Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of King Saud University – Computer and Information Sciences 34 (2022) 7830–7839

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of Covid-19

Prasetyo Wibowo, Chastine Fatichah *

Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

ARTICLE INFO

Article history:
Received 1 May 2021
Revised 28 July 2021
Accepted 25 September 2021
Available online 30 September 2021

Keywords:
Oversampling
Smoothed bootstrap resampling
Imbalanced data
Machine learning
COVID-19

ABSTRACT

The Coronavirus Disease (COVID-19) was declared a pandemic disease by the World Health Organization (WHO), and it has not ended so far. Since the infection rate of the COVID-19 increases, the computational approach is needed to predict patients infected with COVID-19 in order to speed up the diagnosis time and minimize human error compared to conventional diagnoses. However, the number of negative data that is higher than positive data can result in a data imbalance situation that affects the classification performance, resulting in a bias in the model evaluation results. This study proposes a new oversampling technique, i.e., TRIM-SBR, to generate the minor class data for diagnosing patients infected with COVID-19. It is still challenging to develop the oversampling technique due to the data's generalization issue. The proposed method is based on pruning by looking for specific minority areas while retaining data generalization, resulting in minority data seeds that serve as benchmarks in creating new synthesized data using bootstrap resampling techniques. Accuracy, Specificity, Sensitivity, F-measure, and AUC are used to evaluate classifier performance in data imbalance cases. The results show that the TRIM-SBR method provides the best performance compared to other oversampling techniques.

1. Introduction

The Coronavirus Disease (COVID-19) was declared a pandemic disease by the World Health Organization (WHO) starting from March 11th, 2020, it spreads to 114 countries with a total of more than 118,000 cases and 4,300 deaths. A research report from Chen (Chen et al., 2020) shows that 51% of the patients observed were suffering from chronic diseases, with 11% of them experiencing conditions that getting worse and dying from organ failure.

The article from Mahase (Mahase, 2020) states that COVID-19 has a low mortality rate compared to Severe Acute Respiratory Syndrome (SARS) and the Middle East respiratory syndrome (MERS). However, COVID-19 has resulted in more deaths compared to SARS and MERS. This is because COVID-19 is a disease that has a low mortality rate but spreads faster than SARS and MERS. Hence if detected, better to get treated soon to avoid any complications.

A COVID-19 diagnostic test can be performed in three ways: lateral flow test (LFT), polymerase chain reaction (PCR), and computed tomography (CT) scan. All three diagnostic tests have advantages and disadvantages. The advantage of the LFT test is that it has the lowest price compared to other diagnostic tests but sometimes gives inaccurate results. PCR test results are more reliable than other diagnostic tests but take longer because it requires laboratory analysis than other diagnostic tests. The last is a CT scan that gives accurate results but requires special tools and experienced staff to carry out the test.

LFT and PCR have the same collection method by performing a swab taken from the back of the nose or throat. These two diagnostic tests have different working methods, for PCR tests by detecting viral RNA (genetic material) while LFT detects virus-specific proteins contained in patient samples. The use of LFT has become a hot topic of discussion due to poor sensitivity results compared to PCR (Deeks and Raffle, 2020; Armstrong, 2020; Kmietowicz, 2021). Poor sensitivity occurs because LFT cannot detect in asymptomatic patients and is corroborated by a study report from Ferguson (Ferguson et al., 2021), which concluded that LFT could not detect very early or very late stages of infection. WHO recommends that initial testing can use LFT, while for confirmation

* Corresponding author.
E-mail address: chastine@if.its.ac.id (C. Fatichah).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.09.021
1319-1578/

testing can use the PCR test (World Health Organization (WHO), 2020).

High consumption of PCR tests makes CT scans an alternative in diagnosing patients infected with COVID-19. The way a CT scan works is to take images using X-rays in the chest area. A radiologist will examine the images to identify abnormalities that are associated with COVID-19 disease. Studies that have been carried out show that CT scans have better results than PCR (Ai et al., 2020; Long et al., 2020; Zheng et al., 2020). However, researchers do not recommend CT scanning as the primary diagnostic test (Dennie et al., 2020; Dickson et al., 2020; Hope et al., 2020; Laghi, 2020) and recommend using PCR instead. Another option is to combine the two diagnostic tests for better detection performance. While PCR and CT scans have yielded the desired results, a computational approach is required to speed up diagnosis and minimize human error compared to traditional diagnostics.

The main challenge when using imbalanced datasets is that the number of negative test results is unbalanced compared to positive test results or vice versa. This issue also often occurs in computer vision (Wang et al., 2019; Oksuz et al., 2020; Wang et al., 2020) and big data (Rendón et al., 2020; Wibowo and Fatichah, 2021), where the class imbalance problem can severely impact the classification method. When the classification method is used on an imbalanced dataset, the machine learning algorithm results will become biased against the majority class because the results are usually focused on accuracy. It leads to a phenomenon where the overall results are given very high accuracy but poor generalization of data against minority data. In general, there are two main strategies to address data imbalances, namely the cost-sensitive method and the resampling method (Leevy et al., 2018). The cost-sensitive method is a method that considers the cost of misclassification to minimize the total cost so that it will be more appropriate when applied to complex datasets. The downside of this cost-sensitive strategy is that it is unstable when applied to small datasets or has significantly skewed data (Zhang et al., 2011; Lu et al., 2019), so a resampling approach will be utilized in this study to alleviate the problem of data imbalance. (Longadge and Dongre, 2013; Nanni et al., 2015).

Resampling is a common practice for solving class imbalance problems in data sets (Estabrooks et al., 2004). This resampling process tries to change the training data so that the data distribution is balanced. There are two groups in resampling, namely undersampling and oversampling. Oversampling flattens class data by creating new samples, duplicating minority data classes, and undersampling flattens class data by subtracting or removing samples from the majority class. This study focuses on the oversampling techniques since the majority of class data has relevant information for processing and using the oversampling technique will result in more balanced minority class data. However, it will increase the risk of overfitting (Santos et al., 2018).

One state of the art of oversampling techniques is Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). This method generates sample data on the feature space between the original minority classes by taking each minority class and introducing the sample data along the line of the closest neighbors selected from the minority class. SMOTE's advantage is to make the decision area larger and less specific, causing the area that is often studied is the minority class rather than the surrounding majority class (He and Garcia, 2009). However, SMOTE has a drawback, namely overgeneralization, where the sample data is generated into the majority of the area (Bunkhumpornpat et al., 2009). This happens because SMOTE does not consider the distribution of the majority class area.

This study proposes a new oversampling technique for the diagnostic classification of patients infected with COVID-19. The main contributions are outlined as follows:

- TRIM-Smoothed Bootstrap Resampling (TRIM-SBR) is a proposed method to reduce the overgeneralization problem that usually occurs when synthetic data is formed into a majority class region with evenly distributed synthetic data effects. Our method is based on pruning by looking for a particular minority area while maintaining the generality of the data so that it will find the minority data set while filtering out irrelevant data. The pruning results will produce a minority data seed that is used as a benchmark in duplicating data. To ensure the duplication of data is evenly distributed, the bootstrap resampling technique is used to create new data.
- Conduct extensive experiments with various learning algorithms so that they can be compared with the current state of the art research.
- This study will also show how to preprocess data and create new features to capture essential features in COVID-19 datasets.

This paper is organized into 5 sections. Section 1 introduces the highlighted issues. Section 2 presents the details of the material and supporting theories. Section 3 outlines the methodology for the proposed work, and Section 4 describes the experimental results, followed by discussion and analysis in Section 5.

2. Related works

During the pandemic, especially in developing countries, the need of PCR as the primary test diagnosis increases and requires a long diagnosis time to delay the initial steps to prevent the spread of COVID-19. The first researchers who tried to make a computational approach were (Yan et al., 2020), who modelled the prediction of survival rates in critical patients with Covid-19. This study utilized 404 blood samples of infected patients in Wuhan using three features: lactate dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP). The prediction model used is XGBoost because it can make tree-based interpretations by giving significant values to each feature during the tree creation process. The results show an accuracy rate of 93% with a predictive value of 100% mortality and a predictive value of survival of 90%. Although this study yields good accuracy, the authors suggest more than 3,000 patient data and 80 clinical data explored further.

Furthermore, a research by (Vaid et al., 2020) predicts the mortality rate for positive COVID-19 patients in New York City. This study used 3,055 positive patient data for COVID-19 obtained from five different hospitals using a decision tree. The predictive result obtained was 84% for AUC by concluding that age features, inflammatory markers, coagulation parameters, and D-Dimer are essential features of the model. This analysis's drawback is that it only used the AUC metric value in the classification assessment to demonstrate how accurate the classification model was (Batista et al., 2020) tried to predict the diagnosis of COVID-19 in emergency care patients using a machine learning approach. This study used 235 data with 15 features derived from patient data at Hospital Israelita Albert Einstein using five machine learning algorithms: neural networks, random forest, gradient boost, logistic regression, and support vector machine (SVM). This study obtained good results with an average AUC value of 84%, a sensitivity of 74%, and an F1 score of 75%. Furthermore (Soares et al., 2020), conducted an analysis focusing on blood testing based on 599 patient results. The model used SVM and added a SMOTE to produce synthetic data. The results show a sensitivity value of 70% and a specificity of 85%. Compared to Batista, this study's strength is the use of resampling techniques to resolve class data imbalance. However, the biggest downside is the lack of exploration of resampling methods, which could have better results than the main algorithm.

COVID-19 data has the characteristics of data imbalance and medical data such as cancer, diabetes, and hepatitis data (Dua and Graff, 2019). Therefore, the data balancing method is needed to produce good evaluation results. In data imbalance, the overlap is a common problem faced by researchers and becomes a crucial problem when the data is highly skewed (Vuttipittayamongkol and Elyan, 2020). The problem of overlapping is also supported by the results of the paper, which indicate that there is a very strong relationship between class imbalance and class overlap that influences the performance of the classification (García et al., 2006; Almutairi and Janicki, 2020; Stefanowski, 2013). One way to overcome overlap is to select instances to be sampled (Fernández et al., 2018). This strategy aims to reduce overlap and noise in the dataset by choosing the closest sample or not duplicating data depending on the number of minority classes in the area. In the case of COVID-19, the current data is very valuable to process. Hence, the undersampling technique is not a good choice in this study for its major weakness that can throw away potentially valuable data in the classification process (Batista et al., 2005). The oversampling technique is the best choice in this study because it does not lose data when processed.

Several studies prove that oversampling techniques can improve test evaluation of unbalanced datasets, such as research conducted by (Akbani et al., 2004) which tried to add oversampling techniques to medical data using the SMOTE method by taking into account the cost of each minority data. The results show a significant increase in data using oversampling techniques from 36% to 70% in the sensitivity evaluation metric. Some studies compare several oversampling methods, such as SMOTE, Borderline-SMOTE, and random oversampling (Douzas et al., 2018). The results show a significant increase in the F1 score and AUPRC results compared to the initial data.

Several oversampling techniques are often used to overcome data imbalance problems. The random oversampling method (ROS) operates by randomly replicating a set of selected minority classes (Batista et al., 2004). Because the sampling process is carried out randomly, the selection function will find it difficult to find the difference between the two classes. The drawbacks of the random oversampling method are the increased training time for classification and the possibility of overfitting (Ganganwar, 2012) in duplicating minority class data and making class imbalance worse. The Synthetic Minority Oversampling (SMOTE) technique creates artificial samples based on features rather than data based on similarities between minority classes of the sample (Chawla et al., 2002; He and Garcia, 2009). This synthetic example will create a segment line based on the portion or whole of the K nearest neighbors of the minority class. Depending on the amount of oversampling data required, neighbors are randomly selected.

There are many extensions made using SMOTE as a technique to balance class distribution. Borderline-SMOTE is used to take minority classes near the boundary line and the same class's surrounding area (Han et al., 2005). When compared to the original SMOTE, borderline-SMOTE does not synthesize data but focuses on data that is in the border area so that it will help in creating areas between classes. The Adaptive Synthetic Sampling for Imbalanced Learning (ADASYN) approach uses the density distribution in the minority class as a criterion for synthesizing data in each minority sample (He et al., 2008). This approach can differentiate the density distribution in each minority sample and add as many minority samples as needed to balance the majority class. This approach helps focus minority class centers depending on modelling difficulties. Safe-Level-SMOTE is a method that creates a safe level for minority sample data before generating synthetic data (Bunkhumpornpat et al., 2009). Each synthesized data will be approached with the highest security level so that all synthesized data will only be in a safe area. The safety level ratio depends on each sample data set and each sample data area. DBSMOTE uses a density-based clustering approach and generates samples synthesized along with each minority data (Bunkhumpornpat et al., 2012). DBSMOTE can work in overlapping areas, such as Borderline-SMOTE, but the difference from Borderline-SMOTE is that it can maintain the accuracy of the minority and majority classes. The next approach is to perform class decomposition by finding the similarities of the majority class instances and grouping them into one called CDSMOTE (Elyan et al., 2021). This approach tries to reduce the dominance of the majority class without eliminating the information on the majority class.

A distance-based approach is known as the Mahalanobis distance-based oversampling (MDO) technique, a multiclass approach based on Mahalanobis (Abdi and Hashemi, 2015). MDO synthetic data was created using the same Mahalanobis distance in each class average based on other minority classes. Vuttipittayamongkol (Vuttipittayamongkol and Elyan, 2020) uses the fuzzy C-Means method approach to eliminate overlapping classes to identify negative and positive classes accurately. After identifying between negative and positive classes, data duplication was carried out using the concept of Borderline-SMOTE so that the duplicated data was in the minority area.

3. Methodology

3.1. Diagram system of oversampling data for imbalanced dataset

The system design is prepared to obtain a good predictive model, as shown in Fig. 1.

Before the data modeling from the dataset, data preprocessing is applied to get the ideal dataset. The preprocessing data consist of data cleaning, data reduction, and data transformation. Data cleaning and reduction will be carried out to eliminate data deemed not to significantly affect the pattern data or interfere with the classification evaluation results. In the data transformation, feature correlation is used to find correlations between features and find any features with almost the same characteristics to make the data processed simpler. The data preprocessing is very influential in the evaluation results. Afterward, the data modeling is applied to get the best predictive model, that is focused on using machine learning. First, the splitting dataset using the Holdout method into two parts, namely training data and testing data. An oversampling technique is carried out to strengthen the value between classes in training data. The hyperparameter optimization is applied to produce the most optimal model evaluation in each classification model.

3.2. The proposed oversampling method

The proposed method is called TRIM-Smoothed Bootstrap Resampling (TRIM-SBR), which aims to reduce the overgeneralization problem that usually occurs when synthetic data is formed into the majority class region with the effects of synthetic data being evenly distributed. As in Fig. 2, the proposed method is divided into two parts: pruning and data duplication. The TRIM method is used in the pruning section, a preprocessed method to avoid overgeneralizing data using a greedy approach in finding minority data sets while filtering irrelevant data. However, this method does not guarantee the best global results but offers a reasonable estimate for the optimal set (Puntumapon and Waiyamai, 2012). The pruning results will produce a minority data seed used as a benchmark in duplicating data. A smoothed bootstrap technique is used to create new data to ensure the duplication of data is evenly distributed. This technique will reduce overfitting by test-

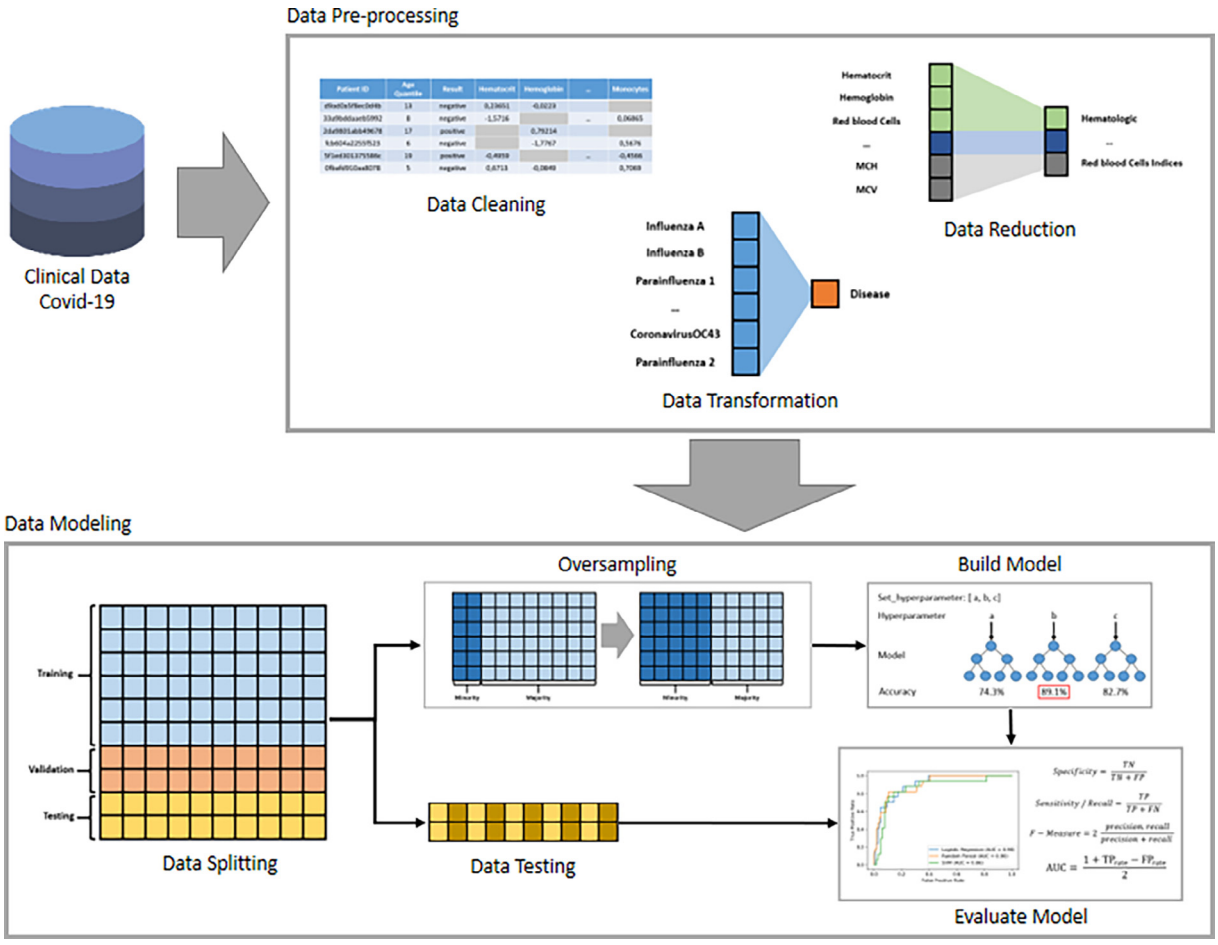

Fig. 1. The system designs.

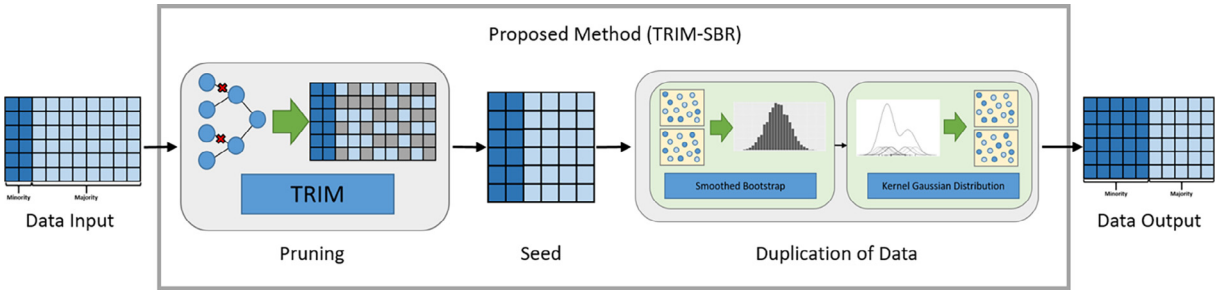

Fig. 2. Block diagram for the proposed method.

ing the dimensional estimation of observations belonging to the data duplication class to minimize data redundancy.

3.3. Pruning method

TRIM is a method that aims to avoid overgeneralization of data. The basic idea is to identify the collection of minority data with the best compromise between generalizability and precision between data (Puntumapon et al., 2016). Equation (1) is used to measure the precision and generalization of data. The higher the TRIM criterion value, the more precise and general the seed data will be.

$$TRIM = \frac{|minority|^2}{N}$$

The value of $|minority|$ is the measure of the minority data. To get the TRIM Gain $(T - Gain)$ seed data, two separate datasets are evaluated and compared with TRIM. If $(T - Gain) > TRIM$, the seed data is obtained by performing a binary separator operation. $|minority_{left}|$ and $|minority_{right}|$ are the sum of the left and right minority data subsets; N is the total number of sample data, N_{left} and N_{right} are the numbers of left and right subset data. As a result, Equation (2) will be formulated as follows:

$$T - Gain = max\left(\frac{|minority_{left}|^2}{N_{left}}, \frac{|minority_{right}|^2}{N_{right}}\right)$$

Eq. (2) is designed to capture the characteristics of the resampling data. The first characteristic is to create new synthetic data based on several samples from the minority data to evaluate the precision of the minority data. Furthermore, the second characteristic is that synthetic data are always generated in the convex hull of the minority data. The purpose of $(T - Gain)$ is to identify the most irrelevant data that lie on the outside of the convex hull and filter it. The pseudocode of TRIM is presented in Algorithm 1.

Algorithm 1 TRIM (N)

Input: data (N)
Output: seed $(Seed)$
Method:
1. $D = \{\}$
2. Add data N to D
3. **while** (D is not empty)
4. $Trim$ = ComputeTRIM (D)
5. $DataSplit$ = splitting point in data $|majority_{left}|$ or $|majority_{right}|$
6. $TrimSplit = max_{DataSplit}$ (ComputeTrimSplit$(D, DataSplit)$)
7. **if** $(TrimSplit > Trim)$
8. Split data D to D_{left} and D_{right}
9. **if** ($|majority_{left}| == 0$),
10. $D_{new} = D_{right}$
11. **else**
12. $D_{new} = D_{left}$
13. **end if**
14. $D = D_{new}$
15. **end if**
16. $Trim$ = ComputeTRIM (D)
17. $DataSplit$ = splitting point in data (D)
18. $TrimSplit = max_{DataSplit}$ (ComputeTrimSplit$(D, DataSplit)$)
19. **if** $(TrimSplit > Trim)$
20. Split data D to D_{left} and D_{right}
21. Add data D_{left} to D
22. Add data D_{right} to D
23. **end if**
24. **end while**
25. **return** $Seed$

3.4. Data duplication method

Random Over Sampling Examples (ROSE) is a data duplication method with a smoothed bootstrap approach (Menardi and Torelli, 2014). This method has three main procedures in forming new synthetic data, namely:

- Choose $y = y_j \in y$ with probability $\frac{1}{2}$
- Select (x_i, y_i) on $T_n$as $y_i = y$ with probability$p_i = \frac{1}{n_j}$
- Sample x of $K_{H_j}(., x_i)$ where K_{H_j} is the middle probability distribution on x_i and depends on the scale of the matrix parameter H_j

Operationally, ROSE requires a H_j matrix to duplicate data. In theory, the selection of a smoothing matrix will affect the size of K_{H_j}. Previous studies discussed this in selecting smoothing parameters (Bowman and Azzalini, 1999; Silverman, 1986). From the number of alternatives available, the Gaussian Kernel was chosen with a diagonal smoothing matrix $H_j = diag\left(h_1^{(j)}, \cdots, h_d^{(j)}\right)$as in Equation (3) where $\hat{\sigma}_q^{(j)}$is a sample estimate of the standard deviation of the $q - th$ dimension whose observations belong to the class y_j. After getting the $h_q^{(j)}$matrix smoothing, data distribution is carried out through the Gaussian Kernel through Equation (4) where μ is the mean, σ is the value of the standard deviation, and σ^2 is the variance.

$$h_q^{(j)} = \left(\frac{4}{(d+2)n}\right)^{\frac{1}{d+4}} \hat{\sigma}_q^{(j)} (q = 1, \cdots, d = 1, 2)$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Pseudocode smoothed bootstrap resampling presented in Algorithm 2.

Algorithm 2 SMB (N)

Input: data (N)
Output: The synthetic minority class $(Samples)$
Method:
1. X = transform N stack arrays in sequence vertically
2. Y = transform N stack arrays in sequence horizontally
3. X_{min} = add X and Y
4. $CalcStd$ = ComputeStdDev(X_{min})
5. $Value_{Data}$ = ReturnValue(N)
6. $ValueX_{min}$ = ReturnValue(X_{min})
7. H_{matrix} = ComputeMatrix$(CalcStd, Value_{Data}, ValueX_{min})$ // using Eq. 3
8. $Samples$ = {}
9. **for each** $index$ in N
10. $Rand$ = random number
11. $Value_{Xmin}$ = ReturnValue(X_{min})
12. H_{index} = ComputeRandomIndex $(Rand, Value_{Xmin})$
13. $Value_{Gauss}$ = ComputeGaussianDistrib $(X_{min}[H_{index}], H_{matrix})$ // using Eq. 4
14. Add $Value_{Gauss}$ to $Samples$
15. **end for**
16. **return**$Samples$

4. Experiment and result discussion

4.1. Dataset description

This study's dataset uses data collected and processed by Hospital Israelita Albert Einstein through the results of reverse transcription-polymerase chain reaction (RT-PCR) (Data4u, 2020) the source code can be access with link https://github.com/praswibowo/TRIM-SBR. The dataset is public access and anonymous, with 111 features containing 5,644 patient data derived from patient examination results and laboratory data. The summary of the dataset can be seen in Table 1.

As shown in Table 1, a lot of missing data is due to medical personnel's decision-making, which requires a complex process such as taking each patient's medical records, complaints suffered, and laboratory results. This dataset's most crucial aspect is the positive dataset's imbalanced features for COVID-19, respectively 10% and 90% for negative COVID-19 patients.

Table 1
Characteristics of the COVID-19 dataset.

Characteristic	Value
Total data	5,644
Number of features	111
Data comparison negatives: positives	90:10
Data frequency null on feature	65%–100%

4.2. Preprocessing and data transformation

The initial step is to investigate how much the distribution of negative and positive data for COVID-19 patients is based on the variable SARS-COV-2 exam result. The results show that the positive data value is 0.09886, and the negative data value is 0.90113. Therefore, it concluded that the data is imbalanced. Afterward, the percentage of null data in each feature variable to be processed is determined. Removing features that contain more than 50% null data is the right choice to get a reliable dataset (Salgado et al., 2016). About 65% of the data from 111 features contained null data, so the data cleaning was done by removing variable data containing null data. The data cleaning process resulted in 39 variable data that can be processed further.

In this dataset, some variables can be combined into new variables based on simplifying the patient's disease variables. Eighteen disease features can be used as a new variable. This feature is categorical, which contains whether or not a person is affected by a disease. One Hot Encode is a way to convert categorical features to binary variables by creating additional variables to differentiate between the various feature categories. This technique has been used in the medical world to simplify datasets that are often too complex when data is not transformed (Wollenstein-Betech et al., 2020; Dickson et al., 2020; Schwab et al., 2020). The result of a transformation of 18 disease features became a new feature named *has_disease*.

The next step is to check the null data in each row of the dataset. Line-based data checking occurs because the data experiences Missing Completely at Random (MCAR) and Missing at Random (MAR) (Salgado et al., 2016). MCAR occurs when observational data is missing and is not related to specific values obtained through observational data. MAR occurs when probability data is missing on observational data, which still has a dependence on observational data but is not related to specific feature data. Data from MCAR or MAR can be deleted to simplify the features that will be used for modelling. One of the techniques often used in overcoming the problem of missing data is listwise or case deletion (Kang, 2013; Newman, 2014). This technique will eliminate all existing data on the row. The first check was carried out on 5,644 rows of data, resulting in about 3,596 rows of data with 32 null feature variables. The data were then cleaned with the rule that if there are less than 26 feature variables filled in the data, the data will be deleted. The final result left 1,588 data ready to be processed to look for important variables using the correlation function.

The correlation feature used the Spearman technique since it can handle categorical and numeric features simultaneously (Khamis, 2008). Several variables have a strong relationship with each other. In the SARS-COV-2 exam result variable, crucial variables can be seen in this dataset. The variables Leukocytes, Platelets, Eosinophils, and *has_disease* have values close to the minus area, in contrast to the Monocytes and age variables showing a positive correlation. From the resulting correlation, two feature variables are highly correlated with each other such as Hematocrit, Hemoglobin, and Red Blood Cells. Also Mean Corpuscular Hemoglobin (MCH) and Mean Corpuscular Volume (MCV) so that both can be reduced to decrease the number of feature variables processed by the data set.

Based on the rule of thumb from feature correlation (Hinkle et al., 2003), the Hematocrit, Hemoglobin, and Red blood cells have a high correlation with each other. A high assessment was obtained because the features of Hematocrit, Hemoglobin, and Red blood cells were included in the Hematologic parameters so that the highest value was taken based on the SARS-COV-2 exam result, i.e., Red blood cell (Gligoroska et al., 2020).

There are also highly correlated features: Mean corpuscular hemoglobin (MCH) and Mean corpuscular volume (MCV). These two variables are included in the Red blood Cell Indices, so the highest value was chosen based on the SARS-COV-2 exam result, i.e., the Mean corpuscular volume (MCV) used in this test (Von Tempelhoff et al., 2016). Fig. 3 is the final result of the preprocessing using Principle Component Analysis (PCA) to reduce the dataset's dimensions to visualize data efficiently.

4.3. Modeling

The proposed method was compared with four different resampling techniques: random oversampling (ROS), SMOTE, Borderline-SMOTE, and ADASYN. In addition, this study used two different types of models, namely: random forest (RF), Logistic Regression (LR), and support vector machine (SVM). To minimize bias towards the model, hyperparameter selection and optimization were used (Wong et al., 2019; Schwab et al., 2020). For each prediction model, hyperparameter optimization was carried out by selecting based on a predetermined list ranges, as shown in Table 2. Each hyperparameter optimization performance was evaluated, and the best candidate was chosen in each model of the test set.

4.4. Evaluation metrics

There are four characteristics of values commonly used in matrices, namely: the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). TP and TN mean the number of samples from the test set that are correctly classified as positive and negative. In contrast, FN and FP represent the number of samples from the test set that are mistakenly classified as negative and positive. The usual evaluation criterion for classification is to use accuracy. This metric provides a comprehensive picture when the dataset in size is relatively balanced between classes. In the metric accuracy section, the percentage of sample data is calculated correctly as defined in Eq. (5)

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

Specificity aims to measure how much negative actual sample data is predicted to be negative as defined in Eq. (6)

$$Specificity = \frac{TN}{TN + FP}$$

Sensitivity/Recall intends to measure how much positive actual sample data is predicted to be positive as defined in Eq. (7)

$$Sensitivity/Recall = \frac{TP}{TP + FN}$$

Afterward, measurement of how much positive predictive data is a positive prediction called precision as defined in Eq. (8)

$$Precision = \frac{TP}{TP + FP}$$

In the case of imbalanced data, getting high precision and recall is very difficult and often happens where one of the models gets a high value on one metric, but the other is very low. The F-Score is a metric that interprets the average weights of precision and recall to get a balanced result from the two metrics as defined in Eq. (9)

$$F - Measure = 2\frac{precision.recall}{precision + recall}$$

To achieve good results for both classes, sufficient positive and negative class indicators are combined using the ROC curve. ROC is a chart type that demonstrates the TP rate against the FP rate. The ROC also shows that each classifier can increase true positives without increasing false positives. The area under the ROC (AUC) curve is a value that tries to tell how well the model is at differen-

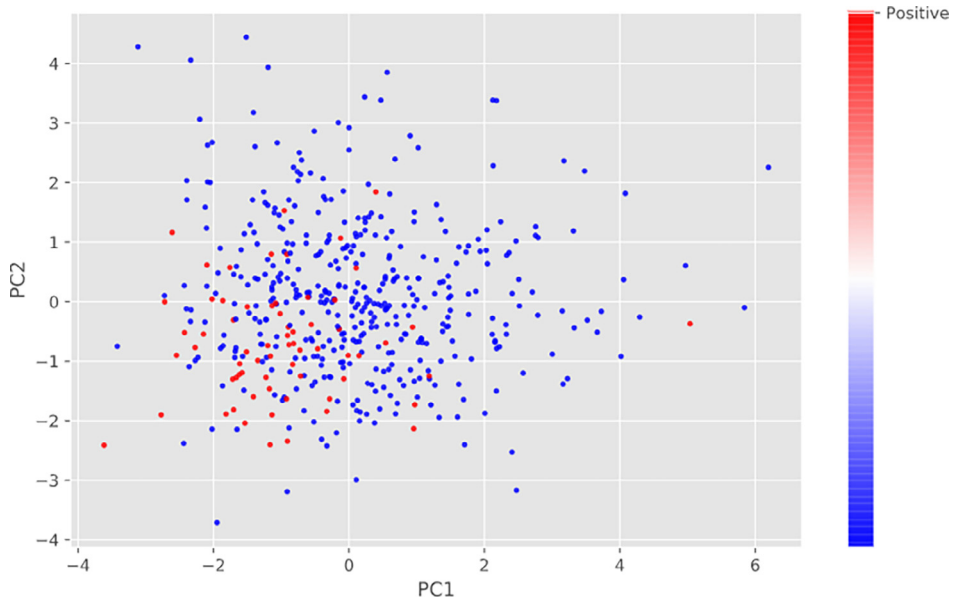

Fig. 3. The distribution of COVID-19 data set with the red is minority class and the blue is majority class.

Table 2
Hyperparameter ranges used for the experiment.

Model and hyperparameter	Choice
Random forest	
Number of trees	10, 50, 100, 200, 500
Features	auto, sqrt, log2, 0.5, 0.1, 0.3
Depth of the tree	2, 8, 16, 32, 64, 128
Number of samples	2, 4, 8, 16, 24
Number of leaf	1, 2, 5, 10, 15, 30
Logistic Regression	
Penalty	l1, l2
Regularization strength C	100, 10, 1, 0.1, 0.01, 0.001
Support vector machine	
Regularization C	0.1, 1, 10, 100, 1000
Kernel coefficient	auto, 1, 0.1, 0.01, 0.001, 0.0001
Kernel type	linear, poly, rbf, sigmoid

tiating labelling between classes. The higher the value obtained, the better the model will differentiate between class labels. Calculated using Eq. (10) with TP rate is the percentage of 'positive' TP instances classified as 'positive', and FP rate is the percentage of 'negative' instances classified as 'positive'.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

4.5. Result discussion

The results of the distribution of the TRIM-SBR class are shown in Fig. 4. TRIM-SBR uses the Gaussian distribution in duplicating data so that the distribution results are given more equally. This data distribution increases the data variance so that the modelling assessment can be improved. The interesting thing from TRIM-SBR is that it shows a series of minority data duplications in the outer areas of the data distribution center resulting in a strengthening of the minority data region.

Table 3 shows the composition of the majority and minority data on the training data processed by oversampling. In the original data, the number of class data is not balanced with one another, with the majority data percentage of 86.27% and minority data of 13.72%. This is a problem because unbalanced data can cause bias in the model being created; in this case, the oversampling techniques used to provide satisfactory results in balancing class data. It can be seen that TRIM-SBR, ROS, SMOTE, and Borderline-SMOTE give balanced results with the percentage of majority and minority data was 50%. In contrast to the results given by ADASYN, it turned out to be more duplicated by the minority than the majority's results. This happened because ADASYN made duplications based on the density of minority data. When the minority data still did not enter into the balanced criteria by ADASYN, it would be duplicated until it was felt that the minority data were balanced based on the density distribution. Table 3 shows that all oversampling techniques have succeeded in balancing minority data to be equivalent to majority data.

Table 4 displays the results of classification Accuracy, Specificity, Sensitivity, F-measure, and AUC of the five oversampling methods on the COVID-19 dataset. Accuracy has a role in identifying the predictive value that has a relationship with the predictive response. The overall model results in 83%–91% with the best oversampling method held by Borderline-SMOTE with 91.74% results. The comparison between Borderline-SMOTE and TRIM-SBR results has a difference of about 3.3%, which shows that the TRIM-SBR is very competitive with the best results from this experiment.

Specificity tries to find out how many relevant values are predicted to be correct for all people who are, in fact, healthy. The results obtained are based on evaluations between 82%–97% with the best oversampling method held by ROS and Borderline-SMOTE, i.e., 97.12%. Compared to the TRIM-SBR, there was a difference of 7.70%, indicating that the proposed method can be compared with other oversampling methods. High Borderline-SMOTE results were obtained from duplicating data using the concept of nearest neighbors, thereby increasing the similarity of data between neighbors.

Sensitivity is a critical evaluation metric in this study because it minimizes false negatives in the model used. The results obtained were between 47%–82%, with the best oversampling method owned by TRIM-SBR and ADASYN, i.e., 82.35%. The high sensitivity value of TRIM-SBR was due to the varied and even replication of the data.

Precision and recall have a characteristic that is interdependent with each other. The F1-score value is a metric that interprets the average weights of precision and recalls to obtain a balanced result for the two metrics. The experimental results were between 57%–66% with the best oversampling method held by TRIM-SBR, SMOTE,

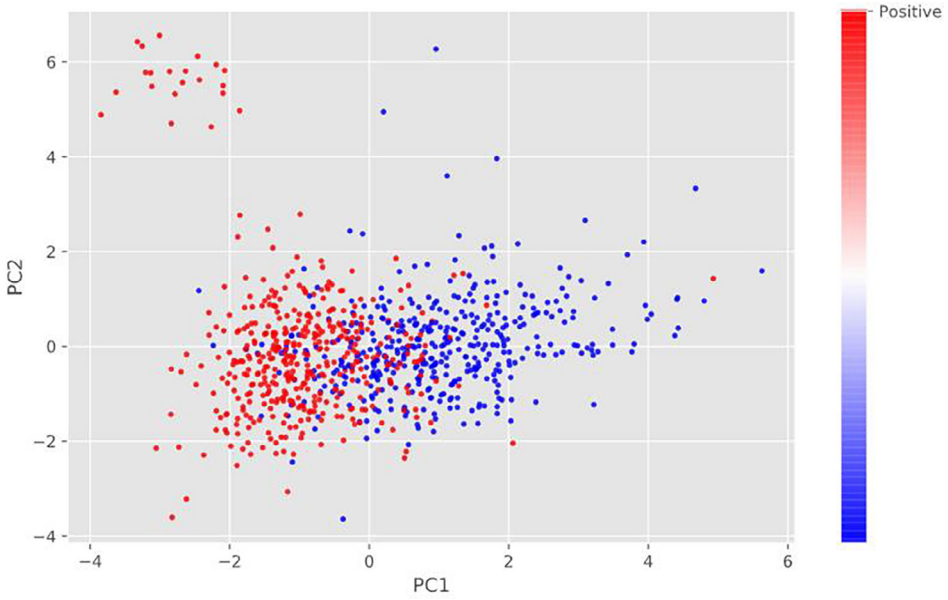

Fig. 4. The data class distribution of TRIM-SBR.

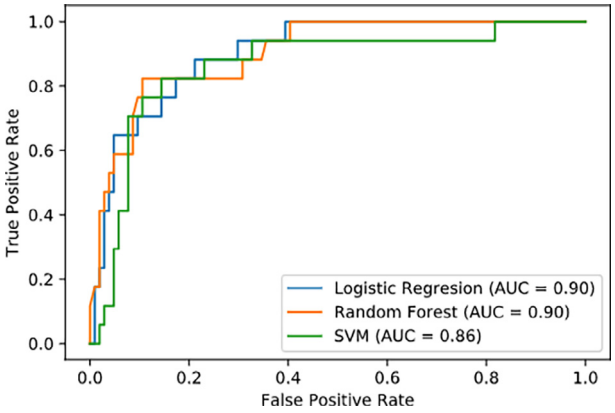

Fig. 5. ROC curve comparison TRIM-SBR.

Table 3
The results of the composition of the training data oversampling.

Oversampling	Attributes	#Class (maj; min)	%Class (maj; min)
Original data	13	(415; 66)	(86.27; 13.72)
TRIM-SBR	13	(415; 415)	(50.00; 50.00)
ROS	13	(415; 415)	(50.00; 50.00)
SMOTE	13	(415; 415)	(50.00; 50.00)
Borderline-SMOTE	13	(415; 415)	(50.00; 50.00)
ADASYN	13	(415; 417)	(49.87; 50.12)

and Borderline-SMOTE, i.e., 66.67%. This means that the TRIM-SBR is very competitive with all oversampling methods. As in the previous explanation, TRIM-SBR tries to generalize all minority data so that the trade-off effect between precision-recall is very influential on the evaluation results of this method. Because the characteristics of the imbalance dataset with very high spread data, causing all results from machine learning models to be below 67%.

The AUC value tries to find out how well the model distinguishes between class labels. The higher the value obtained, the better the model will be in distinguishing between class labels. The experimental results of the AUC ranged between 82%–90%, with the best results being held by TRIM-SBR, i.e., 90.41%. Fig. 5 shows the AUC classifier of TRIM-SBR, which is between 86%–90%, demonstrating that TRIM-SBR can consistently maintain data generalization. The TRIM-SBR method enriches the diversity of synthetic data with data distribution to improve class differentiation.

The subsequent comparison in this experiment compares the proposed model with the state-of-the-art from other researchers, which can be seen in Table 5. Using the same dataset from Hospital Israelita Albert Einstein, the main difference is the preprocessing method and the model used. The proposed method gave the best results compared to the models produced by other researchers, with the results of Accuracy, Specificity, Sensitivity, and AUC were 88.43%, 89.42%, 82.35%, 90.41%, respectively. This high result was obtained because it preprocessed data by cleaning data that did not significantly impact the model, and the addition of TRIM-SBR strengthened the minority class, which often if there is data imbalance it can cause modeling to tend to detect the majority class.

The proof that oversampling can improve modeling results is also shown in the second model results using SMOTEBoost with the results of Specificity, Sensitivity, and AUC were 85.98%, 70.25%, 86.78%, respectively. However, when a comparison is made with the proposed technique, it shown the Sensitivity value is the biggest weakness in this modeling because the generalization of the data generated by SMOTEBoost is very minimal. Even though oversampling can improve modeling results, choosing an incorrect technique can result in an inadequate evaluation. For example, the third model that uses the SMOTE technique shows a Sensitivity value of 43.00%. Sensitivity detects how many predicted patients labelled positive for COVID-19, so a low value indicates the difficulty to distinguish between patients infected with COVID-19 or not for the model.

When discussing preprocessing data, a method chooses direct data sampling, as shown in the fourth model. The results are based on the evaluation of the Specificity, Sensitivity, and AUC metrics 80.00%, 80.60%, and 84.20%, respectively. This modeling method has a weakness in the minimum variation of data which can cause a minimal scope of the data to be processed. However, with too much data and features processed, it also affects the used modeling. The fifth model tries to process all the data contained in the dataset. The evaluation results of Specificity, Sensitivity, and AUC were 49.00%, 75.00%, and 66.00%, respectively. This result was obtained because there was no oversampling technique to balance the imbalanced data to affect the specificity of the model being processed.

Table 4
Experiment result comparison.

Model	Oversampling	Accuracy (%)	Specificity (%)	Sensitivity (%)	F1-score (%)	AUC (%)
Random Forest	TRIM-SBR	88.43	89.42	82.35	66.67	90.41
	ROS	90.08	97.12	47.06	57.14	90.27
	SMOTE	90.91	95.19	64.71	66.67	90.36
	Borderline-SMOTE	91.74	97.12	58.82	66.67	89.99
	ADASYN	90.08	96.15	52,94	60.00	89.17
Logistic Regression	TRIM-SBR	83.47	84.62	76.47	66.52	90.38
	ROS	85.12	86.54	76.47	59.09	89.2
	SMOTE	85.12	87.50	70.59	57.14	89.2
	Borderline-SMOTE	83.47	85.58	70.59	54.55	89.88
	ADASYN	82.64	82.69	82.35	57.14	89.08
Support Vector Machine	TRIM-SBR	86.78	88.46	76.47	61.9	86.37
	ROS	87.6	91.35	64.71	59.46	86.99
	SMOTE	88.43	91.35	70.59	63.16	83.6
	Borderline-SMOTE	86.78	90.38	64.71	57.89	82.41
	ADASYN	85.95	88.46	70.59	58.54	85.61

Table 5
Comparison of the proposed model with other state-of-the art researchers.

Model	Dataset Size (Selection)	Total Feature (Selection)	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC (%)
TRIM-SBR with RF Hyperparameter	5644 (481)	111 (13)	88.43	89.42	82.35	90.41
SMOTEBoost with SVM (Soares et al., 2020)	5644 (599)	111 (16)	–	85.98	70.25	86.78
SMOTE with artificial neural networks (ANN) (Banerjee et al., 2020)	5644 (598)	111 (14)	87.00	91.00	43.00	80.00
Gradient Boost Tree (Batista et al., 2020)	256 (256)	111 (15)	–	80.00	80.60	84.20
Gradient Boost Hyperparameter (Schwab et al., 2020)	5644 (5644)	111 (106)	–	49.00	75.00	66.00

5. Conclusion

This paper proposed a new oversampling method called TRIM-SBR that reduces over-generalization, which usually occurs when synthetic data is formed into the majority class area with even distribution of synthetic data. The proposed method consists of three steps. First, the calculation uses a greedy approach called TRIM to search for minority data sets while filtering irrelevant data. Second, TRIM's optimal set forms a seed to be processed at the data duplication stage. Third, smoothed bootstrap approach is used to establish new synthetic data. The proposed method was evaluated using five different evaluation metrics, namely: Accuracy, Specificity, Sensitivity, F-measure, and AUC. The experimental results show that the proposed method succeeded in obtaining a sensitivity value of 82.35% and an F1-score of 66.67%. In addition, the proposed method achieved a value of AUC 90.41% when compared with other oversampling techniques. Although TRIM-SBR successfully resolved data imbalance issues, some drawbacks need to be strengthened in the future. By incorporating an algorithm to concentrate on a particular area while duplicating data to boost this process evaluation results. Furthermore, it also shows comparative evidence with other studies to prove the reliability of the proposed model. Selection of the appropriate preprocessing and oversampling technique can improve the metric evaluation results of the modelling. The experiments show that the proposed model results are superior to the results of another models. There are still many ways to do the preprocessing, one of which is by performing statistical imputation of data. The missing data is input based on the statistical calculation of the variable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdi, L., Hashemi, S., 2015. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. Soft Comput. 19, 3369–3385. https://doi.org/10.1007/s00500-014-1291-z.

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L., 2020. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology 2019, 200642. https://doi.org/10.1148/radiol.2020200642.

Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying Support Vector Machines to Imbalanced Datasets. Eur. Conf. Mach. Learn., 39–50

Almutairi, W.A., Janicki, R., 2020. On relationships between imbalance and overlapping of datasets. Epic Ser. Comput. 69, 141–150. https://doi.org/10.29007/h71z.

Armstrong, S., 2020. Covid-19: Tests on students are highly inaccurate, early findings show. BMJ 371, m4941. https://doi.org/10.1136/bmj.m4941.

Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., Baker, M., Mackenzie, L.S., 2020. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. Int. Immunopharmacol. 86, 106705. https://doi.org/10.1016/j.intimp.2020.106705.

Batista, A.F. de M., Miraglia, J.L., Donato, T.H.R., Filho, A.D.P.C., 2020. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv 2020.04.04.20052092. 10.1101/2020.04.04.20052092

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2005. Balancing strategies and class overlapping. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 3646 LNCS, 24–35. 10.1007/11552253_3

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6, 20–29. https://doi.org/10.1145/1007730.1007735.

Bowman, A.W., Azzalini, A., 1999. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations. J. Am. Stat. Assoc. 94, 982. https://doi.org/10.2307/2670015.

Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C., 2009. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 5476 LNAI, 475–482. 10.1007/978-3-642-01307-2_43

Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C., 2012. DBSMOTE: Density-based synthetic minority over-sampling technique. Appl. Intell. 36, 664–684 https://doi.org/10.1007/s10489-011-0287-y.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., Zhang, L., 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan,

China: a descriptive study. Lancet 395, 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7.
Data4u, E., 2020. Diagnosis of COVID-19 and its clinical spectrum AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr).
Deeks, J.J., Raffle, A.E., 2020. Lateral flow tests cannot rule out SARS-CoV-2 infection. BMJ 371, 1–2. https://doi.org/10.1136/bmj.m4787.
Dennie, C., Hague, C., Lim, R.S., Manos, D., Memauri, B.F., Nguyen, E.T., Taylor, J., 2020. Canadian Society of Thoracic Radiology/Canadian Association of Radiologists Consensus Statement Regarding Chest Imaging in Suspected and Confirmed COVID-19. Can. Assoc. Radiol. J. 0846537120924606. https://doi.org/10.1177/0846537120924606.
Dickson, J., Griffin, M., Alderson, D., Taylor, J., Mealy, K., Allum, B., 2020. Guidelines for pre-operative COVID-19 testing for elective cancer surgery.
Douzas, G., Bacao, F., Last, F., 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf. Sci. (Ny) 465, 1–20. https://doi.org/10.1016/j.ins.2018.06.056.
Dua, Dheeru, Graff, C., 2019. UCI Machine Learning Repository [WWW Document]. URL http://archive.ics.uci.edu/ml
E. Hinkle, D., Wiersma, W., G. Jurs, S., 2003. Applied statistics for the behavioral sciences.
Elyan, E., Moreno-Garcia, C.F., Jayne, C., 2021. CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural Comput. Appl. 33, 2839–2851. https://doi.org/10.1007/s00521-020-05130-z.
Estabrooks, A., Jo, T., Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20, 18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x.
Ferguson, J., Dunn, S., Best, A., Mirza, J., Percival, B., Mayhew, M., Megram, O., Ashford, F., White, T., Moles-Garcia, E., Crawford, L., Plant, T., Bosworth, A., Kidd, M., Richter, A., Deeks, J., McNally, A., 2021. Validation testing to determine the sensitivity of lateral flow testing for asymptomatic SARSCoV-2 detection in low prevalence settings: Testing frequency and public health messaging is key. PLoS Biol. 19, 1–9. https://doi.org/10.1371/journal.pbio.3001216.
Fernández, A., García, S., Herrera, F., Chawla, N.V., 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. J. Artif. Intell. Res. 61, 863–905. https://doi.org/10.1613/jair.1.11192.
Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. Int. J. Emerg. Technol. Adv. Eng. 2, 42–47.
García, V., Alejo, R., Sánchez, J.S., Sotoca, J.M., Mollineda, R.A., 2006. Combined effects of class imbalance and class overlap on instance-based classification. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 4224 LNCS, 371–378. 10.1007/11875581_45
Gligoroska, J.P., Gontarev, S., Maleska, V., Efremova, L., Stojmanova, D.S., Manchevska, S., 2020. Red blood cell variables and correlations with body mass components in boys aged 10-17 years. Turk. J. Pediatr. 62, 53–60. 10.24953/turkjped.2020.01.008
Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Adv. Intell. Syst. Comput., 878–887 https://doi.org/10.1007/11538059_91.
He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proc. Int. Jt. Conf. Neural Networks 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969.
He, H., Garcia, E.A., 2009. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
Hope, M.D., Raptis, C.A., Shah, A., Hammer, M.M., Henry, T.S., 2020. A role for CT in COVID-19? What data really tell us so far. Lancet 395, 1189–1190. https://doi.org/10.1016/S0140-6736(20)30728-5.
Kang, H., 2013. The prevention and handling of the missing data. Korean J. Anesthesiol. 64, 402–406. https://doi.org/10.4097/kjae.2013.64.5.402.
Khamis, H., 2008. Measures of association: How to choose? J. Diagnostic Med. Sonogr. 24, 155–162. https://doi.org/10.1177/8756479308317006.
Kmietowicz, Z., 2021. Covid-19: Controversial rapid test policy divides doctors and scientists. BMJ 372, n81. https://doi.org/10.1136/bmj.n81.
Laghi, A., 2020. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. Lancet Digit. Heal. 2, e225. https://doi.org/10.1016/S2589-7500(20)30079-0.
Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N., 2018. A survey on addressing high-class imbalance in big data. J. Big Data 5. https://doi.org/10.1186/s40537-018-0151-6.
Long, C., Xu, H., Shen, Q., Zhang, X., Fan, B., Wang, C., Zeng, B., Li, Z., Li, X., Li, H., 2020. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? Eur. J. Radiol. 126, 108961. https://doi.org/10.1016/j.ejrad.2020.108961.
Longadge, R., Dongre, S., 2013. Class Imbalance Problem in Data Mining Review. Eur. J. Intern. Med. 24, e256.
Lu, H., Xu, Y., Ye, M., Yan, K., Gao, Z., Jin, Q., 2019. Learning misclassification costs for imbalanced classification on gene expression data. BMC Bioinformatics 20, 1–10. https://doi.org/10.1186/s12859-019-3255-x.
Mahase, E., 2020. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ 368, m641. https://doi.org/10.1136/bmj.m641.
Menardi, G., Torelli, N., 2014. Training and assessing classification rules with imbalanced data. Data Mining Knowl. Discov. https://doi.org/10.1007/s10618-012-0295-5.
Nanni, L., Fantozzi, C., Lazzarini, N., 2015. Coupling different methods for overcoming the class imbalance problem. Neurocomputing 158, 48–61. https://doi.org/10.1016/j.neucom.2015.01.068.
Newman, D.A., 2014. Missing Data: Five Practical Guidelines. Organ. Res. Methods 17, 372–411. https://doi.org/10.1177/1094428114548590.
Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E., 2020. Imbalance Problems in Object Detection: A Review. IEEE Trans. Pattern Anal. Mach. Intell. 1–1. https://doi.org/10.1109/tpami.2020.2981890.
Puntumapon, K., Rakthamamon, T., Waiyamai, K., 2016. Cluster-based minority over-sampling for imbalanced datasets. IEICE Trans. Inf. Syst. E99D, 3101–3109. https://doi.org/10.1587/transinf.2016EDP7130.
Puntumapon, K., Waiyamai, K., 2012. A pruning-based approach for searching precise and generalized region for synthetic minority over-sampling. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 7301 LNAI, 371–382. 10.1007/978-3-642-30220-6_31
Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F.J., Granda-Gutiérrez, E.E., 2020. Data sampling methods to dealwith the big data multi-class imbalance problem. Appl. Sci. 10. https://doi.org/10.3390/app10041276.
Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M., 2016. Missing Data, in: Secondary Analysis of Electronic Health Records. Springer International Publishing, Cham, pp. 143–162. 10.1007/978-3-319-43742-2_13
Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H., Santos, J., 2018. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. IEEE Comput. Intell. Mag. 13, 59–76. https://doi.org/10.1109/MCI.2018.2866730.
Schwab, P., DuMont Schütte, A., Dietz, B., Bauer, S., 2020. Clinical Predictive Models for COVID-19: Systematic Study. J. Med. Internet Res. 22, e21439. https://doi.org/10.2196/21439.
Silverman, B.W., 1986. Density estimation for statistics and data analysis. CRC Press.
Soares, F., Villavicencio, A., Fogliatto, F.S., Rigatto, M.H.P., Anzanello, M.J., Idiart, M., Stevenson, M., 2020. A novel specific artificial intelligence-based method to identify {COVID}-19 cases using simple blood exams. medRxiv 2020.04.10.20061036. 10.1101/2020.04.10.20061036
Stefanowski, J., 2013. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. Smart Innov. Syst. Technol. 13, 277–306. https://doi.org/10.1007/978-3-642-28699-5_11.
Vaid, A., Somani, S., Russak, A.J., De Freitas, J.K., Chaudhry, F.F., Paranjpe, I., Johnson, K.W., Lee, S.J., Miotto, R., Zhao, S., Beckmann, N., Naik, N., Arfer, K., Kia, A., Timsina, P., Lala, A., Paranjpe, M., Glowe, P., Golden, E., Danieletto, M., Singh, M., Meyer, D., Reilly, P.F., Huckins, L.H., Kovatch, P., Finkelstein, J., Freeman, R.M., Argulian, E., Kasarskis, A., Percha, B., Aberg, J.A., Bagiella, E., Horowitz, C.R., Murphy, B., Nestler, E.J., Schadt, E.E., Cho, J.H., Cordon-Cardo, C., Fuster, V., Charney, D.S., Reich, D.L., Bottinger, E.P., Levin, M.A., Narula, J., Fayad, Z.A., Just, A., Charney, A.W., Nadkarni, G.N., Glicksberg, B.S., 2020. Machine Learning to Predict Mortality and Critical Events in COVID-19 Positive New York City Patients. medRxiv 2020.04.26.20073411. 10.1101/2020.04.26.20073411
Von Tempelhoff, G.F., Schelkunov, O., Demirhan, A., Tsikouras, P., Rath, W., Velten, E., Csorba, R., 2016. Correlation between blood rheological properties and red blood cell indices(MCH, MCV, MCHC) in healthy women. Clin. Hemorheol. Microcirc. 62, 45–54. https://doi.org/10.3233/CH-151944.
Vuttipittayamongkol, P., Elyan, E., 2020. Improved Overlap-based Undersampling for Imbalanced Dataset Classification with Application to Epilepsy and Parkinson's Disease. Int. J. Neural Syst. 30. https://doi.org/10.1142/S0129065720500434.
Wang, S., Sun, J., Mehmood, I., Pan, C., Chen, Y., Zhang, Y.D., 2020. Cerebral micro-bleeding identification based on a nine-layer convolutional neural network with stochastic pooling. Concurr. Comput. 32, 1–16. https://doi.org/10.1002/cpe.5130.
Wang, S.H., Xie, S., Chen, X., Guttery, D.S., Tang, C., Sun, J., Zhang, Y.D., 2019. Alcoholism identification based on an Alexnet transfer learning model. Front. Psychiatry 10, 1–13. https://doi.org/10.3389/fpsyt.2019.00205.
Wibowo, P., Fatichah, C., 2021. An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset. Regist. J. Ilm. Teknol. Sist. Inf. 7, 63–71. https://doi.org/10.26594/register.v7i1.2206.
Wollenstein-Betech, S., Cassandras, C.G., Paschalidis, I.C., 2020. Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator. medRxiv. 10.1101/2020.05.03.20089813
Wong, J., Manderson, T., Abrahamowicz, M., Buckeridge, D.L., Tamblyn, R., 2019. Can Hyperparameter Tuning Improve the Performance of a Super Learner?: A Case Study. Epidemiology 30, 521–531. https://doi.org/10.1097/EDE.0000000000001027.
World Health Organization (WHO), 2020. Antigen-detecting rapid diagnostic tests.
Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Yang, Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Ying, Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Yaru, Huang, S., Tan, X., Huang, N., Jiao, B., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H., Yuan, Y., 2020. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. medRxiv 2020.02.27.20028027. 10.1101/2020.02.27.20028027
Zhang, J., Lu, H., Chen, W., Lu, Y., 2011. A comparison study of cost-sensitive learning and sampling methods on imbalanced data sets. Adv. Mater. Res. 271–273, 1291–1296. https://doi.org/10.4028/www.scientific.net/AMR.271-273.1291.
Zheng, Z., Yao, Z., Wu, K., Zheng, J., 2020. The Diagnosis of Pandemic Coronavirus Pneumonia: A Review of Radiology Examination and Laboratory Test. J. Clin. Virol. 104396. https://doi.org/10.1016/j.jcv.2020.104396.