

Covid-19 detection using dominant SMOTE in imbalance classification

Saiful Islam

*Department of Computer Science and Engineering
International Islamic University Chittagong (IIUC), Bangladesh*

Article info

Keywords

Dominant SMOTE
COVID-19
Nearest neighbor
Imbalanced dataset
Active case.

Abstract

Global healthcare systems have faced difficulties since the start of the COVID-19 epidemic. For overburdened hospitals, identifying positive patients is a simple and effective fix. The disproportionate distribution of classes poses a significant challenge in identifying the positive case of COVID-19, leading to biased prediction outcomes favoring dominant classes. Consequently, classifiers struggle to learn from imbalanced datasets, resulting in reduced performance. Various techniques, such as oversampling, undersampling, and hybrid sampling, have been proposed to mitigate this issue. However, the Synthetic Minority Oversampling Technique (SMOTE) remains a commonly utilized resampling method despite its limitations, including class mixture. To address these shortcomings, I introduce Dominant SMOTE, a modified version of SMOTE. The proposed method comprises of developing a dominant sample selection approach based on numerical attribute values from the minority class, and selecting the nearest neighbors from the majority class for each minority class sample based on dominance values to achieve balanced dataset. The proposed method is compared with traditional SMOTE and Out-Layer SMOTE, evaluating accuracy, precision, recall, and F1-score on two benchmark datasets. The results indicate that the proposed model outperforms than both the traditional SMOTE and Out-Layer SMOTE.

Corresponding author

Saiful Islam

Department of Computer Science and Engineering, International Islamic University
Chittagong, Kumira, Chattogram, Bangladesh - 4318
Email: saifulcse@iiuc.ac.bd

1. Introduction

The novel coronavirus “severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2)” belongs to the Coronaviridae family in the Nidovirales Order. Its ability to survive on various surfaces, ranging from one hour to

up to nine days at room temperature, is a critical property that facilitates rapid community transmission (Van Doremalen et al., 2020). This virus, causing COVID-19, can trigger Acute Respiratory Distress Syndrome (ARDS) and dysfunction in various human organs, leading to physiological decline and mortality in infected individuals (Gibson, Qin, & Puah, 2020). The ongoing pandemic continues to strain global medical systems; resulting in surges in demand for hospital beds and critical shortages in medical supplies, all while many healthcare workers themselves have contracted the virus. Therefore, the ability to make immediate clinical decisions and utilize healthcare resources effectively is paramount. The most reliable diagnostic test for COVID-19, Reverse Transcription Polymerase Chain Reaction (RT-PCR), has long faced shortages in developing countries, contributing to increased infection rates and delays in implementing crucial preventive measures. Examples of biomedical use of machine learning techniques encompass diagnosis of diseases such as malaria and typhoid, classification of vascular diseases, risk assessment for diabetes, and analysis of genomic and genetic data (Huang et al., 2020). These machine learning techniques can be categorized in four main ways.

1. Supervised learning techniques are machine learning methods or algorithms that use labeled data to link the current and historical datasets in order to forecast future events (Mathkunti & Rangaswamy, 2020). With the right training procedure, this technology may provide outcomes in input data. It can then compare those results with expectations and actual results to find mistakes and adjust the model accordingly (Lavanya, Katyayani, Jancy, Latha, MaryA, Sujihelen, & Selvan, 2023).
2. Labeled and unlabeled datasets are utilized in the training process of semi-supervised learning techniques, which fall between supervised and unsupervised learning approaches (Liu et al., 2020). The learning strategies are better when a labeled dataset requires qualified and suitable resources for training or learning in it, and they can be modified to attain more accuracy (Genoud, Gao, Williams, & Thomas, 2020; Ahmad, Bakar, Yaakub, & Muhammad, 2020).
3. In order to identify faults, reinforcement learning strategies engage with the learning environment through actions (Liu et al., 2020). These methods are employed to pinpoint the optimal conduct within a particular setting in order to enhance the modal (Asadi, Dowling, Yan, & Mitchell, 2014; Schrider & Kern, 2018; Saravanan & Sujatha, 2018).

In December 2019, it was initially discovered in Wuhan, China to be COVID-19, a serious health issue that causes severe acute respiratory illnesses in people. The principal signs and symptoms of COVID-19 are

fever (98%), diarrhea (3%) and cough (76%) that are more dangerous for the adults than the young men with chronic illness (Huang et al., 2020) and many patients have been identified with the shortness of breaths that appears as symptoms of the flu (Gralinski & Menachery, 2020). The pandemic covid-19 that was first recognized in late 2019 is spreading out exponentially over the world by which about 209 nations has been affected (Muhammad & Sharif, 2020; Muhammad, Algehyne, & Usman, 2020; Muhammad & Sharif, 2020; Muhammad, Algehyne, & Usman, 2020). The World Health Organization (WHO) issued a Public Health Emergency of International Concern due to the COVID-19 virus spreading through direct contact with infected individuals through sneezing or other respiratory secretions. COVID-19 has neither effective treatment nor vaccine though various types of vaccines have been being invented and applied to the human body to boost up antibody and even though the scientists and scholars have been trying to invent both cure and prevention (Morens, Daszak, & Taubenberger, 2020). As a result, a number of people of the world have become jobless and business less that poses a big treat to the health and economic sector of the country.

It is a well-established fact that non-medical methods, including data mining and machine learning, are crucial to predicting the likelihood of contracting COVID-19. One of the most popular and cutting-edge areas of artificial intelligence is machine learning (ML), which offers a variety of systematic methods for multimodal, multidimensional, mathematical data analysis and prediction. Machine learning has already demonstrated superior performance and potential in the diagnosis, detection, prediction, and therapeutic monitoring of a wide array of diseases (Haruna, Muhammad, Yahaya, Garba, Oye, and Jung (2019); Hussain, Muhammad, Ishaq, Yakubu, and Mohammed (2018); Ishaq, Muhammad, Z, and Abubakar (2020)). Most of the world datasets contains disproportion of classes such as one class has maximum number of samples than the other class that leads to the bias result of the classification whereas classification is the root task in machine learning. Let's consider a dataset whereas one class that belongs to the minimum number of samples that is 5% of data samples and other class has 95% data samples that belongs to the maximum number of observations is called majority class. Many industries, including robotics, traffic management, engineering, and medical, use machine learning techniques (Huang, Zhu, & Siew, 2006).

The main contribution of this paper over other existing methods is described as follows:

1. I've proposed a newer version of a resampling method called Dominant SMOTE for imbalanced COVID-19 dataset to enhance the unbalanced

COVID-19 dataset's classification performance during the identification of COVID-19 active and inactive cases.

2. I've introduced an approach that selects the closest neighbors belonging to the dominant class for each and every sample of the minority class that is equal to the dominance value of that minority sample to make the dataset balance.

3. I've applied the proposed model on two benchmark datasets with skewed distributions to assess its effectiveness.

The following is a description of the remaining sections of this paper: The previous and relevant work on the imbalance dataset is briefly described in Section 2. The specifics of the Dominant-SMOTE algorithm and its operation are covered in Section 3. Section 4 states the experimental result and discussion that is compared with the SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), and Out-Layer SMOTE (Turlapati & Prusty, 2020). Section 5 describe conclusion of this paper.

2. Related works

The three main approaches for imbalance learning that are now available are data-level approaches, algorithm-level methods, and ensemble methods. The data-level strategies suggest preparing the original data in order to rebalance the class distribution. The algorithmic approaches, mostly grounded in cost-sensitive processes, aim to modify current classifiers to accommodate unbalanced datasets and introduce bias towards the minority class. Ensemble techniques combine the aforementioned strategies especially with ensemble classifiers with efficient ensemble learning solutions. In this work, I concentrate on data level solutions because they are more flexible and may be applied with any classifier. The most logical techniques for rebalancing the class distribution are random oversampling (ROS) and random under-sampling (ROU), which involve randomly duplicating minority class examples or randomly removing majority class examples (RUS). These two approaches address imbalance, but they have drawbacks: oversampling can quickly result in overfitting, while under-sampling might cause valuable data to be discarded, missing important details. Synthetic Minority over Sampling Technique is an improvement on random oversampling (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). During pre-processing, data-based approaches use various sampling strategies to rebalance the distribution of class. This involves under-sampling instances of the majority class or oversampling examples of the minority class. Oversampling is the process of duplicating instances from a minority class at random (Hlosta, Stríž, Kupčík, Zendulka, & Hruška, 2013; Krawczyk &

Woźniak, 2015; Zou, Xie, Lin, Wu, & Ju 2016). Under-sampling entails the removal of cases from majority classes at random. Borderline-SMOTE (Han, Wang, & Mao, 2005) and ADASYN (He, Bai, Garcia, & Li, 2008), SMOTE can be improved by assuming that samples near the borderline are more important. The concept of mixing under-sampling and oversampling to create a balanced training set with the appropriate number of positive and negative samples. A full discussion of ensemble approaches for the problem of class imbalance may be found in Over-Bagging (Wang & Yao, 2009) and (Barandela, Rosas, & Sánchez, 2003) Using bagging, combine random under-sampling and oversampling. The problem of data imbalance has gained prominence due to the increasing use of deep learning, particularly in the data mining industry (Maheshwari, Jain & Jadon, 2017), medical diagnosis (Song, Morency, & Davis, 2013), the detection of fraudulent calls (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2012), risk management (Zięba & Świątek, 2013; Birla, Kohli, & Dutta, 2016; Khalilia, Chakraborty, & Popescu, 2011), fault diagnosis (Li, Sun, & Zhu, 2010; Zhang, Li, Wang, & Wen, 2018) anomaly detection (Zhu & Song, 2010; hang, Jiang, Han, Wang, Wei-Guo, & Yang, 2017) and face recognition (Khreich, Granger, Miri, & Sabourin, 2010). In the investigation of class imbalance, traditional machine learning models have been widely used.

Naseriparsa, Al-Shammari, Sheng, Zhang, and Zhou (2020) introduced RSMOTE to enhance the performance of classification of the medical imbalance dataset; a modified version of RSMOTE called Region Based Simple Minority Oversampling Technique was presented. The RSMOTE is based on the density of minority class samples, and it divides them into four categories: normal areas, also known as safe regions, semi-normal regions, semi-critical regions, and critical regions. Then, to efficiently predict diseases, the base machine learning classification algorithms, such as Decision Trees (DT), Support Vector Machines (SVM), and Nave Bayes (NV), were applied. Because of this, the classification outcome for the minority class could be manipulated. In the safe region, the majority class samples are the fewest, while the minority class samples are the most abundant. The majority class has the most samples in the key zone, whereas the minority class has the fewest. As a result, the unbalanced samples of both classes affect all regions. Koziarski (2020) developed the Radial Based Oversampling method, which uses the mutual class potential and assigns a Gaussian Radial Basis Function (RBF) to each and every sample in the dataset. The polarity of the samples from the majority class and minority class is classified as positive and negative, respectively. To determine the position of the region that minimizes the absolute value of potentiality, an iterative optimization is used. Finally, to

eliminate the bias of the majority class, new synthetic samples are created from these kinds of places. (Han, Wang, & Mao, 2005) suggested Borderline SMOTE, in which only the dataset's minority observations that are at a safe distance from the majority sample's border line are over sampled. When the samples from the minority are kept a safe distance from the samples from the majority, oversampling is avoided, making the classifier's job easier in extracting the key features. When the results of the experiment are compared to SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), ADASYN performs better in terms of True Positive (TP) rate and F-score. Because just a small number of samples are oversampled, the same samples may appear multiple times, causing an over-occurrence problem and obviating the need to sample other significant samples.

Turlapati and Prusty (2020) suggested Outlier SMOTE, a strategy for improving covid-19 prediction, was suggested. Each data point in the outlier SMOTE is oversampled based on its distance from the other data samples, with data samples with a greater distance from the other samples being given more relevance and being oversampled than data samples with a smaller distance. The distance between each data sample and the other data samples is determined using the Euclidean Distance technique. It attempts to limit the likelihood of minority data samples overlapping, which is a problem with the classic SMOTE method. During the over sample, all of the samples are given equal priority according to the oversampling rate in regular SMOTE, however Outlier-SMOTE prioritizes the samples with the highest oversampling rate. This method may cause the dataset to contain several instances of the same sample while ignoring the other samples. As a result, the classification result may be skewed toward the most often seen sample, affecting the total classification result. Feng, Zhao, and Fu (2020) proposed a data level solution; it addressed a cluster-based hybrid method. SMOTE is employed in the creation of artificial minority class samples, while samples of the majority class are selected from the different clusters. The minority class subset is a cluster in and of itself, whereas the entire dataset is divided into clusters. When there are the most majority class samples in a cluster, the majority class samples contribute more and more representatively. On the other side, minority class samples contribute more than majority class samples in one cluster, which has fewer majority class samples than minority class samples. The majority class instances are chosen from each cluster via under sampling. Finally, the balance dataset is created by combining the minority and majority class datasets. Many critical samples from each cluster can be dropped due to under-sampling,

resulting in a minority categorization result. Yen and Lee (2009) introduced cluster-based under-sampling, which involves distributing all dataset samples into K clusters, was examined. The number of selected majority class samples is determined using the normalized rule implementation on all clusters, and the ratio of majority class and minority class samples is calculated. Ultimately, all minority class samples are combined with the calculated number of majority class samples from each cluster to form a balancing dataset. Vuttipittayamongkol and Elyan (2020) offered a neighborhood-based under-sampling strategy to address the imbalance dataset's class bias. The potential overlapped data points are removed by precisely detecting and removing the majority class instances from the overlapped zone, which optimizes the visibility of minority class instances while reducing formation loss. Many important samples may be omitted from the dataset as a result of under-sampling, which may have an impact on the prediction system's classification result. According to Lin and Nguyen (2020), the Boosting of Minority Class Prediction System is addressed, whereas an early deep network handles the segmentation of point cloud data from the imbalance dataset. In order to resolve between-class data point imbalances, the technique uses a hybrid sampling strategy that instructs the deep network to give more weight to data samples from the minority class.

Based on the comprehensive review of various approaches to address class imbalance in datasets, it becomes apparent that the emphasis on oversampling minority data has not been a focal point in existing literature. This observation highlights a gap and a potential novel direction in imbalanced classification methodologies. While many studies discuss different techniques such as SMOTE, Borderline SMOTE, RSMOTE, and Outlier SMOTE, the specific importance of oversampling the minority class is not extensively addressed. This gap indicates an opportunity for further research and innovation in the field of imbalanced classification, especially concerning the development and refinement of oversampling techniques tailored to minority class representation.

3. Method

Every data sample from the minority class is equally oversampled in the classic SMOTE. In the practical, all samples can't be over sampled equally due to less dominance in the feature space. The author strongly believes that the most dominant sample has to be over sampled more than the less dominant samples. On this idea, Dominant SMOTE operates. The Dominant SMOTE's workflow is depicted in Figure 1.

3.1 Data preprocessing

The dataset contains a few unnecessary attributes, such as ID numbers, which are not relevant to the users' data. Hence, all kinds of unnecessary attributes are removed from the dataset. To preprocess data for feature importance in machine learning, begin by addressing the missing values in the dataset and loading it, if any, by filling them with the mean or median. Next, encode categorical variables using techniques like one-hot encoding. Split the data into features (X) and the target variable (y). Then, use a model like Random Forest, creating a Random Forest Regressor instance, fitting it to the data, and obtaining feature importance. To determine which features have the most influence, sort this importance. For example, to select the top 5 features, create a new dataset with only these features. Remember to scale features if needed and validate the selected features' performance with cross-validation. This approach helps identify and prioritize features that have the most impact on the target variable, aiding in dimensionality reduction and improving model efficiency. As a result, the most important features are selected and used in the model training and testing.

3.2 Dominant SMOTE method

There are three crucial aspects to the proposed dominant SMOTE approach. First, the most important features from all of the dataset's attributes are selected using an important feature selection approach. Second, out of all the dataset's features, the consecutive dominant features are chosen. Third, the nearest majority samples from the majority class are chosen based on the dominance rate value. After cleaning the dataset, majority and minority classes are separated into two groups.

Table I
Showing the operation of the Dominant SMOTE algorithm.

N	1	2	3	4	5	6	7	8	9	10
Sum List	30	15	18	12	9	14	10	12	15	16
Normalized sum	3.97	1.98	2.38	1.58	1.19	1.85	1.32	1.58	1.98	2.11
Round of Sum	4	2	2	1	1	2	1	1	2	2
Oversampling Rate	4	2	2	1	1	2	1	1	2	2
Nearest samples of majority class	4	2	2	1	1	2	1	1	2	2

Then the minority class dataset is feed to the dominant algorithm. Let's say, minority class dataset is consisting of N samples and F features. Now, the sum of all features of each sample of the minority dataset is calculated and stored in the list.

$Sum_{I=0} = \sum_{J=0}^M F$, where M =Total number of the features in the dataset and $J \in [0, M]$, F = Feature.

$Sum_{I=1} = \sum_{J=0}^M F, \dots, Sum_{I=N} = \sum_{J=0}^M F$, where N =Total quantity of minority class samples and $I \in [0, N]$.

$SumList = [Sum_0, Sum_1, \dots, Sum_N]$.

$$NormalizationRule = \frac{Sum}{\sum_{i=1}^N Sum} \quad (1)$$

Equation 1 describes the normalization rule that is used in the proposed model. From the normalized sum list, the rounded sum list is formed by the use of ceiling function. The oversampling rate and the number of nearest samples from the majority class are calculated for each value in the rounded sum list. For every minority class sample in the existing SMOTE, $N = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ have been given equal priority to make over sampling 5 times when the over sampling rate is 500%. As a result, a less dominant sample whose sum of the all features is less than the other samples whose dominance is greater gets equal priority to be over sampled. To overcome the pitfall of the synthetic oversampling, the dominant SMOTE gives less priority to the less dominant samples and more priority to the greater dominant samples. Let's examine the oversampling of both conventional SMOTE and dominant SMOTE using ten samples from the minority class dataset. In the Dominant SMOTE, the dominance value of 10 samples is 4, 2, 2, 1, 1, 2, 1, 1, 2 and 2. The highest dominant sample whose dominance value is 4 will be over sampled maximum four times whereas second and third highest dominant samples whose dominance value are 2 and 1 will be over sampled twice and once respectively. After oversampling these 10 samples according to Dominant SMOTE, the total samples will be as follows, $4+2+2+1+1+2+1+1+2=20$. In the existing SMOTE, these 10 samples will be sampled according to equal priority of each sample as $2+2+2+2+2+2+2+2+2=20$.

As a result, over sampled minority dataset may be biased due to equal priority of lower dominance samples as the greater dominant samples of the dataset. So, the Dominant SMOTE priorities each sample according to the magnitude of features and generates the synthetic over sampled dataset. Most dominant sample has more nearest samples in the synthetic over sampled dataset according to the proposed method. So, equal numbers of nearest samples of the dominant samples are chosen from the majority dataset using K-Nearest neighbor machine learning algorithm. Let's consider the above Table I, the first sample has most dominance value of 4 and it is over sampled four times. So, 4 nearest neighbors of the first samples are selected using K-Nearest neighbor from the majority

class dataset. The most dominant samples have the most impact on the dataset. Thus, data samples for the majority class are selected according to the dominance value of each sample of the synthetic over sampled dataset to make a complete balanced dataset. As a result, less important data samples can be avoided in the classification problem. The Euclidean Distance is the most widely used popular default system that is used in SK-learn library of Python for the K-Nearest neighbor. It can be represented by the value of Minkowski Distance Metric.

$DistanceList_N[M,1] = [Dist_i, Dist_1, Dist_M]$.

$$Dist(a, b) = \sqrt{\sum_0^F (a_i - b_i)^2} \quad (2)$$

In the equation 2, $i \in [M,1]$, where N represents all of the samples from the minority class and M represents all of the samples from the majority class, the DistanceList is a list that contains all of the distances between each minority and majority class samples. The DistanceList is used to calculate and store the distance between a minority class sample and a majority class sample. The matching numbers of majority class samples that are equal to the dominance value are selected according to the sequential minimum value of the distance matrix. Based on their dominance value, this process is being repeated for all minority class samples. Figure 1 shows the flowchart of proposed Dominant SMOTE model. If the total count of samples from minority classes is $N=10$ and the total number of samples from the majority class is $M=50$, then the separation between every sample from the minority class and every sample from the majority class can be represented as follows:

$DistanceList_1[N_1M_1, N_1M_2, N_1M_3, N_1M_4, N_1M_5, N_1M_6, ..., N_1M_{50}]$
 $= [2.5, 3.5, 1.2, 4.6, 3.1, ... 12];$

$DistanceList_2[N_2M_4, N_2M_5, N_2M_7, N_2M_8, N_2M_9, N_2M_{10}, ..., N_2M_{50}] =$
 $[3.6, 2.4, 1.5, 1.6, 3.2, 5, ..., 1.5];$

$DistanceList_3[N_3M_4, N_3M_7, N_3M_8, N_3M_9, N_3M_{10}, N_3M_{11}, ..., N_3M_{49}] =$
 $[3.8, 6.8, 7.5, 11, 5.6, ... 13];$

$DistanceList_4[N_4M_7, N_4M_8, N_4M_9, N_4M_{10}, N_4M_{12}, N_4M_{13}, ..., N_4M_{49}] =$
 $[5.8, 3.9, 4.9, 7.5, 8.6, 4.6, 3.1, 11.2];$

$DistanceList_5[N_5M_7, N_5M_9, N_5M_{10}, N_5M_{12}, N_5M_{14}, N_5M_{15}, ..., N_5M_{49}] =$
 $[5.4, 13, 11, 8.9, 6.9, 4.6];$

$DistanceList_6[N_6M_7, N_6M_9, N_6M_{10}, N_6M_{12}, N_6M_{14}, N_6M_{15}, ..., N_6M_{48}] =$
 $[7.6, 4.8, 5.9, 6.3, 14, 16, 6.7];$

$DistanceList_7[N_7M_7, N_7M_{10}, N_7M_{12}, N_7M_{14}, N_7M_{15}, N_7M_{16}, ..., N_7M_{48}] =$
 $[9.5, 8.6, 9.5, 6.5, 8.4, 11.5, 12];$

$DistanceList_8[N_8M_7, \mathbf{N_8M_{10}}, N_8M_{12}, N_8M_{15}, N_8M_{16}, N_8M_{17}, \dots, N_8M_{48}] = [5.4, \mathbf{3}, 7.8, 8.7, 8.4, 5., 8.6];$
 $DistanceList_9[N_9M_7, N_9M_{12}, N_9M_{15}, N_9M_{16}, \mathbf{N_9M_{17}}, N_9M_{18}, \dots, N_9M_{48}] = [8.9, 13.2, 9.8, 6, \mathbf{4.8}, 20., 12];$
 $DistanceList_{10}[N_{10}M_7, N_{10}M_{12}, N_{10}M_{15}, \mathbf{N_{10}M_{16}}, N_{10}M_{18}, N_{10}M_{19}, \dots, N_{10}M_{48}] = [5.7, 4.9, 8.5, \mathbf{4.8}, 13.5, 8., 15.6];$

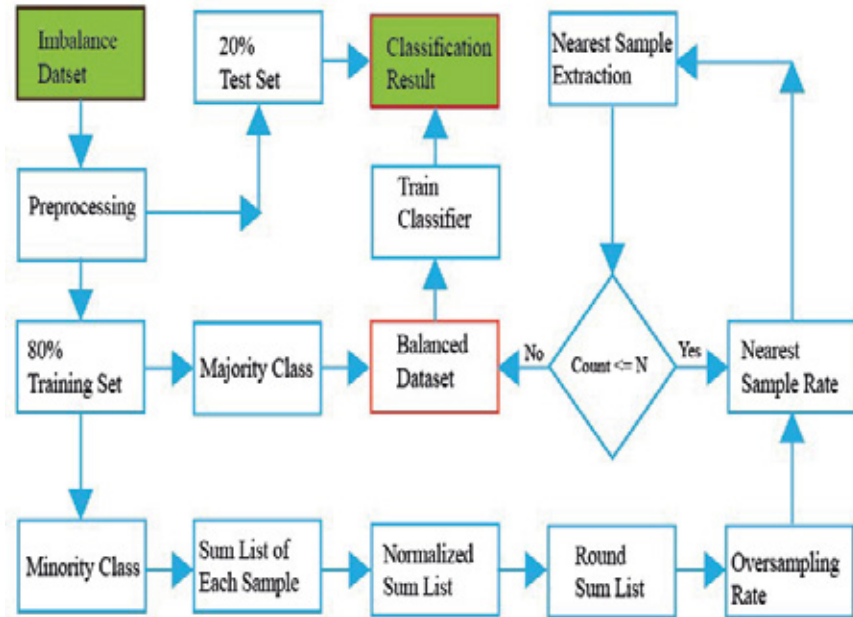


Figure 1
Flow chart of dominant SMOTE

The bold text indicates the minimum distance and their corresponding samples of the majority class. In the DistanceList1, according to dominance value 4 of the first sample of the minority class, the 4 minimum distances are 2.5, 3.5, 1.2, 3.1 and their samples such as M1, M2, M3, and M6 are selected to make balanced dataset with first sample of the minority class. The samples (M5, M50), (M4, M11) and (M8, M13) are selected from the majority class according to the DistanceList2, DistanceList3, DistanceList4 respectively. Thus, from the rest of distance lists such as DistanceList5, DistanceList6, DistanceList7, DistanceList8, DistanceList9, DistanceList10, the samples M49, M9, M14, M10, M17, M16 of majority class are selected according the minimum distance respectively. As a result, a complete balanced dataset is created according to the proposed algorithm. Algorithm 1 demonstrates the overall

procedure of the suggested dominant SMOTE model that is applied to identify the active and inactive cases of COVID-19.

Algorithm 1: DominantSMOTE

Input: Imbalanced Dataset
Output: Balanced Dataset
Data: Testing Imbalanced Dataset

```
1  $M_1CS \leftarrow \text{MajorityClassSamples}$ 
2  $M_2CS \leftarrow \text{MinorityClassSamples}$ 
3  $N \leftarrow \text{TotalNumberOfSamplesInMinorityClass}$ 
4  $NSFMC \leftarrow \text{NearestSampleFromMajorityClass}$ 
5  $OSRL \leftarrow \text{OverSamplingRateList}$ 
6  $NSL \leftarrow \text{NormalizedSumList}$ 
7 while  $\text{OuterCount} \in [1, N]$  do
8    $\text{SumOfFeatures} \leftarrow 0$ 
9   while  $\text{InnerCount} \in [1, M]$  do
10     $\text{SumOfFeatures} \leftarrow \text{SumOfFeatures} + M_1CS[\text{InnerCount}]$ 
11     $\text{SumList}[\text{OuterCount}] \leftarrow \text{SumOfFeatures}$ 
12 while  $\text{OuterCount} \in [1, N]$  do
13    $\text{Summation} \leftarrow 0$ 
14   while  $\text{InnerCount} \in [1, M]$  do
15     $\text{Summation} \leftarrow \text{Summation} + \text{SumList}[\text{InnerCount}]$ 
16     $NSL[\text{OuterCount}] \leftarrow \text{Summation} / \text{SumList}[\text{OuterCount}]$ 
17 while  $\text{OuterCount} \in [1, N]$  do
18    $\text{RoundSumList}[\text{OuterCount}] \leftarrow \text{Round}(NSL[\text{OuterCount}])$ 
19    $OSRL[\text{OuterCount}] \leftarrow \text{Round}(NSL[\text{OuterCount}])$ 
20    $\text{Append} \leftarrow M_1CS, NSFMC(M_1CS, OSRL[\text{OuterCount}])$ 
21  $\text{BalanceDataset} \leftarrow M_2CS$ 
22 return  $\text{BalanceDataset}$ 
```

4. Result and discussion

The datasets, evaluation methods, and test results are presented in this section.

4.1 Dataset

The authors ran this technique on two prominent benchmark datasets with, all of which were retrieved using the Scikit-learn Python 3.5 package from the University of California at Irvine (UCI) repository. The credit card dataset was acquired as part of collaboration between ULB's

Worldline and Machine Learning Group (ULB Universite' Libre de Bruxelles) on big data processing and fraud detection research. In September 2013, cardholders in Europe conducted 284,807 transactions, according to the dataset. Since, this work is concentrated on Binary classification. Since, this research is highlighted on binary classification; this work takes one class as a minority. Firstly, the author has tested COVID-19 dataset to detect the active and inactive cases of COVID-19. After that, the credit card dataset is tested using the algorithm to proof the robustness of the proposed algorithm. Table II depicts the all imbalanced dataset description that has been used in the experiment.

Table II
Characteristics, Illustration, and overview of the covid-19 and credit card dataset.

Dataset Name	Description	Minority: Majority	Samples	Features
COVID-19	COVID-19 patient active and inactive cases	558:5086	→1:10	5645 20
Credit Card	Credit Card fraud detection	492:284315	→1:577284,807	31

4.2 Experiment setup

To evaluate the effectiveness of our addressed algorithm on imbalanced dataset domain, I'm giving the answer the following inquiries:

Question 1: Is the addressed machine learning-based forecasting algorithm can predict the active cases of COVID-19 imbalanced dataset from the various active and inactive cases significantly and accurately?

Question 2: How efficient and effective is our addressed algorithm to predict the active cases of COVID-19 dataset from various active and inactive cases than other existing prediction algorithm in imbalanced dataset domain?

For answering the above mention questions, I have executed the experiment on the COVID-19 dataset that is made up of binary classes 'majority and minority. All the experiments and tests have been done in Python Programming Language, in which the most popular machine learning library Scikit-Learn has been used on Windows 2010 PC for predictive data analysis.

4.3 Evaluation metric

The proposed approach is compared to SMOTE, and Out Layer SMOTE, which are all traditional resampling methods. The accessible dataset was arbitrarily split into a 20% test set and an 80% training set. To create a new balanced training set, the dominating SMOTE approach was used. The models used in the analysis, including Decision Tree (DT), Naive Bayes (NV), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), were trained on an unbalanced training set

and then evaluated on the same test set, allowing us to compare the results and evaluate our proposed method. For each method, a comprehensive set of performance indicators was calculated, including precision in Equation 4, recall in Equation 5, accuracy in Equation 3, F1-Score in Equation 6, and confusion matrix. The majority of research in the imbalance dataset has concentrated on majority and minority issues. Traditionally, it is known as the negative class to refer to the minority class and the positive class to refer to the dominant class.

Figure 2 displays the confusion matrix for a two-class imbalance problem in the domain. The two class imbalance dataset's real two class labels are shown in the first column of the table in Figure 2. True Positive (TP) & True Negative (TN) represent the quantity of accurately identified anticipated true negative and positive samples, whereas False Positive (FP) & False Negative (FN) indicate the quantity of false positive and false negative samples that were misclassified. Accuracy is typically more likely to anticipate the majority class accurately and to act poorly toward the minority class during the performance evaluation of a research project for an unbalanced dataset. Because there are far more majority class samples than minority class samples in a highly imbalanced dataset, the accuracy value is 100% high if the classifier correctly classifies all majority class samples and misclassifies the minority class samples. In this condition, the accuracy cannot make trustworthy prediction for the minority class. For this reason, more reliable and authentic metrics for evaluation are required.

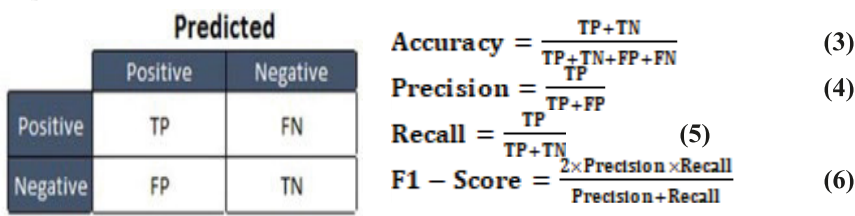


Figure 2
Example of a confusion matrix.

4.4 Result analysis

Figure 3 shows the importance of values of the all attributes of the COVID-19 dataset where there are twenty attributes after the cleaning of all unnecessary attributes. The important attribute selection approach has been applied on the COVID-19 dataset. The importance value of the few attributes is almost zero that indicates that these attributes don't impact the overall final classification result. As a result, these attributes have been

removed from the main dataset during the experiment based on the proposed algorithm. Tables III, IV, and V display the comparative results of SMOTE, Out Layer SMOTE, and the proposed Dominant SMOTE using the COVID-19 dataset. It is observed that the precision, accuracy, recall, and F1-score of Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF) are higher than those of SMOTE and Out Layer SMOTE. This serves as evidence of the robustness of the proposed Dominant SMOTE method. To assess the novelty and robustness of the proposed Dominant SMOTE method, I've applied the model to a credit card dataset. Tables VI, VII, and VIII present the experimental results of Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF) using the credit card dataset using SMOTE, Out Layer SMOTE, and Dominant SMOTE. It is observed that the results of Dominant SMOTE are superior to those of existing SMOTE and Out Layer SMOTE methods, confirming the robustness of the proposed Dominant SMOTE approach. Figures 4 and 5 show the comparison of accuracy for DT, NB, SVM, KNN, LR, and RF using the COVID-19 and credit card datasets with SMOTE, Outlier SMOTE, and the proposed Dominant SMOTE. The accuracy of DT, NB, SVM, KNN, LR, and RF using Dominant SMOTE is higher than that using the existing SMOTE and Outlier SMOTE methods. Figures 8 and 9 present the confusion matrices of DT, NB, SVM, KNN, RF, and LR using the COVID-19 dataset with the proposed Dominant SMOTE. It is observed that the True Positive (TP) and False Negative (FN) values are higher, while the False Positive (FP) and True Negative (TN) values are lower. These results indicate the better performance of the proposed Dominant SMOTE model. As a result, the proposed model achieves higher and more accurate detection rates for both active and inactive cases of COVID-19. Figures 8 and 9 present the confusion matrices of DT, NB, SVM, KNN, RF, and LR using the credit card dataset with the proposed Dominant SMOTE. It is observed that the True Positive (TP) and False Negative (FN) values are higher, while the False Positive (FP) and True Negative (TN) values are lower. These results indicate the better performance of the proposed Dominant SMOTE model.

The study compares three oversampling techniques SMOTE, Out Layer SMOTE, and the proposed Dominant SMOTE using both COVID-19 and credit card datasets. In the COVID-19 dataset, Dominant SMOTE consistently outperforms SMOTE and Out Layer SMOTE across various classifiers like Decision Tree, Naive Bayes, SVM, KNN, Logistic Regression, and Random Forest. This pattern is reflected in

higher precision, accuracy, recall, and F1-scores, showcasing the robustness of Dominant SMOTE. Similar results are seen in the credit card dataset, with Dominant SMOTE demonstrating superior performance in detecting both fraud and non-fraud cases compared to traditional SMOTE and Out Layer SMOTE. Figures depicting the accuracy comparisons and confusion matrices further illustrate the effectiveness of Dominant SMOTE, with higher True Positive (TP) and lower False Positive (FP) rates in both datasets. These findings suggest that the proposed Dominant SMOTE model offers improved detection rates for both COVID-19 cases and credit card fraud, highlighting its potential for enhancing predictive modeling in these domains.

The Dominant SMOTE method improves the traditional SMOTE algorithm for handling imbalanced classification by focusing on the most important features and dominant samples within the minority class. Instead of treating all minority samples equally, this method prioritizes those with higher dominance values, leading to more representative synthetic samples. By carefully selecting nearest neighbors from the majority class based on dominance, it reduces noise and prevents bias, ensuring that the synthetic data is more aligned with the most informative instances. This approach results in a better-balanced dataset, which can enhance the accuracy and reliability of classification models, particularly in scenarios where rare events or critical minority classes need precise identification.

Figure 3

Importance value of attributes of COVID-19 dataset.

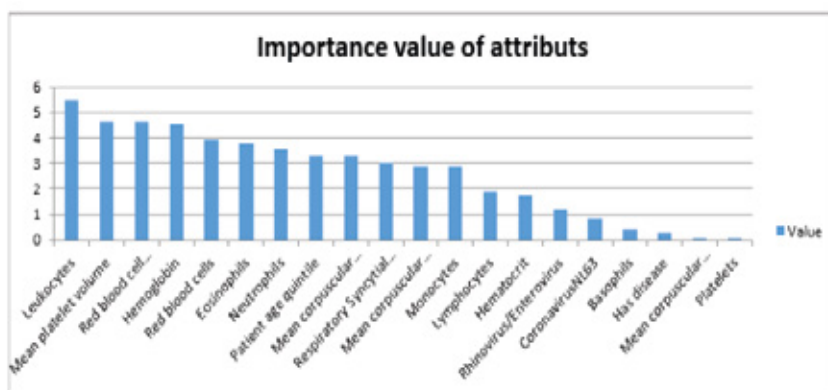


Table III*Experiment Results of the COVID-19 dataset using SMOTE.*

Model	Precision	Accuracy	Recall	F1-Score
DT	0.92	0.92	0.92	0.92
NV	0.91	0.91	0.91	0.92
SVM	0.92	0.92	0.92	0.92
KNN	0.92	0.92	0.92	0.92
RF	0.65	0.92	0.95	0.771
LR	0.94	0.96	0.95	0.94

Table IV*Experiment Result of COVID-19 dataset using Out Layer SMOTE.*

Model	Precision	Accuracy	Recall	F1-Score
DT	0.95	0.98	0.93	0.94
NV	0.94	0.95	0.95	0.94
SVM	0.95	0.91	0.96	0.97
KNN	0.91	0.96	0.93	0.92
RF	0.92	0.98	0.91	0.91
LR	0.95	0.97	0.94	0.95

Table V*Experiment Result of COVID-19 dataset using proposed Dominant SMOTE.*

Model	Precision	Accuracy	Recall	F1-Score
DT	0.97	0.99	0.97	0.97
NV	0.99	0.99	0.98	0.98
SVM	0.99	0.97	0.99	0.99
KNN	0.97	0.97	0.98	0.97
RF	0.98	0.99	0.98	0.98
LR	0.99	0.99	0.98	0.98

Table VI*Experiment Result of credit card dataset using SMOTE.*

Model	Precision	Accuracy	Recall	F1-Score
DT	0.92	0.93	0.95	0.93
NV	0.93	0.95	0.95	0.94
SVM	0.93	0.96	0.92	0.92
KNN	0.94	0.94	0.93	0.94
RF	0.92	0.94	0.93	0.95
LR	0.97	0.95	0.96	0.96

Table VII*Experiment Result of credit card dataset using Out Layer SMOTE.*

Model	Precision	Accuracy	Recall	F1-Score
DT	0.95	0.96	0.94	0.94
NV	0.95	0.93	0.93	0.94
SVM	0.95	0.92	0.95	0.95
KNN	0.93	0.94	0.94	0.93
RF	0.93	0.92	0.93	0.93
LR	0.97	0.98	0.98	0.97

Table VIII
Experiment Result of credit card dataset using Dominant SMOTE.

Model	Precision	Accuracy	Recall	F1-Score
DT	0.99	0.99	0.98	0.98
NV	0.97	0.99	0.98	0.97
SVM	0.98	0.97	0.99	0.98
KNN	0.98	0.98	0.98	0.98
RF	0.99	0.99	0.98	0.98
LR	0.99	0.99	0.99	0.99

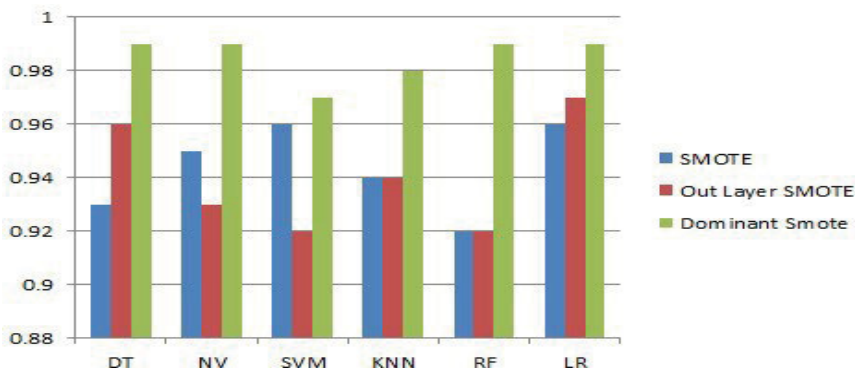


Figure 4
Comparison of accuracy using SMOTE, Outlier SMOTE, and Proposed Dominant SMOTE on Covid-19 dataset.

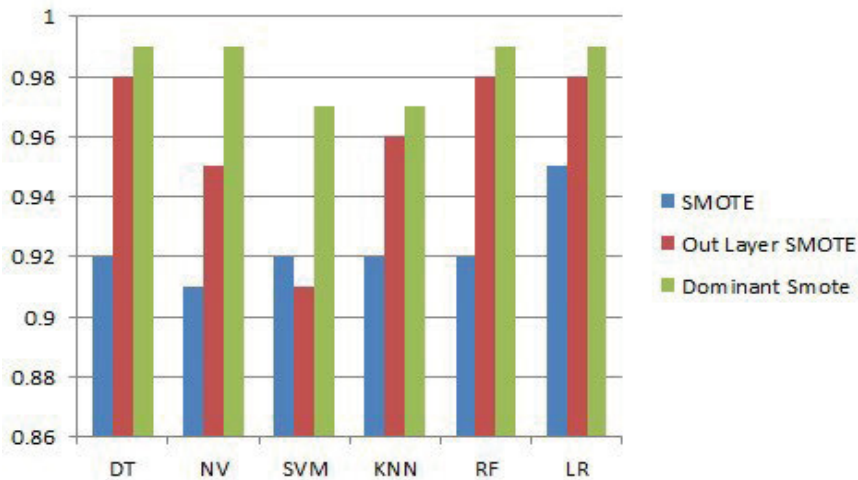


Figure 5
Comparison of accuracy using SMOTE, Outlier SMOTE, and Proposed Dominant SMOTE on credit card dataset.

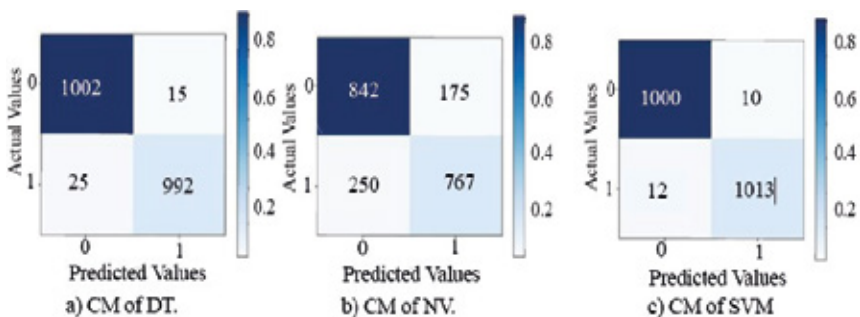


Figure 6

Confusion matrix of DT, NV, and SVM using covid-19 dataset and proposed dominant SMOTE.

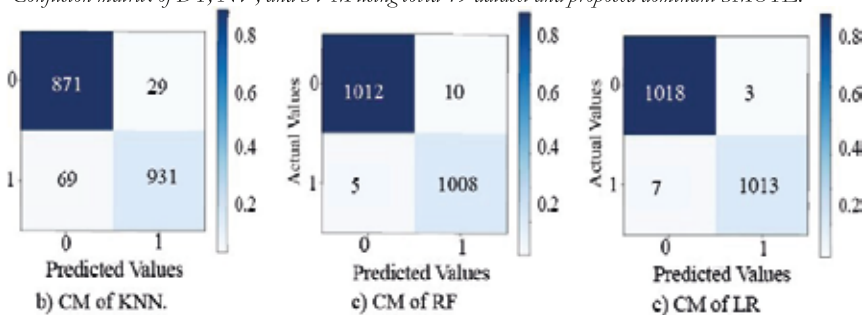


Figure 7

Confusion matrix of KNN, RF and LR using covid-19 dataset and proposed dominant SMOTE.

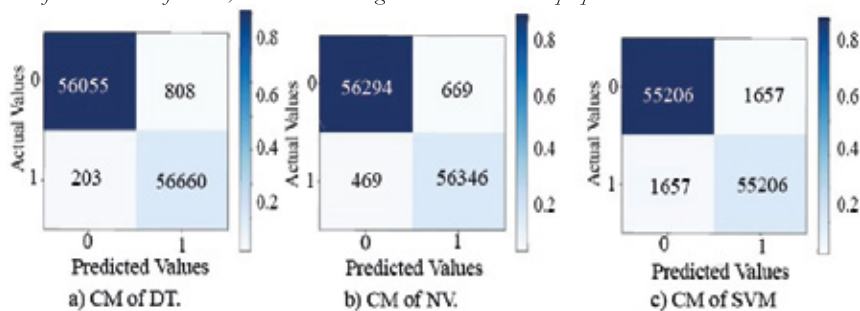


Figure 8

Confusion matrix of DT, NV, and SVM using credit card dataset and proposed dominant SMOTE.

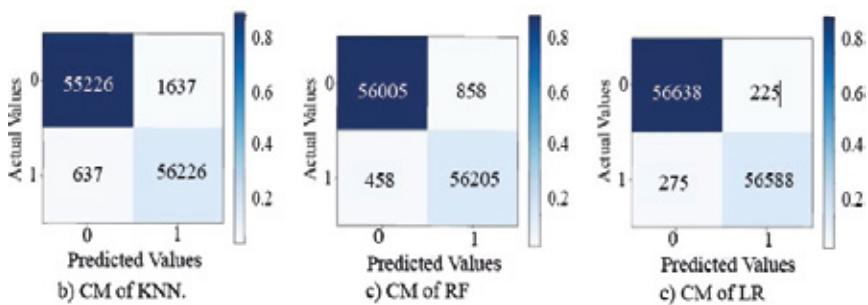


Figure 9
Confusion matrix of KNN, RF and LR using credit card dataset and proposed dominant SMOTE.

5 Conclusion

Due to the skewed result of prediction, learning with an imbalanced dataset has garnered considerable attention in both theoretical and practical areas of machine learning. The use of an imbalanced dataset often diminishes the performance of a classifier in a machine learning prediction system. This work introduces Dominant SMOTE, a method based on key features that selectively oversamples only dominant attributes, with the more dominant attributes being oversampled to a greater extent than the less dominant ones. This approach significantly enhances the accuracy of COVID-19 patient classification. The addressed machine learning-based prediction algorithm, specifically the proposed Dominant SMOTE method, demonstrates the ability to predict active cases of COVID-19 significantly and accurately from the various active and inactive cases. The study indicates that the proposed Dominant SMOTE method consistently outperforms other methods (SMOTE and Out Layer SMOTE) across multiple classifiers (DT, NB, SVM, KNN, LR, and RF) when applied to the COVID-19 dataset. Results such as higher precision, accuracy, recall, and F1-scores for the proposed method compared to others indicate its effectiveness in handling imbalanced data, specifically in predicting active cases of COVID-19. The addressed algorithm, which is the proposed Dominant SMOTE method, proves to be more efficient and effective in predicting active cases of the COVID-19 dataset compared to other existing prediction algorithms, especially in handling imbalanced datasets. The confusion matrices further support the efficiency of the proposed Dominant SMOTE method, showing higher True Positive (TP) rates and lower False Positive (FP) rates, which are crucial for accurately predicting active cases. The comparison figures, such as accuracy comparisons and confusion matrices, clearly demonstrate the superiority of the proposed method over existing techniques when it

comes to predicting active cases in the COVID-19 dataset. These results emphasize the robustness and superior performance of the Dominant SMOTE method across all datasets, showcasing its enhanced predictive capabilities in detecting minority classes compared to algorithms trained on imbalanced datasets.

The Dominant SMOTE method, while effective in enhancing class balance by focusing on dominant features, has limitations across different classifiers. It can lead to overfitting, especially in models like k-nearest neighbors, and its complexity may increase computational demands. The method's reliance on accurate feature selection and its potential to distort class balance can also hinder generalization and performance, particularly in classifiers less sensitive to class imbalance, like decision trees and random forests.

In future work, there are several avenues to explore for enhancing the Dominant SMOTE method's effectiveness in COVID-19 patient classification. One direction is to investigate additional relevant features, such as demographic information, which could provide more comprehensive insights. Optimizing the selection of dominant attributes dynamically and considering weighted oversampling strategies could further refine the sampling approach.

References

- Ahmad, I. S., Bakar, A. A., Yaakub, M. R., & Muhammad, S. H. (2020). A survey on Machine Learning Techniques in movie Revenue Prediction. *SN Computer Science/SN Computer Science*, 1(4). <https://doi.org/10.1007/s42979-020-00249-1>
- Asadi, H., Dowling, R., Yan, B., & Mitchell, P. (2014). Machine Learning for outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial therapy. *PloS One*, 9(2), e88225. <https://doi.org/10.1371/journal.pone.0088225>
- Barandela, R., Rosas, R. M. V., & Sánchez, J. (2003). New applications of ensembles of classifiers. *Pattern Analysis and Applications/Pattern Analysis & Applications*, 6(3), 245–256. <https://doi.org/10.1007/s10044-003-0192-z>
- Birla, S., Kohli, K., & Dutta, A. (2016). Machine Learning on imbalanced data in Credit Risk. *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. <https://doi.org/10.1109/iemcon.2016.7746326>
- Chawla, N. V., Bowyer, K. W., Hall, L., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research/the Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

- Feng, S., Zhao, C., & Fu, P. (2020). A cluster-based hybrid sampling approach for imbalanced data classification. *Review of Scientific Instruments Online/Review of Scientific Instruments*, 91(5). <https://doi.org/10.1063/5.0008935>
- Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and Hybrid-Based approaches. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, 42(4), 463–484. <https://doi.org/10.1109/tsmcc.2011.2161285>
- Genoud, A. P., Gao, Y., Williams, G. M., & Thomas, B. A. (2020). A comparison of supervised machine learning algorithms for mosquito identification from backscattered optical signals. *Ecological Informatics*, 58, 101090. <https://doi.org/10.1016/j.ecoinf.2020.101090>
- Gibson, P. G., Qin, L., & Puah, S. H. (2020). COVID -19 acute respiratory distress syndrome (ARDS): clinical features and differences from typical pre- COVID -19 ARDS. *Medical Journal of Australia*, 213(2), 54. <https://doi.org/10.5694/mja2.50674>
- Gralinski, L. E., & Menachery, V. D. (2020). Return of the coronavirus: 2019-NCOV. *Viruses*, 12(2), 135. <https://doi.org/10.3390/v12020135>
- Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: a new Over-Sampling method in imbalanced Data sets learning. In *Lecture notes in computer science* (pp. 878–887). https://doi.org/10.1007/11538059_91
- Haruna, A. A., Muhammad, L. J., Yahaya, B. Z., Garba, E. J., Oye, N. D., & Jung, L. T. (2019). An Improved C4.5 Data Mining Driven Algorithm for the Diagnosis of Coronary Artery Disease. *2019 International Conference on Digitization (ICD)*. <https://doi.org/10.1109/icd47981.2019.9105844>
- He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. <https://doi.org/10.1109/ijcnn.2008.4633969>
- Hlosta, M., Stríž, R., Kupčík, J., Zendulka, J., & Hruška, T. (2013). Constrained classification of large imbalanced data by logistic regression and genetic algorithm. *International Journal of Machine Learning and Computing*, 214–218. <https://doi.org/10.7763/ijmlc.2013.v3.305>
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., . . . Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, 395(10223), 497–506. [https://doi.org/10.1016/s0140-6736\(20\)30183-5](https://doi.org/10.1016/s0140-6736(20)30183-5)

- Huang, G., Zhu, Q., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3), 489–501. <https://doi.org/10.1016/j.neucom.2005.12.126>
- Hussain, S. P., Muhammad, L. J., Ishaq, F. S., Yakubu, A., & Mohammed, I. A. (2018). Performance evaluation of various data mining algorithms on road traffic accident dataset. In *Smart innovation, systems and technologies* (pp. 67–78). https://doi.org/10.1007/978-981-13-1742-2_7
- Ishaq, F. S., Muhammad, L. J., Z, Y. B., & Abubakar, A. (2020). Fuzzy based Expert System for Diagnosis of Diabetes Mellitus. *International Journal of Advanced Science and Technology*, 136, 39–50. <https://doi.org/10.33832/ijast.2020.136.04>
- Khalilia, M., Chakraborty, S. B., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1). <https://doi.org/10.1186/1472-6947-11-51>
- Khreich, W., Granger, É., Miri, A., & Sabourin, R. (2010). Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs. *Pattern Recognition*, 43(8), 2732–2752. <https://doi.org/10.1016/j.patcog.2010.03.006>
- Koziarski, M. (2020). Radial-Based Under-sampling for imbalanced data classification. *Pattern Recognition*, 102, 107262. <https://doi.org/10.1016/j.patcog.2020.107262>
- Krawczyk, B., & Woźniak, M. (2015). Cost-Sensitive Neural Network with ROC-Based Moving Threshold for Imbalanced Classification. In *Lecture notes in computer science* (pp. 45–52). https://doi.org/10.1007/978-3-319-24834-9_6
- Lavanya, B., Katayani, V., Jancy, S., Latha, S. P., MaryA, V. A., Sujihelen, L., & Selvan, M. P. (2023). Covid-19 Future Forecasting Using Supervised Machine Learning Models. *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, <https://doi.org/10.1109/iconstem56934.2023.10142413>
- Li, Y., Sun, G., & Zhu, Y. (2010). Data Imbalance Problem in Text Classification. *2010 Third International Symposium on Information Processing*. <https://doi.org/10.1109/isip.2010.47>
- Lin, H., & Nguyen, M. C. (2020). Boosting minority class prediction on imbalanced point cloud data. *Applied Sciences*, 10(3), 973. <https://doi.org/10.3390/app10030973>
- Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., ...Santillana, M. (2020). A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet

- searches, news alerts, and estimates from mechanistic models. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2004.04019>
- Mathkunti, N. M., & Rangaswamy, S. (2020). Machine learning techniques to identify dementia. *SN Computer Science/SN Computer Science*, 1(3). <https://doi.org/10.1007/s42979-020-0099-4>
- Maheshwari, S., Jain, R. C., & Jadon, R. S. (2017). A review on class imbalance problem: Analysis and potential solutions. *International journal of computer science issues (IJCSI)*, 14(6), 43-51
- Morens, D. M., Daszak, P., & Taubenberger, J. K. (2020). Escaping Pandora's Box — another novel Coronavirus. *New England Journal of Medicine/the aNew England Journal of Medicine*, 382(14), 1293–1295. <https://doi.org/10.1056/nejmp2002106>
- Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. *SN Computer Science/SN Computer Science*, 1(5). <https://doi.org/10.1007/s42979-020-00250-8>
- Muhammad, L. J., & Sharif, U. S. (2020). Power of artificial intelligence to diagnose and prevent further COVID-19 outbreak: A short communication. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2004.12463>
- Naseriparsa, M., Al-Shammari, A., Sheng, M., Zhang, Y., & Zhou, R. (2020). RSMOTE: improving classification performance over imbalanced medical datasets. *Health Information Science and Systems*, 8(1). <https://doi.org/10.1007/s13755-020-00112-w>
- Saravanan, R., & Sujatha, P. (2018). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. <https://doi.org/10.1109/iccons.2018.8663155>
- Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A new paradigm. *Trends in Genetics*, 34(4), 301–312. <https://doi.org/10.1016/j.tig.2017.12.005>
- Song, Y., Morency, L., & Davis, R. (2013). Distribution-sensitive learning for imbalanced datasets. *2013 10th IEEE In- Ternational Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. <https://doi.org/10.1109/fg.2013.6553715>
- Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-based Medicine*, 3–4, 100023. <https://doi.org/10.1016/j.ibmed.2020.100023>

- Van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, ... Munster, V. J. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *New England Journal of Medicine/ the æNew England Journal of Medicine*, 382(16), 1564–1567. <https://doi.org/10.1056/nejmc2004973>
- Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based under-sampling approach for handling imbalanced and overlapped data. *Information Sciences*, 509, 47–70. <https://doi.org/10.1016/j.ins.2019.08.062>
- Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. *2009 IEEE Symposium on Computational Intelligence and Data Mining*. <https://doi.org/10.1109/cidm.2009.4938667>
- Yen, S., & Lee, Y. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3), 5718–5727. <https://doi.org/10.1016/j.eswa.2008.06.108>
- Zhang, X., Jiang, D., Han, T., Wang, N., Wei-Guo, Y., & Yang, Y. (2017). Rotating machinery fault diagnosis for imbalanced data based on fast clustering algorithm and support vector machine. *Journal of Sensors*, 2017, 1–15. <https://doi.org/10.1155/2017/8092691>
- Zhang, Y., Li, X., Wang, L., & Wen, L. (2018). Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning. *Journal of Manufacturing Systems*, 48, 34–50. <https://doi.org/10.1016/j.jmsy.2018.04.005>
- Zhu, Z., & Song, Z. (2010). Fault diagnosis based on imbalance modified kernel Fisher discriminant analysis. *Process Safety and Environmental Protection/Transactions of the Institution of Chemical Engineers. Part B, Process Safety and Environmental Protection/Chemical Engineering Research and Design/Chemical Engineering Research & Design*, 88(8), 936–951. <https://doi.org/10.1016/j.cherd.2010.01.005>
- Zięba, M., & Świątek, J. (2013). Ensemble SVM for imbalanced data and missing values in postoperative risk management. *2013 IEEE 15th International Conference on e-Health Networking*. <https://doi.org/10.1109/healthcom.2013.6720646>
- Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, 2–8. <https://doi.org/10.1016/j.bdr.2015.12.001>

