# Gradually Generative Adversarial Networks Method for Imbalanced Datasets

Muhammad Misdram[1]*, Muljono[2]*, Purwanto[3], Edi Noersasongko[4]
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 51031 Indonesia[1, 2, 3, 4]
Faculty of Information Technology Merdeka University Pasuruan 67129 Indonesia[1]

***Abstract*—Imbalanced dataset can cause obstacles to classification and result in a decrease in classification performance. There are several methods that can be used to deal the data imbalances, such as methods based on SMOTE and Generative Adversarial Networks (GAN). These methods are used for overcoming data oversampling so that the amount of minority data can increase and it can reach a balance with the majority data. In this research, the selected dataset is classified as a small imbalanced dataset of less than 200 records. The proposed method is the Gradually Generative Adversarial Network (GradGAN) model which aims to handle data imbalances gradually. The stages of the GradGAN model are adding the original minority dataset gradually so that it will create new minority datasets until a balance of data is created. Based on the algorithm flow described, the minority data is multiplied by the value of the variable that has been determined repeatedly to produce new balanced minority data. The test results on the classification of datasets from the GradGAN model produce an accuracy value of 8.3% when compare to that without GradGAN.**

***Keywords*—Classification; imbalance; GAN model; GradGAN model; significant oversampling***

## I. INTRODUCTION

The data mining solving on classification is a research topic that still needs contribution [1]. Because every use the dataset often result in data imbalances which can reduce the performance both of images and classification accuracy [2], [3]. In research on data imbalance, many methods have been offered but the method's robustness has not been satisfactory. This phenomenon is shown by existence of several data classification results that are less than [4], [5]. There are several cases of imbalanced data, for example, in the medical field, regarding disease prediction [6] and hepatitis diseases detection [7]. Imbalanced data classification is a vital problem [8], and the key is to create a flexible and correct method of classifying minority and majority data. This raises an urgent need for a better solution to the data imbalance problem so that the classification process can be more optimal.

The sampling method can be carried out at the preprocessing stage, for example, in the case of missing values, which has less significant impact on machine learning outcomes [9]. In the majority class, the use of under-sampling technique is relatively easier to do to balance the data [10]. Whereas in the minority class, the over-sampling method that commonly uses Synthetic Informative Minority Over-Sampling (SMOTE) algorithm to find synthetic data, is currently effective [11], [12]. Another over-sampling method for dealing with synthetic data is Borderline-SMOTE [13]. There is a drawback of the SMOTE algorithm, namely there is no consideration of neighboring information from minority class samples, resulting in over generation. The Modified Fuzzy-Neighbor Weighted Algorithm has also been proposed, the classification results are also better [14]. Although existing classification methods have been able to overcome data imbalances, a new approach is still needed to overcome the challenges of the problem of heavy imbalanced data flows [15]. The online ensemble learning algorithm has also been used for unbalanced data streams [16]. Recently, a relatively new method has been developed to overcome over-sampling techniques for imbalanced data, namely the Generative Adversial Network (GAN) method; the goal is to overcome the difficulties of minority data samples to be more balanced with the majority data [4], [17]. Initially, the GAN model was implemented in deep learning machines, namely about human faces, adopting images so as to produce better images [18]. Another implementation of GAN, is that it models complex real-world image data and normalizes data imbalances [19]. There are several developments of GAN models, including Triple Generative Adversarial Nets (TripleGANs) [20], and Senti Generative Adversarial Networks (SentiGAN) [21]. The development of GANs is not only for deep learning but also addressing imbalances of machine learning in data mining. For example, to classify public data of Bank Marketing, Credit, and Lending Club from the UCI repository, the GAN model and the random forest method produce several sub-classes of sample data and these sub-classes will be combined with the original dataset to form a new minority dataset [4]. Generative Adversarial Networks (GAN) model is also used to create new text from the scarcity of the Dialectal Arabic dataset which generates annotations (notes or comments) with automatic generation [21].

Learning methods that are commonly used and have good performance are Naïve Bayes (NB), k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), and Decision Tree (D-Tree). NB has the advantage that it is simple and works well in the real world. The weakness is that the dataset to be tested must be in numerical form [22], [23]. K-NN has advantages, which are effective and robust against noisy training data [24]. SVM has the advantage of being able to generalize and is known to produce high accuracy values but is less than optimal when applied to imbalanced data [25]. Meanwhile, the drawback is that it is difficult to implement in large cases and is developed for two-class problems [26]. The D-Tree method has the advantage of a simpler and more straightforward, specific, and flexible way of making decisions, while the

weaknesses are frequent overlap and difficulty in designing a more optimal one [27]. The random forest ensemble applied to predict heart disease produces a better accuracy value than the other methods [28]; besides that, it is applied to the classification of Sarcastic Tweet, resulting in the highest accuracy [22].

Based on the above explanation, to overcome the heavy flow of imbalanced data, this research proposes a Gradually Generative Adversarial Network (GradGAN) model. The GradGAN model works to handle oversampling data with a resampling technique, namely adding minority data gradually. This technique is expected to be optimally applied to all dataset sizes. In future research, the GradGAN model can be developed and relied upon to deal with data imbalance problems. To test the quality of the data resulting from the GradGAN model, data can be classified by using the NB, SVM, k-NN, D-Tree and RF methods because they are common methods and have good classification performance. The dataset used in this research is a small imbalanced dataset that is taken from the UCI repository.

## II. Related Work

So far, the obstacle faced by researchers in handling datasets in machine learning and deep learning applications is data imbalance. The reason is because the data imbalance makes the performance of the machine not optimal. In fact, the results of research on classification show that classification performance is less significant. There are several methods in research to handle the classification and imbalance of datasets.

TABLE I. List of Research Supporting Paper References

| Papers | Classification method | | | | | Imbalanced Method | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| [2] | | | | | | | | | | | | | | ✔ | |
| [3] | | | | | | ✔ | | | | ✔ | | | | | |
| [4] | | | | | ✔ | | | | ✔ | | | | | | |
| [5] | ✔ | ✔ | | ✔ | | | | ✔ | | | | | | | |
| [6] | ✔ | | ✔ | | | ✔ | | | | | | | | | |
| [7] | | | | ✔ | | | | | | | ✔ | | | | |
| [8] | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | | | | | | | | | |
| [9] | ✔ | | ✔ | | | ✔ | | | | | | | | | |
| [10] | | | | | ✔ | | | | ✔ | | | | | | |
| [11] | | | | | | | | | ✔ | | | | | | |
| [12] | | | | | | | | | ✔ | | | | | | |
| [13] | | ✔ | | | | | | | | | | | ✔ | | |
| [14] | | | | | | | | | ✔ | | | ✔ | | | |
| [15] | ✔ | ✔ | ✔ | | | | | | | | | | | | |
| [16] | ✔ | | | | | ✔ | | | | | | | | | |
| [17] | | | | ✔ | | ✔ | | | | | | | | | |
| [18] | | ✔ | ✔ | ✔ | | | | | | | | | | | |
| [19] | | ✔ | ✔ | ✔ | | | | | | | | | | | |
| [20] | | | ✔ | | ✔ | | | | | | | | | | |
| [21] | | | | | ✔ | | | | | | | | | | |
| [22] | | ✔ | | ✔ | ✔ | | | | | | | | | | |
| [23] | | | | | | | | | ✔ | | | | | | |
| [24] | ✔ | | ✔ | | ✔ | ✔ | | | ✔ | | | | | | |
| [25] | | | | | | ✔ | ✔ | | | | | | | | |
| [26] | ✔ | | ✔ | ✔ | | | ✔ | | | | | | | | |
| [27] | | | | | | | | | | | | | | | |
| [28] | | | | | | | | | | | | | | | ✔ |
| Proposed | ✔ | ✔ | ✔ | ✔ | ✔ | GradGAN | | | | | | | | | |

Abbreviations: a)NB; b)SVM ; c)D-Tree; d)k-NN; e)RF; f)SMOTE; g)Borderline-SMOTE; h)Radial-Based Undersampling; i)GAN; j)PWIDB; k)Ada- NN; l)SentiGAN; m)TripleGANs; n)BAGAN;o) ) *LSTM*

The development model in this research refers to the literature research in Table I. The GAN model used in this paper is developed into a Balancing GAN (BAGAN). The BAGAN model requires a large amount of data to function properly [2]. In order to achieve automatic balancing, Piece-Wise Incremental Data Re-Balancing (PWIDB) is combined with the Racing Algorithm (RA) technique and incremental iterative balancing techniques. According to [3], PWIDB outperforms other balancing techniques. The GAN model serves to overcome data imbalance and handle minority data samples. Several data samples based on a random forest model were broken down into sub-sections and then combined with the original data to form new minority data. The most ideal selected dataset is large size. The advantage in classification has produced good AUC and F1 [4]. The Radial-Based Under sampling method is proposed to handle imbalanced data to get good results, by combining the NB, SVM, and k-NN classification [5]. The Synthetic Informative Minority Oversampling (SMOTE) Technique method is used to balance training data by creating artificial minority data. The test results of the Naïve Bayes, SVM-RBF, C4.5, and RIPPER classification methods produce the best accuracy with a value of 89.5%, 90% precision, 89.4% recall, 89.5% F-score and 83.5% Kappa [6]. In this paper, the problem to be solved is to classify the hepatitis dataset that is not balanced.

To increase the accuracy of the k-NN method due to the influence of data imbalance, an Adaptive-Condensed NN (Ada-CNN) method has been developed and then the accuracy value can reach a maximum [10]. In this paper, it is explained that to improve accuracy, it must balance the data, the method used is SMOTE. To find higher accuracy, data has been tested by the common classification methods, namely SVM, Random Forest, KNN, Naïve Bayes, and Decision Tree. Then the general classification methods are compared with the proposed method using a voting multiclassifier; the results are better than the general method [11]. To overcome the lack of ROC values due to imbalance data, this research used the SMOTE method. In order to know the ROC value, the NB and D-Tree methods are used for data classification [12]. To overcome the decline in classification performance due to imbalanced data, this paper applies a Generative Adversarial Network (GAN) model. The workflow begins with creating artificial data so that minority sub-data are formed. Furthermore, from some of the minority sub-data, they are formed together to form a new minority data. Classification algorithm used is the Random Forest (RF) method [17]. Generative Adversarial Network (GAN) model was first proposed by Goodfellow et al. [18]. Invoked as a model to bridge between supervised and unsupervised learning in 2014, it has been hailed as the most exciting ideas in machine learning in the last ten years. To overcome imbalances level in classification, object detection and pixels in segmentation, this paper applies the Generative Adversarial Neural Networks (GANs) model [19]. The GAN model is promising in its application in handling image forms, but there are weaknesses between the generator and discriminator which are not optimal and cannot control the resulting sample. To overcome this problem, research is applied using a generator, discriminator and classifier called the triple Generatif Adversarial Net (Triple-GAN), which has produced a good classification [20]. The GAN model collaborated with sentimental to form sentimental GAN (SentiGAN) has been succeeded in text creation domain. The modified SentiGAN utility will amplify a small data set and produce a variety of high-quality sentences in different Arabic dialects obtained from the MADAR data set. And it can produce a higher number of sentences than the original data, and this method can reduce vocabulary and only use common words. Even though there are slight deficiencies in the resulting text, the key features detected can help the classification remain strong. So the Generator process is not only effective and consistent but can also improve classification when used from the original dataset and the resulting dataset from the model process [21]. To find the accuracy value of the 14 datasets, in this paper, the classification methods used include Naive Bayes (NB), SVM, D-Tree, and k-NN. And from the test results, the NB method has produced the best accuracy value [22]. To overcome class imbalance, in this paper, the algorithm chosen is SMOTE. Meanwhile, the NB method is used for classification. The test results have resulted in an accuracy value of 88.5%, more significant than the R algorithm of 87.5% [23]. To overcome the imbalance problem, the popular SMOTE algorithm is used while the k-NN method is chosen for classification. Imbalanced data test results produce lower accuracy, while data that is balanced has greater accuracy [24]. The case of credit assessment can not be just any method that can be applied because it can hinder the assessment work itself. A good method to use is the SVM method in collaboration with Least Squares SVM (LS-SVM), It is shown from the test results with eight data set that it produces better performance when compared to other methods such as D-Tree and k-NN. But, in this research, SVM is also supported by meta-heuristics to be better [25]. Another research on the treatment of breast cancer uses several classification methods, the best results is the SVM method with an accuracy of 80.4% [26]. In implementing the algorithm, it is very vulnerable to the existence of data, and sometimes it is found that the data is imbalanced. To overcome that case, in this research has utilized decline tree with the D-Tree and Random Forest (FR) methods for the data balance process. The test results show that all classifications are strongly influenced by the balance of the data to achieve maximum results [27]. Detecting heart disease with feature selection and Random Forest Ensemble methods, the resulting accuracy is better by 99% [28]. Comparison of SVM, k-NN, Maximum Entropy and Random Forest methods for classification, the best accuracy value is Random Forest [22]. The GAN model can be applied in several cases to handle dataset imbalances. For example, in the field of financial anti-fraud with smaller data samples, the GAN model can be collaborated with the Long Short-Term Memory (LSTM) network algorithm so that the problem of data imbalance can be taken seriously [28]. Both of these models will share roles to process data in a time sequence completed by the Long Short-Term Memory network, while the GAN is to distribute selected real data which will produce data that is similar to the original data [23]. In this research, the SMOTE algorithm was used to overcome data imbalance, while the best method was chosen to classify, namely NB, D-Tree and RF, for example in the case of breast disease prediction [24]. The class with a small number of observations is called the minority class, while the class with the largest number of observations is called the

majority class. In this paper, to handle data imbalance, the SMOTE and Borderline-SMOTE algorithms were selected, while to measure the accuracy value, they relied on the Safe Level Graph [13]. In the end, the GAN model will produce data similar to the original [23]. In this paper, to overcome data imbalances, the SMOTE algorithm is used, while for classification, the method chosen is D-Tree [26]. The preprocessing process is a step for preparing the dataset to be tested. The new original dataset is usually still not perfect, there are still deficiencies, for example an empty record is found, it is necessary to handle imputation to perfect the dataset. The goal of imputation is to keep the number of datasets ideal. There are several imputation methods to choose from, including Zero, Mean/Median, k-NN, Multivariate Imputation with Chained Equations (MICE), Deep Learning (Datawig) [26].

In the previous research, the GAN model algorithm can be described in Eq. (1), and the algorithm flow [4], [18] as follows:

- Collect m noise samples {z (1).....z (m )} from the noise chamber Pg(z).
- Collect m data samples (x (1).....x (m)) from the Pdata(x) data set.
- Update the discriminator by promoting a random gradient. The specific formula is as follows:

$$\Delta\theta \frac{1}{m} = \sum_{1=0}^{m} \left[ log D^{(i)} + log \left( 1 - D \left( G\left(Z^{(i)}\right) \right) \right) \right] \quad (1)$$

The next step:

- Group m noise samples {z(1)...z(m)} from the noise space Pg(z).
- Refine the generator by reducing the random gradient. The specific formula is as Eq. (2):

$$\Delta\theta \frac{1}{m} = \sum_{1=0}^{m} log \left( 1 - D \left( G\left(Z^{(i)}\right) \right) \right) \quad (2)$$

The GAN flow diagram can be described as Fig. 1.

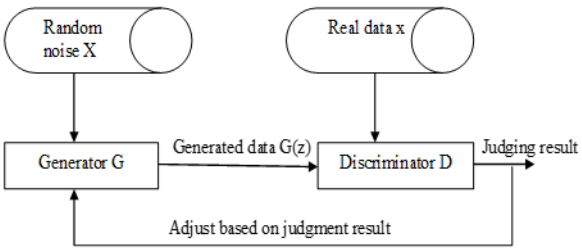


Fig. 1. GAN algorithm flowchart.

The first GAN produces two sub-models namely Generator G (generative model) and Discriminator D (discriminative model). In the initial algorithm, the generative model G based on available noise generates some data. Then Discriminator D will assess whether the data obtained is in the form of real data or data generated by the generative model G. The purpose of the generative model G is to make the resulting data as close as possible to the actual data and cannot be easily identified by the discriminatory model D. The purpose of the discriminator the discriminative model D is to distinguish between real data and data generated by the generative model G [4].

## III. Proposed Methodology

The steps of the research methodology process are carried out, starting from the dataset collection stage to the process stage of producing the expected output from the research. In this research, there are several steps as follows.

### A. Data Collection

The selected dataset are Hepatitis, Immunotherapy, and Echocardiogram medical data, all of are taken from the UCI repository, where all of them are datasets with a relatively small number of less than 200 records and are not balanced [11].

### B. Pre-Processing

The preprocessing process is a step for preparing the dataset to be tested. The new original dataset is usually still not perfect, and there are still deficiencies; for example, if an empty record is found, it is necessary to handle imputation to perfect the dataset. The goal of imputation is to keep the sum of the datasets perfect. There are several imputation methods to choose from; and in this research the imputation method used was the k-NN method. [26].

### C. Proposed Model GradGAN

The GradGAN method is a development and modification of the GAN model. The working pattern of the Gradually Generative Adversarial Network (GradGAN) model is to overcome imbalanced data gradually until it gets balance. The process in GradGAN will involve a generator function to generate random datasets which will create majority and minority data from random samples of the original random dataset. And the discriminator that functions knows that the majority data are original data from the generator results. In this research, the main concern is minority data. Minority data will be processed with the GradGAN model to produce a new minority data sample. The next step is that the minority data will be multiplied by the variable value gradually so that new minority values will be created until a balance value is formed with the original majority data. Then classification is carried out where the new minority data serves as test data and the original majority data as training data. This GradGAN model is a new discovery model and has never been used by other researchers in dealing with imbalanced data. Following is a Fig. 2 of a flowchart and its explanation.

The process in GradGAN will involve a generator function to generate random datasets, which will create majority and minority data from random samples of the original random dataset. Moreover, the discriminator that functions knows that the majority and minority data are original data from the generator results. In this research, the main concern is minority data. Minority data will be processed with the GradGAN model to produce minority data samples. The next step is that the minority data will be multiplied by a variable value gradually so that new minority values will create until a balance value forms with the original or original majority data. Then the two data, namely the new minority data, serve as test

data and the original majority data as training data for classification. This GradGAN model is a new discovery model and has never been used by other researchers in dealing with imbalanced data.

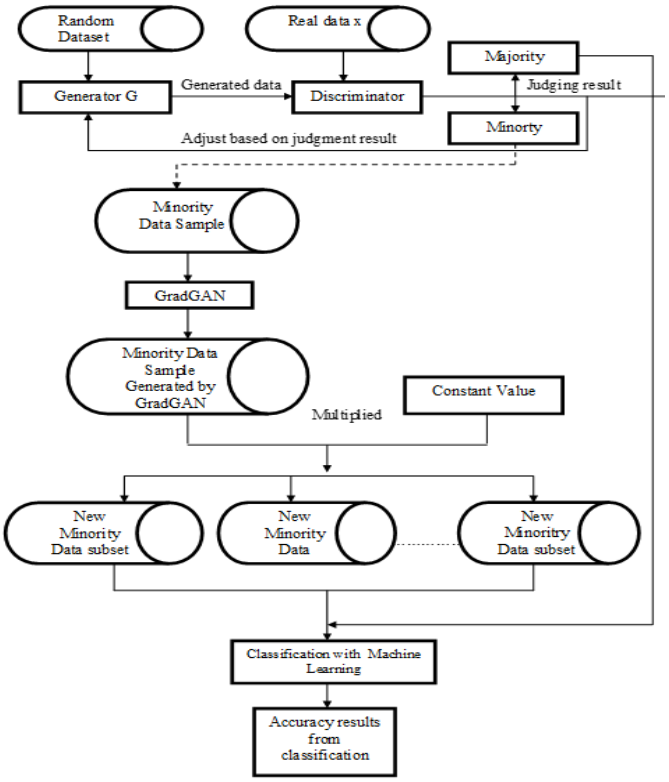


Fig. 2. Flowchart of the GradGAN algorithm.

Based on the above explanation, the GadGRAN model can be explained by calculating the equations in detail as follows:

- A sample set of x random datasets ($d^{(1)}$……$d^{(x)}$)with a total dataset of the $Dx^{(d)}$ variable will be generated to produce a total of majority and minority data.
- Determine the sample set x majority data class ($a^{(1)}$…….$a^{(x)}$) with the number of the $Mx^{(a)}$ variable.
- Determine the sample set x minority data class ($i^{(1)}$……$i^{(x)}$) with the number of the $Mx^{(i)}$ variable.

In this research, the primary concern of data oversampling is to resample minority data with the Mx(i) variable multiplied by a variable value to add minority data called new minority data. The goal is to create new minority data so that there is a balance with the original majority data. A mathematical formula can calculate with the following Eq. (3).

- Determine the sample x new minority data that is close to the balance ($i2^{(1)}$…..$i2^{(x)}$) with the amount of data the Mx(i2) variable.
- Determine the sample x variable value ($k^{(1)}$…..$k^{(x)}$) with the value of the Kx(k) variable.
- Calculate the number of new minority data using Eq. (3) or it can be written with Eq. (4).

$$Mx(i2) = Mx(i)(\sum_{i=0}^{x} Kx(k) \quad (3)$$

$$\text{Mx(i2)} = \{(\text{Mx(i)})\ (\text{Kx(k)})\} \quad (4)$$

The Mx(i2) variable describes the number of new minority data, which results from the calculation of the constant the Kx(k) variable multiplied by the number of original minority data in the Mx(i) variable. This calculation is carried out in stages until there is a balance between the number of the Mx(i2) and Mx(a) variables' original majority data.

The GradGAN model, which has been described mathematically, can be described in the form of a Pseudo-code algorithm as follows:

1. *Start*
2. *Input minority dataset $M_x(i)$ ;*
3. *Input Dataset majority $M_x(a)$;*
4. *Initialize $M_x(i2) = 0$;*
5. *Initialize $K_x(k) = 1$;*
6. *Compute the new minority value $M_x(i2) = (M_x(i))*K_x(k)$*
7. *Is the value of $M_x(a) >= M_x(i2)$ then*
8. *$K_x(k) = K_x(k) + 1$ ;*
9. *Goto 5*
10. *Else*
11. *Print $M_x(i2)$*
12. *Classification process*
13. *End*

In the process of the algorithm for measuring balance, the Kx(k) variable is multiplied by the minority data Mx(i) on gradually, where Kx(k) is variable value (1, 2,…..k). Mx(i2) or new minority data is the product of multiplying the original minority Mx(i) with Kx(k) variable, where k variable represents the number of variable values. In the next process, the minority data will be multiplied by a variable value gradually to create another new minority data until it finds data that is close to class balance.

This algorithm can be applied to all imbalanced data of all types of data sizes. In this research has been tested on three imbalanced datasets, which have different attributes and number of records. All data can be implemented well in this algorithm and produce a significant average accuracy value. The limitations of this algorithm cannot be applied to imbalanced datasets with missing values or with empty attributes.

## IV. Results and Discussion

The dataset is described and shown in Table II. In the description of the Hepatitis dataset, there are two classes representing death, "Die" and life, "live" In the Echocardiogram dataset, there are two classes representing death, "Dead" and life, "Alive". In the Immunotherapy dataset, there are two classes represented "No" and "Yes"; the purpose is to assist treatment. Treatments can be stopped if a positive "Yes" or a negative "No" is treated.

TABLE II. ORIGINAL DATASET DESCRIPTION

| Dataset | Number of records | Number of features | Imbalanced ratio | Minority class | Majority class | Number of classes |
|---|---|---|---|---|---|---|
| Hepatitis | 155 | 20 | 80 : 20 | "Die" | "Live" | 2 |
| Echocardiogram | 131 | 13 | 70 : 30 | "Alive" | "Dead" | 2 |
| Immunotherapy | 90 | 8 | 80 : 20 | "No" | "Yes" | 2 |

Fig. 3 compares the majority class and the minority class between the original dataset and the dataset that has been processed with the GradGAN model in stages with the final value. In the original dataset, the imbalance can obtain from the training data, which describes the majority class, and the test data, which describes the minority class. Furthermore, the GradGAN model functions to gradually increase the number of minority class data to become new minority data that is close to balance. By generating the original data, the calculation results can be seen in Fig. 3.

The three datasets in the GradGAN model to achieve balance can be seen in the comparison of the majority and the minority as follows: hepatitis dataset 124: 124, echocardiogram 107: 96, immunotherapy 71: 57.

In Fig. 3, the data balance stage process has been carried out. It starts from balancing the original data to form majority and minority data. Then the next process is applying the Gradually Generative Adversarial Network (GradGAN) model, balancing the original data gradually, forming a comparison of the original majority data with the new minority data gradually. Each comparison of the datasets formed by majority and minority data will be carried out with a classification experiment. The majority of the experimental results resulted in increase in the accuracy value of each dataset in the five classification methods. The result is shown in Table III.

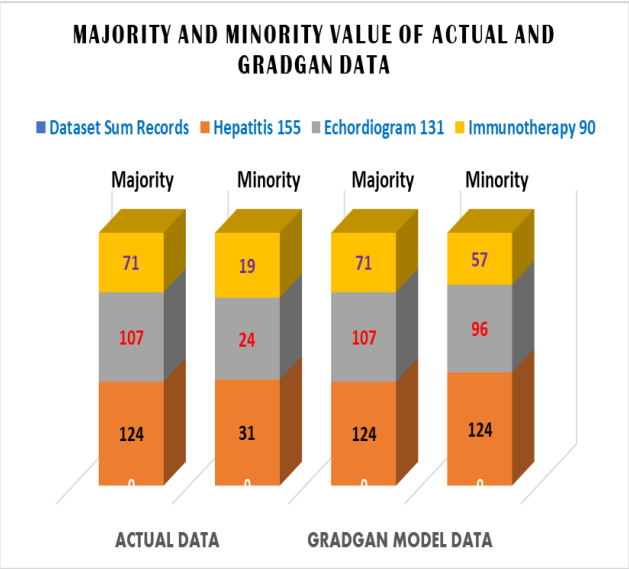


Fig. 3. Comparison chart of majority class and minority class original data and GradGan model.

In Table III, the results of the experimental accuracy of the two datasets are the original dataset and the dataset that already has the influence of the Gradually Generative Adversarial Network (GradGAN) model. First, the lowest classification value for the original dataset is 66%, obtained from the hepatitis dataset, while the highest is 83% from the echocardiogram dataset. Second, is the classification of the dataset that has been balanced with the GradGAN model. Classification produces an accuracy value of at least 68% and a maximum of 94%. Of the three original datasets and datasets that already have influence of the GradGAN model that It has been tested on five classification methods, and there is an average increase in accuracy of approximately 8,3%.

TABLE III. THE RESULTS OF CALCULATING THE ACCURACY VALUE OF THE DATASET

| Dataset Name | Methods | Accuracy Values (%) | | |
|---|---|---|---|---|
| | | (a) | (b) | (c) |
| Original Datasets | NB | 73 | **75** | 73 |
| | SVM | 81 | 79 | 81 |
| | D-Tree | 66 | 82 | 81 |
| | k-NN | 77 | 65 | 75 |
| | RF | 79 | **83** | 82 |
| GradGAN Model Influence Dataset | NB | 80 | 85 | 69 |
| | SVM | 84 | 84 | 70 |
| | D-Tree | 84 | 87 | **94** |
| | k-NN | 71 | 71 | 75 |
| | RF | 88 | 89 | 93 |

Abbreviations:(a)Hepatitis;(b)Echocardiogram;(c)Immunotherapy

A description of the dataset is a way to find out the character of the data. From the description of the selected datasets in this research, it can be seen that there is an imbalance in data. The evidence of data imbalance is an unequal comparison between the majority class and the minority class. An imbalance will affect the resulting accuracy value in the classification. To overcome the imbalance of the dataset in this research, a new innovation was developed from the GAN model, namely the GradGAN model. The flow of the GradGAN algorithm is shown in Fig. 2. The generation process starts from a random dataset and a discriminator to find out the original dataset so as to form an imbalanced majority and minority data. The goal of the GradGAN model is to form some new, larger minority data gradually. In this research, to increase the number of minority data gradually, by multiplying the number of variables (1, 2, ... k) that has been determined. So that it will produce several data sets which eventually form new minority data that are balanced or close to balance, as shown in the Fig. 3. The dataset that will be tested is the original imbalanced dataset and the balanced dataset using the five selected classification methods. The results of the classification accuracy can be seen in Table III. Experiments from three balanced datasets have been carried out to prove that the GRadGAN model has good classification performance by producing a better increase in accuracy values. Table IV shows a collection of classification results from the original dataset and the GradGAN model influence dataset. The results show that a balanced dataset has a significant increase in accuracy values, namely hepatitis by 88%, echocardiogram 89%, and immunotherapy 94%. The increased accuracy of the

original dataset and the GradGAN model for each dataset, namely hepatitis 7%, echocardiaogram 6% and immunotherapy 12% with an average increase in accuracy of 8.3%. Hypothesis analysis can be described in that, the influence of the GradGAN model can increase the accuracy value significantly and in the future it can be applied to data other than the data that has been tested.

TABLE IV. THE INCREASE IN THE RESULTS OF THE ACCURACY VALUE ON THE ORIGINAL DATASET AND THE GRADGAN MODEL

| Datasets | (a) | (b) | (c) |
|---|---|---|---|
| Hepatitis | 81% | 88% | 7% |
| Echocardiogram | 83% | 89% | 6% |
| Immunotherapy | 82% | 94% | 12% |
| Average Accuracy Improvement | 82% | 90,3% | 8,3% |

Abbreviations:(a)Originaldatasets;(b)GradGANmodel;(c)Accuracy improvement

## V. CONCLUSION

Based on the results of the research discussion above, it can be concluded that the dataset is getting closer to being balanced and the results are getting better. In this research, accuracy values were generated from the classification of three datasets, both of original and influence datasets from the GradGAN model. Experimental results involving five classification methods and three datasets, the application of the GradGAN model resulted in an increase in the accuracy value of 8,3%, compared to the original dataset. Thus the hypothesis can be concluded that the GradGAN model is very influential in the process of handling imbalanced datasets. Evidence of increased accuracy from imbalanced datasets and balanced datasets resulting from the resampling process of the GradGAN model is shown in Table IV. This means that by applying the GradGAN method the accuracy results are superior to those without GradGAN. In this research, the model was only tested on small imbalanced datasets, but other trials need to be carried out on large imbalanced datasets.

## VI. CONFLICTS OF INTEREST

The author has verified that they do not have any competing interests, as mentioned in their affirmation.

## VII. AUTHORS' CONTRIBUTION

Conceptualization of paper topics, M. Misdram and Muljono; research methodology, M. Misdram and Muljono; validation of research results, E. Noersasongko, Purwanto, Muljono; the formal analysis and the research investigation, M. Misdram and Muljono; the resources, M. Misdram, E. Noersasongko, Purwanto and Muljono; accuration spatial datasets, M. Misdram, and Fandi Yulian Pamuji; writing—original draft preparation, M. Misdram; writing—review and editing, E. Noersasongko, Purwanto and Muljono; visualization data and the research results, M. Misdram; supervision, E. Noersasongko, Purwanto, and Muljono; spatial and attribute data collector, M. Misdram, and Fandi Yulian Pamuji.

## ACKNOWLEDGMENTS

These research findings are part of a thesis at Universitas Dian Nuswantoro, Indonesia. Development of research results is supported by Universitas Merdeka Pasuruan, Indonesia.

## REFERENCES

[1] M. Bramer, Principles of Data Mining, vol. 53, no. 9. 2019.

[2] G. Huang and A. H. Jafari, "Enhanced balancing GAN: minority-class image generation," Neural Comput. Appl., vol. 8, 2021, doi: 10.1007/s00521-021-06163-8.

[3] R. A. Mohammed, K. W. Wong, M. F. Shiratuddin, and X. Wang, "Pwidb: A framework for learning to classify imbalanced data streams with incremental data re-balancing technique," Procedia Comput. Sci., vol. 176, pp. 818–827, 2020, doi: 10.1016/j.procs.2020.09.077.

[4] Q. Shu, T. Hu, and S. Liu, "Random Forest Algorithm Based on GAN for Imbalanced Data Classification," J. Phys. Conf. Ser., vol. 1544, no. 1, 2020, doi: 10.1088/1742-6596/1544/1/012014.

[5] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," Pattern Recognit., vol. 102, 2020, doi: 10.1016/j.patcog.2020.107262.

[6] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," Artif. Intell. Med., vol. 104, no. January, p. 101815, 2020, doi: 10.1016/j.artmed.2020.101815.

[7] N. G. Siddappa and T. Kampalappa, "Adaptive condensed nearest neighbor for imbalance data classification," Int. J. Intell. Eng. Syst., vol. 12, no. 2, pp. 104–113, 2019, doi: 10.22266/IJIES2019.0430.11.

[8] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Improving classifiaction performance of fetal umbilical cord using combination of SMOTE method and multiclassifier voting in imbalanced data and small dataset," Int. J. Intell. Eng. Syst., vol. 13, no. 5, pp. 441–454, 2020, doi: 10.22266/ijies2020.1031.39.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, no. February 2017, pp. 321–357, 2002, doi: 10.1613/jair.953.

[10] T. Zhou, W. Liu, C. Zhou, and L. Chen, "GAN-based semi-supervised for imbalanced data classification," 2018 4th Int. Conf. Inf. Manag. ICIM 2018, pp. 17–21, 2018, doi: 10.1109/INFOMAN.2018.8392662.

[11] I. Goodfellow et al., "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[12] V. Sampath, I. Maurtua, J. J. Aguilar Martín, and A. Gutierrez, A survey on generative adversarial networks for imbalance problems in computer vision tasks, vol. 8, no. 1. Springer International Publishing, 2021.

[13] C. Li, K. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," Adv. Neural Inf. Process. Syst., vol. 2017-Decem, pp. 4089–4099, 2017.

[14] X. A. Carrasco, A. Elnagar, and M. Lataifeh, "A Generative Adversarial Network for Data Augmentation: The Case of Arabic Regional Dialects," Procedia CIRP, vol. 189, pp. 92–99, 2021, doi: 10.1016/j.procs.2021.05.072.

[15] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 787–795, 2013, doi: 10.2478/amcs-2013-0059.

[16] T. Sajana and M. R. Narasingarao, "Classification of imbalanced malaria disease using naïve bayesian algorithm," Int. J. Eng. Technol., vol. 7, pp. 786–790, 2018, doi: 10.14419/ijet.v7i2.7.10978.

[17] A. M. De Carvalho and R. C. Prati, "Improving kNN classification under Unbalanced Data. A New Geometric Oversampling Approach," Proc. Int. Jt. Conf. Neural Networks, vol. 2018-July, no. Cmcc, pp. 1–6, 2018, doi: 10.1109/IJCNN.2018.8489411.

[18] R. Y. Goh and L. S. Lee, "Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches," Adv. Oper. Res., vol. 2019, 2019, doi: 10.1155/2019/1974794.

[19] T. A. Khan, K. A. Kadir, S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease classification," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 560–569, 2020, doi: 10.14569/IJACSA.2020.0111170.

[20] C. O. Truică and C. A. Leordeanu, "Classification of an imbalanced data set using decision tree algorithms," UPB Sci. Bull. Ser. C Electr. Eng. Comput. Sci., vol. 79, no. 4, pp. 69–84, 2017.

[21] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," Int. J. Pharm. Res., vol. 12, no. 4, pp. 56–66, 2020, doi: 10.31838/ijpr/2020.12.04.013.

[22] R. Kumar and J. Kaur, Random forest-based sarcastic tweet classification using multiple feature collection, vol. 163. Springer Singapore, 2020.

[23] Z. Zhang et al., "A generative adversarial network–based method for generating negative financial samples," Int. J. Distrib. Sens. Networks, vol. 16, no. 2, 2020, doi: 10.1177/1550147720907053.

[24] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting breast cancer via supervised machine learning methods on class imbalanced data," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 54–63, 2020, doi: 10.14569/IJACSA.2020.0110808.

[25] C. Bunkhumpornpat and S. Subpaiboonkit, "Safe level graph for synthetic minority over-sampling techniques," 13th Int. Symp. Commun. Inf. Technol. Commun. Inf. Technol. New Life Style Beyond Cloud, Isc. 2013, no. September 2013, pp. 570–575, 2013, doi: 10.1109/ISCIT.2013.6645923.

[26] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," Appl. Nanosci., no. 0123456789, 2022, doi: 10.1007/s13204-021-02063-4.

[27] W. Badr, "6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)," Towar. Data Sci., 2019, [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

[28] A. Muslim, A. B. Mutiara, R. Refianti, C. M. Karyati, and G. Setiawan, "Comparison of accuracy between long short-term memory-deep learning and multinomial logistic regression-machine learning in sentiment analysis on twitter," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 2, pp. 747–754, 2020, doi: 10.14569/ijacsa.2020.0110294.