Prediction of Forest Fire Risk for Artillery Military Training using Weighted Support Vector Machine for imbalanced data

Ji Hyun Nam
Department of Statistics and Data Science, Inha University, Incheon, Republic of Korea

Jongmin Mun *
Marshall School of Business, University of Southern California

Seongil Jo†
Department of Statistics and Data Science, Inha University, Incheon, Republic of Korea

Jaeoh Kim‡
Department of Statistics and Data Science, Inha University, Incheon, Republic of Korea

January 24, 2024

Abstract

Since the 1953 truce, the Republic of Korea Army (ROKA) has regularly conducted artillery training, posing a risk of wildfires—a threat to both the environment and the public perception of national defense. To assess this risk and aid decision-making within the ROKA, we built a predictive model of wildfires triggered by artillery training. To this end, we combined the ROKA dataset with meteorological database. Given the infrequent occurrence of wildfires (imbalance ratio $\approx$ 1:24 in our dataset) achieving balanced detection of wildfire occurrences and non-occurrences is challenging. Our approach combines a weighted support vector machine with a Gaussian-mixture based oversampling, effectively penalizing misclassification of the wildfires. Applied to our dataset, our method outperforms traditional algorithms (G-mean=0.864, sensitivity=0.956, specificity= 0.781), indicating balanced detection. This study not only helps reduce wildfires during artillery trainings but also provides a practical wildfire prediction method for similar climates worldwide.

Keywords: Forest fire, Imbalanced data, Risk prediction, Weighted support vector machine

*Co-first author.
†Co-corresponding author.
‡Corresponding author.

1 Introduction

Wildfires are a significant natural disaster that can have a profound impact on the ecosystem and economy, with the majority of such disasters being human-caused. Common causes include unattended campfires, burning debris, equipment malfunctions, inadvertently discarded cigarettes, and intentional acts of arson. In addition to these causes, in recent years, South Korea has experienced several wildfires that have been attributed to military artillery training. Conventional artillery firing training remains necessary due to the country's geopolitical location and the ongoing technical state of war between South and North Korea, which began a truce in 1953. To mitigate the risk of wildfires, the military training is meticulously planned in advance, taking into account the meteorological conditions on the day of training. Despite these precautions, multiple forest fires still occur annually during training sessions, causing substantial environmental damage. An illustrative instance of the damage caused by wildfires is the Goseong wildfire. In April 2019, one of the most intense forest fires in South Korea took place in the mountainous Goseong area of Gangwon province, situated close to the North Korean border. The fire quickly spread due to strong winds, high temperatures, and a lack of rain, ultimately burning over 1,200 hectares of forest and residential areas. The fire forced tens of thousands of residents to evacuate, resulting in several injuries and significant economic losses. The National Forestry Cooperative Federation estimates that the fire caused approximately KRW 78 billion in economic damages (approximately USD 70 million). The impact of the fire was not limited to the immediate region; the air quality in Seoul, located approximately 150 kilometers away from the fire, was severely affected, with PM2.5 levels reaching hazardous levels. Furthermore, the fire had a negative impact on the local ecosystem, destroying vast areas of forests and wiping out numerous plant and animal species.

In addition to economic and environmental impacts, wildfires caused by military training in South Korea contribute to negative public perception of the military's role in national defense. Even though the training sessions are planned ahead of time and take into account weather and other factors that could increase the risk of fires, the public believes that the military is to blame for the fires. This negative public perception of the military has led to increased scrutiny of its activities and calls for greater accountability for environmental damage resulting from military training. Additionally, there have been protests and demands for the military to assume a greater role in preventing forest fires during training exercises. Therefore, the military must actively

minimize forest fires and other environmental harm during training exercises, and convey its efforts to the public in a transparent and effective manner to regain the public's trust in its role as a national protector.

Numerous studies investigates the causes of forest fires from diverse perspectives. These viewpoints encompass environmental factors, sociological factors, and forest fire management systems and policies designed to reduce forest fires through quantitative research (Halofsky et al., 2020; Krueger et al., 2022; Shaw et al., 2017; Belloi et al., 2022). Our focus is specifically on understanding the meteorological factors associated with wildfires triggered by military artillery training exercises, a significant concern in South Korea. We combine recent meteorological big data with training session data of the Republic of Korea Army (ROKA), and leverage advanced computational capabilities to construct a data-driven predictive model based on classification algorithms in machine learning.

In our dataset, the number of military artillery training sessions involving forest fires is very small (approximately 4% of the total sessions). This skewed class distribution is a prevalent characteristic observed in datasets across various fields, such as geoscience (Ahmadlou et al., 2022) and artificial intelligence (Gao and Li, 2022). The presence of such class imbalance make it more challenging to construct a data-driven prediction model through conventional classification algorithms in machine learning, due to the following reason. These algorithms adjust the classifier function to maximize overall classification accuracy. Since small number of minority class samples have little impact on overall classification accuracy, minority class is poorly classified by conventional classification algorithms. This is problematic because benefits from proper prediction of minority class usually outweigh the costs of misclassification of majority class. One immediate example is diagnosis of rare cancers in diagnostic tests. Another example is military alert systems for detecting enemy infiltrations through various sensor systems such as visual, auditory, radar or other sensors. In order to address this imbalanced data problem, extensive research has been conducted both from the general algorithmic perspective (He and Garcia, 2009; Japkowicz and Stephen, 2002) and domain-specific perspective (Ahmadlou et al., 2022; Gasparin et al., 2022; Gao and Li, 2022; Ahmadlou et al., 2023; Zhang et al., 2023).

We view the problem of forest fire prediction during artillery fire training as classification of imbalanced data, where the minority class (or positive class) refers to the artillery training sessions

that triggered forest fires. From this perspective, we propose a two-step methodology that incorporates the two commonly used approaches for addressing the imbalanced classification problems: oversampling and cost-sensitive learning. In the first step, we alleviate the class imbalance by generating synthetic samples from the learned probability distribution of the minority class. In the second step, we utilize weighted support vector machine (SVM) that puts larger misclassification cost to the minority class in order to improve the prediction accuracy for the minority class. The contribution of our study is as follows.

1. This study can substantially reduce the likelihood of forest fires triggered by the ROKA's artillery military training, mitigating both forest fire catastrophes and potential adverse effects on the ROKA.

2. This study promotes the development of practical methods for predicting forest fires in climates comparable to the Korean Peninsula's moderate temperate climate zone, allowing for worldwide research applications.

3. Our proposed method can facilitate the study of classification models for various imbalanced datasets. Furthermore, the dataset utilized in this paper will be made publicly available (upon researchers' request), in order to facilitate the advancement of methodologies for addressing class imbalance, as well as promoting further research into environmental matters.

The study is structured as follows. We present previous studies for predicting forest fire risk and applying machine learning algorithms to imbalanced data in Section 2. Section 3 details the datasets used in this study, and Section 4 proposes a suitable analysis approach for this dataset. Section 5 presents the data analysis results obtained from applying our proposed method to our dataset. We provide a discussion of our findings and their implications in Section 6, and conclusion in Section 7.

2 Background

This section offers crucial background for predicting forest fires with machine learning methods. Section 2.1 gives a concise overview of traditional methods and recent machine-learning based

approaches for predicting forest fires. Given the rarity of forest fires, it is common for the relevent datasets to exhibit class imbalance. Section 2.2 provides a literature review on machine-learning based classification techniques specifically designed for imbalanced datasets.

2.1 Forest fire risk prediction methods

Methods for predicting forest fire risk are roughly categorized into two: traditional fire indices and modern machine-learning based methods. One illustrative example within the formal category is the Canadian Forest Fire Weather Index (FWI). Developed in the early 1970s by the Canadian Forest Service, the FWI is among the most widely used index for wildfire prediction and has been adopted by many other countries around the world. The FWI assesses the likelihood and severity of forest fires by analyzing various weather factors, including temperature, humidity, wind speed, and rainfall. To eleborate, the FWI consists of three primary components: the Fuel Moisture Code (FMC), the Fire Behavior Index (FBI), and the Fire Danger Rating (FDR). The FMC gauges the moisture content of fuels, indicating their ignition readiness and burn speed. It incorporates daily weather observations and adjusts for precipitation through a correction factor. The FBI measures the speed of fire spread and its intensity, taking into account wind speed, slope, and fuel type. Finally, the FDR summarizes the overall risk of forest fires by integrating the FMC and FBI into a single number. Besides FWI, there are many other fire risk index systems such as National Fire Danger Rating System (NFDRS), Forest Fire Danger Index (FFDI), and Grassland Fire Danger Index (GFDI).

Although the FWI and other index systems have proven to be effective prediction tools for over 50 years, there exist some drawbacks. First, the index number may not be directly comparable across different regions or countries. Second, the set of independent variables is fixed and thus may not fully capture the unique fire conditions found in different parts of the world. Finally, they have the limitations inherent in empirical-formula based approaches. Empirical formulas are relatively simple mathematical equations derived from relationships between variables observed in fixed historical data. It is not based on an universal physical mechanisms that govern fire behavior, and does not adjust to new dataset. Therefore, empirical-formula based approaches may fail to accommodate shifts in existing variables, such as climate change and new fuel conditions (Van Wagner, 1987; Stocks et al., 1989; Agriculture, 2015). In addition, empirical formulas

may not account for complex factors that are not included in the current model but influence the spread and intensity of forest fires—such as topography and human activity.

Consequently, several machine-learning based approaches for predicting forest fires have been suggested. Based on any given dataset, these methods approximate the underlying relationship between variables and fire risk in a data-driven way, which fits into the recent trend in forest fire prediction; Approximation, rather than precise derivation of mathematical model, is frequently used in forest fire prediction (Kloprogge et al., 2011), and data-driven approaches have proven to be extremely useful in forest fire predictions and vast environmental applications (Yu et al., 2022; Crowley et al., 2019; Ngoc Thach et al., 2018; Jafari Goldarag et al., 2016; Rodrigues and de la Riva, 2014; Al-Fugara et al., 2021). Machine-learning based methods are capable of learning from data of large volumes and various modalities, such as weather data, historical fire records, and satellite images. Moreover, the learned models can be continually updated as new data becomes available, allowing for more accurate and up-to-date fire risk assessments. There exist vast literature on predicting forest fire risk through standard machine learning techniques such as logistic regression and neural networks (Jafari Goldarag et al., 2016), random forest, boosting regression trees, and SVMs (Rodrigues and de la Riva, 2014), deep-learning based methods (Kim et al., 2016; Jiao et al., 2019; Xu et al., 2021). However, these studies do not take into account the extreme class imbalance in the forest fire records, which can significantly impede the model-building process.

2.2 Machine learning methods for imbalanced data

This section explains machine-learning based approaches for tackling the issue of classification on imbalanced datasets. We begin by summarizing previous approaches and algorithms in Section 2.2.1. Subsequently, we introduce performance metrics that can effectively demonstrate their superiority over conventional methods.

2.2.1 Approaches for imbalanced data

Classifcation on imbalanced data is a pervasive problem across various fields and tasks. Examples include protein classification (Zhao et al., 2008), disease prediction (Khalilia et al., 2011), odor impression prediction (Debnath and Nakamoto, 2022), urban gain modeling (Ahmadlou et al., 2022, 2023), electric load forecasting (Gasparin et al., 2022), and computer game AI modeling (Gao

and Li, 2022). The prevalence of imbalanced data in a variety of disciplines has piqued the interest of the machine learning community, which has developed the problem into a major area of machine learning research (see He and Garcia (2009); López et al. (2013); Haixiang et al. (2017); Krawczyk (2016) for comprehensive reviews). As stated in Section 1, conventional classification algorithms in machine learning result in biased classifier functions that poorly detect the minority class. Solutions addressing this bias typically involve data preprocessing, modification of learning algorithms, or hybrid combination of both. We now briefly introduce these three approaches.

Data preprocessing methods directly adjust the class imbalance by undersampling the majority class and (or) oversampling the minority class. Undersampling tends to be effective on moderately imbalanced datasets, while oversampling usually shows efficacy under high degree of class imbalance (Barandela et al., 2004). Since our forest fire dataset is highly imbalanced, our review focuses on oversampling methods. Oversampling may lead to longer training time and overfitting to the noise in the minority class samples (Chawla et al., 2004). In addition, the synthetic samples may contain physically implausible feature values, such as geographic coordinates corresponding to regions with no fire risk. Therefore, the goal of oversampling techniques is to generate synthetic samples that are distinct from the original samples but still accurately represent their fundamental properties. One widely used technique is Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), which linearly interpolates between original minority class samples. Its success has prompted numerous subsequent modifications. The examples include Boarderline SMOTE (Han et al., 2005) that generates synthetic minority class samples near the borderline between the two classes, and safe-level-SMOTE (Bunkhumpornpat et al., 2009) which creates synthetic minority class samples in regions where many original minority class samples exist. Undersampling methods, in contrast, do not pose a danger of overfitting or generation of implausible values. However, undersampling inevitably results in information loss. Therefore, undersampling techniques aim to identify the most representative samples of the majority class. For example, Near-Miss algorithm (Mani, 2003) removes majority class samples that are close to k-nearest neighbor minority class samples. KNN-Und (Beckmann et al., 2015), on the other hand, discards majority class samples that have many minority class samples within its k-nearest neighbor. One-sided sampling (Kubát and Matwin, 1997) removes majority class samples participating in Tomek links (Tomek, 1976), which refers to the pair exhibiting the minimum distance between

its elements among all feasible combinations of majority class samples and minority class samples.

Approaches involving modifications to learning algorithms are designed to enhance the algorithm's focus on the minority class. Usually, these approaches magnify the consequences of misclassifying minority class samples in the loss function, where the resulting model of the algorithm is defined as the minimizer of the loss function given the dataset. For neural network, Anand et al. (1993) suggest calculating the gradient vector in the weight-space to accelerate the back-propagation when the data is imbalanced. For SVM, Veropoulos et al. (1999) propose weighted SVM which puts higher misclassification penalty on the hinge loss function of SVM.

We finish our review by listing hybrid methods that combines data preprocessing and modifications of learning algorithms. Sun et al. (2007) incorporate data preprocessing and cost-sensitive learning for boosting algorithms. **?** combine SMOTE and weighted SVM, where the misclassification penalties are set to the reciprocal of the original imbalance ratio. Bang and Jhun (2014) implement undersampling by selectively preserving the centers from the clusters generated using the k-means, and the misclassification costs for the weighted SVM are set proportionally to the reciprocal of the cluster size. Fernández et al. (2009) incorporates SMOTE and hierarchical fuzzy rule based classification systems. In addition to these methods, there exists a line of works that combine ensemble methods of machine learning with undersampling or oversampling methods (Chawla et al., 2003; Mease et al., 2007; Liu et al., 2009; Gao et al., 2011; Ramentol et al., 2012).

It is imperative to recognize that, apart from the machine learning community, domain specialists across various scientific disciplines are actively engaged in conducting research on imbalanced data using their own tools and approaches. In the field of geoscience, Ahmadlou et al. (2022) proposes clustering and ensemble model to treat class imbalance problem in urban gain modeling. Also, advanced deep neural network based methods were proposed to solve class imbalance problem in various fields such as image denoising Zhang et al. (2023), time series forecasting Gasparin et al. (2022) and game playing strategy Gao and Li (2022). Ahmadlou et al. (2023) utilize maximum entropy to build a strategy for sampling and building training dataset when the dataset is imbalanced.

2.2.2 Performance metrics for imbalanced data

As stated in Section 1, under severe class imbalance, a machine-learning model still achieves high classification accuracy even if it classifies all samples as the majority class and ignore the minority class samples. As a solution we employee specialized evaluation metrics that consider the significance of both classes. There is no universally accepted performance metric, and this section reviews widely employed options among those available, focusing on binary classification. For more discussions about metrics, see Bekkar et al. (2013); He and Garcia (2009). For more comprehensive list of metrics, see Drummond and Holte (2006); Hand (2009); Prati et al. (2011); Walter (2005); Winkler (1969).

From now on, we follow the convention to refer to the majority class as negative and minority class as positive. Based on real labels and predicted labels, each new data point can be categorized into one of four groups: true positive, true negative, false positive, and false negative. We refer to the number of data points in each group as TP, TN, FP, and FN, respectively. These four values are bases of three fundamental performance metrics: specificity, precision, and sensitivity. Each metric captures the model's strengths and weaknesses in different contexts. Specificity is TN / (TN + FP) and indicates the model's ability to correctly identify negative cases. Precision is TP / (TP + FP) and reflects the model's ability to correctly identify positive cases. Sensitivity (or recall) is TP / (TP + FN) and also represents the model's ability to correctly identify positive cases. Compared to precision, sensitivity is more useful when the cost of false negatives is high.

From these basic metrics, we now define two metrics for highly imbalanced data: F_β-score and G-mean. The F_β-score (Baeza-Yates and Ribeiro-Neto, 1999) is the weighted harmonic mean of precision and recall. In a more formal way, it is defined as follows:

$$F_\beta := (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Typically, β is set to 0.5, 1, or 2. The F_1-score is an equilibrium between precision and recall and presents an overall estimate of the model's performance. Compared to F_1-score, the $F_{0.5}$-score gives more weight to precision than recall, making it more sensitive to false positives than false negatives. In contrast, the F_2-score gives more emphasis to recall than precision, which makes it more sensitive to false negatives than false positives. The G-mean is the geometric mean of

specificity and sensitivity. Since there exists trade-off between specificity and sensitivity, G-mean evaluates the model's performance across both majority and minority classes. A high G-mean score indicates that the model performs well on both minority and majority classes, while a low score suggests that the model may exhibit bias towards one of the classes. G-mean is a widely used metric for evaluating the effectiveness of classification models, particularly in imbalanced data (Kubát and Matwin, 1997). This paper employees G-mean as a performance metric.

3 Dataset

This section elucidates our dataset employed in constructing a predictive model for forest fire triggered by ROKA artillery training. Our dataset is a combination of two datasets with separate origins. The first dataset consists of meteorological data sourced from the South Korea Meteorological Administration (SKMA). The second contains information regarding artillery training, obtained from the ROKA. Table 1 and Table 2 list and explains the variables in datasets from SKMA and ROKA, respectively. From the ROKA dataset, we generated an indicator variable for the occurrence of wildfires and designated it as the response variable for our prediction model. For the predictor variables, we chose temperature, precipitation, wind speed, and relative humidity from the SKMA dataset. Two datasets are linked via temporal information variable. Section 3.1 explains the SKMA dataset in detail and Section 3.2 describes the ROKA dataset and the combined dataset.

We believe that the collected dataset can be used for a variety of future research and applications, including developing practical approaches for forest fire prediction in countries with a climate similar to that of the Korean Peninsula and developing classification algorithms for highly imbalanced data. The dataset and source code have been authorized for public release (upon request) by the Republic of Korea Military Security.

3.1 Meteorological database

Here we describe the SKMA dataset in detail. Utilizing weather observation data provided by the SKMA, this study collected source data from 612 observation stations located nationwide, spanning from January 1, 2011 to January 31, 2021. These stations have two categories: the

Table 1: Meteorological variables in the SKMA. Except for wind direction, which is difficult to measure consistently, these variables are used as predictor variables. Exploratory analysis of these variables are presented in Figure 1 and Figure 2.

Meteorological variable	Detailed Variable Description	Unit
Temperature	Instantaneous temperature measured at a height of 2m above the ground	Degree Celsius
Precipitation	Accumulative precipitation over last 60 minutes	mm
Wind speed	Average of 10 minute at a height of 10m above the ground	m/s
Wind direction	Average of 10 minute at a height of 10m above the ground	Degree
Relative humidity	-	%

Automated Synoptic Observing System (ASOS) and the Automatic Weather System (AWS), both of which provide meteorological data such as temperature, precipitation, wind, and humidity. The geographical location of each station is identified by its elevation, latitude, and longitude, which are later employed to normalize the meteorological variables. The temperature and humidity data are available at time intervals of 1 minute, 1 hour, and 1 day. The temperature is expressed as an instantaneous value at a height of 2 meters above the ground, measured in degrees Celsius. Wind velocity is averaged over a 10-minute interval, and the wind direction is indicated in degrees, with a value of 0 degrees corresponding to the north and 360 degrees to the north again. Precipitation data are calculated as cumulative precipitation over 1 hour. For humidity readings, relative humidity at the given time is employed. Table 1 provides a description of the meteorological variables obtained from SKMA.

Observation stations are sparsely dispersed across the nation, resulting in a lack of stations in close proximity to some artillery training sites. Also, the altitude of stations may differ from that of training sites, which leads to differences in temperature. Therefore, the weather conditions at these training sites may not be accurately reflected in the raw measurements within the SKMA dataset. To enhance precision of meteorological variables, this study utilizes interpolation and normalization. First, a uniformly gridified map of weather measurements is created using Barnes interpolation Barnes (1964) with a 1km resolution. From this gridified map, we pick the weather measurement that is in close proximity to the training site. Next, a high-resolution (30m) Digital Elevation Model (DEM) of the artillery training sites is acquired from the Shuttle Radar Topography Mission (SRTM). DEM is used to normalize the meteorological variables based on the

Table 2: Four sample forest fires reported during the artillery firing training, from the ROKA dataset. The ROKA dataset also records training sessions that did not result in forest fire, and our predictive modeling utilizes both of the fire and non-fire cases. As for the response variable, we create an indicator variable for the occurrence of wildfire in each session. We do not use the variables for location and shooting type as predictor variables. Instead, we link the SKMA dataset (explained in Section 3.1) to the ROKA dataset and use meterological variables as predictor variables.

No	Date	Location	Shooting type	Damaged Area
1	31.Oct.2019, 20:44	Goseong-gun, Gangwon-do	Red parachute flare	2,500
2	8. Nov.2019, 12:15	Goseong-gun, Gangwon-do	Star shell	3,000
3	13.Apr.2017, 20:10	Paju-si, Gyeonggi-do	60mm trench mortar	150,000
4	22.Mar.2018, 13:51	Paju-si, Gyeonggi-do	Panzerfaust3	3,300

elevations. Specifically, the temperature was adjusted from the altitude of the observation station to 0m using the lapse rate (0.65/100m) and then adjusted back to the altitude of training site in the high-resolution DEM.

3.2 Dataset for predictive modeling

Here we describe the ROKA dataset in detail and elucidate the combination of SKMA and ROKA dataset that is used in our analysis. Each observation in the ROKA dataset corresponds to one artillery training session. For each session, the following information is recorded: date, time, location, shooting type, and total damaged area. Note that the total damaged area is recorded only if a forest fire occurred. The ROKA dataset comprises 984 artillery training sessions of the past five years, with only 40 of them leading to forest fires. Out of the 40 wildfires resulting from these sessions, 39 took place during the February-April and October-November periods, with only one occurrence in June. Four samples of reported forest fire cases are presented in Table 2; It should be noted that the remaining 36 cases are not included due to space limitations, but are available upon request.

Now we explain in detail the combined dataset which is used for our analysis. To focus on meteorological conditions, we exclude location and shooting type in the ROKA dataset from the predictor variables set. As for the response variable, we ignore the damaged area variable in ROKA dataset since we view the predictive model as a classification task. Instead, we create an indicator variable for the occurrence of wildfire in each session (-1 = no wildfire, 1 = wildfire

occurred) and use it as the response variable. Future research can involve regression modeling of the damaged area. Note that non-occurrence is coded as -1 instead of 0, following the convention of SVM explained in Section 4. From the SKMA dataset, we exclude wind direction variable due to its challenging nature to be precisely gauged; SKMA also does not make effective use of this variable. We combine the ROKA dataset with the SKMA dataset by matching the date and time variable in ROKA dataset with the time variable in SKMA dataset. As a result, we obtain the dataset where each observation represents one training session and contains five variables: one response variable (the indicator variable from ROKA dataset) and four meteorological predictor variables (from SKMA dataset), namely precipitation, temperature, wind speed, and humidity. In summary, our dataset is a 984×5 matrix with the first column consisting of -1 (no wildfire occurrence, referred to as negative class samples) and 1 (wildfire occurrence, referred to as positive class samples), and the rest of the columns correspond to precipitation, temperature, wind speed, and humidity. We emphasize that this dataset is highly imbalanced, with 944 non-occurrence and 40 occurrence in the response variable. The imbalance ratio is approximately 1:24.

Now we provide exporatory data analysis results of our dataset. Figure 1 provides a graphical summary of each predictor variable. We observe differences in the distribution of negative and positive samples for each variable. Additionally, we observe distinct distribution ranges among meteorological variables, a phenomenon inherent to the nature of these variables. Specifically, temperature ranges from 20 to 40, precipitation varies from 0 to 80, wind speed ranges from 0 to 20, and humidity ranges from 0 to 100. Since SVMs are heavily sensitive to the range differences between predictor variables, we normalize the predictor variables to have a zero mean and a standard deviation of one. As some temperature values are negative, they are converted to absolute temperature prior to normalization. The normalization parameters (means and standard deviations) are learned from training dataset and subsequently utilized during the validation and testing phases.

Next, Figure 2 visualizes the overall structure of the predictor variables through a 3D scatter plot. In the plot, each axis corresponds to humidity, temperature, and wind speed. Also, the size of point denotes precipitation, and the color indicates the response variable (grey = non-occurrence, red = occurrence). It is worth noting that a small number of forest fire cases form several clusters, suggesting that the probability distribution of the minority samples can be

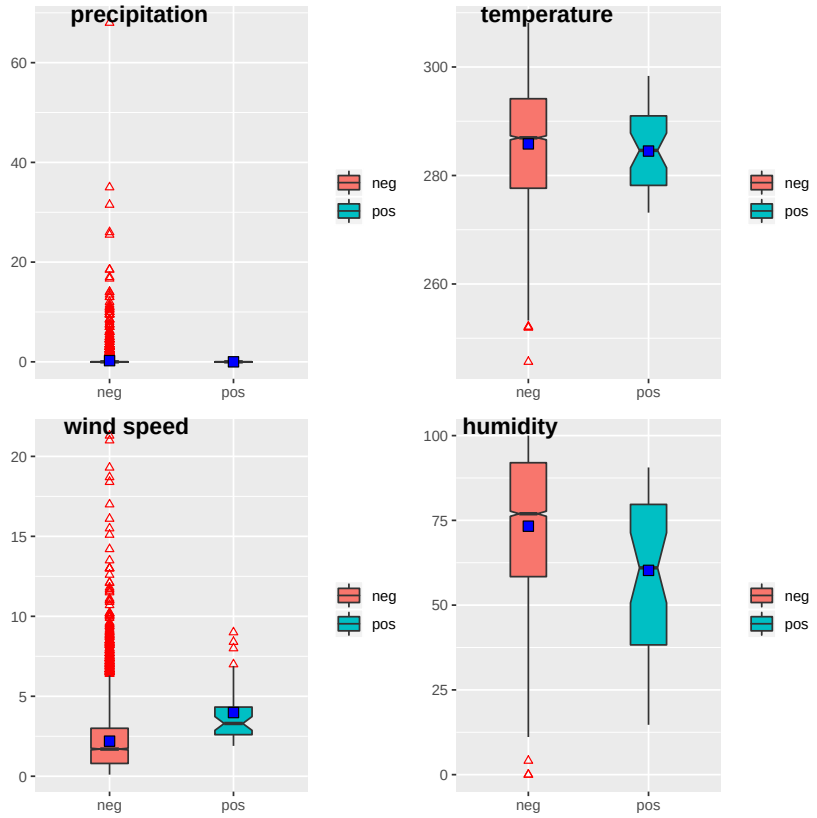

Figure 1: Distributions of meteorological predictor variables for wildfire occurrence (pos) and non-occurrence (neg) groups.

effectively approximated through mixture models.

To validate the clustered structure of the minority class samples, Figure 3 illustrates a 2D projection of the predictor variables. For the dimension reduction, we utilize a classical multidimensional scaling technique (Cox and Cox, 2000, also also known as principal coordinates analysis or Torgerson–Gower scaling). From the locations of the red dots representing the 40 wildfire occurrences, we still observe clustered structure of the minority class.

4 Proposed method

This section formally defines the problem setting and explains our proposed method, namely Gaussian mixture clustering weighted SVM (GC-WSVM). Let n_{maj} and n_{min} indicate the number of samples in the majority class and the minority class, respectively. Let $n := n_{\text{maj}} + n_{\text{min}}$ and R be the desired imbalance ratio after the oversampling. We observe samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in$

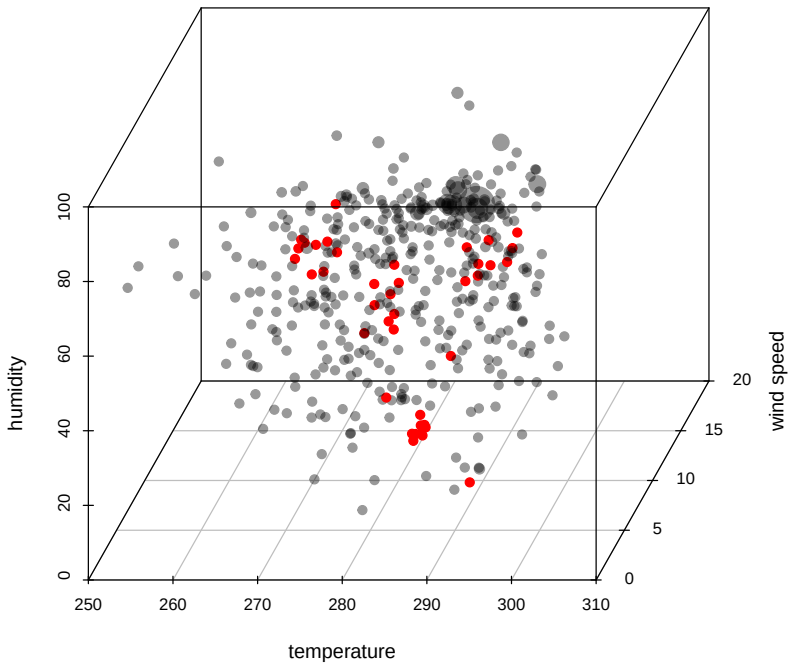

Figure 2: Scatterplot of meteorological predictor variables, where each point represents one observation. Each axis represents humidity, temperature, and wind speed. The size of the point denotes precipitation. There exist 40 wildfire occurrences (red points) and 944 non-occurrences (grey points).

$\mathbb{R}^4 \times \{-1, 1\}$. GC-WSVM consists of two steps. In the first step, we estimate the probability distribution of the minority class through Gaussian mixture model (GMM) and generate synthetic samples from the learned distribution. Let n^*_{min} denote the number of minority class samples after oversampling. Synthetic minority class samples are generated so that $n_{\text{maj}} : n^*_{\text{min}}$ equals $R : 1$. Let us denote $n^* = n^*_{\text{min}} + n_{\text{maj}}$. In the second step, we apply weighted Gaussian kernel SVM (Veropoulos et al., 1999) to the modified dataset containing n^* samples. In particular, we set the weight (misclassification cost) of the minority class sample and majority class sample as $1/n^*_{\text{min}}$ and $1/n_{\text{maj}}$, respectively.

It is worth noting that in general, some kinds of predictor variables (e.g. geographic coordinates) may not be suitable for oversampling approach because oversampling might produce samples with physically infeasible values. Therefore, prior to deciding between oversampling and

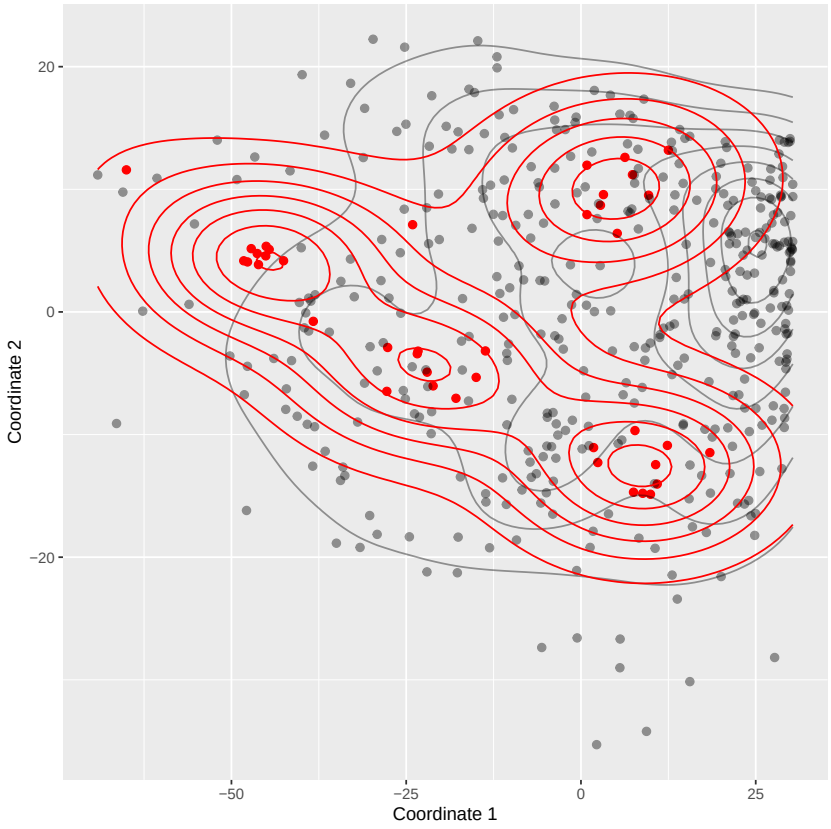

Figure 3: The dimension of data is reduced to two dimensions using a multidimensional scaling technique. Red dots are 40 events in which forest fires have occurred, and gray dots represent the absence of wildfire outbreaks.

undersampling approach, it is necessary to conduct a thorough exploratory data analysis. In the exploratory analysis in Section 3.2, our predictor variables —precipitation, temperature, wind speed, and humidity —do not exhibit the potential to produce synthetic samples with improbable values.

We examine the rationale and assumptions underlying the use of GMM for oversampling. In order to oversample the minority class, we assume that it follows a Gaussian mixture distribution. This assumption implies that data points are divided into several small groups, each of which follows a normal distribution. We estimate the parameters of the Gaussian mixture distribution and generate new samples from that mixture distribution. This assumption is often consistent with reality since rare events, despite their low occurrence rate, are often divided into several small subgroups (which is true for our dataset; see Figure 2 and Figure 3). One explanation for this clustering phenomenon is that when the number of predictors is small, different values of predictors

can produce the same result (in our case, wildfire). Different values of predictors correspond to different clusters. Even if the minority class does not consist of subgroups, most continuous distributions with sufficiently smooth densities can be accurately approximated with Gaussian mixture distributions with large enough number of components (Goodfellow et al., 2016). We do not employee the conventional SMOTE algorithm (Chawla et al., 2002), despite its simplicity and popularity, due to the following reasons. First, SMOTE only creates new samples from a straight line between existing samples, and thus it may not effectively broaden the area of the minority class. Second, if the minority class consists of subgroups, SMOTE may incorrectly generate new samples in the area between two subgroups.

Now we formally explain the GMM in our context. We have observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_{\min}} \in \mathbb{R}^4$ sampled from the probability distribution of minority class with density $f(\boldsymbol{x})$. The GMM assumes that $f(\boldsymbol{x})$ is represented as a linear combination of K Gaussian density functions as follows:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k),$$

where $\mathcal{N}(\cdot \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k)$ denotes a four-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma}_k$, and π_k are mixing coefficients satisfying the constraints that $0 \leq \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Note that the mixing coefficients indicate how much of each Gaussian density contributes to the total distribution. The parameters $\{\boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k, \pi_k\}$ of the GMM are estimated by maximizing the log-likelihood $\mathcal{L}(\pi_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_{\min}}) = \sum_{i=1}^{n_{\min}} \log\big(\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k)\big)$ through the expectation-maximization (EM) algorithm (Dempster et al., 1977). EM algorithm estimates the maximum likelihood estimator thorough iterative procedure.

Now we formally introduce the weighted SVM that we utilize in the second step of GC-WSVM. Weighted SVM adjusts the weight (misclassification cost) by setting the weight of minority class (wildfire occurrence) larger relative to that of the majority class (non-occurrence). In particular, we set the weight to be proportional to the inverse of the number of samples. This weighting scheme effectively increases the weight on the minority class. In a more formal way, for $i = 1, \ldots, (n^*_{\min} + n_{\text{maj}})$, the weight is $c_i = 1/n_{\text{maj}}$ if $Y_i = -1$ and $c_i = 1/n^*_{\min}$ if $Y_i = 1$. This misclassification cost setting aligns with the real-world scenario; A single wildfire in an unprepared setting incurs greater costs than numerous wildfire preparations that never materialize. With these

weights, weighted SVM solves the optimization problem with the following objective function:

$$\text{minimize}_{\boldsymbol{w},\boldsymbol{b},\xi_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n^*} c_i\xi_i, \tag{1}$$

under constraints $y_i(\langle \boldsymbol{w}, \phi(\boldsymbol{x_i})\rangle + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1, ..., n^*$. Here, $\phi(\cdot)$ is a non-linear feature mapping onto the infinite-dimensional space corresponding to the Gaussian kernel, and $\langle \cdot, \cdot \rangle$ represents the corresponding inner product. This mapping onto infinite-dimensional space allows SVM to capture the non-linear relation between predictors and response. Squared Euclidean norm $\|\boldsymbol{w}\|^2$ in the objective function represents the reciprocal of "margin", which is the distance between decision boundary and each class samples. By maximizing the margin, SVM achieves small generalization error for future data. Variable ξ_i is related to how generous we are to the misclassification of $\boldsymbol{x}_i$. Parameter C controls the relative weight between margin maximization (regularization) and correct classification (fit), and is usually determined through cross-validation. Problem (1) is solved through transformation to its Lagrangian dual form given by

$$\text{maximize}_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}),$$

under the constraints $\sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq c_i C$. Here, α_i's are the Lagrangian multipliers and $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j})\rangle$ is the kernel function that gives the inner product of the feature mapping of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. The solution α_i^* of the above problem gives the separating hyperplane $g(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i^* y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b$. The SVM classifies the input vector $\boldsymbol{x}$ as positive class (wildfire) if $g(\boldsymbol{x}) > 0$.

Regarding the selection of the modeling approach, our suggested methodology is an effort to mitigate the uncertainty arising from oversampling. Since weighted SVM increses the influence of each minority class sample, it effectively mitigates the need for excessive synthetic oversampling when the class imbalance is very severe. Between oversampling and undersampling, it is known that both approaches generally have a favorable impact on the classification performance on imbalance data (Shamsudin et al., 2020; Koziarski, 2021). However, we find the oversampling approach more suitable for our algorithm (SVM) and our dataset which demonstrates an extreme class

imbalance (imbalance ratio of 1:23.6). Combining undersampling and oversampling is another attractive option, and is known to yield outstanding performance when the class imbalance is severe. However, the implementation is virtually infeasible since the combination requires careful adjustment of the oversampling and undersampling ratios through substantial amount of grid search. There are specific scenarios in which the application of undersampling surpasses the effectiveness of oversampling—for example, when the number of predictors is comparable to the sample size (Blagus and Lusa, 2013). Our dataset, comprising four predictors and a sample size of around 1000, does not exhibit known circumstances that would exclude the application of oversampling techniques.

5 Result

This section reports on the results of applying GC-WSVM introduced in Section 4 to the artillery training dataset described in Section 3. The performance of GC-WSVM, measured in terms of G-mean, sensitivity and specificity, is compared to linear SVM and Gaussian kernel SVM. The experiments utilize R packages, including `mclust` for GMM, `mvtnorm` for generating multivariate Gaussian random samples, `e1071` for linear and Gaussian kernel SVM, `WeightSVM` for weighted SVM, and `caret` for 5-fold cross-validation. All experiments were conducted on a 2.8 GHz Intel Core i7 quad-core processor (4558U) with 16GB RAM.

We briefly introduce the baseline methods and discuss their strengths and weaknesses. Linear SVM (Cortes and Vapnik, 1995) corresponds to the solution of problem (1) with $c_i = 1$ for all $i = 1, \ldots, n^*$ and where ϕ is the the identity mapping. Linear SVM is fundamentally constrained by its ability to generate only linear classifier functions. However linear SVM has computational advantage: for sample size of n, linear SVM has $O(n^2)$ time complexity in its modern implementation, while kernel SVM has time complexity is $O(n^3)$. Since SVMs are extremely sensitive to hyperparameters and thus need to be fitted hundreds of times for hyperparameter tuning, linear SVM has computational advantage over kernel SVM when the the sample size is large. Gaussian kernel SVM (Cortes and Vapnik, 1995) corresponds to the solution of problem (1) with $c_i = 1$ for all $i = 1, \ldots, n^*$. The method gained its popularity due to its nice performance in nonlinear classification in various types of data. Gaussian kernel is considered almost de facto when using the kernel SVM, in both research and practical data analysis.

Table 3: Performance comparison for GC-WSVM, linear SVM and Kernel SVM.

metric	GC-WSVM	Linear SVM	Kernel SVM
G-mean	0.864 (0.011)	0.00 (0.00)	0.480 (0.124)
sensitivity	0.956 (0.021)	0.00 (0.00)	0.230 (0.146)
specificity	0.781 (0.014)	1.00 (0.00)	1.00 (0.00)

The numbers in parentheses are standard errors.

We employ 5-fold cross-validation to assess the performance metrics and tune the hyperparameters. This choice is driven by the scarcity of samples in the minority class, as excluding the entire test data during training could result in inadequate training of the SVM. In each fold, 80% of the data is used for training, and 20% was used for evaluating the performance metrics. Within the 80% training data, we again conduct 5-fold cross-validation for hyperparameter tuning, using G-mean as the performance metric. All of three methods share the hyperparemter C, which is explained in the formal introduction of weighted SVM (1). Kernel SVM and GC-WSVM have additional hyperparemter γ, which is the bandwidth of the Gaussian kernel and specifies the extent to which the influence of each sample can be reached. For both of C and γ, we perform grid search on values of $\{2^{-10}, 2^{-9}, \ldots, 2^{0}, \ldots, 2^{9}, 2^{10}\}$. For GC-WSVM, we treat the desired imbalance ratio R as a hyperparameter and perform greed search on values of $\{1, 2, 5, 10, 15, 20\}$. We repeat this two-fold cross-validation procedure one hundred times, resulting in one hundred different independent splits. This Monte Carlo simulation allows for more reliable assessment of performance. Additionally, it gives an estimate of the standard error of the algorithms, which is a proxy for their generalization ability to new data. We report the mean and standard deviation of the resulting G-means, specificities, and sensitivities in Table 3.

We now discuss the implication of the results in Table 3. GC-WSVM outperforms all competitors in G-mean and sensitivity. For specificity, GC-WSVM has lower value than Linear SVM and Kernel SVM. Nevertheless, the competitors' exceptionally high specificity comes at the cost of notably low sensitivity. In particular, Linear SVM completely fails to predict the minority class (0 sensitivity), which leads to 0 G-mean. Kernel SVM produces slightly better specificity than the linear SVM, but its overall performance falls short of GC-WSVM. GC-WSVM, on the other hand, keeps balance between sensitivity and specificity, which implies that it successfully predicts both of wildfire and non-wildfire cases. Failure of linear SVM implies that relationship between meteo-

rological conditions and occurrence of wildfire is non-linear. This phenomenon is expectable from visual representations in Figure 2 and Figure 3, where subgroups of the positive class are dispersed among the negative class samples, which would require non-linear decision boundary. This data structure in conjunction with the class imbalance lead to failure of linear SVM. Failure of Gaussian kernel SVM implies that nonlinear classifying ability granted by Gaussian kernel is not enough for highly imbalanced classification, and some other tricks such as oversampling is required for successful classification. It should be noted that despite the random nature of oversampling, the GC-WSVM has a lower standard error of sensitivity and G-mean than kernel SVM. Since there are only a few minority training samples in the highly imbalanced dataset, random splitting by 5-fold cross-validation greatly changes the distribution of the training dataset, resulting in high standard error. Oversampling from GMM recovers the distribution of the minority class, filling in the blank. Consequently, it results in a more stable performance when predicting the minority class.

6 Discussion

This study provides insight into machine-learning based approach for forest fire prediction, addressing the inherent challenges posed by imbalanced data. The proposed method, GC-WSVM employs a two-phased strategy involving oversampling of the minority class and weighted SVM. It demonstrates a substantial improvement in performance compared to traditional classification methods such as linear SVM and kernel SVM. One particularly noteworthy aspect is the stability of the GC-WSVM approach. Despite the inherent randomness in oversampling techniques, GC-WSVM exhibits a lower standard error in sensitivity and G-mean compared to kernel SVM. This enhanced stability can be attributed to GMM based oversampling. It effectively restores the distribution of the minority class, reducing the sensitivity to dataset fluctuations. This, in turn, leads to a more consistent and reliable performance when predicting the minority class.

It is important to acknowledge that oversampling may not always be the optimal strategy for preprocessing imbalanced datasets. Certain predictor variables, such as geographic coordinates, have the potential to generate synthetic samples that exhibit physically implausible values. In the dataset under consideration, the synthetic samples are devoid of implausible values as a result of the inherent physical characteristics of the predictor variables, namely precipitation, temperature,

wind speed, and humidity. However, it is imperative to perform a comprehensive evaluation of the attributes of predictor variables prior to executing our proposed oversampling technique on a customized dataset. We have also considered and explored undersampling techniques, specifically random undersampling and K-means center-based undersampling. However, these approaches fell short in performance, exhibiting a significant drop in sensitivity compared to the proposed GC-WSVM methodology. The reduction in sensitivity is particularly worrisome, as overlooking even one wildfire can lead to disastrous consequences.

The findings presented in this study avenues for future research in forest fire prediction and imbalanced data classification. Researchers can build upon the GC-WSVM approach and our dataset, exploring further enhancements and adaptations for different geographic regions and environmental factors. One further research direction is using the fire damage area variable, which is a continuous variable, to perform the regression modeling. Additionally, the findings have broader applications in related fields, such as natural disaster prediction and management. It can also offer a valuable solution for other fields grappling with imbalanced datasets, such as medical diagnosis, fraud detection, and anomaly detection.

7 Conclusion

This study proposes a machine-learning based approach to predict forest fire risk for artillery training in ROKA, recognizing the devastating effects of wildfires on the economy, human lives, and the environment. The task is viewed as a classification problem of imbalanced data, which is addressed through a two-phased method involving oversampling of the minority class from GMM-estimated distribution and the application of weighted SVM to the balanced dataset. Compared to classical classification methods like linear SVM and kernel SVM, the proposed method shows a significant improvement in performance.

The subjective decision-making processes of dozens of divisions and artillery brigades of the ROKA can lead to relatively high risks of forest fires and insufficient military training. This study can contribute to providing reasonable and objective guidelines on the conduct of artillery military training. Furthermore, the open-sourced dataset and the proposed method can contribute to the study of imbalanced data and the design of various classification models for forest fire prediction in countries with climates similar to the Korean Peninsula.

To sum up, in order to provide a more comprehensive understanding of the causes and consequences of forest fires caused by military training in South Korea, this study examines the factors contributing to these fires and proposes a two-step method for predicting and mitigating the risk of such fires. Additionally, this study aims to raise awareness about the negative impact of these fires on the environment and public perception of the military. By identifying the root causes of these fires and proposing a practical solution, this study can contribute to more effective forest fire prevention and mitigation strategies, leading to a safer and more sustainable environment in South Korea.

Code availability

The code that support the findings of this study are available from the corresponding author upon request.

Data availability

The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

Disclosures

The authors declare no competing interests.

Funding

Seongil Jo was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Korea government(MSIT) (RS-2023-00209229). Ji Hyun Nam was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00278691).

References

Agriculture, U. S. D. o. (2015), *FARSITE: Fire Area Simulator - Model Development and Evaluation*, CreateSpace Independent Publishing Platform.

Ahmadlou, M., Karimi, M., and Al-Ansari, N. (2023), "The use of maximum entropy and ecological niche factor analysis to decrease uncertainties in samples for urban gain models," *GIScience & Remote Sensing*, 60, 2222980, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15481603.2023.2222980.

Ahmadlou, M., Karimi, M., and Pontius Jr, R. G. (2022), "A new framework to deal with the class imbalance problem in urban gain modeling based on clustering and ensemble models," *Geocarto International*, 37, 5669–5692, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10106049.2021.1923826.

Al-Fugara, A., Mabdeh, A. N., Ahmadlou, M., Pourghasemi, H. R., Al-Adamat, R., Pradhan, B., and Al-Shabeeb, A. R. (2021), "Wildland Fire Susceptibility Mapping Using Support Vector Regression and Adaptive Neuro-Fuzzy Inference System-Based Whale Optimization Algorithm and Simulated Annealing," *ISPRS International Journal of Geo-Information*, 10, 382, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Anand, R., Mehrotra, K., Mohan, C., and Ranka, S. (1993), "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Transactions on Neural Networks*, 4, 962–969, conference Name: IEEE Transactions on Neural Networks.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Harlow: Addison Wesley, first edition ed.

Bang, S., and Jhun, M. (2014), "Weighted Support Vector Machine Using k-Means Clustering," *Communications in Statistics - Simulation and Computation*, 43, 2307–2324, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610918.2012.762388.

Barandela, R., Valdovinos, R. M., Sánchez, J. S., and Ferri, F. J. (2004), in *Structural, Syntactic, and Statistical Pattern Recognition*, eds. Fred, A., Caelli, T. M., Duin, R. P. W., Campilho,

A. C., and de Ridder, D., Berlin, Heidelberg: Springer, Lecture Notes in Computer Science, pp. 806–814.

Barnes, S. L. (1964), "A Technique for Maximizing Details in Numerical Weather Map Analysis," *Journal of Applied Meteorology and Climatology*, 3, 396–409, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Beckmann, M., Ebecken, N., and Lima, B. (2015), "A KNN Undersampling Approach for Data Balancing," *Journal of Intelligent Learning Systems and Applications*, 7, 104–116.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013), "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, 3, 27.

Belloi, A. P., Campesi, S., Nieddu, C., Tola, F., Deiana, S., Zizi, M., Muntoni, G., Tesei, G., Delitala, A., and Dessy, C. (2022), "Strategies and Measures for Wildfire Risk Mitigation in the Mediterranean Area: The MED-Star Project," *Environmental Sciences Proceedings*, 17, 124, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Blagus, R., and Lusa, L. (2013), "SMOTE for high-dimensional class-imbalanced data," *BMC bioinformatics*, 14, 106.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2009), in *Advances in Knowledge Discovery and Data Mining*, eds. Theeramunkong, T., Kijsirikul, B., Cercone, N., and Ho, T.-B., Berlin, Heidelberg: Springer, Lecture Notes in Computer Science, pp. 475–482.

Chawla, N., Lazarevic, A., Hall, L., and Bowyer, K. (2003), "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," vol. 2838, pp. 107–119.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, 321–357.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004), "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, 6, 1–6.

Cortes, C., and Vapnik, V. (1995), "Support-vector networks," *Machine Learning*, 20, 273–297.

Cox, T. F., and Cox, M. A. A. (2000), *Multidimensional Scaling*, CRC Press, 2nd ed., google-Books-ID: SKZzmEZqvqkC.

Crowley, G., Kwon, S., Ostrofsky, D. F., Clementi, E. A., Haider, S. H., Caraher, E. J., Lam, R., St-Jules, D. E., Liu, M., Prezant, D. J., and Nolan, A. (2019), "Assessing the Protective Metabolome Using Machine Learning in World Trade Center Particulate Exposed Firefighters at Risk for Lung Injury," *Scientific Reports*, 9, 11939, number: 1 Publisher: Nature Publishing Group.

Debnath, T., and Nakamoto, T. (2022), "Predicting individual perceptual scent impression from imbalanced dataset using mass spectrum of odorant molecules," *Scientific Reports*, 12, 3778, number: 1 Publisher: Nature Publishing Group.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Drummond, C., and Holte, R. C. (2006), "Cost curves: An improved method for visualizing classifier performance," *Machine Learning*, 65, 95–130.

Fernández, A., del Jesus, M. J., and Herrera, F. (2009), "Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets," *International Journal of Approximate Reasoning*, 50, 561–577.

Gao, M., Hong, X., Chen, S., and Harris, C. J. (2011), "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, 74, 3456–3466.

Gao, S., and Li, S. (2022), "Bloody Mahjong playing strategy based on the integration of deep learning and XGBoost," *CAAI Transactions on Intelligence Technology*, 7, 95–106, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12031.

Gasparin, A., Lukovic, S., and Alippi, C. (2022), "Deep learning for time series forecasting: The electric load case," *CAAI Transactions on Intelligence Technology*, 7, 1–25, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12060.

Goodfellow, I., Bengio, Y., and Courville, A. (2016), *Deep Learning*, Cambridge, Massachusetts: The MIT Press.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017), "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, 73, 220–239.

Halofsky, J. E., Peterson, D. L., and Harvey, B. J. (2020), "Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA," *Fire Ecology*, 16, 4.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005), in *Advances in Intelligent Computing*, eds. Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Huang, D.-S., Zhang, X.-P., and Huang, G.-B., Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 3644, pp. 878–887, series Title: Lecture Notes in Computer Science.

Hand, D. J. (2009), "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning*, 77, 103–123.

He, H., and Garcia, E. A. (2009), "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263–1284, conference Name: IEEE Transactions on Knowledge and Data Engineering.

Jafari Goldarag, Y., Mohammadzadeh, A., and Ardakani, A. S. (2016), "Fire Risk Assessment Using Neural Network and Logistic Regression," *Journal of the Indian Society of Remote Sensing*, 44, 885–894.

Japkowicz, N., and Stephen, S. (2002), "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, 6, 429–449.

Jiao, Z., Zhang, Y., Xin, J., Mu, L., Yi, Y., Liu, H., and Liu, D. (2019), "A Deep Learning Based Forest Fire Detection Approach Using UAV and YOLOv3," in *2019 1st International Conference on Industrial Artificial Intelligence (IAI)*, pp. 1–5.

Khalilia, M., Chakraborty, S., and Popescu, M. (2011), "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, 11, 51.

Kim, S., Lee, W., Park, Y.-s., Lee, H.-W., and Lee, Y.-T. (2016), "Forest fire monitoring system based on aerial image," in *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pp. 1–6.

Kloprogge, P., van der Sluijs, J. P., and Petersen, A. C. (2011), "A method for the analysis of assumptions in model-based environmental assessments," *Environmental Modelling & Software*, 26, 289–301.

Koziarski, M. (2021), "CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, iSSN: 2161-4407.

Krawczyk, B. (2016), "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, 5, 221–232.

Krueger, E. S., Levi, M. R., Achieng, K. O., Bolten, J. D., Carlson, J. D., Coops, N. C., Holden, Z. A., Magi, B. I., Rigden, A. J., and Ochsner, T. E. (2022), "Using soil moisture information to better understand and predict wildfire danger: a review of recent developments and outstanding questions," *International Journal of Wildland Fire*, 32, 111–132, publisher: CSIRO PUBLISHING.

Kubát, M., and Matwin, S. (1997), "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," .

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009), "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39, 539–550, conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013), "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, 250, 113–141.

Mani, I. (2003), "knn approach to unbalanced data distributions: A case study involving information extraction," in *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Mease, D., Wyner, A. J., and Buja, A. (2007), "Boosted Classification Trees and Class Probability/Quantile Estimation," *Journal of Machine Learning Research*, 8, 409–439.

Ngoc Thach, N., Bao-Toan Ngo, D., Xuan-Canh, P., Hong-Thi, N., Hang Thi, B., Nhat-Duc, H., and Dieu, T. B. (2018), "Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study," *Ecological Informatics*, 46, 74–85.

Prati, R. C., Batista, G. E. A. P. A., and Monard, M. C. (2011), "A Survey on Graphical Methods for Classification Predictive Performance Evaluation," *IEEE Transactions on Knowledge and Data Engineering*, 23, 1601–1618, conference Name: IEEE Transactions on Knowledge and Data Engineering.

Ramentol, E., Verbiest, N., Bello, R., Caballero, Y., Cornelis, C., and Herrera, F. (2012), "Smote-frst: a new resampling method using fuzzy rough set theory," in *Uncertainty Modeling in Knowledge Engineering and Decision Making*, WORLD SCIENTIFIC, vol. Volume 7 of *World Scientific Proceedings Series on Computer Engineering and Information Science*, pp. 800–805.

Rodrigues, M., and de la Riva, J. (2014), "An insight into machine-learning algorithms to model human-caused wildfire occurrence," *Environmental Modelling & Software*, 57, 192–201.

Shamsudin, H., Yusof, U. K., Jayalakshmi, A., and Akmal Khalid, M. N. (2020), "Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset," in *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pp. 803–808, iSSN: 1948-3457.

Shaw, J. D., Goeking, S. A., Menlove, J., and Werstak, Jr., C. E. (2017), "Assessment of Fire Effects Based on Forest Inventory and Analysis Data and a Long-Term Fire Mapping Data Set," *Journal of Forestry*, 115, 258–269.

Stocks, B. J., Lawson, B. D., Alexander, M. E., Wagner, C. E. V., McAlpine, R. S., Lynham, T. J.,

and Dubé, D. E. (1989), "The Canadian Forest Fire Danger Rating System: An Overview," *The Forestry Chronicle*, 65, 450–457, publisher: Canadian Institute of Forestry.

Sun, Y., Kamel, M. S., Wong, A. K. C., and Wang, Y. (2007), "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, 40, 3358–3378.

Tomek, I. (1976), "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, 769–772, conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

Van Wagner, C. E. (1987), "Development and structure of the Canadian Forest Fire Weather Index System," *Forestry Technical Report*, 35, 35, iSSN: None.

Veropoulos, K., Campbell, C., and Cristianini, N. (1999), "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, Stockholm, vol. 55, p. 60.

Walter, S. D. (2005), "The partial area under the summary ROC curve," *Statistics in Medicine*, 24, 2025–2040, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2103.

Winkler, R. L. (1969), "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association*, 64, 1073–1078, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Xu, R., Lin, H., Lu, K., Cao, L., and Liu, Y. (2021), "A Forest Fire Detection System Based on Ensemble Learning," *Forests*, 12, 217, number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Yu, Y., Mao, J., Wullschleger, S. D., Chen, A., Shi, X., Wang, Y., Hoffman, F. M., Zhang, Y., and Pierce, E. (2022), "Machine learning–based observation-constrained projections reveal elevated global socioeconomic risks from wildfire," *Nature Communications*, 13, 1250, number: 1 Publisher: Nature Publishing Group.

Zhang, Q., Xiao, J., Tian, C., Chun-Wei Lin, J., and Zhang, S. (2023), "A robust deformed convolutional neural network (CNN) for image denoising," *CAAI Transactions on Intelligence Technology*, 8, 331–342, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/cit2.12110.

Zhao, X.-M., Li, X., Chen, L., and Aihara, K. (2008), "Protein classification with imbalanced data," *Proteins: Structure, Function, and Bioinformatics*, 70, 1125–1132, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21870.