bring in the element of mobility and localized data collection. Historical fire records, analyzed in research by [38] and [43], offer a valuable resource for calibrating and validating predictive models. Lastly, crowd-sourced data, also investigated by [40], represents an innovative approach, leveraging citizen participation to augment fire-related datasets. This multifaceted approach to data collection not only provides a comprehensive view of current conditions but also offers multiple avenues for future research and technological development in forest fire management. Table 2.1 summaries these approaches.

Rather than collecting data from scratch, many studies have utilized pre-existing datasets. One such dataset is the UCI ML dataset, which is composed of 517 instances and 13 attributes from the Northeast region of Portugal [8][10]. Another dataset, SaskFire, was employed for time-series classification in a study cited as [22]. Additionally, Kaggle datasets were used by researchers in [32] to predict forest fires using ML techniques.

2.2.2 Imbalance Handling Techniques

Within the specialized area of ML for predicting forest fires, addressing the imbalanced nature of data remains a critical yet often overlooked issue. A majority of existing research [4][25][17] seems to proceed with an equal distribution of fire and non-fire data points, maintaining a 1:1 ratio. This overlooks the naturally imbalanced distribution that one finds in actual fire incidents, casting doubts on the real-world utility of the predictive models thus derived [19].

Table 2.1: Summary of Forest Fire Data Collection Methods

Paper Title	**Method**	**Description**
[31] [13]	Satellite Imagery	Utilizing satellite sensors to capture images of forests and analyze them for fire detection
[2] [20] [16]	Remote Sensing	Using remote sensing technologies, such as LiDAR or infrared sensors, to gather data on vegetation health, temperature, and other relevant factors
[30]	Weather Stations	Deploying weather stations in or near forested areas to collect real-time weather data, including temperature, humidity, wind speed, and precipitation
[41] [49] [16]	IoT Sensors	Deploying Internet of Things (IoT) sensors in forested areas to collect environmental data, such as temperature, humidity, and air quality
[40]	Unmanned Aerial Vehicles (UAVs)	Using drones equipped with cameras and sensors to capture high-resolution images and collect data on fire-prone areas
[40]	Ground-Based Sensors	Installing ground-based sensors, such as temperature and moisture sensors, to monitor forest conditions and detect anomalies
[38] [43]	Historical Fire Records	Analyzing historical fire records and incorporating them into the dataset for model training and validation
[40]	Crowd-Sourced Data	Gathering data from crowd-sourcing platforms where volunteers contribute fire-related information, including fire incidents, burned areas, and fire severity assessments

Certain studies are starting to fill this void. For example, research denoted by [21] incorporates a deep learning framework to assess wildfire areas, accounting for the unequal distribution between large and small fire incidents. Yet, there are others, like the one indicated by [11], that preset their model with an imbalance ratio of 1.4:1, seemingly dismissing the data imbalance issue. Further, some researchers, marked by [46] and [47], do recognize this problem but opt for modest predefined imbalance ratios like 10:1 or 3:1.

SMOTE-ENN was used by [27] for handling imbalanced data related to cardiac arrhythmia classification. They coupled SMOTE-ENN with XGBoost and various ML algorithms to achieve a high accuracy of 97.48%, substantiating the effectiveness of SMOTE-ENN in imbalanced scenarios [27]. Another study by [28] applied SMOTE-ENN in predicting diarrhoea cases among children in developing countries. The authors demonstrated that models trained on data balanced with SMOTE-ENN exhibited high precision, recall, and F1-score, further validating the potential of SMOTE-ENN in various applications [28].

GANs have also been explored for their potential to augment minority classes in imbalanced datasets. [42] specifically applied tabular GANs to handle time-to-event data related to survival prediction. Their findings suggested that although GANs can sometimes outperform traditional methods, they are generally less effective than classical oversampling techniques such as SMOTE [42]. In a meta-analysis involving 18 GitHub repositories, [36] examined the application of GANs in imbalanced class scenarios. They highlighted the flexibility and broad applicability of GANs but also pointed out specific limitations [36].

This body of literature indicates that techniques like SMOTE-ENN and GANs hold promise for addressing class imbalance but also come with their own set of challenges and limitations. As such, their integration into forest fire prediction models warrants careful consideration and empirical evaluation to ascertain their efficacy in this specific domain.

2.2.3 ML in Forest Fire Domain

The application of ML techniques in the forest fire domain has witnessed a paradigm shift from traditional statistical methods to more complex algorithms. Early models primarily employed logistic regression and decision trees, but their limitations in capturing the non-linear relationships between variables paved the way for more sophisticated methods like neural networks, deep learning, and ensemble methods. For instance, neural networks have been found to offer more accurate predictions of fire occurrence, spread, and intensity compared to traditional models [35]. Ensemble methods like Random Forests and Gradient Boosting combine multiple weak learners to improve predictive performance and have

shown promise in several studies [34]. Time-series models such as Long Short-Term Memory (LSTM) networks have also been effective, especially when dealing with the temporal nature of forest fire data [39].

Real-time prediction and alert systems are becoming increasingly feasible due to advancements in IoT and sensor networks. These developments enable the integration of ML models with real-time sensor data for early detection and timely alerts [48].

A significant number of studies have been conducted to investigate ML applications in forest fire prediction, with 38 relevant papers published between 2014 and 2022 identified in a comprehensive review on IEEE Xplore [33]. This growing body of work frequently utilizes Random Forest and SVMs as initial models for classification tasks. These algorithms are often stepping stones for researchers who later venture into more advanced decision tree models and neural network architectures, such as gradient boosting and LSTM networks [33].

However, despite this substantial progress in applying ML to the forest fire domain, a notable research gap exists. The existing literature lacks standardization across studies, making it challenging to establish best practices and identify the most effective models for predicting forest fires. Moreover, an additional challenge is the issue of data imbalance. Forest fire datasets often exhibit a significant imbalance between the classes, with the majority of instances being non-fire events and a minority being actual fire occurrences. This data imbalance can lead to biased model performance and may result in the neglect of the minority class. The development of appropriate evaluation metrics that account for class imbalance, such as precision-recall curves and F1-score, should be emphasized. Additionally, exploring anomaly detection techniques to identify rare fire events within a predominantly non-fire dataset could also be beneficial.

Chapter 3

Methodology

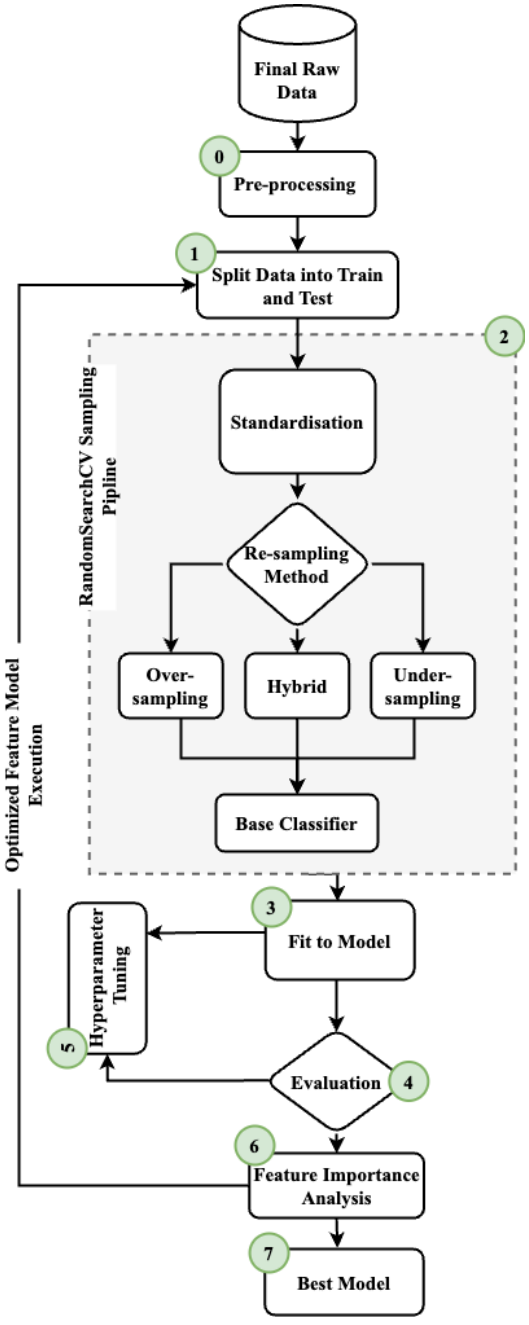
This section initially provides an overview of our modeling framework and outlines how each component has been utilized in this research. In 3.2 subsection, we elucidate the fundamental methodologies that have guided our focus on forest fire prediction. These methodologies are integral to various aspects of our work. For instance, the Haversine formula serves as a cornerstone of our data collection framework. Its definition is detailed in 3.2 subsection of this chapter, while its application is further discussed in Chapter 4. We also offer a comprehensive overview of how ML techniques can be specialized for forest fire prediction. In terms of ML algorithms, we employ models from both the random forest and gradient boosting families, specifically XGBoost, LightGBM, and the more recently developed CatBoost. During the modeling stage, we address the issue of data imbalance by implementing various strategies such as under-sampling with NearMiss3, over-sampling through SMOTE and GANs, and a hybrid method using SMOTE-ENN. Finally, to assess the performance of our models, we explore the evaluation metrics most suited to our research problem. By the end of this section, readers will have acquired the foundational technical knowledge necessary to understand this research.

3.1 Overview

The methodology employed in this research encompasses a multi-faceted, systematic approach designed to robustly model the raw data and optimize predictive performance.

Figure 3.1 illustrates the steps taken in the modeling framework. Initially, the final raw data, sourced from the Dataset Creation Framework (refer to figure 4.1, chapter 4),

Figure 3.1: Fire Classification Framework



undergoes a series of pre-processing steps, labeled as Box 0, to enhance its quality and usability. The pre-processed dataset is subsequently partitioned into 80% training and 20% test sets, depicted in Box 1, serving the dual purpose of model training and performance evaluation. Before addressing the class imbalance in the training data (Box 2), the data is standardized, ensuring a consistent scale across features, which in turn bolsters model stability and performance. Three re-sampling techniques - over-sampling, under-sampling, and hybrid - are evaluated.

The performance of the model is rigorously evaluated using predefined metrics (accuracy, sensitivity, specificity and ROC-AUC), as shown in Box 4, to provide an initial comparative analysis. The research includes hyperparameter tuning as an additional optimization layer, represented in Box 5, which allows the algorithms to be finely adjusted for maximum predictive accuracy. The concluding stage of the methodology focuses on identifying the features that have a significant impact on the performance of the chosen model, as indicated in Box 6. A carefully selected list of these important features is compiled, and the entire modeling process is rerun with this condensed feature set. The aim is to determine whether a simplified feature space can maintain or even improve model performance without losing predictive power. Since the study utilizes three distinct ML algorithms, the whole process is repeated four times to find the most effective model, represented as the last step in Box 7 of the framework.

This exercise is especially useful in achieving a streamlined model that is both efficient and easy to interpret. As a result, the methodology not only aims to achieve high predictive performance, but also emphasizes model interpretability and feature selection. This multifaceted approach strengthens the research's thoroughness and practicality.

3.2 Notion

Having established an overview of the modeling framework, attention must now be directed towards understanding the technical components deployed in this research. Subsequent sections will explain the specific models and algorithms that have been utilized. This will not only provide the necessary technical background but also clarify the rationale behind the selection of these particular methodologies. This comprehensive approach aims to enhance the reliability and depth of the research findings.

3.2.1 Haversine Formula

To add the "fire" points to the feature data set, we needed to calculate the distance When dealing with points on a sphere, in order to calculate the shortest distance between two points using the latitude and longitude, haversine can come in handy. The Haversine formula to find the distance d between two points with coordinates $(\text{lat}_1, \text{lon}_1)$ and $(\text{lat}_2, \text{lon}_2)$ is given by:

$$\begin{aligned} a &= \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta\text{lon}}{2}\right), \\ c &= 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right), \\ d &= R \cdot c, \end{aligned} \tag{3.1}$$

Where:

$$\begin{aligned} \Delta\text{lat} &= \text{lat}_2 - \text{lat}_1, \\ \Delta\text{lon} &= \text{lon}_2 - \text{lon}_1, \\ R &\text{ is the Earth's radius (mean radius = 6,371 km).} \end{aligned}$$

The Haversine formula offers several benefits that make it an excellent choice for use in our research. By approximating the Earth as a sphere, the formula delivers highly accurate distance measurements. Its straightforward nature and simplicity make it easy to implement, even for those who may not be experts in computational geometry. Furthermore, it is computationally less demanding compared to methods that model the Earth as an ellipsoid, a crucial advantage in our case. Given that we are dealing with a large volume of data, the ability to perform rapid calculations across large datasets is of paramount importance. Thus, the Haversine formula's combination of accuracy, ease of use, and computational efficiency aligns well with the objectives of our study.

3.2.2 ML Algorithm

Prior to delving into the models that are picked for our thesis, it is best to grasp some of the ML concepts and algorithms that are the building blocks of Random Forest, XGBoost, LightGBM and CatBoost.

Supervised Learning

In supervised learning, algorithms aim to find patterns within a dataset that includes both features and labels by training a model on it. The trained model is then used to predict the labels for new, unseen data based on its features, Figure 3.2

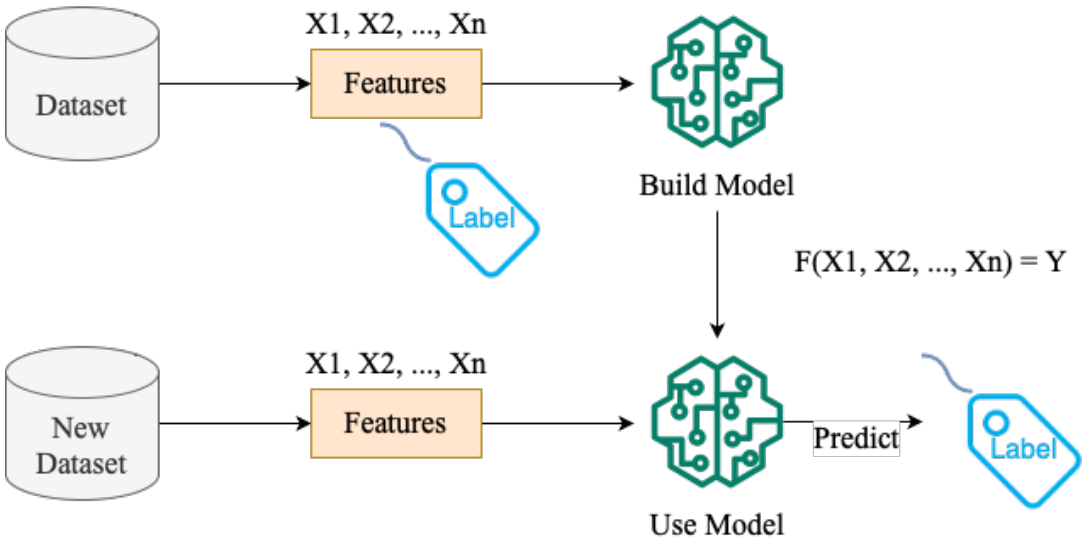


Figure 3.2: Supervised Learning

DTs

DTs are non-parametric supervised learning. It can be used for both classification and regression. Decision trees construct a model that predicts the label by navigating through a hierarchical tree of true/false feature questions. The goal is to minimize the number of questions needed to accurately determine the probability of making a correct decision.

Ensemble Learning

Ensemble methods integrate multiple decision trees to create predictive models that outperform a standalone decision tree. Ensemble learning serves as the umbrella term for both averaging (bagging) and boosting methods. This approach combines the predictions

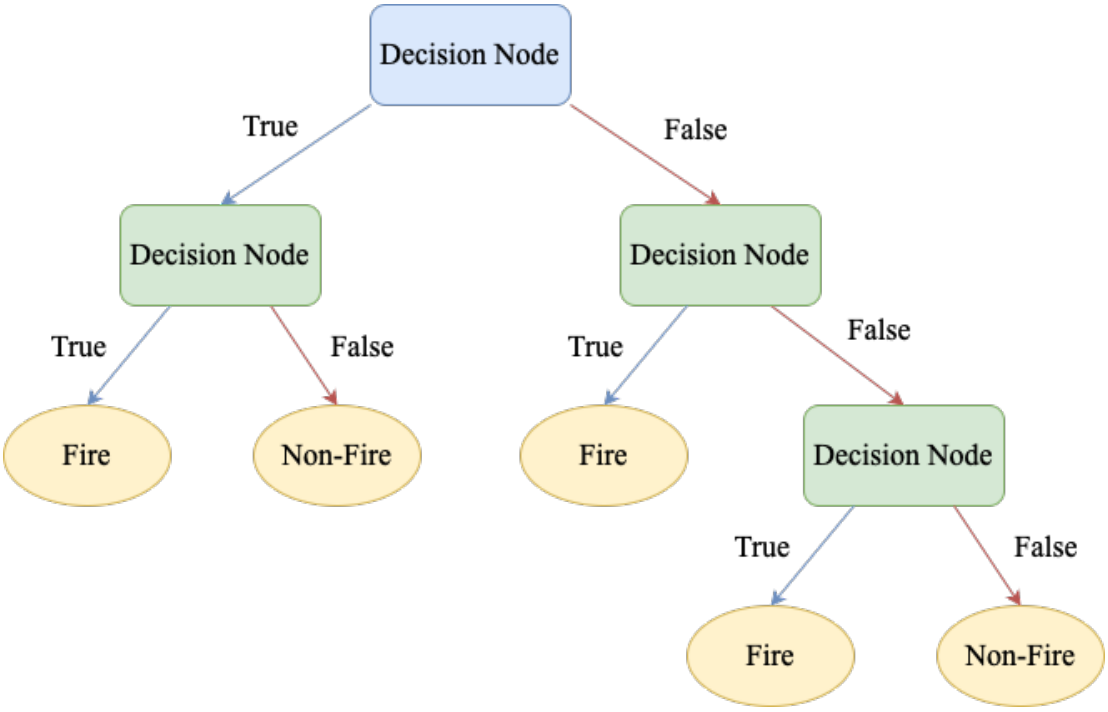


Figure 3.3: DTs

of multiple base estimators using a given learning algorithm to enhance generalizability and robustness.

In averaging methods, as the name suggests, the predictions from several independent estimators are averaged, particularly when dealing with regression problems. In the case of classification, majority voting is employed. Random Forest is one example of this approach.

On the other hand, boosting methods construct estimators sequentially, with each one aiming to reduce the bias of the combined model. This method merges several weak models to create a single, more powerful model. Gradient Boosting is an example of this method.

To achieve a better model, ensemble learning algorithms combine multiple ML algorithms.

Gradient Boosting

Gradient boosting serves as a powerful team-building strategy for weaker predictive models. Essentially, the method starts with a base model that isn't particularly strong in its

predictive power. It then iteratively adds more models into the ensemble, each one focused on correcting the mistakes of the collective group that came before it. This iterative process is structured around a mathematical framework known as gradient descent, which systematically adjusts the model to minimize errors.

In practice, gradient boosting deploys a series of shallow decision trees, trained in a sequence. Each successive tree is fit based on the error residuals, essentially the "mistakes", of the previous ensemble of trees. The end result is a robust model, its final prediction being a weighted sum of the individual trees' predictions. Gradient boosting decision tree "boosting" minimizes the bias and underfitting.

3.2.3 Random Forest Algorithm

In Random Forest, bagging is employed to construct complete decision trees concurrently, each generated from random bootstrap samples of the dataset. The ultimate prediction is derived by averaging the predictions from all these trees. The Random Forest algorithm is a popular tree-based algorithm commonly used in supervised ML tasks involving labeled data. Random Forest, also known as Random Decision Forest, is versatile and can be used for both classification and regression tasks. In our study, we focus on the classification of the "fire" class. The Random Forest classifier constructs multiple decision trees by randomly selecting subsets of both features and data points from the training set. The final prediction is determined by aggregating the individual predictions of these trees, typically through a majority-vote mechanism. Random forest "bagging" minimizes the variance and overfitting.

3.2.4 Extreme Gradient Boosting (XGBoost) Algorithm

XGBoost represents an efficient and highly accurate extension of gradient-boosted trees, optimized for both computational speed and model performance. Unlike traditional Gradient Boosted Decision Trees (GBDT), where trees are built sequentially, XGBoost employs a parallel construction method. The algorithm adopts a level-wise strategy, systematically examining gradient values to swiftly assess the quality of potential data splits at each level of the training set.

XGBoost has gained prominence as one of the most popular ML models for both classification and regression tasks. One of the key reasons for its popularity is its capacity to handle large datasets efficiently, thanks to its built-in support for parallel processing. This allows for training the model in a reasonable timeframe without compromising on

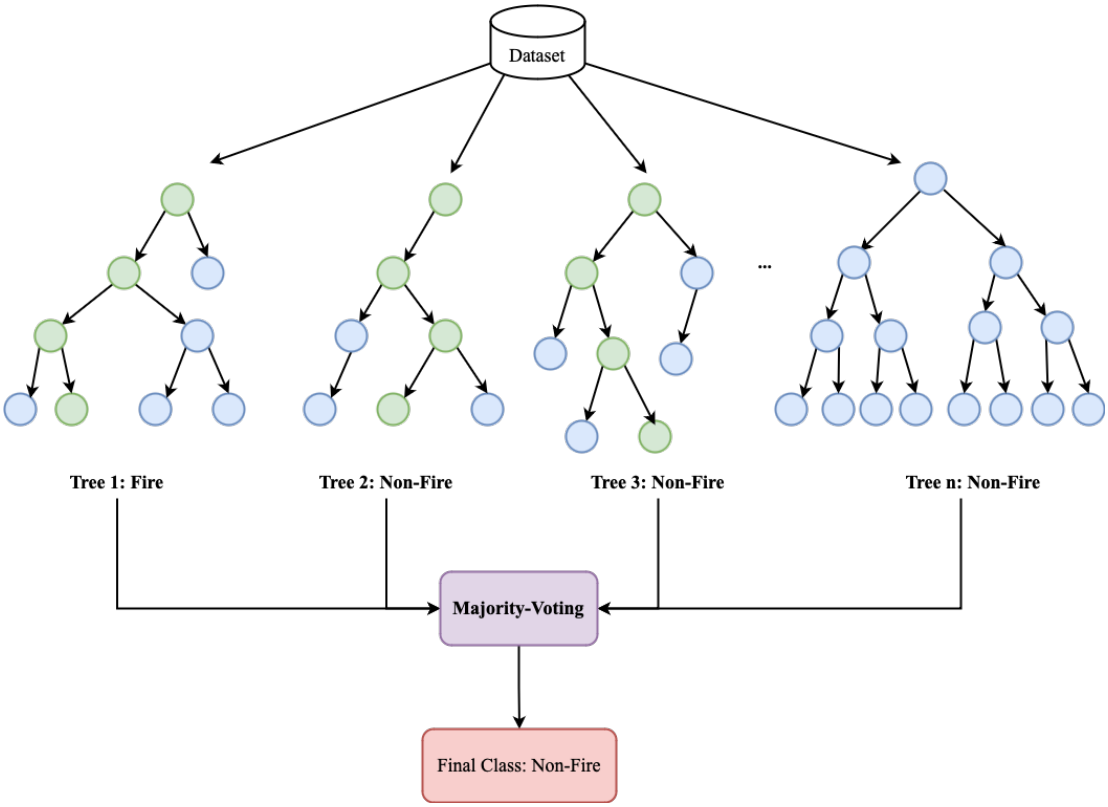


Figure 3.4: Random Forest Classifier

its state-of-the-art performance. In the context of forest fire prediction and management, XGBoost's capabilities are particularly valuable. Its robustness and ability to work with complex datasets make it ideal for handling the varied and often non-linear factors that contribute to forest fires, such as weather conditions, vegetation type, and human activities.

3.2.5 LightGBM Algorithm

LightGBM is a distributed, high-performance framework for gradient-boosted tree algorithms. Developed to optimize both computational speed and memory usage, LightGBM offers particular advantages in handling large and complex datasets. Unlike traditional gradient boosting algorithms, which grow trees depth-wise, LightGBM employs a histogram-based algorithm that grows trees leaf-wise. This results in a more accurate model, as LightGBM focuses on the leaves with large errors. Figure 3.5 explains the implementation

of LightGBM and other boosting algorithms.

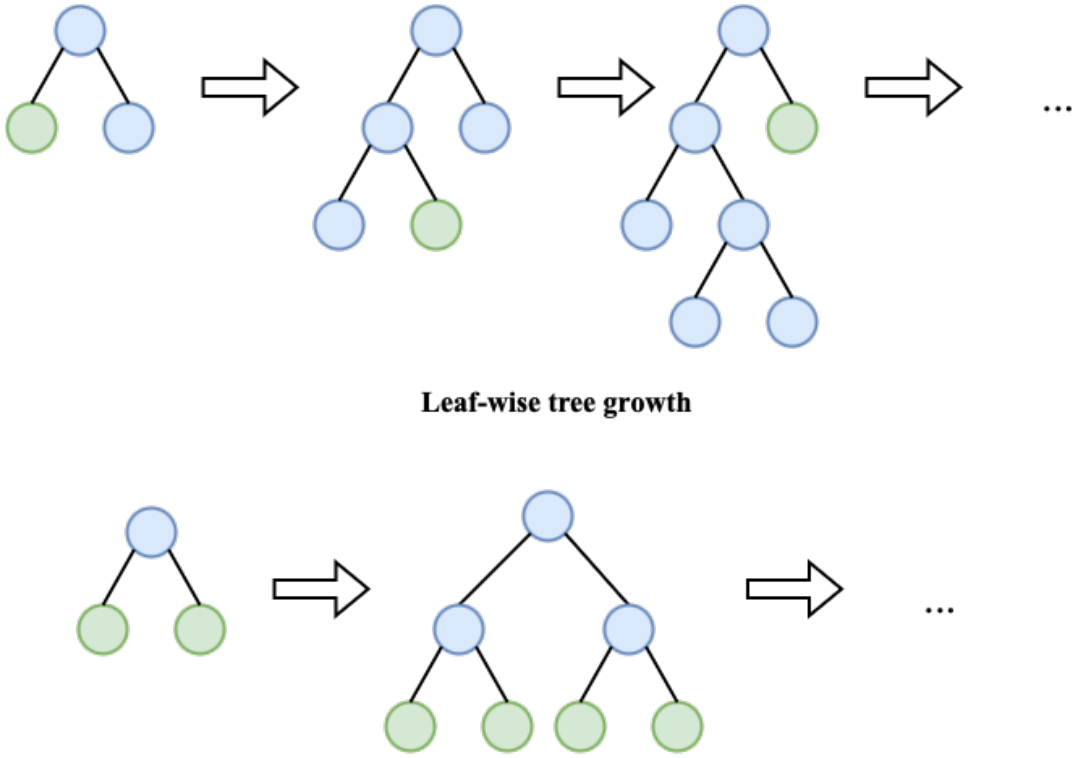


Figure 3.5: Leaf-Wise vs. Level-Wise: Tree Growth in LightGBM and Other Boosting Algorithms

The model also incorporates advanced regularization techniques, like Gradient-based One-Side Sampling (GOSS) and , to prevent overfitting. GOSS retains the data points with large gradients, ensuring the model learns effectively from the minority "fire" class. Exclusive Feature Bundlin (EFB), on the other hand, bundles mutually exclusive features, reducing the dimensionality of the data and making the model faster and more efficient. LightGBM's histogram-based training approach further enhances speed and computational efficiency. In the context of forest fire prediction, this means the model can rapidly assimilate real-time data—such as sudden changes in weather conditions, like temperature and

humidity, or the status of other contributing factors like vegetation dryness. Consequently, LightGBM can provide timely, reliable fire risk assessments, allowing for quicker response and potentially reducing the devastation caused by wildfires.

3.2.6 CatBoost Algorithm

Figure 3.6: Evolution of Tree-Based Boosting Algorithms: From XGBoost to CatBoost

A few years after the release of LightGBM by Microsoft, Yandex, a Russian tech company, introduced a more robust algorithm for gradient boosting on decision trees. Figure 3.6 illustrates the timeline during which these boosting trees were released. Three main characteristics led this project to experiment with this algorithm.

1. Its exceptional quality without the need for parameter tuning significantly reduces the time spent on hyperparameter optimization. The default parameters yield excellent results.

2. Its fast prediction capabilities lead to quicker and more efficient model training.

3. Its ability to handle categorical features without any prior encoding is noteworthy. Although this feature may not be directly testable with the dataset used in this research, it's highly relevant in the forest fire domain where we deal with satellite images. It offers a valuable initiative for learning how to manage categorical and numerical features concurrently.

CatBoost, which stands for "Categorical Boosting" is an ensemble learning method based on gradient boosting. What sets CatBoost apart is its robust handling of categorical features without requiring any explicit preprocessing steps like one-hot encoding. The model internally applies mean encoding to categorical features, utilizing a novel scheme to prevent target leakage. In addition to ordered boosting, CatBoost implements an Oblivious Trees model, a more symmetric and balanced tree compared to conventional decision trees.

This structure results in consistent prediction rules for each leaf, speeding up both the training and prediction phases. CatBoost is also engineered for parallel and distributed computing, which is crucial for handling large datasets.

The algorithm incorporates various regularization techniques, including a Bayesian method for tree-structure optimization and L2 leaf regularization, to control overfitting effectively. Furthermore, it uses oblivious trees combined with oblivious randomized trees during the model boosting process, improving model accuracy by adding an element of randomness that helps in generalizing well on unseen data.

3.2.7 Imbalance Handling

Data imbalance is a prevalent issue when dealing with large datasets. In such imbalanced datasets, the classes are not equally represented, as visually depicted in Figure 3.7. This inequality makes it challenging for ML algorithms to learn and effectively predict the minority class. The ratio between the majority class and the minority class is so significant that the minority class is often dismissed or ignored. As a result, the algorithm tends to favor the majority class in its predictions.

Various techniques have been proposed to handle imbalanced data, and this section discusses some of the popular approaches. We present an overview of three such techniques: NearMiss3 Under-sampling, SMOTE and GANs Over-sampling, and SMOTE-ENN Hybrid. While under-sampling methods such as NearMiss3 are computationally efficient, they risk discarding potentially useful information. Over-sampling methods such as SMOTE and GANs can generate high-quality synthetic data but come at a high computational cost and are sometimes unreliable. Hybrid methods like SMOTE-ENN offer a middle ground but are not free from drawbacks, such as the risk of overfitting. Choosing the right technique often depends on the specific requirements of the project, and a combination of methods may be the most effective approach.

Since imbalanced data makes our model more prone to scenarios where "Fire" (the minority class) has negligible or extremely low recall, we aim to evaluate the impact of these techniques on our dataset and assess their contribution to our modeling efforts.

NearMiss3 Under-sampling

NearMiss serves as an under-sampling methodology, devised to rectify class imbalances by meticulously eliminating instances from the majority class. The core rationale is to augment the decision space between both classes by excluding specific majority-class samples

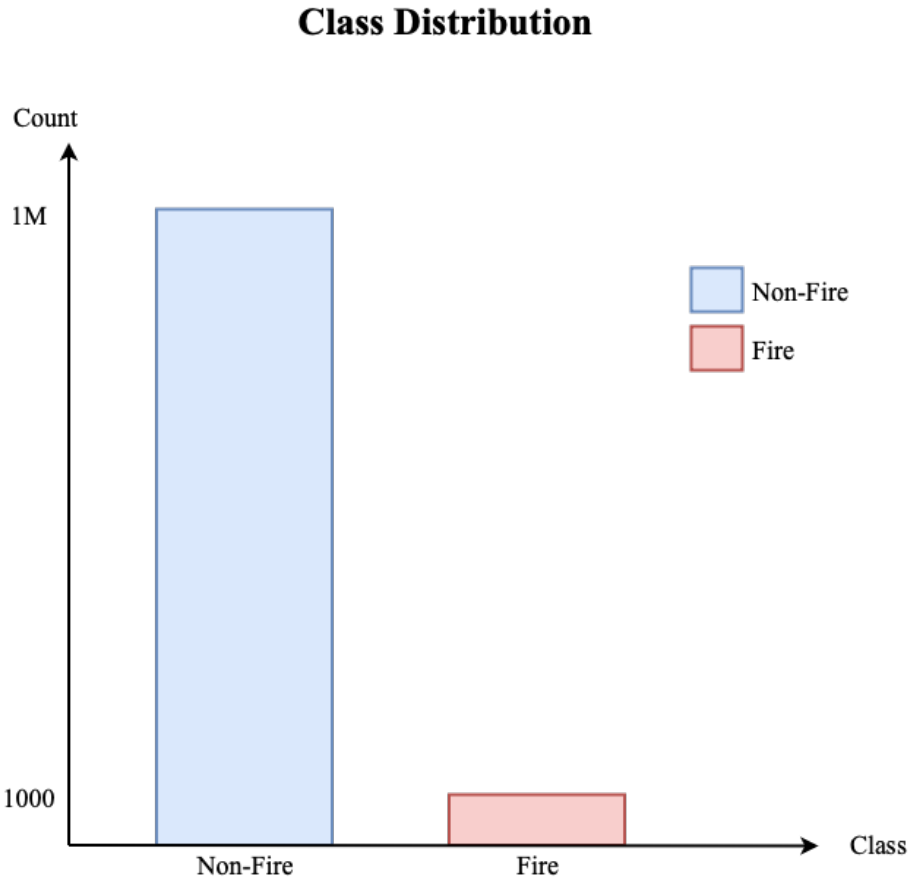


Figure 3.7: Imbalanced Data Representation

that are in close proximity to those of the minority class. This optimization enhances the performance of subsequent classifiers.

A recurrent issue in under-sampling methods is the forfeiture of vital data. To counter this, NearMiss incorporates a near-neighbor strategy. The operational mechanics of this approach are outlined in a sequential fashion:

1. Initially, the algorithm calculates the distances between each minority class instance and all instances belonging to the majority class, which is the targeted class for under-sampling.

2. Subsequently, n instances from the majority class, having the least distances to the minority class instances, are pinpointed.

3. Assuming k instances are present in the minority class, this approach results in $k \times n$ instances retained from the majority class.

Different incarnations of the NearMiss algorithm offer varied methodologies for identifying these n closest majority-class instances:

- **NearMiss-Version 1**: This variant opts for majority-class instances with the smallest average distance to the k nearest minority-class instances.
- **NearMiss-Version 2**: This iteration chooses majority-class instances based on their smallest average distance to the k most distant minority-class instances.
- **NearMiss-Version 3**: Employing a two-tier process, it initially retains the M nearest majority-class neighbors for each minority class instance. Subsequently, it selects those majority-class instances that have the largest average distance to their N nearest minority-class neighbors.

We chose NearMiss3 over versions 1 and 2 because it offers several advantages, such as better preservation of class boundaries and reduced risk of information loss. Unlike its predecessors, which focus solely on distance metrics, version 3 incorporates more contextual information by considering multiple nearest neighbors from both the majority and minority classes. This approach tends to be more adaptable to complex data distributions and is generally less sensitive to outliers. NearMiss3 improves the classifiers' ability to generalize, making them more adaptable to variations in new and unseen data. As a result, it stands out as an excellent option for us seeking more sophisticated under-sampling approach.

SMOTE Over-sampling

SMOTE is a data augmentation method used in machine learning to address class imbalance. It creates synthetic samples for the minority class by interpolating between existing minority class instances. SMOTE aims to balance class distributions, improving model performance on underrepresented classes.

GANs Over-sampling

GANs can also be used to tackle the problem of imbalanced datasets by generating synthetic samples for the minority class. GANs consist of two neural networks: the Generator and

the Discriminator, which work against each other. The Generator tries to create data instances that are indistinguishable from real instances, while the Discriminator tries to differentiate between the real and synthetic samples. By training the network, high-quality synthetic samples are generated that can be used to balance the class distribution. The major drawback of using GANs is the computational cost, which can be significantly higher compared to other techniques.

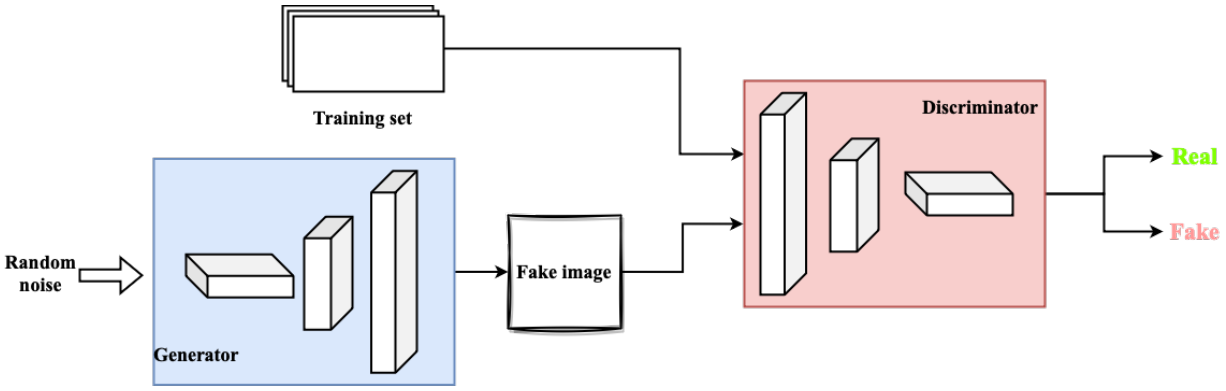


Figure 3.8: GANs Implementation

SMOTE-ENN Hybrid

SMOTE-ENN is a hybrid technique that combines SMOTE and Edited Nearest Neighbors (ENN) algorithms. SMOTE generates synthetic samples in the feature space, while ENN removes instances of the majority class that are misclassified by their nearest neighbors. This approach tries to balance the benefits of both under-sampling and over-sampling. While it can be more effective than applying either SMOTE or ENN alone, the computational cost can be high, and the risk of overfitting may increase due to the synthetic samples.

Sampling Ratio

The main objective of leveraging re-sampling techniques is to refine class distributions, thereby bolstering model performance in scenarios where instances of the minority class are of critical importance. A fundamental aspect of these methods is the notion of the sampling strategy. This term and the phrase "sampling ratio" are used interchangeably in our context. When the `sampling_strategy` parameter is defined as a float, it signifies

the intended proportion of the number of samples from the minority class relative to the majority class post re-sampling [18].

Mathematically, if α represents the sampling strategy expressed as a float and N_{minority} and N_{majority} denote the counts of samples in the minority and majority classes, respectively, the relationship can be articulated as:

$$\alpha = \frac{N_{\text{minority}}}{N_{\text{majority}}}$$

3.2.8 ML Evaluation Metrics

In the context of forest fire classification using ML algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost, a multifaceted evaluation approach is crucial for a comprehensive understanding of model performance. Four primary metrics, namely, accuracy, sensitivity, specificity, weighted F1 score, and ROC-AUC, offer distinct yet complementary insights.

Accuracy

This metric provides a general overview of how well the model is performing. It is the ratio of the number of correct predictions to the total number of predictions. While accuracy can offer a quick snapshot of performance, it can be misleading when the classes are imbalanced, as is often the case in rare events like forest fires. Thus, we consider additional metrics for a nuanced assessment.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

Where:

$$TP = \text{True Positives}$$
$$TN = \text{True Negatives}$$
$$FP = \text{False Positives}$$
$$FN = \text{False Negatives}$$

Sensitivity (Recall)

Sensitivity measures the model's ability to correctly identify positive cases, in this context, the occurrence of a forest fire. A high sensitivity is crucial for public safety, as failing to predict a forest fire could result in significant damages and loss of life. Sensitivity and recall are terms that are often used interchangeably in this context

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.3}$$

Where:

$$TP = \text{True Positives}$$
$$FN = \text{False Negatives}$$

Specificity

This metric gauges the model's ability to correctly identify negative cases or non-occurrence of forest fires. High specificity is equally important to prevent false alarms, which could lead to unnecessary evacuations and resource allocation, inducing panic and economic loss.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.4}$$

Where:

$$TN = \text{True Negatives}$$
$$FP = \text{False Positives}$$

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

This metric provides a more comprehensive evaluation as it considers various thresholds. It shows the trade-off between sensitivity and specificity. A high ROC-AUC value implies that the model is capable of distinguishing between the occurrence and non-occurrence of forest fires effectively. By looking at each of these metrics, we can derive a well-rounded understanding of the model's strengths and weaknesses. Sensitivity ensures we capture as many real incidents as possible, while specificity minimizes the costs associated with false alarms. Accuracy gives us a baseline for general performance, and ROC-AUC offers a nuanced evaluation that considers a range of scenarios. Therefore, employing all these metrics ensures that our chosen model is not only accurate but also reliable and efficient for forest fire classification.

Weighted F1 Score

The weighted F1 score is a metric used to evaluate the performance of a classification model, taking into account both precision and recall for each class and then calculating a weighted average based on class frequencies. It provides a single score that balances the trade-off between precision and recall across multiple classes, with greater weight given to classes with more instances.

$$\text{Weighted F1 Score} = \frac{\sum_{i=1}^{N} w_i \cdot \text{F1 Score}_i}{\sum_{i=1}^{N} w_i} \tag{3.5}$$

Where:

$$\begin{aligned} N &= \text{Number of classes} \\ w_i &= \text{Weight for class } i \\ \text{F1 Score}_i &= \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \end{aligned}$$

Chapter 4

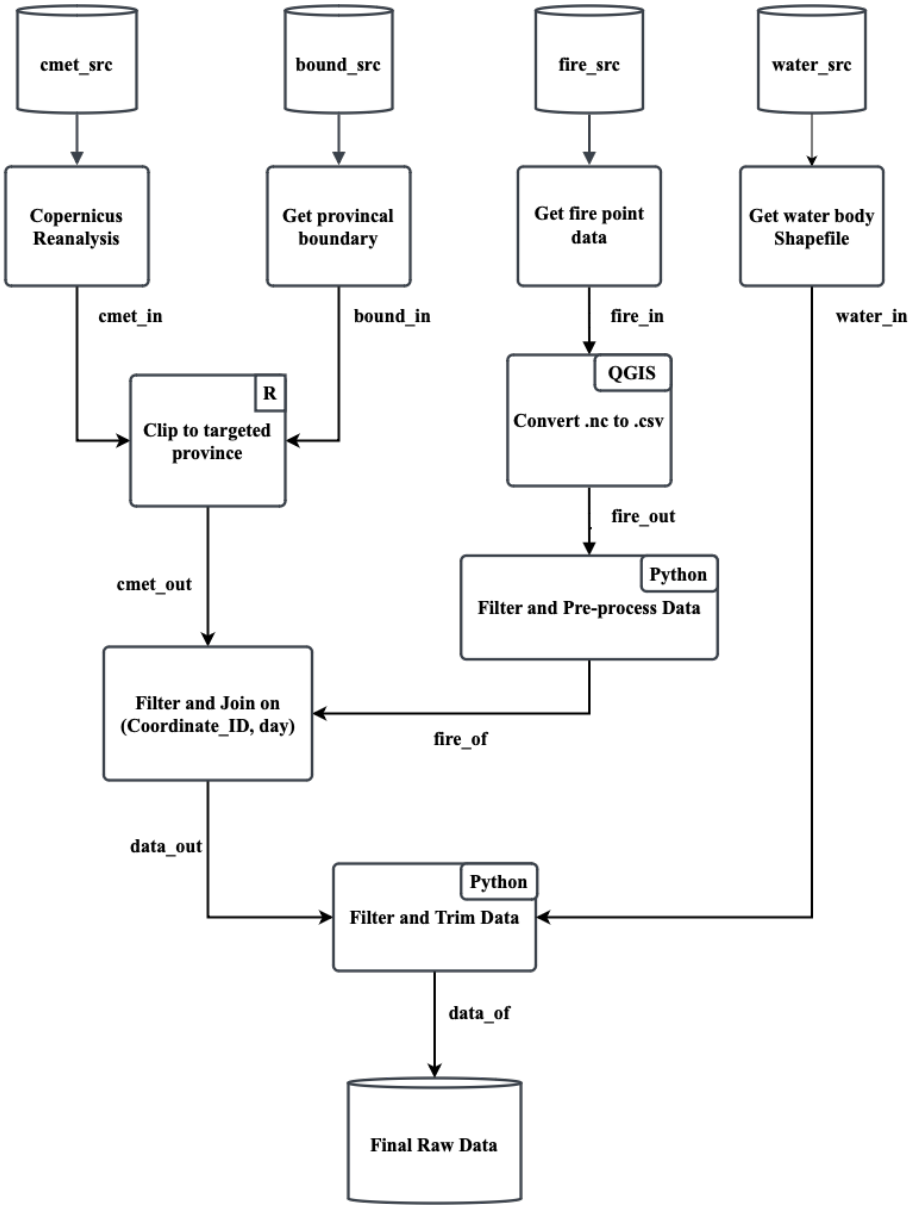
Data-set Creation Framework

This chapter introduces the methodology employed to collect a novel dataset, specifically designed for the prediction and detection of forest fires using ML and image processing techniques. Data gathering serves as a critical initial step in predicting ignition points within forests. To achieve this goal, we have considered various factors that are believed to be correlated with the inception or propagation of forest fires. These commonly considered factors include humidity, temperature, surface pressure, and precipitation. While several academic papers have proposed datasets based on one or two of these factors, our project aims to provide a more comprehensive view by encompassing a broader array of factors. This approach aims to offer a more nuanced understanding of the interrelationships among fire-related variables for the development of the ML model.

4.1 Setting Up the Data Framework

Figure 4.1 outlines the framework creation process followed by table 4.1 and table 4.2. We delve into the details of its implementation and discuss the generation of the Saskatchewan dataset specifically designed for our forest fire prediction research.

Figure 4.1: Dataset collection Framework



There are four main data sources for this framework, as illustrated in Figure 4.1: meteorological data, an area-of-interest boundary file, historical fire data, and water bodies files. Both the meteorological data and the historical fire data are expected to include

Table 4.1: Nomenclature of Input and Output of Data

Name	Meaning
x_src	Source of data of type x, x $\in$ cmet, bound, water, fire
x_in	Inputs from data source x
x_out	Intermediate outputs after processing of x type of data
x_of	Final output after processing of x type of data

Table 4.2: Data Input and Output Files with Formats

Name	Format	File Name
cmet_in	NetCDF	netCDFxxxx.nc
cmet_out	csv	feature_xxxx.csv
bound_in	Shape File	province_bound.shp
fire_in	Shape File	canada_firepoint.shp
fire_out	Shape File	canada_firepoint.csv
fire_of	csv	province_firepoint.csv
water_in	Shape file	province_water.shp
data_out	csv	xxxx-xxxx_dataset.csv
data_of	csv	xxxx_dataset_filtered.csv

coordinate and date information. Additionally, the historical fire dataset should contain details about the size of the fire and its cause. The boundary file serves a specific purpose: to trim the dataset to the designated area of interest. It acts as a ruler, allowing us to clip the data to the specified region. Once the data for this specific area are isolated, it is time to integrate the fire-related information. To combine the meteorological data with the historical fire data, several factors must be considered, such as the size of the area impacted by the fire. The "fire" column will be populated based on the size of the fire in that particular range. Further refinement is required by eliminating data that fall within the water bodies in the area. These water bodies can include lakes, seas, oceans, large rivers, and dry salt flats. The final raw data set will contain the following column titles as demonstrated in Table 4.3.

Table 4.3: Summary of Data Variables for Forest Fire Prediction

Coordinate_ID	Lon	Lat	Feature 1,2,...,n	Day	Year	Fire

4.2 Saskatchewan Data Framework Deployment

Saskatchewan, one of the Canadian provinces, was selected as the study area. Composing an area of 651,900 km^2, Saskatchewan had a population of 1.195 million in 2022. The province's highest elevation stands at approximately 1390.7 meters, while its lowest point measures around 212.8 meters.

Boasting a diverse ecosystem, Saskatchewan transitions from grasslands in the south to forests in the north. These varying ecological zones offer a rich platform for meticulously examining the behavior and characteristics of the fire. The availability of historical fire data, coupled with the diversity in fire incidents concerning frequency, intensity, and causes, underscores the rationale behind selecting this region for data collection. Saskatchewan's remoteness from substantial water bodies, combined with its northerly latitude, ensures it receives more sunlight hours than many other Canadian provinces. This geography results in warmer summers, elevating the risk of drought. Historically, the fire season begins in early June and lasts approximately 13 weeks. Between 2001 and 2021, fires have consumed around 6.13Mha of Saskatchewan's tree cover.

Given all these characteristics of Saskatchewan, it is the area of study for this thesis. In the following sections, we examine the particular data sources (presented in table 4.4) employed during the initialization of the data collection framework. We then detail the methods used for consolidating the gathered data.

Table 4.4: Saskatchewan Data Sources

Name	Type	Links
bound_src	shape	ArcGIS GIS Shapefile Map Layers
cmet_src	netcdf	ERA5 hourly data on single levels
water_src	shapefile	ArcGIS REST Services Directory
fire_src	csv	Natural Resources Canada

4.2.1 ERA5 Data

The primary source for retrieving meteorological data is Copernicus Climate reanalysis data [5]. Climate reanalysis integrates past observations with models to generate consistent time series for various climate variables. These reanalyses are among the most widely used datasets in the geophysical sciences and offer a comprehensive description of the observed climate, covering its evolution over recent decades. The data are organized in 3D grids and provided at sub-daily intervals. Specifically, we employ ERA5 hourly data on single levels, spanning from 1940 to the present[15]. ERA5 represents the fifth-generation reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), and it covers global climate and weather patterns for the past eight decades. This data source is referred to as "cmet_src" in Figure 4.1. To download the necessary data, certain preferences and selections must be set. Table 4.5 outlines the categories of data targeted for retrieval from this source.

Table 4.5: Variable Preferences

Preferences
Temperature
Soil Water
Evaporation
Runoff
Wind
Pressure
Precipitation
Vegetation

Spacial Resolution

The horizontal resolution of the fixed grid is 0.25° x 0.25° n a regular lat-lon grid projection. To isolate data specific to the targeted province, the subregion coordinates provided in table 4.6 are taken into account.

The retrieved files consist of 19 files in the Network Common Data Form (NetCDF) with a ".nc" extension, referred to as `cmet_in`. NetCDF files are multidimensional scientific

Table 4.6: Saskatchewan Sub-Region Coordinates

	xmin	ymin	xmax	ymax
Saskatchewan	-109.99	48.99	-101.36	60.00

data files. Each layer stores information about one of the retrieved features, such as temperature, humidity, pressure, and wind speed. These files have been analyzed and converted to tabular data with ".csv" extensions using *R* programming language [**?**]. The output from the "Clip to targeted province" box in Figure. 4.1 is named `cmet_out`. The algorithm 1 describes the steps.

Algorithm 1 Converting raster data from NetCDF to CSV

Import essential libraries
Initialize year, day offset, and output filenames ▷ Mask raster data
Load NetCDF file into raster format ▷ Convert raster data to dataframe
Extract variables from raster data
Create an empty dataframe to store results
for each day in year **do**
    Extract daily feature slices from raster data
    Convert slices to vectors
    Create new dataframe from vectors and coordinates
    Append new dataframe to results dataframe
end for
Write results dataframe to CSV

Temporal Resolution

ERA5 provides data with hourly temporal resolution, spanning from January 1950 up to the present. For the purposes of this project, however, we focus only on the data collected at 12 noon. This decision is guided by the CFFWI, which indicates that noon is a critical time for predicting wildfire risk levels.

The data set we have gathered covers the period from 2000 to 2018 and includes observations for every day of each month. During the data processing phase, any months with zero reported fires have been excluded from the analysis.

4.2.2 Historical Fire Data Point

The historical fire data has been sourced from the CWFIS Datamart [37]. Fire point data from the National Fire Database consist of a collection of forest fire locations, provided by various Canadian fire management agencies, including provinces, territories, and Parks Canada.

The National Fire Database's fire point data shapefile, referred to as `fire_in` in Figure. 4.1, has been downloaded. This shapefile contains historical fire data for all of Canada, spanning the years 1946 to 2021. We imported this shapefile as a vector layer into the Quantum Geographic Information System (QGIS) and subsequently saved it as a Comma Separated Values (CSV) file. The exported file, named `fire_out`, was filtered based on province and year, selecting only records pertaining to Saskatchewan and covering the years 2000–2018. During this process, the dates in the YYYY-MM-DD format were converted to the day of the year. Some preliminary data cleaning was also carried out to remove data points outside the provincial boundaries. These steps were performed in the box labeled "Filter and Pre-processing Data" in Figure. 4.1, and the output file is called `fire_of`.

The `cmet_out` and `fire_of` files from the previous steps were merged, and the target column "fire" was added using Algorithm 2. This is highlighted in the box titled "Filter and Join on (Coordinate_ID, day)" in Figure. 4.1.

To populate the "fire" column, we considered both the spatial resolution of the meteorological data, which was set at 0.25 degrees and the size of the fires. The "fire" column is populated based on two conditions: 1) whether there are any historical fires within a given bounding box; and 2) whether there are any within a radius calculated based on the fire's size. If a fire meets either of these conditions, the column "fire" for that particular location is set to 1; otherwise, it remains at 0. This approach helps us identify locations that are in close proximity to historical fires.

Algorithm 2 serves two primary functions:

1. It employs the *haversine formula*, as detailed in Algorithm 3, to calculate the distance between fire locations and meteorological data points based on their latitude and longitude. The Haversine formula is specifically designed to compute distances on a sphere, making it ideal for calculating distances on the Earth's surface given its curvature. This ensures a more accurate distance measurement compared to simpler Cartesian calculations.

2. The algorithm checks whether each data point lies within a square bounding box of a given resolution, as demonstrated in Algorithm 4.

Algorithm 2 Generalized Fire Proximity Check

```
Prepare the dataframes for checking
Initialize necessary columns and flags
for each location in all locations do
    Extract location details (latitude, longitude, day, year)
    Filter relevant fire data based on day and year
    for each fire incident in filtered fire data do
        Extract fire details (latitude, longitude, size)
        Check if fire is inside location buffer
        Calculate distance from fire to location
        if fire is inside location OR distance is within threshold then
            Update location as affected by fire
            Exit inner loop since fire is found near location
        end if
    end for
    if fire found for current location then
        Continue to next location
    end if
end for
```

Algorithm 2 gives the full overview of the process. The objective is to find out if a certain location (given by the latitude and longitude coordinates) is within the vicinity of a fire event from historical data. During this data preprocessing phase, we filtered out rows corresponding to periods outside of the fire seasons to ensure that our dataset primarily captures the relevant timeframes when forest fires are most likely to occur. The `data_out` file is the final output of this step.

4.2.3 Provincial Boundary

To isolate data specific to our targeted province, Saskatchewan, we require a separate source for provincial boundary information. These boundary files provide geographic coordinates in terms of latitude and longitude and portray the full extent of the area, including any adjacent coastal water regions. This data source is represented as `bound_src` in workflow Figure 4.1.

For the purposes of this study, `bound_in` file, the 2021 census boundary shapefile provided by Statistics Canada was used to determine the Saskatchewan provincial boundary.

Algorithm 3 Calculate Distance Using Haversine Formula

function HAVERSINE(lon1, lat1, lon2, lat2)
    Convert lon1, lat1, lon2, lat2 to radians
    $dlon \leftarrow lon2 - lon1$
    $dlat \leftarrow lat2 - lat1$
    $a \leftarrow \sin^2(dlat/2) + \cos(lat1) \times \cos(lat2) \times \sin^2(dlon/2)$
    $c \leftarrow 2 \times \text{asin}(\sqrt{a})$
    $R \leftarrow 6371$ ▷ R: Radius of Earth in kilometers
    return $c \times R$
end function

Algorithm 4 Check if a Coordinate is Inside a Bounding Box

function IS_COORDINATE_INSIDE(lat1, lon1, lat2, lon2, resolution_degrees)
    $lat\_diff \leftarrow |lat1 - lat2|$
    $lon\_diff \leftarrow |lon1 - lon2|$
    if $lat\_diff \leq resolution\_degrees$ and $lon\_diff \leq resolution\_degrees$ **then**
        return True
    else
        return False
    end if
end function

4.2.4 Water Body Shapefile

To refine the quality and relevance of our dataset, we performed a second round of data extraction specifically designed to exclude water bodies from the geographical locations studied. Since water bodies such as lakes, rivers, and oceans are not susceptible to fires, their inclusion in the dataset would not provide any meaningful insights for our predictive fire model. These extraneous data could even introduce noise or bias, thereby affecting the model's accuracy.

To introduce this supplementary layer of data cleaning, a specific shapefile, termed `water_in`, which contains detailed geographic information about water bodies, was employed. Using this shapefile, data points in `data_out` corresponding to water bodies were effectively removed. This optimization enhances the accuracy and relevance of analysis in subsequent research stages. `data_of` represents the output file containing the finalized raw data.

4.2.5 Saskatchewan Data-set Summary

The final data set includes the features described in Table 4.8. Data collection and processing were conducted on an annual basis for the years 2000 through 2018, ensuring a comprehensive data set. The ERA5 data retrieval was executed 19 times, once for each year in the range. All other steps were completed using a single Python script. Table 4.7 illustrates the modifications made during the data collection framework process, as depicted in Figure 4.1.

Table 4.7: Saskatchewan Data-set Summary

Data-set	Total Rows	Feature	Class	Fire	Non-Fire
cmet_out	5,655,825	26	0	0	0
data_out	5,655,825	27	1	30,548	5,625,277
data_of	4,381,020	27	1	28,256	4,352,764

Table 4.8: Feature Description

Feature	Unit	Description
lon	degrees	Upper left hand corner longitude of a cell.
lat	degrees	Upper left hand corner latitude of a cell.
u10	m/s	Eastward component of wind at 10m above the earth's surface.
v10	m/s	Northward component of wind at 10m above the earth's surface.
d2m	Kelvin (K)	The dewpoint temperature at 2m above the Earth's surface. It is a measure of humidity.
t2m	Kelvin (K)	Air temperature at 2m above the Earth's surface.
lai_hv	m^2/m^2	One-half of the total green leaf area per unit horizontal ground surface area for high vegetation.
lai_lv	m^2/m^2	One-half of the total green leaf area per unit horizontal ground surface area for low vegetation.
ro		Runoff
src	m of water	The skin reservoir content is the amount of water in the vegetation canopy and/or in a thin layer on the soil.
skt	Kelvin (K)	Skin temperature is the temperature at the Earth's surface.
stl1	Kelvin (K)	Temperature of the soil in layer 1 (0 -7 cm) of the ECMWF Integrated Forecasting System.
stl2	Kelvin (K)	Temperature of the soil in layer 2 (7 -28 cm) of the ECMWF Integrated Forecasting System.
stl3	Kelvin (K)	Temperature of the soil in layer 3 (28 -100 cm) of the ECMWF Integrated Forecasting System.
stl4	Kelvin (K)	Temperature of the soil in layer 1 (100 -289 cm) of the ECMWF Integrated Forecasting System.
ssr		surface net solar radiation
str		surface net thermal radiation
sp	Pa	Surface pressure (force per unit area)is the atmospheric pressure at the earth's surface.
e	m of water	The amount of water evaporation from bare soil.
tp	m	total precipitation.
swvl1	m^3/m^3	Volume of water in soil layer 1 (0 -7 cm) of the ECMWF Integrated Forecasting System.
swvl2	m^3/m^3	Volume of water in soil layer 1 (7 -28 cm) of the ECMWF Integrated Forecasting System.
swvl3	m^3/m^3	Volume of water in soil layer 1 (28-100 cm) of the ECMWF Integrated Forecasting System.
swvl4	m^3/m^3	Volume of water in soil layer 1 (100 -289 cm) of the ECMWF Integrated Forecasting System.
d	day	The day of the year for the given year.
y	year	The year.
Coo_ID	Int	A unique identifier representing a specific pair of latitude and longitude coordinates.
fire	Boolean	1 if any ignition occurred in the cell on the day, otherwise 0.

Chapter 5

Experiments and Result Analysis

In this section, we focus on the evaluation of three state-of-the-art ML algorithms—CatBoost, XGBoost, and LightGBM—for their effectiveness in classifying forested areas as fire-affected (labeled as "1") or non-fire-affected (labeled as "0"). Utilizing a range of metrics such as Accuracy, Sensitivity, Specificity, and ROC-AUC, we aim to provide a holistic view of each model's performance in predicting these crucial environmental states. Given the critical nature of timely and accurate forest fire detection, these metrics serve as key indicators for the reliability and efficacy of the models in real-world applications. The ensuing discussion will elucidate the nuances of these results, emphasizing both their successes and limitations in the realm of forest fire classification.

5.1 Initial Result Before Re-sampling

As shown in Table 5.1, the models—CatBoost, XGBoost, and LightGBM—display accuracy rates of around 99%. Detailed analysis reveals sensitivity values ranging from 0.01 to 0.04, indicating challenges in classifying the minority class. Specificity is consistent at 1.00, denoting the models' effectiveness in recognizing the majority class. These metrics highlight a class imbalance issue in the dataset.

The ROC-AUC scores are relatively high but are not aligned with the low sensitivity, suggesting the models, while potentially discriminative, are biased towards the majority class. To address this class imbalance and achieve a balanced and generalizable model, resampling techniques such as oversampling the minority class or undersampling the majority class are necessary in subsequent steps.

Table 5.1: Initial Results

Model	Accuracy	Sensitivity	Specificity	ROC-AUC
CatBoost	0.99	0.04	1.00	0.93
XGBoost	0.99	0.01	1.00	0.92
LightGBM	0.99	0.03	1.00	0.88

When plotting the class distribution, as shown in Figure 5.1, we observe that our dataset suffers significantly from class imbalance.

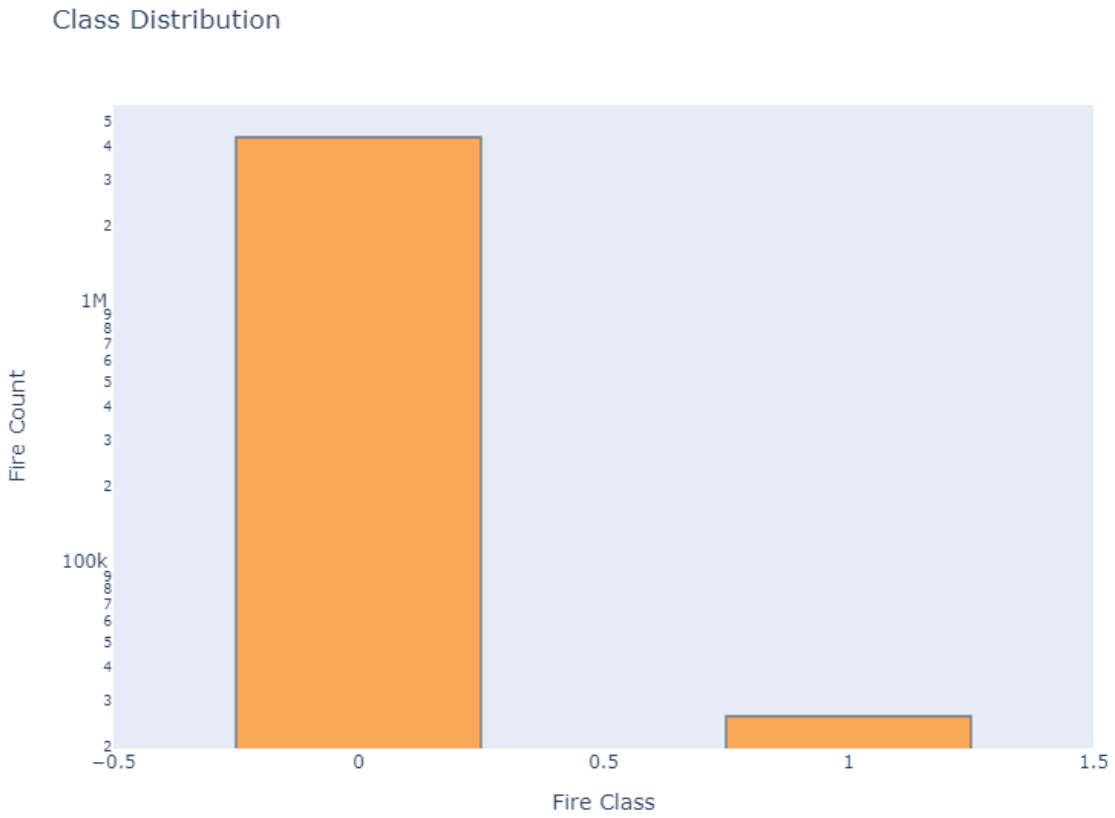


Figure 5.1: Class Distribution Prior Applying Re-sampling Techniques

5.2 Re-sampling

To address the class imbalance in the fire detection dataset, several strategies were adopted. The primary step involved data pruning, removing rows associated with locations without a fire history. Further refinement was achieved by including only days that fall within recognized fire seasons. Furthermore, data points located within water bodies were removed. These measures significantly reduced the gap between the "non-fire" class (4,684,435 rows) and the "fire" class (30,548 rows).

When employing different re-sampling techniques and using varying ratios, it was observed that Random Forest was computationally more costly compared to XGBoost and LightGBM. Consequently, the experiment was structured in two separate runs. The first run solely utilized Random Forest, while the second combined XGBoost and LightGBM. Among the re-sampling techniques, SMOTE-ENN proved to be the most time-consuming, requiring approximately 1000 minutes to determine the model's performance. In comparison, SMOTE took around 30 minutes, and NearMiss3 approximately 12 minutes. The results of these methodologies are elucidated in subsection 5.2.1.

To further counteract the imbalance, NearMiss3, an undersampling technique designed to preserve the intrinsic distribution of the minority class, was used. This method is especially valuable for classifying rare but crucial events such as fires. The empirical evaluation of the NearMiss3 model is presented in Table 5.2. The table demonstrates that different sampling ratios provide consistent performance in key metrics such as sensitivity, specificity, and ROC-AUC. For instance, with a sampling strategy of 0.08, an overall accuracy rate of 0.8 was achieved.

For example, when employing a sampling strategy of 0.08, the model attained a specificity of 0.76, a sensitivity of 0.72, and an ROC-AUC of 0.86, culminating in an overall accuracy rate of 0.8. In this specific configuration, the model accurately classified 16,660 "fire" instances and 208,250 "non-fire" instances. It is noteworthy that the performance metrics, particularly ROC-AUC and accuracy, exhibit a consistent range across different sampling strategies, thereby indicating stable model behavior.

While undersampling techniques, such as NearMiss3, demonstrated efficacy in balancing our dataset for fire detection tasks, our experiments with GANs for oversampling yielded less promising results.

Oversampling is generally considered advantageous because it allows for the introduction of synthetic instances that can help the model learn complex decision boundaries. This

Sampling Ratio	Specificity	Sensitivity	ROC-AUC	fire	non-fire
0.05	0.71	0.68	0.82	16000	200500
0.06	0.72	0.69	0.81	16100	199500
0.07	0.74	0.70	0.83	16300	201000
0.08	0.76	0.72	0.86	16660	208250
0.09	**0.76**	**0.75**	**0.85**	16660	185111
0.10	0.77	0.73	0.87	16700	209000

Table 5.2: NearMiss3 Sampling Ratio Analysis

could be particularly useful in scenarios where the minority class is not just rare but also carries complex features that are hard to generalize from the limited number of instances.

In the case of GANs, the potential to generate high-quality synthetic instances that resemble the minority class makes them an attractive choice for oversampling. However, our empirical findings indicate that the GANs-augmented model did not outperform the NearMiss3 model across key performance metrics such as sensitivity, specificity, and ROC-AUC.

The less effective performance of GANs-based oversampling in our study may be attributed to several factors: the complexity of generated instances that diverge from the true data distribution, the risk of model overfitting due to synthetic data, the high computational costs involved, and the reduced model interpretability which is especially crucial in safety-critical applications like fire classification.

In our quest to address class imbalance, we also tested the SMOTE-ENN hybrid technique. This method combines SMOTE and ENN with specific parameters: a sampling strategy of 0.09, k-neighbors set to 50 for SMOTE, and n-neighbors set to 100 for ENN, all with a random state of 42.

The classification report revealed a precision of 1.00 and a recall of 1.00 for the majority class ("non-fire"). For the minority class ("fire"), the precision was 0.34, and the recall was 0.37, yielding an F1-score of 0.36. The confusion matrix reported 865,863 true negatives, 3,699 false positives, 3,277 false negatives, and 1,929 true positives. The ROC-AUC was notably high at 0.9527, indicating a good discriminative power. 5.3 summeries the results and 5.2 demonstrates the ROC-AUC graph.

Despite its computational complexity and the promise of generating a balanced dataset, SMOTE-ENN did not outperform the NearMiss3 method in terms of sensitivity for the minority class. However, the high ROC-AUC score suggests that the model has the potential for excellent discriminatory performance between classes.

Several factors might contribute to the modest sensitivity performance. Like the GANs

Table 5.3: Classification Report for SMOTE-ENN Method

Metric	Non-fire Class	Fire Class
Recall (Sensitivity)	1.00	0.37
Specificity	0.996	0.34
ROC-AUC	0.9527	

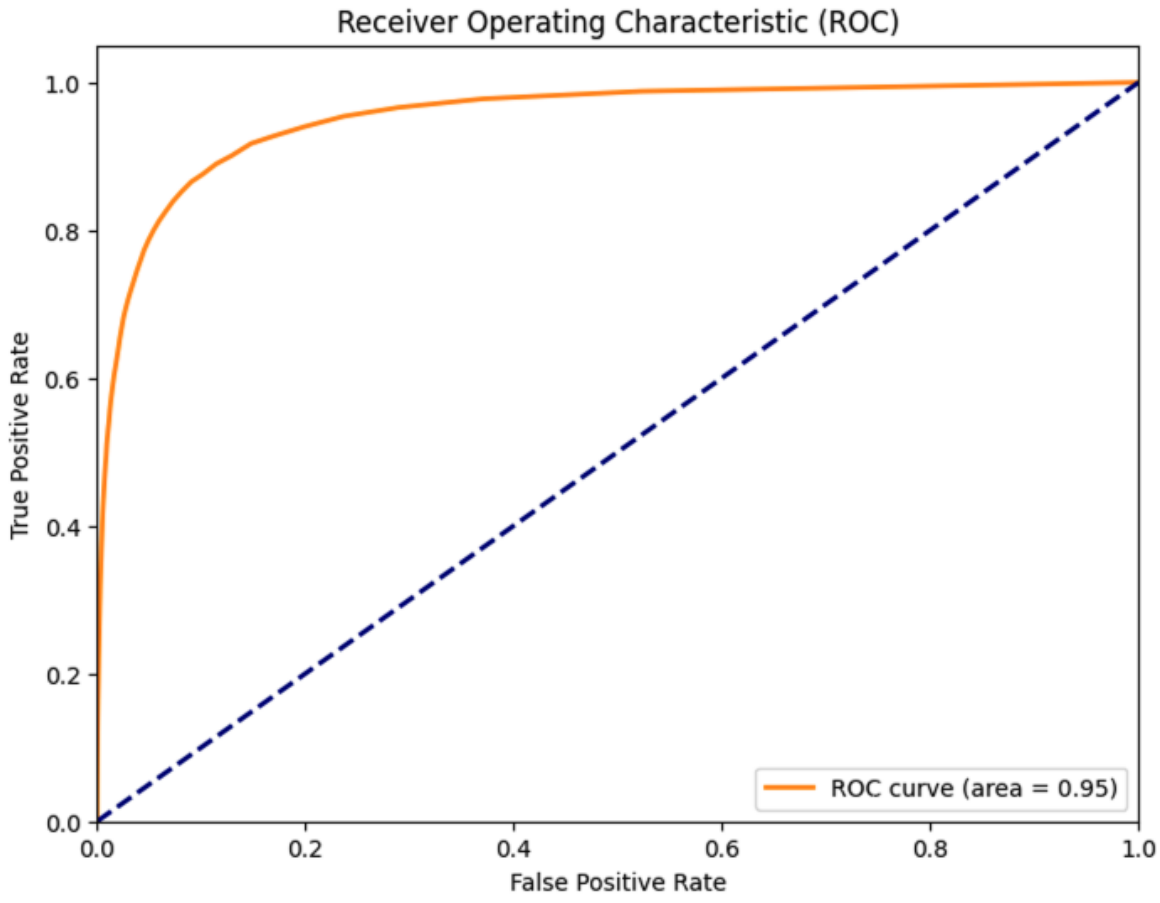


Figure 5.2: ROC-AUC for SMOTE-ENN

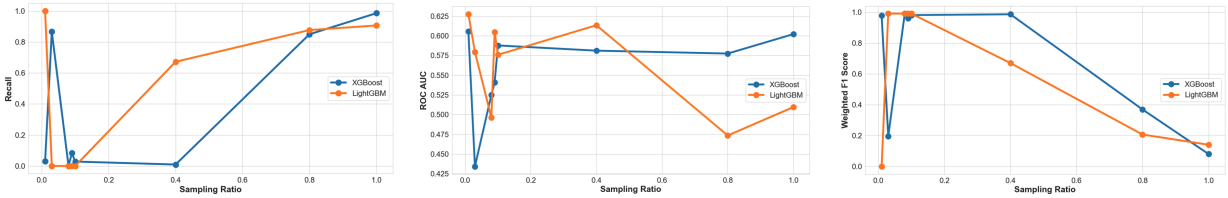
approach, the synthetic instances generated by SMOTE could make the model susceptible to overfitting, while the editing function of ENN may not adequately refine the class boundaries.

In summary, although SMOTE-ENN showed promise in aspects such as ROC-AUC, its performance on other key metrics suggests that it may not be the optimal solution for this specific fire detection task when compared to simpler methods such as NearMiss3.

5.2.1 Performance Analysis of Gradient Boosting Algorithms

This section examines the effects of sampling ratios on XGBoost and LightGBM, two prominent gradient boosting algorithms. XGBoost employs a depth-first approach suitable for sparse data, while LightGBM, using a histogram-based method, excels with large datasets due to its unique growth and pruning mechanisms. Both offer advanced regularization. The performance of three re-sampling techniques, NearMiss3, SMOTE, and SMOTE-ENN, is presented in subsequent figures.

Figure 5.3 elucidates the results with the NearMiss3 approach, where both models exhibit fluctuations in recall as undersampling intensifies, but maintain commendable ROC-AUC and weighted F1 scores. NearMiss3 showed a wide variance in performance based on the sampling strategy. For XGBoost, a notable recall of approximately 82.13% was observed for the 0.03 sampling ratio. In contrast, the specificity was consistent across most strategies, peaking at 100%. However, the precision was significantly low, resulting in a diminished weighted F1 score. LightGBM, at a 0.01 sampling ratio, exhibited perfect recall but zero specificity, essentially classifying all instances as positive.

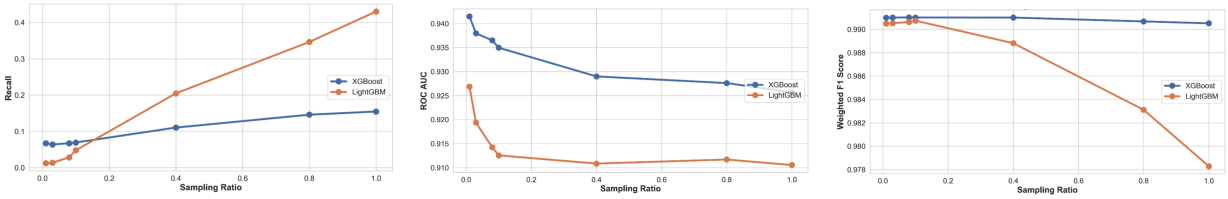


(a) Recall (b) ROC-AUC (c) Weighted F1 Score

Figure 5.3: Performance metrics of XGBoost and LightGBM with NearMiss re-sampling.

SMOTE re-sampling results, shown in Figure 5.4, indicate consistently high specificity across the techniques, with both models demonstrating a notable ability to distinguish between classes as evident in the ROC-AUC scores. The Weighted F1 Score further signifies the balance between precision and recall achieved by the models. Employing the SMOTE technique resulted in improved outcomes, especially in terms of recall. XGBoost's recall peaked at the 0.400, 0.800, and 1.000 ratio. LightGBM, on the other hand, exhibited its maximum recall of 0.43 at the 1.000 sampling ratio. Specificity remained consistently

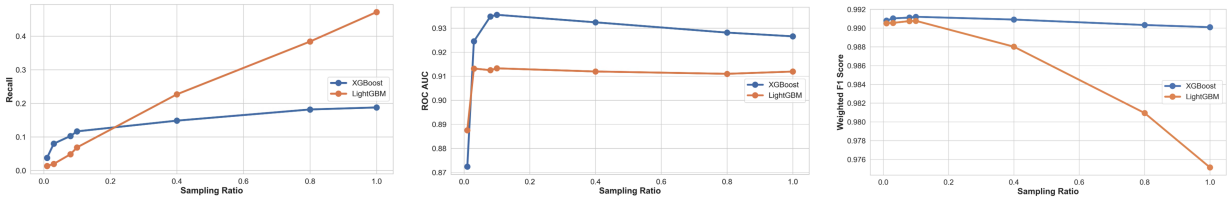
high across both models and ratios, with a marginal drop observed for LightGBM at the most aggressive oversampling levels. The ROC-AUC metric indicated strong discriminative power for both models across all strategies.



(a) Recall (b) ROC-AUC (c) Weighted F1 Score

Figure 5.4: Performance metrics of XGBoost and LightGBM with SMOTE re-sampling.

Moving to the SMOTE-ENN technique, as illustrated in Figure 5.5, both algorithms maintain a robust class-distinguishing capability, even as the recall rate witnesses variance across the sampling ratios. The SMOTE-ENN technique provided an improvement in recall, especially for XGBoost at higher sampling strategies. For LightGBM, the recall significantly increased with more aggressive sampling strategies, peaking at 0.43 for the 1.00 ratio. This, however, resulted in a decrease in specificity. ROC-AUC values remained consistently high for both models, suggesting good discriminative capabilities.



(a) Recall (b) ROC-AUC (c) Weighted F1 Score

Figure 5.5: Performance metrics of XGBoost and LightGBM with SMOTE-ENN re-sampling.

In terms of recall, the SMOTE method provided superior results compared to NearMiss3 and SMOTE-ENN, especially when aggressive oversampling strategies were adopted. Specificity was consistently high in all techniques and strategies, suggesting minimal compromise in accurately identifying negative instances.

It's essential to highlight that while NearMiss3 provided high recall in specific instances, it often came at the cost of precision. Such a scenario is not ideal, especially when the consequences of false positives are significant.

The high ROC-AUC values across models and techniques underline the efficacy of the models in distinguishing between the two classes.

5.3 Feature Importance

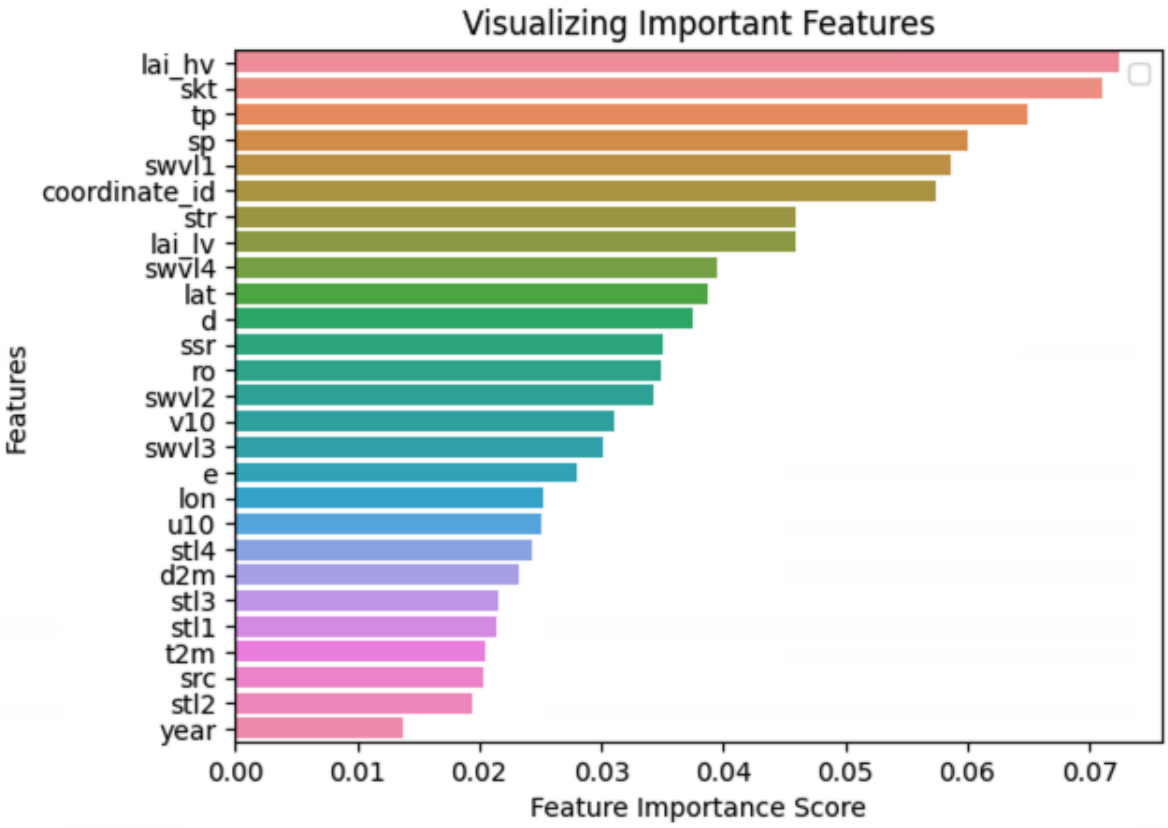


Figure 5.6: Random Forest Feature Importance Analysis

we also conducted feature importance analyses utilizing both Random Forest and CatBoost algorithms to better understand the contribution of individual features to the model's predictive performance. Random Forest provides an intuitive, average-based measure of feature significance derived from the ensemble of decision trees, capturing the average reduction in impurity caused by each feature. On the other hand, CatBoost employs a

more sophisticated approach that accounts for the potential interactions between features, offering a more nuanced interpretation.

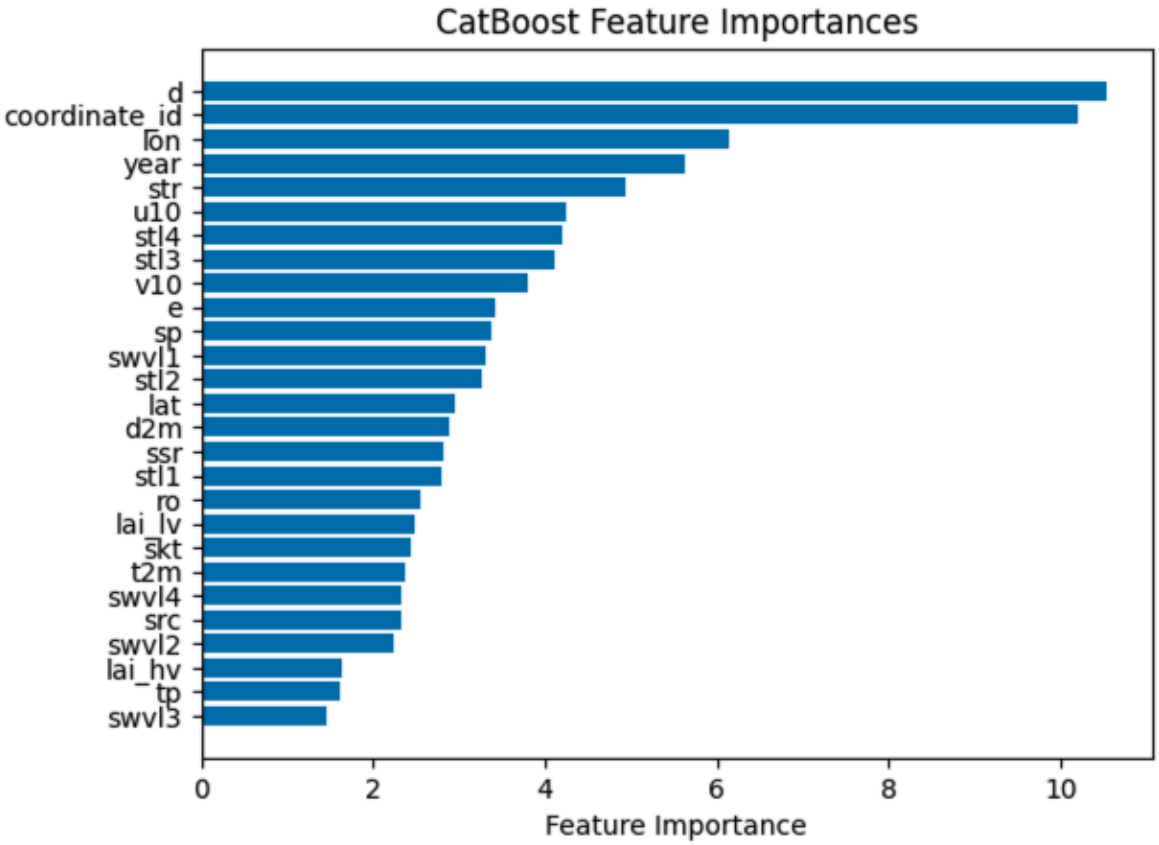


Figure 5.7: CatBoost Feature Importance Analysis

5.4 Best Model

The NearMiss3 method, set with a 0.09 sampling ratio, was implemented to address the class imbalance in fire classification, adjusting the majority class ("non-fire") in proportion to the minority class ("fire"). The performance metrics of four ML models, namely, Random Forest, XGBoost, LightGBM, and CatBoost, were initially considered for analysis. However, the CatBoost model was discarded since our dataset does not contain any categorical data.

Random Forest achieved an accuracy of 78.32%, a sensitivity of 74.78%, and a specificity of 78.34%. On the contrary, XGBoost excelled, recording an accuracy of 98.08%, a sensitivity of 86.06%, and a specificity of 93.03%. LightGBM's metrics were 72.38% for accuracy, 76.03% for sensitivity, and 72.36% for specificity.

In summary, XGBoost demonstrated superior results. The synergy between XGBoost and undersampling arises from the former's gradient boosting mechanism which inherently handles bias towards the majority class. When combined with undersampling, which reduces the volume of the majority class, XGBoost is better equipped to discern patterns in the minority class, thereby enhancing model performance on imbalanced datasets. This underscores the importance of an optimized undersampling technique when dealing with such datasets.

Table 5.4: Summary of Best Performance Results

Model	Accuracy	Sensitivity	Specificity
Random Forest	0.7832	0.7478	0.7834
XGBoost	**0.9808**	**0.8606**	**0.9303**
LightGBM	0.7238	0.7603	0.7236

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Forest fires, particularly in the context of climate change, are an increasing concern for both ecological systems and human communities. Traditional methods for their prediction and management have increasingly shown their limitations. This thesis successfully addresses these challenges by developing a ML-based predictive framework specifically tailored for forest fire classification in the Canadian context. Using Copernicus reanalysis data, the study focused on employing four advanced ML algorithms, namely, Random Forest, XGBoost, LightGBM, and CatBoost, alongside a comprehensive set of features to deliver a robust and scalable solution.

Through testing, NearMiss3 was the standout re-sampling method. Metrics recorded were: Random Forest (78.3% accuracy, 74.8% sensitivity, 78.3% specificity), XGBoost (98.08% precision, 86.06% sensitivity, 93.03% specificity), and LightGBM (72.38% accuracy, 76.03% sensitivity, 72.36% specificity).

During this research, the potential of GANs was explored as a potential method for data augmentation and class balancing. However, their application in this domain remains a work in progress and warrants further investigation. Furthermore, the CatBoost algorithm was eventually dropped from the main experimentation due to the absence of categorical data in our dataset, deeming it less suitable compared to the other algorithms employed.

Research contributions range from the creation of a feature-rich dataset, and the efficient handling of class imbalances, to the development of a ML framework specifically designed for forest fire classification. Therefore, the findings not only have academic sig-

nificance but are also pivotal for practical applications in proactive fire prevention and management strategies.

6.2 Future Work

While this research serves as a strong foundation, several avenues for future work are evident:

1. **Extending the Dataset**: The current dataset, although comprehensive, is geographically limited. Future work could include expanding the dataset to include other regions and countries, and potentially integrating additional variables such as human activity metrics, to refine the model's generalizability.

2. **Algorithmic Enhancements**: Although Random Forest emerged as the superior algorithm in this study, the ever-evolving field of ML provides opportunities for testing newer algorithms and ensemble methods that could potentially improve prediction accuracy.

3. **Real-time Prediction**: The present framework is focused on historical data. Adapting the model for real-time predictions could provide valuable insights for immediate forest fire interventions.

4. **Investigate Alternative Re-sampling Methods**: While the NearMiss method showed promising results, future work could explore other re-sampling techniques or perhaps even custom-developed algorithms for performance improvements.

5. **Deep Learning Approaches**: The study's initial foray into using GANs for handling class imbalance looks promising. Further exploration into deep learning techniques could potentially heighten the model's predictive power.

6. **Interdisciplinary Collaboration**: For making the model more actionable, partnerships with governmental agencies, ecologists, and fire management experts could be considered. This multi-disciplinary approach would ensure that the technological advancements are efficiently translated into effective forest management policies.

By addressing these areas, future research can build upon the solid foundation laid by this thesis, further contributing to the development of increasingly effective and adaptable models for forest fire prediction and management.

References

[1] *State of Canada's Forests: 2004–2005, The Boreal Forest*. Canadian Forest Service, 2005. p. 43.

[2] S. D. Ali et al. Geoai for disaster mitigation: Fire severity prediction models using sentinel-2 and ann regression. In *2022 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*, pages 1–7, Yogyakarta, Indonesia, 2022.

[3] Mark Baldwin-Smith. Global distribution of the taiga and boreal forest biome, and its ecoregions, 2012. Creative Commons Attribution-Share Alike 3.0 Unported license.

[4] Dieu Tien Bui, Hung Van Le, and Nhat-Duc Hoang. Gis-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. *Ecological Informatics*, 48:104–116, 2018.

[5] Copernicus Climate Change Service (C3S). Era5 hourly data on single levels from 1940 to present, 2023. Accessed on 01-AUG-2023.

[6] Statistics Canada. 2021 standard geographical classification (sgc) - boundaries. https://www12.statcan.gc.ca/census-recensement/2021/geo/sip-pis/boundary-limites/index2021-eng.cfm?year=21, 2021. Accessed: 2022-09-26.

[7] Canadian Interagency Forest Fire Centre. Fire statistics, August 2023.

[8] Paulo Cortez and Aníbal de Jesus Raimundo Morais. A data mining approach to predict forest fires using meteorological data. 2007.

[9] Mario Elia, Marina D'Este, Davide Ascoli, Vincenzo Giannico, Giuseppina Spano, Antonio Ganga, Giuseppe Colangelo, Raffaele Lafortezza, and Giovanni Sanesi. Estimating the probability of wildfire occurrence in mediterranean landscapes using artificial neural networks. *Environmental Impact Assessment Review*, 85:106474, 2020.

[10] Osama Elsarrar, Marjorie Darrah, and Richard Devine. Analysis of forest fire data using neural network rule extraction with human understandable rules. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1917–19176, 2019.

[11] Keke Gao, Zhongke Feng, and Shan Wang. Using multilayer perceptron to predict forest fires in jiangxi province, southeast china. *Discrete Dynamics in Nature and Society*, 2022, 2022.

[12] R. Ghali, M. A. Akhloufi, M. Jmal, W. S. Mseddi, and R. Attia. Forest fires segmentation using deep convolutional neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2109–2114, Melbourne, Australia, 2021.

[13] S. N. Ghate, P. Sapkale, and M. Mukhedkar. Forest wildfire detection and forecasting utilizing machine learning and image processing. In *2023 International Conference for Advancement in Technology (ICONAT)*, pages 1–8, Goa, India, 2023.

[14] Saskatchewan Government. Hydrography map service, 2022. Accessed: 2022-09-26.

[15] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on single levels from 1940 to present, 2018. Accessed on 01-AUG-2023.

[16] N. Hidayanto, A. H. Saputro, and D. E. Nuryanto. Peatland data fusion for forest fire susceptibility prediction using machine learning. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 544–549, Yogyakarta, Indonesia, 2021.

[17] Haoyuan Hong, Paraskevas Tsangaratos, Ioanna Ilia, Junzhi Liu, A-Xing Zhu, and Chong Xu. Applying genetic algorithms to set the optimal combination of forest fire related variables and model forest fire susceptibility based on data mining models. the case of dayu county, china. *Science of the total environment*, 630:1044–1056, 2018.

[18] imbalanced-learn developers. imbalanced-learn: Over-sampling and under-sampling for imbalanced datasets, 2022. Accessed: 2022-09-26.

[19] Parveen Kaur. Forest fire prediction using heterogeneous data sources and machine learning methods. Master's thesis, University of Waterloo, 2023.

[20] B. Kosović et al. Estimation of fuel moisture content by integrating surface and satellite observations using machine learning. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3626–3628, Waikoloa, HI, USA, 2020.

[21] Can Lai, Shucai Zeng, Wei Guo, Xiaodong Liu, Yongquan Li, and Boyong Liao. Forest fire prediction with imbalanced data using a deep neural network method. *Forests*, 13(7):1129, Jul 2022.

[22] Ryan Laube and Howard J. Hamilton. Wildfire occurrence prediction using time series classification: A comparative study. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4178–4182, 2021.

[23] Bruce D Lawson and OB Armitage. Weather guide for the canadian forest fire danger rating system. 2008.

[24] BS Lee, ME Alexander, BC Hawkes, TJ Lynham, BJ Stocks, and P Englefield. Information systems in support of wildland fire management decision making in canada. *Computers and Electronics in Agriculture*, 37(1-3):185–198, 2002.

[25] Yudong Li, Zhongke Feng, Shilin Chen, Ziyu Zhao, and Fengge Wang. Application of the artificial neural network and support vector machines in forest fire prediction in the guangxi autonomous region, china. *Discrete Dynamics in Nature and Society*, 2020:1–14, 2020.

[26] James MacCarthy, Jessica Richter, Sasha Tyukavina, Mikaela Weisse, and Nancy Harris. The latest data confirms: Forest fires are getting worse, August 2023.

[27] B R Manju and Anju R Nair. Classification of cardiac arrhythmia of 12 lead ecg using combination of smoteenn, xgboost and machine learning algorithms. In *2019 9th International Symposium on Embedded Computing and System Design (ISED)*, pages 1–7, 2019.

[28] Elliot Mbunge, Maureen Nokuthula Sibiya, Sam Takavarasha, Richard C Millham, Garikayi Chemhaka, Benhildah Muchemwa, and Tafadzwa Dzinamarira. Implementation of ensemble machine learning classifiers to predict diarrhoea with smoteenn, smote, and smotetomek class imbalance approaches. In *2023 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6, 2023.

[29] N. Nesa, T. Ghosh, and I. Banerjee. Outlier detection in sensed data using statistical learning models for iot. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, Barcelona, Spain, 2018.

[30] N. Omar, A. Al-zebari, and A. Sengur. Deep learning approach to predict forest fires using meteorological measurements. In *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, pages 1–4, Ankara, Turkey, 2021.

[31] T. Preeti, S. Kanakaraddi, A. Beelagi, S. Malagi, and A. Sudi. Forest fire prediction using machine learning techniques. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–6, Hubli, India, 2021.

[32] T Preeti, Suvarna Kanakaraddi, Aishwarya Beelagi, Sumalata Malagi, and Aishwarya Sudi. Forest fire prediction using machine learning techniques. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–6, 2021.

[33] Richard Purcell. Forest fire prediction frameworks using federated learning and internet of things (iot). 2023.

[34] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, 67:93–104, 2012.

[35] Youssef Safi and Abdelaziz Bouroumi. Prediction of forest fires using artificial neural networks. *Applied Mathematical Sciences*, 7(6):271–286, 2013.

[36] Rick Sauber-Cole, Taghi M. Khoshgoftaar, and Justin M. Johnson. Gans for class-imbalanced data: A meta-analysis of github projects. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1419–1424, 2022.

[37] Canadian Forest Service. Canadian wildland fire information system (cwfis) datamart, 2022. Licensed under the Open Government Licence - Canada. Available at http://open.canada.ca/en/open-government-licence-canada.

[38] B. K. Singh, N. Kumar, and P. Tiwari. Extreme learning machine approach for prediction of forest fires using topographical and metrological data of vietnam. In *2019 Women Institute of Technology Conference on Electrical and Computer Engineering (WITCON ECE)*, pages 104–112, Dehradun, India, 2019.

[39] N Srinivashini, M Raveenthini, and R Lavanya. Deep ensemble of texture maps for false positive reduction in mammograms. In *Journal of Physics: Conference Series*, volume 2318, page 012038. IOP Publishing, 2022.

[40] S. Sudhakar et al. Unmanned aerial vehicle (uav) based forest fire detection and monitoring for reducing false alarms in forest-fires. *Computer Communications*, 149:1–16, 2020.

[41] S. Suklabaidya and I. Das. Processing iot sensor fire dataset using machine learning techniques. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 1–7, Silchar, India, 2023.

[42] Huaning Tan, Renxing Chen, Meng Qin, Lining Tang, Zhibing Wu, Qianlin Luo, and Yujuan Quan. Tabular gan-based oversampling of imbalanced time-to-event data for survival prediction. In *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 376–380, 2023.

[43] D. K. Tayal, N. Agarwal, A. Jha, Deepakshi, and V. Abrol. To predict the fire outbreak in australia using historical database. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 1–7, Noida, India, 2022.

[44] Alexandra Tyukavina, Peter Potapov, Matthew C Hansen, Amy H Pickens, Stephen V Stehman, Svetlana Turubanova, Diana Parker, Viviana Zalles, André Lima, Indrani Kommareddy, et al. Global trends of forest loss due to fire from 2001 to 2019. *Frontiers in Remote Sensing*, 3:825190, 2022.

[45] CE Van Wagner et al. *Development and structure of the Canadian forest fire weather index system*, volume 35. 1987.

[46] Zili Wang, Binbin He, and Xiaoying Lai. Balanced random forest model is more suitable for wildfire risk assessment. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 3596–3599, 2022.

[47] Suwei Yang, Massimo Lupascu, and Kuldeep S Meel. Predicting forest fire using remote sensing data and machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14983–14990, 2021.

[48] Liyang Yu, Neng Wang, and Xiaoqiao Meng. Real-time forest fire detection with wireless sensor networks. In *Proceedings. 2005 International Conference on Wireless*

Communications, Networking and Mobile Computing, 2005., volume 2, pages 1214–1217. Ieee, 2005.

[49] V. Zope, T. Dadlani, A. Matai, P. Tembhurnikar, and R. Kalani. Iot sensor and deep neural network based wildfire prediction system. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 205–208, Madurai, India, 2020.