



Big Data Synthesis and Class Imbalance Rectification for Enhanced Forest Fire Classification Modeling

Fatemeh Tavakoli¹^a, Kshirasagar Naik¹^b, Marzia Zaman²^c, Richard Purcell³^d,
Srinivas Sampalli³^e, Abdul Mutakabbir⁴^f, Chung-Horng Lung⁴^g
and Thambirajah Ravichandran⁵^h

¹Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

²Research and Development, Cistel Technology, Ottawa, ON, Canada

³Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

⁴Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

⁵Research and Development, Hegyi Geomatics International Inc., Ottawa, ON, Canada

Keywords: Forest Fire, Classification, Machine Learning, Supervised Learning, Dataset, Big Data, Random Forest, XGBoost, LightGBM, SMOTE, NearMiss, SMOTE-ENN.

Abstract: Forest fires have been escalating in frequency and intensity across Canada in recent times. This study employs machine learning techniques and builds a dataset framework utilizing Copernicus climate reanalysis data combined with historical fire data to develop a fire classification framework. Three algorithms, Random Forest, XGBoost, and LightGBM, were evaluated. Given the pronounced class imbalance of 154:1 between “non-fire” and “fire” events, we rigorously employed two re-sampling strategies: Spatiotemporal, focusing on spatial and seasonal considerations, and Technique-Driven, leveraging advanced algorithmic approaches. Ultimately, XGBoost combined with NearMiss Version 3 in a 0.09 sampling ratio between “non-fire” and “fire” events yielded the best results: 98.08% precision, 86.06% sensitivity, and 93.03% specificity.

1 INTRODUCTION


The increasing prevalence of forest fires poses a substantial threat to both ecological systems and human communities. This challenge is magnified by contributing factors such as climate change, anthropogenic activities, and constrained preventive and management measures. A recent study reveals that the global tree cover loss attributed to forest fires has nearly doubled in the past two decades (Tyukavina et al., 2022). Specifically in Canada, an alarming 9.5 million hectares were burned in just the first seven months of 2023, highlighting the severity of the prob-


lem (Canadian Interagency Forest Fire Centre, 2023).


Amid these challenges, Machine Learning (ML) emerges as a robust solution. With its capability to dissect intricate datasets and adapt to swift environmental changes, it holds promise for transformative forest fire management. Applications of ML in this domain encompass:


- **Early Detection:** Analyzing satellite and sensor data for preliminary fire indications.
- **Predictive Analysis:** Employing historical weather and vegetation data to predict forest fire probabilities in specific regions.
- **Resource Allocation:** Strategically allocating firefighting resources based on past fire incidents and resource availability.
- **Post-Fire Analysis:** Quantifying the environmental ramifications of fires and forecasting the necessary recovery durations.


Building upon the predictive potential of ML, this study introduces an ML-based framework for forest fire classification. Utilizing Copernicus climate data,


^a <https://orcid.org/0009-0008-9734-6417>


^b <https://orcid.org/0000-0002-1064-4905>


^c <https://orcid.org/0000-0002-0610-0470>

^d <https://orcid.org/0009-0005-1526-8338>

^e <https://orcid.org/0000-0002-8742-5786>

^f <https://orcid.org/0009-0004-9850-8239>

^g <https://orcid.org/0000-0002-5662-490X>

^h <https://orcid.org/0000-0002-3579-2832>

we differentiate “fire” and “non-fire” incidents, emphasizing the role of ML in predictive analysis for forest fire management. Figure 1 presents the proposed system model with its key contributions.

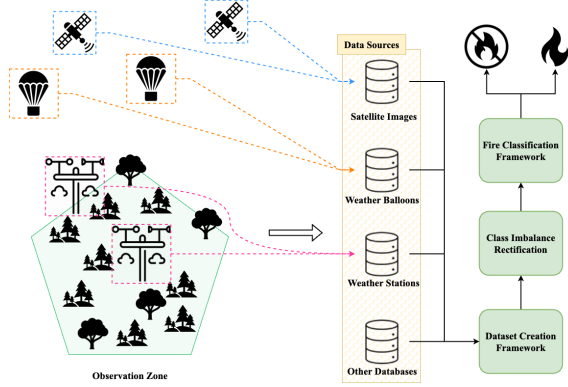


Figure 1: Proposed system model.

This research offers significant contributions to the field of forest fire predictive analysis in the following three primary areas.

C1 Dataset Creation Framework: To address the need for comprehensive forest fire datasets, we assembled a unique dataset comprising 27 variables from various sources. A pivotal source is the Copernicus re-analysis climate data (Hersbach et al., 2023), which integrates multiple data sources to offer a comprehensive record of historical climate conditions, some of which are presented in Figure 1. Other sources include the Canadian Wildland Fire Information System (CWFIS) Datamart (Service, 2022), Statistics Canada (Statistics Canada, 2021), and ArcGIS RESET (Saskatchewan Government, 2022). This multi-source dataset is instrumental in evaluating ML models and bridges knowledge gaps, promoting more nuanced research and practical applications.

C2 Class Imbalance Rectification: Given the 154:1 class imbalance between “non-fire” and “fire” events, we employed two re-sampling strategies. The ratio was calculated by dividing the number of “non-fire” events by the number of “fire” events.

- i. **Spatiotemporal Resampling:** A focus on spatial and seasonal relevance to fires.
- ii. **Technique-Driven Resampling:** Leveraging NearMiss Version 3 (NearMiss3), Synthetic Minority Over-sampling Technique (SMOTE), as well as its variant combined with Edited Nearest Neighbors (ENN) — termed (SMOTE-ENN) — for balanced event representation.

These strategies enhanced our dataset’s robustness against the prevailing class imbalance.

C3 Fire Classification Framework: We devised a specialized ML framework for the classification of forest fires. In this context, forest fire classification refers to the categorization of a given geographical region or 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:

T : Temperature	W : Wind
SW : Soil Water	P : Pressure
E : Evaporation	Pr : Precipitation
R : Runoff	V : Vegetation

Given the feature set \mathcal{F} defined above, our dataset comprises these eight primary categories. They collectively contribute to 22 specific environmental and climatic features. Additional columns in the dataset represent the day of the year, year of fire, latitude, longitude, and a coordinate ID, culminating in a total of 27 columns. In predictive modeling for complex events like forest fires, employing a model with a broad array of features enhances the precision and performance in distinguishing between “fire” and “non-fire” events by providing a comprehensive analysis.

This framework employs widely-used algorithms, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). It has been rigorously tested on unseen data to ascertain its real-world applicability and generalizability.

To conduct experiments and demonstrate contribution C1, a dataset was collected for the province of Saskatchewan, Canada, spanning the years 2000–2018. 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 4,714,983 raw data points, which will be explained further in Section 3. For contributions C2 and C3, a joint methodology is presented in Section 4.

In managing class imbalance in fire detection, the NearMiss3 method with a 0.09 sampling ratio proved pivotal. This led the RF model to achieve 78.3% accuracy, 74.8% sensitivity, and 78.3% specificity.

However, XGBoost, when combined with NearMiss3 at a 0.09 ratio, stood out with 98.08% accuracy, 86.06% sensitivity, and 93.03% specificity. LightGBM reported 72.38% accuracy, 76.03% sensitivity, and 72.36% specificity.

The paper is organized as follows: Section 2 discusses relevant literature on data methods, imbalance techniques, and ML models. Section 3 covers data collection. Section 4 describes the modeling approach. Results are presented in Section 5, and Section 6 concludes with future directions.

2 LITERATURE REVIEW

The domain of forest fire prediction has seen rapid advancements, with researchers adopting varied data collection methods, addressing dataset imbalances, and employing advanced machine learning techniques. This section provides a comprehensive overview of these developments, pinpointing areas of consensus and highlighting future research directions.

2.1 Data Collection Methods

Data collection is pivotal for the development of effective ML models in forest fire detection. Numerous techniques have been explored, ranging from satellite imagery, remote sensing, and Internet of Things (IoT) deployments to the use of weather stations, Unmanned Aerial Vehicles (UAVs), ground sensors, historical records, and crowd-sourced data. A comprehensive overview of these methods and the corresponding studies can be found in Table 1.

Rather than conducting primary data collection, several studies prefer to utilize existing datasets. Prominent examples are the UC Irvine ML repository dataset from Portugal (Cortez and Morais, 2007), (El-sarrar et al., 2019) and the SaskFire for time-series classification (Laube and Hamilton, 2021). Moreover, Kaggle datasets have played a crucial role in research, as seen in studies like (Preeti et al., 2021).

In our study, we utilize the Copernicus reanalysis data, valued for its comprehensive coverage, high temporal resolution, and absence of missing data, ensuring a robust foundation for accurate forest fire predictions with our ML models.

2.2 Imbalance Dataset

ML's application to forest fire prediction often grapples with the challenge of data imbalance. A considerable portion of existing research, such as (Bui et al., 2018), (Li et al., 2020), and (Hong et al., 2018),

employs an equal 1:1 distribution of fire and non-fire data points. This approach, while simplifying analysis, might not accurately represent the natural disparity observed in real-world fire incidents. Such discrepancies can compromise the applicability of derived predictive models in practical scenarios (Kaur et al., 2023).

Some studies acknowledge the imbalance issue but often settle for predefined ratios, like 3:1 or 10:1. Others defer addressing the problem, relegating it to sections on future work. Addressing this gap (Mutakabbir et al., 2023b), the Spatio-Temporal Agnostic Subsampling (STAS) framework has been introduced as an innovative approach to manage data imbalance in forest fire prediction (Mutakabbir et al., 2023a).

In our research, we strive to address this oversight by emphasizing the significance of Spatiotemporal re-sampling, focusing on the spatial and seasonal patterns of fire events. Additionally, our methodology incorporates advanced re-sampling techniques to ensure a balanced and accurate representation of events in the data.

2.3 ML Models in Forest Fire

Forest fire modeling has transitioned from traditional statistical methods to advanced ML techniques. Initial models focused on logistic regression and decision trees, evolving to neural networks and ensemble methods for more effective fire dynamics prediction (Safi and Bouroumi, 2013). Ensemble methods, particularly RF and Gradient Boosting, are noted for their robust predictive capabilities through model aggregation (Rodriguez-Galiano et al., 2012). Time-series analyses, especially with Long Short-Term Memory (LSTM), have proven effective in addressing the sequential nature of forest fire data (Natekar et al., 2021). A survey of literature from 2014 to 2022 identified 38 pertinent studies on IEEE Xplore, reflecting a trend from RF and SVM towards more complex algorithms like gradient boosting and LSTM (Purcell et al., 2023).

Despite deep learning's increasing application for its predictive accuracy as seen in (Mutakabbir et al., 2023a), our study opts for ML strategies that offer computational efficiency and enhanced interpretability in managing imbalanced datasets.

3 DATASET CREATION FRAMEWORK

This section introduces the processes undertaken to gather a dataset suitable for predicting forest fires us-

Table 1: Forest Fire Data Collection Methods.

References	Method	Method Description
(Preeti et al., 2021), (Ghate et al., 2023)	Satellite Imagery	Utilizing satellite sensors to capture images of forests and analyze them for fire detection.
(Ali et al., 2022), (Hidayanto et al., 2021), (Kosović et al., 2020)	Remote Sensing	Using remote sensing technologies, such as LiDAR or infrared sensors, to collect data on vegetation health, temperature, and other relevant factors.
(Suklabaidya and Das, 2023), (Zope et al., 2020), (Hidayanto et al., 2021)	IoT sensors	Deploying IoT sensors in forested areas to collect environmental data, such as temperature, humidity and air quality.
(Omar et al., 2021)	Weather Stations	Deploying weather stations in or near forested areas to collect real-time weather data, including temperature, humidity, wind speed, and precipitation.
(Sudhakar et al., 2020)	UAVs	Using drones equipped with cameras and sensors to capture high-resolution images and collect data in fire-prone areas.
(Sudhakar et al., 2020)	Ground-Based Sensors	Installing ground-based sensors, such as temperature and moisture sensors, to monitor forest conditions and detect anomalies.
(Singh et al., 2019), (Tayal et al., 2022)	Historical Fire Records	Analyzing historical fire records and incorporating them into the dataset for model training and validation.
(Sudhakar et al., 2020)	Crowd-Sourced Data	Gathering data from crowd-sourced platforms where volunteers contribute fire-related information, including fire incidents, burned areas, and fire severity assessments.

ing ML techniques. Data gathering plays a pivotal role as it marks the initial step towards predicting ignition points in forests. Several factors thought to influence forest fire were considered. These common factors include humidity, temperature, surface pressure, and precipitation. Although various papers have proposed datasets considering one or two of these factors, our work encompasses a broader array of features to enhance the relationship between these elements when building ML models.

Figure 2 illustrates the proposed dataset creation framework. Subsequent sections provide a comprehensive insight into its implementation and discuss the generation of the Saskatchewan dataset, tailored for our forest fire prediction research. This framework is versatile enough to accommodate data from various provinces or geographic regions. Within this dataset, “fire” points are represented by 1, and “non-fire” points by 0.

Four primary data sources underpin this framework, as shown in Figure 2: meteorological data (referred to as `cmet_src`), provincial boundary file (referred to as `bound_src`), historical fire data (referred to as `fire_src`), and water bodies file (referred to as `water_src`). Both the meteorological data and the historical fire data are expected to encompass coordinates and date information. Moreover, the historical fire dataset should offer details about the fire’s magnitude and origin.

The provincial boundary file and the water bodies file, both provided in shapefile format, play specific

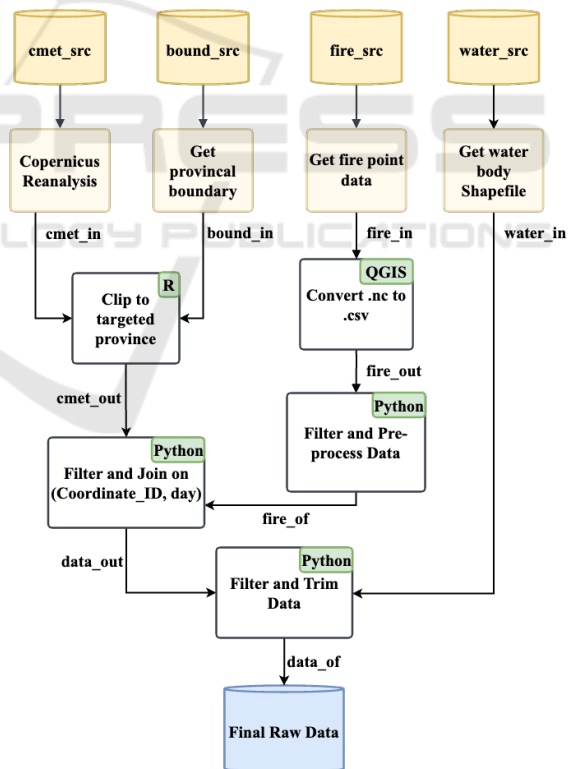


Figure 2: Dataset Creation Framework.

roles in the data processing:

- **Provincial Boundary File:** This file defines the limits of a predetermined area, effectively tailoring the dataset to the desired region. After speci-

fyng this region, it becomes possible to integrate pertinent fire-related information. The process of merging meteorological data with historical fire records involves several factors. One such factor is the scale of the affected area by the fire. The entries in the “fire” column are gauged based on the extent of the fire within a particular spatial range.

- **Water Bodies File:** This file is employed to refine the dataset by filtering out data points that lie within the region’s aquatic zones. Such zones encompass lakes, seas, oceans, substantial rivers, and large salt flats. By excluding these points, the dataset becomes more focused, eliminating regions where fires are less likely to occur.

In the following subsections, the integration of each source into the framework and its role in producing the final raw data is detailed.

3.1 Climate Reanalysis Data

We utilize the high-resolution European Environment Agency (ERA)5 dataset from the Copernicus Climate Change Service (Hersbach et al., 2023) for meteorological data retrieval. Produced by European Centre for Medium-Range Weather Forecasts (ECMWF), this dataset has been operational since 1940, offering a three-dimensional grid of climate variables at sub-daily intervals. Its comprehensive temporal and spatial resolution is particularly suited for analyzing complex interactions in climate patterns, pivotal in forest fire risk modeling. Referred to as `cmet_src` in Figure 2, ERA5’s granularity supports the precise detection of environmental phenomena essential for prediction models.

3.1.1 Spatial Resolution

The horizontal resolution of the fixed grid is $0.25^\circ \times 0.25^\circ$ on a regular lat-lon grid projection. To isolate data specific to the targeted province, the sub-region coordinates provided in Table 2 are taken into account.

Table 2: Sub-Region Coordinates.

	xmin	ymin	xmax	ymax
Saskatchewan	-109.99	48.99	-101.36	60.00

The retrieved files consist of 19 files in the NetCDF format (Network Common Data Form) with a “.nc” extension, referred to as `cmet_in`. NetCDF files are multidimensional scientific 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 (Purcell et al., 2023). The output from the “Clip to targeted province” box in Figure 2 is named `cmet_out`.

3.1.2 Temporal Resolution

ERA5 provides data with an hourly temporal resolution, spanning from January 1950 up to the present. For the purposes of this project, the focus is solely on the data collected at 12 noon. This decision is guided by the Canadian Fire Weather Index, which indicates that noon is a critical time for predicting wild-fire risk levels (Lawson and Armitage, 2008). The dataset covers the period from 2000 to 2018 and includes observations for every day of each month.

3.2 Historical Fire Data Point

The historical fire data has been sourced from the CWFIS Datamart (Service, 2022). 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 2, 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 CSV (Comma Separated Values) 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 2, 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 1. This is highlighted in the box titled “Filter and Join on (Coordinate_ID, day)” in Figure 2. 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.

Data: df1, df2, fire_df_copy, frame
Result: Updated df2 with "fire" column indicating fire proximity
Initialize df1 from fire_df_copy;
Initialize df2 from frame and add "fire" column set to 0;
foreach location in df2 **do**
 Extract location attributes: lat, lon, doy, year;
 Filter fires from df1 by day and year into fire_df;
 foreach fire in fire_df **do**
 Extract fire attributes: LATITUDE, LONGITUDE, SIZE_HA;
 Compute distance using *haversine* between fire and location;
 Determine if fire is within proximity using calculated distance and fire radius;
 if fire is inside or near location **then**
 Update "fire" attribute in df2 and exit loop;
 end
 end
end

Algorithm 1: Check for Fire Proximity.

Data: lon1, lat1, lon2, lat2 (in degrees)
Result: Distance between two points in kilometers
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 $c \times R$;

Algorithm 2: Calculate Distance Using Haversine Formula.

Algorithm 1 serves two primary functions:

1. It employs the haversine formula, as detailed in Algorithm 2, to calculate the distance between fire locations and meteorological data points based on

Data: lat1, lon1, lat2, lon2 (coordinates), resolution_degrees (bounding box resolution)

Result: Boolean indicating if the coordinate is inside the 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

Algorithm 3: Check if a Coordinate is Inside a Bounding Box.

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 3.

Algorithm 1 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.

3.3 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 2.

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.

3.4 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. In Figure 2, `data_of` represents the output file containing the finalized raw data.

3.5 Saskatchewan Dataset Summary

Data was collected annually from 2000-2018 from the ERA5 dataset, as shown in Table 3 for Saskatchewan. While centered on Saskatchewan, the framework, detailed in Section 4, is adaptable for other locations. The Final Raw Data, shown in Figure 2, is used as input for the modeling framework depicted in Figure 3.

Table 3: Summary of Saskatchewan forest fire dataset.

Dataset	Samples	Features	Classes
Sask Forest Fire	4,381,020	27	2

4 MODELING FRAMEWORK

In this section, we detail the modeling methodology employed in this paper. Class imbalance in datasets poses considerable challenges when striving for precise and robust ML models. The approach we adopt to tackle these challenges is illustrated in Figure 3. Within this framework, a fundamental decision centers around choosing the most suitable sampling technique. In this study, we thoroughly examine three techniques, each addressing different facets of imbalance correction: over-sampling, under-sampling, and a hybrid approach. Specifically, the NearMiss3, SMOTE, and SMOTE and ENN (SMOTE-ENN) techniques were assessed to determine their effectiveness in generating balanced datasets, as highlighted in contribution C2. Upon determining the optimal sampling technique using the base classifier, which is RF,

we pivot our attention to model selection and optimization. We employ three classifiers: RF, XGBoost, and LightGBM, aiming to achieve optimal classification, as discussed in contribution C3. Subsequent subsections provide a deeper understanding of each step, illustrating the sophisticated interplay between sampling and modeling within our proposed framework.

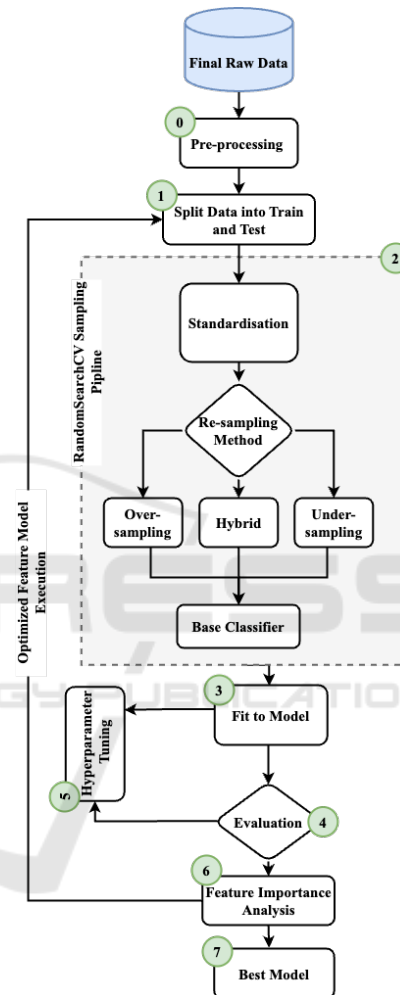


Figure 3: Modeling framework.

4.1 Modeling Overview

Figure 3 displays the modeling framework steps. The final raw data from the Dataset Creation Framework (see Figure 2) is first enhanced in Box 0, then split into training and test sets in Box 1. Data is standardized to ensure a consistent scale across features. Following this, and before addressing the class imbalance in Box 2, three re-sampling techniques - over-sampling, under-sampling, and hybrid - are evaluated.

Using RandomSearchCV within a pipeline architecture helps in determining the most effective pa-

rameters for the chosen re-sampling technique. The RF model serves as a benchmark to evaluate the efficiency of the re-sampling, leading to the selection of the most appropriate re-sampling method for model fitting, as depicted in Box 3.

The model's performance, evaluated using metrics such as, accuracy, sensitivity, specificity, and ROC-AUC, is detailed in Box 4. Hyperparameter tuning, presented in Box 5, refines algorithm accuracy. The final stage of the methodology, in Box 6, identifies important features. By rerunning the modeling with these key features, we assess if a simplified feature space retains or enhances predictive power.

The study applies three ML algorithms, the whole process is repeated three times to identify the optimal model, represented as the last step in Box 7 of the framework. The next subsections briefly describes Technique-Driven re-sampling methods and modeling algorithms.

4.2 Re-Sampling Techniques

Re-sampling techniques, namely, NearMiss3, SMOTE, and SMOTE-ENN, are instrumental for refining class distributions, thereby enhancing model performance, particularly in scenarios where instances of the minority class hold critical significance. An essential aspect of these techniques is the concept of the sampling strategy, interchangeably referred to as the "sampling ratio" in our context.

The sampling strategy, often defined as a floating-point value using the `sampling_strategy` parameter, signifies the intended proportion of samples from the minority class relative to the majority class post-resampling (Lemaître et al., 2017). Mathematically, denoting the sampling strategy as α and the counts of samples in the minority and majority classes as N_{minority} and N_{majority} respectively, this relationship can be expressed as:

$$\alpha = \frac{N_{\text{minority}}}{N_{\text{majority}}} \quad (1)$$

Next, we will offer concise descriptions of each Technique-Driven re-sampling method.

4.2.1 NearMiss3

NearMiss is an under-sampling strategy designed to reduce instances from the majority class. Of its various versions, NearMiss3 stands out. It selects majority class samples based on their distance to distant minority class samples. In essence, it retains majority class instances with the shortest average distance to a specified number of the most distant minority samples (Lemaître et al., 2017).

4.2.2 SMOTE

SMOTE is a prominent over-sampling technique that creates synthetic samples for the minority class. It operates by selecting two or more similar instances in the feature space and producing a new instance as a convex combination of the selected instances. Through interpolation, SMOTE expands the data representation for the minority class, aiming to balance the class distribution (Lemaître et al., 2017).

4.2.3 SMOTE-ENN

SMOTE-ENN integrates principles of both over-sampling and data cleaning. It begins with the SMOTE approach to over-sample the minority class. Subsequently, the ENN method is employed, removing majority class instances that are misclassified by their three nearest neighbors. This dual approach not only adds synthetic instances for balance but also refines the dataset by eliminating ambiguous or extraneous majority class instances (Lemaître et al., 2017).

4.3 Modeling Algorithms

Three state-of-the-art classifiers were employed to achieve optimal classification. A concise overview of these classifiers is as follows:

- **Random Forest (RF):** An ensemble method that uses multiple decision trees. Random subsets of features are chosen for node splits, making it robust and less prone to overfitting.
- **XGBoost:** An optimized gradient boosting algorithm known for its computational efficiency and versatility. It is suitable for large datasets and can capture complex non-linear relationships.
- **LightGBM:** A gradient boosting framework leveraging a histogram-based algorithm. It natively handles categorical attributes and is designed for fast computation, reduced memory use, and scalability with large datasets.

Each classifier provides a unique perspective on the data, enhancing our understanding and predictive capabilities.

5 RESULTS

In this section, we first delve into various re-sampling strategies, including Spatiotemporal, and Technique-Driven approaches. Subsequent analyses then focus on the performance of RF and gradient boosting algorithms, culminating in a discussion on the best model.

Table 4 presents the results from the Final Raw Data post-preprocessing, split into training and testing sets at ratios of 80% and 20%, respectively. All models consistently show high accuracy rates of approximately 99%. However, detailed analysis indicates low sensitivity values, ranging from 0.01 to 0.04, pointing to challenges in classifying the minority class, labeled as 1 or “fire”. On the other hand, specificity is at a consistent 1.00, reflecting the models’ ability to identify the majority class, labeled as 0 or “non-fire”. These metrics confirm a class imbalance in the dataset, emphasizing the necessity of re-sampling techniques for improved model generalization.

Table 4: Initial results before re-sampling.

Model	Accuracy	Sensitivity	Specificity	ROC-AUC
Random Forest	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

5.1 Re-Sampling

To tackle class imbalance in the fire detection dataset, specific strategies were employed. Throughout the process, two types of re-sampling were performed.

5.1.1 Spatiotemporal Re-Sampling

The Spatiotemporal Re-sampling method incorporates both spatial (geographical) and temporal (time-related) dimensions to adapt and streamline data. This ensures consistent data representation across different geographical areas and timeframes. This methodology is especially beneficial for analyzing dynamic geospatial patterns. Within our study, we utilized four distinct data refinement stages:

- i. **Study Area Restriction:** Initially, the data was clipped to match the boundaries of the specified study region, ensuring the exclusion of irrelevant geographical information. This was conducted prior to the inclusion of the “fire” attribute.
- ii. **Seasonal Filtering:** Days not within fire-prone seasons were removed to ensure data relevance.
- iii. **Historical Fire Incidence:** After adding the “fire” column, areas with no historical fire incidents were excluded.
- iv. **Exclusion of Water Bodies:** To elevate the data’s relevance and accuracy, water bodies were systematically eliminated from the dataset.

A visual representation of the data processing flow is available in Figure 2, with a detailed explanation

provided in Section 3. Table 5 presents the changes in the size and composition of the dataset at each stage, delineating both the “fire” and “non-fire” instances.

Table 5: Number of Records Removed in Spatiotemporal Resampling Stages.

Stage	Fire	Non-fire
Study Area Restriction	0	10,930,500
Seasonal Filtering	30,548	5,625,277
Historical Fire Incidence	30,548	4,684,435
Exclusion of Water Bodies	28,256	4,352,764

5.1.2 Technique-Driven Re-sampling

When employing different re-sampling techniques and using varying ratios, it was observed that RF was computationally more costly compared to XGBoost and LightGBM. Consequently, the experiment was structured in two separate runs. The first run solely utilized RF, 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 approximately 30 minutes, and NearMiss3 approximately 12 minutes. The results of these methodologies are elucidated in Subsection 5.3.

5.2 Performance Analysis of RF

To combat the imbalance, NearMiss3, an undersampling technique, was employed, being especially pertinent for classifying critical events such as fire. Table 6 showcases the NearMiss3 model’s empirical results, highlighting consistent performance across metrics.

Table 6: Performance results of RF with NearMiss3 re-sampling.

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

The exploration extended beyond the initial methods. Experiments were conducted using SMOTE for over-sampling and a hybrid method that combines SMOTE with ENN. While the over-sampling techniques have the capability to generate high-quality synthetic data, they did not outperform NearMiss3 in terms of specificity, sensitivity, or ROC-AUC. The associated computational overhead and reduced model interpretability presented further drawbacks.

The SMOTE-ENN technique yielded a ROC-

AUC of 0.9527, but exhibited challenges in sensitivity for the minority class. Refer to Table 7 for a summary.

Table 7: RF classification report for SMOTE-ENN method.

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

To sum up, while SMOTE-ENN showed impressive ROC-AUC scores, it did not provide optimal sensitivity for the minority class. This makes NearMiss3 a more favorable choice for our specific fire detection task, as it demonstrated stable performance across key metrics without the computational complexity associated with the other methods.

5.3 Performance Analysis of Gradient Boosting Algorithms

This section reviews the impact of sampling ratios on XGBoost and LightGBM, gradient boosters optimized for different data scales. Their performance with re-sampling techniques is detailed in upcoming figures.

Figure 4 presents the results of NearMiss3. While both models vary in recall as changes in under-sampling, they maintain notable ROC-AUC and F1 scores. XGBoost excels in recall at a 0.03 sampling ratio, while LightGBM sometimes classifies all instances as “fire”.

In Figure 5, SMOTE re-sampling yields consistently high specificity and balanced precision-recall for both models. XGBoost’s recall peaks at specific ratios, and LightGBM’s recall tops at a 1.000 ratio. Their ROC-AUC scores indicate strong class differentiation.

Figure 6 depicts SMOTE-ENN results. Both models exhibit varying recall with consistent ROC-AUC values. LightGBM’s recall spikes at aggressive sampling, slightly compromising specificity.

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

5.4 Best Model

Utilizing the NearMiss3 method with a 0.09 sampling ratio, the class imbalance in fire detection was addressed by adjusting the majority class (“non-fire”) in relation to the minority class (“fire”). The performance of three ML models, namely, RF, XGBoost, and LightGBM, is summarized in Table 8. Specifically, 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 8: 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

6 CONCLUSION AND FUTURE WORK

Forest fires intensified by climate change emphasize the inadequacies of conventional prediction methods. Addressing the prevalent class imbalance in fire data, our research offers a robust dataset, a tailored ML approach, and effective solution for class imbalance issue for Canadian forest fire classification. Harnessing the Copernicus reanalysis dataset, the framework integrates state-of-the-art algorithms such as RF, XGBoost, and LightGBM. This comprehensive approach not only improves predictive accuracy but also ensures a balanced representation of both fire and non-fire classes, enhancing the model’s reliability in real-world scenarios.

Through testing, NearMiss3 was the standout re-sampling method. Results recorded were: RF (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).

The findings suggest that while the NearMiss3 technique excels in optimizing sensitivity, there is a discernible trade-off with precision as the sampling ratio increases. The optimal range between 0.01 and 0.1 for the sampling ratio was found to strike a bal-

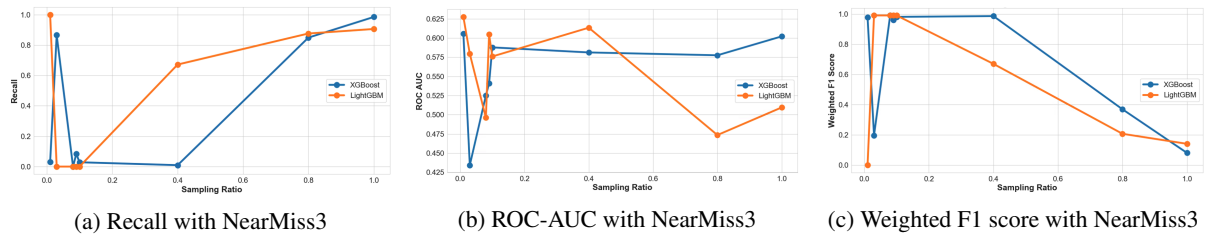


Figure 4: Performance results of XGBoost and LightGBM with NearMiss3 re-sampling.

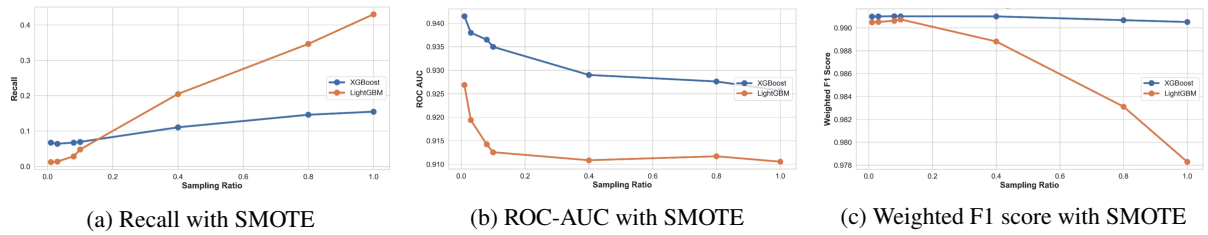


Figure 5: Performance results of XGBoost and LightGBM with SMOTE re-sampling.

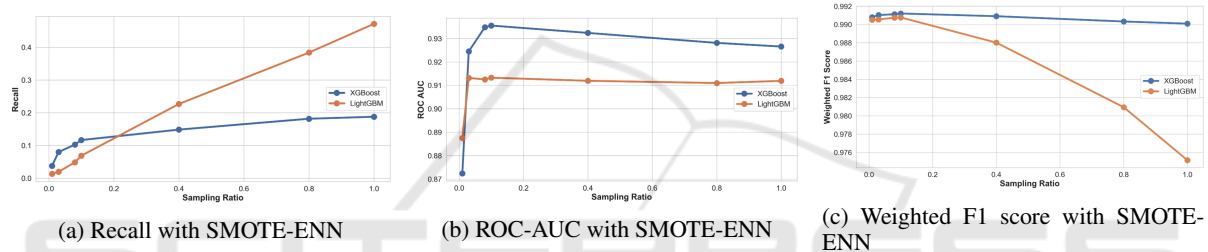


Figure 6: Performance results of XGBoost and LightGBM with SMOTE-ENN re-sampling.

ance by ensuring a diverse enough sample for robust modeling while avoiding over-sampling, which can lead to overfitting and reduce the model’s generalization capabilities.

Future research avenues include expanding datasets and exploring advanced algorithms such as Generative Adversarial Networks (GANs) to handle imbalanced data. Collaborations for real-time predictions can elevate this study into actionable forest management, addressing escalating challenges.

REFERENCES

Ali, S. D., Ridwan, I., Septiana, M., Fithria, A., Rezekiah, A. A., Rahmadi, A., Asyari, M., Rahman, H., and Syarifina, G. A. (2022). 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. IEEE.

Bui, D. T., Van Le, H., and Hoang, N.-D. (2018). Gis-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. *Ecological Informatics*, 48:104–116.

Canadian Interagency Forest Fire Centre (2023). Fire statis-

tics. Available at <https://ciffc.net/statistics>. Accessed: 2023-09-26.

Cortez, P. and Morais, A. d. J. R. (2007). A data mining approach to predict forest fires using meteorological data.

Elsarrar, O., Darrah, M., and Devine, R. (2019). 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. IEEE.

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

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2023). Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Accessed on 01-10-2022.

Hidayanto, N., Saputro, A. H., and Nuryanto, D. E. (2021). 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. IEEE.

- Hong, H., Tsangaratos, P., Ilija, I., Liu, J., Zhu, A.-X., and Xu, C. (2018). 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.
- Kaur, P., Naik, K., Purcell, R., Sampalli, S., Lung, C.-H., Zaman, M., and Mutakabbir, A. (2023). Data integration framework with multi-source big data for enhanced forest fire prediction. Manuscript under review.
- Kosović, B., Jimenez, P., McCandless, T., Petzke, B., Massie, S., Siems-Anderson, A., DeCastro, A., Muñoz-Esparza, D., and Haupt, S. E. (2020). 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. IEEE.
- Laube, R. and Hamilton, H. J. (2021). Wildfire occurrence prediction using time series classification: A comparative study. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4178–4182. IEEE.
- Lawson, B. D. and Armitage, O. (2008). Weather guide for the canadian forest fire danger rating system.
- Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.
- Li, Y., Feng, Z., Chen, S., Zhao, Z., and Wang, F. (2020). 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.
- Mutakabbir, A., Lung, C.-H., Ajila, S. A., Zaman, M., Naik, K., Purcell, R., and Sampalli, S. (2023a). Forest fire prediction using multi-source deep learning. In *EAI BDTA 2023 - 13th EAI International Conference on Big Data Technologies and Applications (BDTA)*.
- Mutakabbir, A., Lung, C.-H., Ajila, S. A., Zaman, M., Naik, K., Purcell, R., and Sampalli, S. (2023b). Spatio-temporal agnostic deep learning modeling of forest fire prediction using weather data. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 346–351. IEEE.
- Natekar, S., Patil, S., Nair, A., and Roychowdhury, S. (2021). Forest fire prediction using lstm. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE.
- Omar, N., Al-zebari, A., and Sengur, A. (2021). Deep learning approach to predict forest fires using meteorological measurements. In *2021 2nd international informatics and software engineering conference (IISEC)*, pages 1–4. IEEE.
- Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., and Sudi, A. (2021). Forest fire prediction using machine learning techniques. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE.
- Purcell, R., Naik, K., Sampalli, S., Lung, C.-H., Zaman, M., Mutakabbir, A., Kaur, P., and Tavakoli, F. (2023). A framework for creating forest fire ignition prediction datasets. Manuscript under review.
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, 67:93–104.
- Safi, Y. and Bouroumi, A. (2013). Prediction of forest fires using artificial neural networks. *Applied Mathematical Sciences*, 7(6):271–286.
- Saskatchewan Government (2022). Hydrography map service. Available at <https://gis.saskatchewan.ca/arcgis/rest/services/Hydrography/MapServer>. Accessed: 2022-09-26.
- Service, C. F. (2022). Canadian wildland fire information system (cwfis) datamart. Licensed under the Open Government Licence - Canada. Available at <http://open.canada.ca/en/open-government-licence-canada>.
- Singh, B., Kumar, N., and Tiwari, P. (2019). 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. IEEE.
- Statistics Canada (2021). 2021 standard geographical classification (sgc) - boundaries. Available at <https://www12.statcan.gc.ca/census-recensement/2021/geo/sip-pis/boundary-limités/index2021-eng.cfm?year=21>. Accessed: 2022-09-26.
- Sudhakar, S., Vijayakumar, V., Kumar, C. S., Priya, V., Ravi, L., and Subramaniaswamy, V. (2020). Unmanned aerial vehicle (uav) based forest fire detection and monitoring for reducing false alarms in forest fires. *Computer Communications*, 149:1–16.
- Suklabaidya, S. and Das, I. (2023). Processing iot sensor fire dataset using machine learning techniques. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 1–7. IEEE.
- Tayal, D. K., Agarwal, N., Jha, A., Abrol, V., et al. (2022). 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. IEEE.
- Tyukavina, A., Potapov, P., Hansen, M. C., Pickens, A. H., Stehman, S. V., Turubanova, S., Parker, D., Zalles, V., Lima, A., Kommareddy, I., et al. (2022). Global trends of forest loss due to fire from 2001 to 2019. *Frontiers in Remote Sensing*, 3:825190.
- Zope, V., Dadlani, T., Matai, A., Tembhurnikar, P., and Kalani, R. (2020). 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. IEEE.