IJCI V9-01 (2022) 15–28
International Journal of Computers and Information
(IJCI)
Available online at *https://ijci.journals.ekb.eg/*

Imbalanced Data Oversampling Technique Based on Convex Combination Method

Mohammed Elnahas, Mahmoud Hussein, Arabi Keshk

Computer Science Department, Faculty of Computers and Information,Menoufia University, Shebin Elkom 32511, Egypt
m.moustafa.elnahas@gmail.com, mahmoud.hussein@ci.menofia.edu.eg, arabikeshk@yahoo.com

Abstract

Classification process is the predicting a label for a specific set of inputs. In such process, it is difficult to classify given inputs when a dataset is imbalanced. Most of existing machine learning classifiers suffer from dealing with the imbalanced data, because it makes the classifiers highly biased towards the majority class. This bias may lead to less accuracy in minority class prediction. Data oversampling is one of the most important solutions used to balance the data particularly when dataset is small and/or imbalanced dataset. Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE, Adaptive Synthetic (ADASYN) and Weighted SMOTE(W-SMOTE) are the most popular techniques used for data oversampling. However, the main drawback of SMOTE and ADASYN techniques is they increase the overlapping between classes and then the produced samples are not representative of the original data distribution. The Borderline-SMOTE may neglect some important samples to produce new samples. To overcome, the problems in the existing over-sampling techniques, in this paper, we propose a new data over-sampling method that depends on the convex combination method to generate new samples of the minority class. The convex combination allows us to produce new samples that have the same original data distribution. We evaluated our approach over four standard imbalanced datasets (Yeast, Glass Identification, Paw, and Wisconsin Prognosis Breast Cancer (WPBC)). The experimental results show that our proposed method gives better performance in terms of accuracy, precision, recall. F1-measure and Area under the curve (AUC).

Keywords: Imbalanced dataset; Oversampling; SMOTE; ADASYN; Borderline-SMOTE;

1. **Introduction**

In machine learning, the classification is a predictive process that assigns a class label to a given data sample. A classification process requires training data examples with a given class label. It learns from this data how to make the best decision to assign a class label to other samples that are not included in the training data. To get a good classification performance, the dataset itself plays an important role because the dataset may suffer from some problems such as missing value, irrelevant features, and data imbalance. All of these problems mislead the classifiers [1]. Dealing with imbalanced datasets is imperative because imbalanced data exist in most real-world applications such as medical diagnosis, fraud detection, and network intrusion detection [2]. The imbalanced problem occurs when the distribution of instances to classes is not fair. A class that contains a large number of instances is known as a majority class, and a class that contains fewer numbers of instances is known as a minority class [3]. Most machine learning classifiers have been designed to learn from balanced data. This makes dealing with imbalanced data is very challenging, because the classifiers are biased towards the majority class, but in most imbalanced data the minority class is most important.

Different solutions are proposed to deal with the imbalanced dataset one of these solutions is to resample the dataset to be balanced using over-sampling [4], under-sampling [5] ,and hybrid-sampling [6].First, data over-sampling methods are used to make new samples of the minority class to produce a balanced dataset [4].

Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE**,** and Adaptive Synthetic (ADASYN) are the most popular algorithms that are used to make over-sampled datasets. Introducing new samples may increase overlapping between classes if the distribution of majority is not taken into consideration, where some of the new samples are produced in the majority class space. This problem known as class overlapping [7]. To avoid data overlapping problems, suitable samples need to be determined as a seed to generate new samples. The over-sampling techniques are suitable for small datasets, but if the given dataset has high dimensionality it takes more computation cost [4].Second, data under-sampling methods are used to balance datasets by eliminating some samples from the majority class to be balanced with the minority class. It is not preferred with small datasets because in small datasets all of the samples are important and eliminating one sample will lead to information loss. Tomek Link (T-Link) [24] and random under-sampler are the most popular techniques to create a balanced dataset by under-sampling the majority class. Third, the hybrid-sampling method is using both under-sampling and over-sampling to create a balanced dataset. A hybrid sampling SVM [25] and CSMOUTE [6] are examples to approaches that are proposed to resample the imbalanced datasets by combining the over-sampling techniques and the under-sampling techniques

The existing approaches of data over-sampling have some limitations. First, SMOTE is developed to reduce the overfitting problem because it creates new synthetic samples instead of replication of samples. But SMOTE does not consider the distribution of classes, which means it may increase overlapping between classes. Also, SMOTE does not reproduce the distribution of the original dataset. Thus, the produced data will contain information that does not reflect the original dataset [8]. Second, Borderline-SMOTE is a new version of SMOTE. However, it focuses on producing new samples from the minority samples that belonging to danger area that have some majority samples in its k-nearest neighbor. This can lead to the neglect of some unrelenting samples to the danger area [9]. Finally, ADASYN [10] is considered a modified version of SMOTE, but it focuses on generating samples for the minority samples that harder to learn. These samples are very close to majority samples, and then the newly generated samples which causes a low classification accuracy.

In this paper, to solve the existing approaches' problems, we propose an oversampling approach that divide the minority sample space into regions based on the *k*-nearest neighbors' algorithm (KNN) [11]. Then, the new synthetic samples are produced in the best regions based on convex combination method [12]. To evaluate our approach, we compare it with the existing approaches (SMOTE, Borderline-SMOTE and ADASYN) on four standard datasets (Yeast, Glass Identification, Paw, and Wisconsin Prognosis Breast Cancer (WPBC)). The results show that our proposed approach gives the highest classification performance in terms accuracy, precision, recall. F1-measure and Area under the curve (AUC).

The rest of this paper is organized as follow. Section 2 presents existing approaches that used to solve imbalanced data problem. Section 3 presents our proposed approach. Section 4 presents the datasets, evaluation metrics used to evaluate our proposed method and the results. The conclusion of this paper is presented in Section 5.

2. **Related Work**

The most popular solutions to solve imbalanced data problem is data resampling which is known as data-driven approaches (i.e. data over-sampling, under-sampling, and hybrid-sampling). Such approaches adjust the classes distribution. To improve the accuracy of classification process, in this paper, we will focus on the data over-sampling because it is the most suitable approach that is used with the small imbalanced dataset [13].

2.1. Synthetic Minority Over-sampling Technique (SMOTE).

SMOTE is an over-sampling approach proposed by “Nitesh V. Chawla” [14]. It was the first approach that was used to add new synthetic samples to the dataset. Its idea can be summarized as follows: finding the K nearest neighbors of each minority instance then, for each minority sample (I) randomly select one of the k nearest neighbors (I′) and calculate the difference between each feature from I and I′; to produce the new synthetic sample, add the original sample I to the difference between I and I′ multiplied by a random number between 0 to 1(R (0,1)) where the new synthetic sample (I_new) is produced by :

$$I\_new=I+(I'-I)\times R\,(0,1) \tag{1}$$

SMOTE uses a linear combination to produce new synthetic samples so the new samples will lie on the line between (I) and (I′) as shown in Fig. 1 that shows how the SMOTE balances the dataset that contain 120 sample for majority class and 20 sample for minority class with K= 3. One of the defects found in SMOTE is that the class distribution is not taken into consideration while producing the new synthetic samples. It also does not differentiate the neighbor of the new sample is belonging to the minority or majority class, which will increase the possibility of overlapping between classes [15]. Another defect in SMOTE is the new synthetic samples does not represent the distribution of the original seed samples, because the new synthetic samples are produced on the line between two minority samples, so that the produced synthetic samples will not cover all the minority class space.

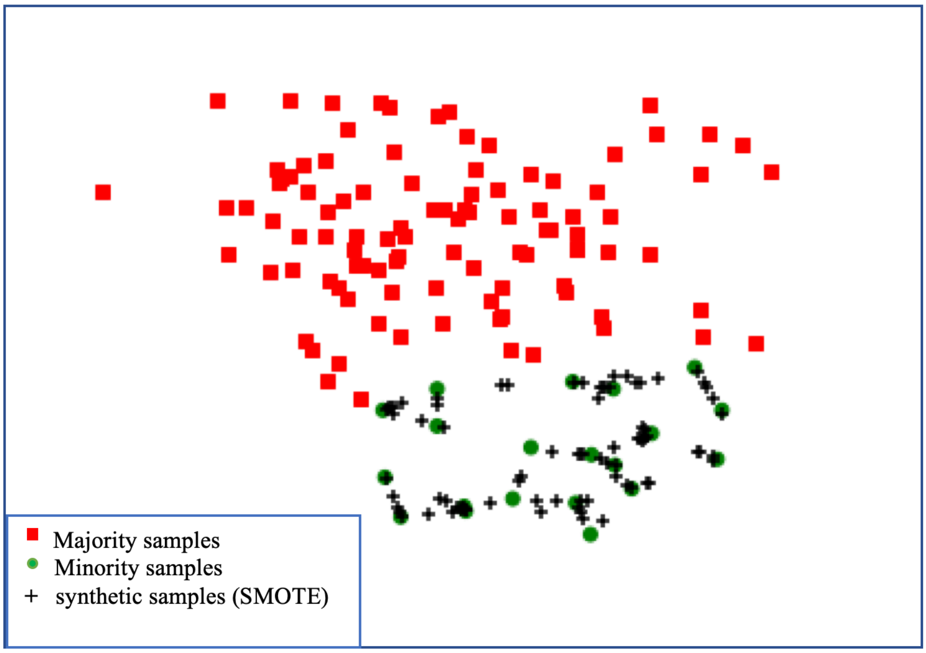

Fig. 1. The use of SMOTE to balance the dataset

2.2. Borderline-SMOTE.

Borderline-SMOTE [16] is an improvement of the SMOTE. The main idea of Borderline-SMOTE is to divide the minority samples into three categories or regions (safe, noise, and danger). By determining the K nearest-neighbor of each minority instance and determine the numbers of the majority samples (m) that found in K nearest-neighbor of this instance. The three regions are defined according to Table 1. The authors exclude the noise region and the safe region from producing new samples and they are depending on the danger region to produce new samples. Borderline-SMOTE produces the new samples as same as SMOTE on the line between two minority samples from the danger region. Although the Borderline-SMOTE focuses on danger region that has high misclassification rate to produce the new samples, it also the produced samples that do not take the distribution of original seed sample. As the result of only considering the samples that are nearest to the majority class that may leads neglecting some important samples of minority class. Fig. 2 represent how Borderline-SMOTE balances the dataset.

Table 1. *Borderline-SMOTE regions*

Region	Definition
Noise	$m=k$
Safe	$0 \leq m < k/2$
Danger	$k/2 \leq m < k$

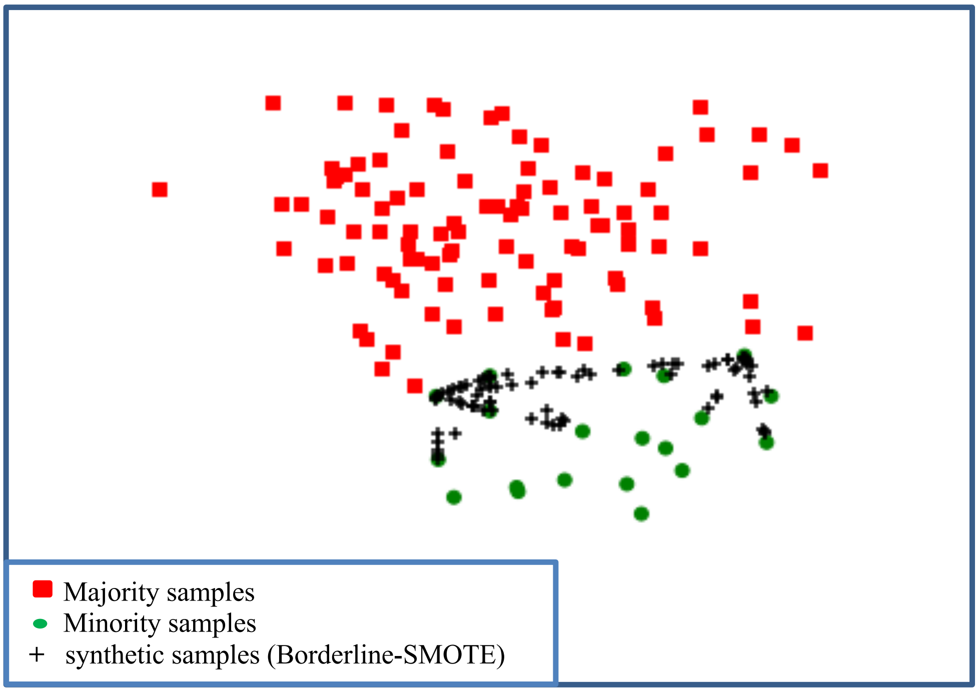

Fig. 2.The use of Borderline-SMOTE to balance the dataset

2.3. Adaptive Synthetic (ADASYN).

Another modification of the SMOTE is ADASYN [9]. It is an over-sampling technique that uses weighted distribution for each sample of minority class based on its level of learning. Most new synthetic samples are generated from samples that are harder to learn. It firstly calculates the numbers of samples that are needed to balance the dataset (G). Then for each minority sample (i), calculate the ratio (ri) by Equation 2:

$$\mathrm{ri} = \Delta \mathrm{i}/\mathrm{K} \tag{2}$$

where Δi is the number of majority samples found in K nearest neighbor of minority instance. Then, to calculate the density distribution (ri′), normalize ri for each sample by dividing it by the summation of ri for all minority samples (mn) according to Equation 3:

$$\mathrm{ri}' = ri/\textstyle\sum_{i=1}^{mn} \mathrm{ri} \tag{3}$$

The density distribution (ri′) is used to calculate the numbers of new synthetic samples (gi) that are produced from each minority sample (i) by: Equation4:

$$\mathrm{gi} = \mathrm{ri}' \times \mathrm{G} \tag{4}$$

In order to produce new samples from minority instance (i) it uses Equation 1 as the same SMOTE. But, the key difference between ADASYN and SMOTE is that it uses density distribution to decide the number of new synthetic samples that must be produced for each minority sample. Although ADASYN was developed to overcome the SMOTE limitations, many of literatures had compared the performance of SMOTE and ADASYN and all agreed that ADASYN does not improve the classification performance compared by SMOTE, but sometimes SMOTE may give high classification performance than ADASYN [17,18,19]. Fig. 3 represents the result of ADASYN for balancing the dataset.

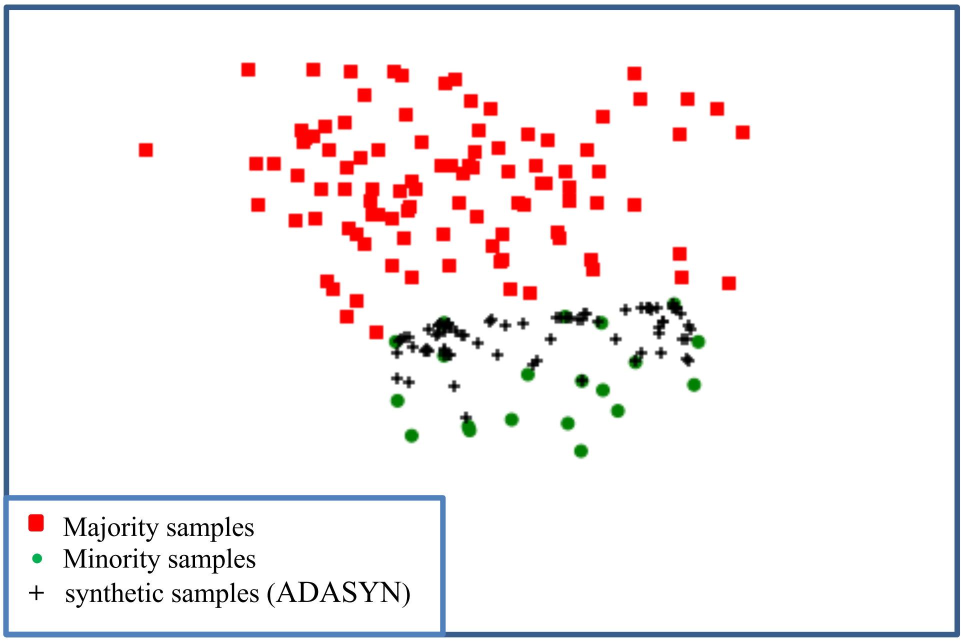

Fig. 3. The use of ADASYN to balance the dataset

2.4. Weighted SMOTE(W-SMOTE).

Another modification of the SMOTE is W-SMOTE [28]. It is an oversampling technique that assigns a weight to each minority sample, these weights are then used to determine the number of new synthetic samples that are produced from each minority samples. In W-SMOTE, each minority sample produces different number of new synthetic samples according to it is weight. To produce these weights, the Euclidean distance between minority samples with respect to all other minority samples is calculated. After getting the weight matrix, the W-SMOTE produce the new synthetic samples on the line between minority samples.

2.5. Related Work Relevant to Image Domain.

To overcome the imbalanced data problem in the image domain, an approach has been proposed [29]. This approach uses perceptual two-layer image fusion by Deep Learning (DL) to get a new image that more informative than the input images. The approach also uses Nonsubsampled Contourlet Transform (NSCT) to analyze the input image and get the high frequency and low-frequency images then uses convolutional neural networks (CNN VGG19 to extract the low-pass and high-pass feature vector. As a fusion rule for high-pass and low-pass, temporal consistency of extracted features, Euclidean distance, and weights sub-band calculations are used. Then, the inverse of NSCT is used to get the fused image.

2.6. Related Work Summary.

We can summarize the limitations found in the mentioned approaches as follows. First, SMOTE produces new samples on the line between any minority samples and its K nearest neighbor. Such technique increases the possibility of classes overlapping, because it does not restrict which samples is used to produce a new synthetic sample. Second, Borderline-SMOTE restricts the minority samples that are used to produce new samples, It includes some majority samples into its K nearest neighbors and uses the same method that used in SMOTE “Equation 1” to produce new samples from the candidate minority samples. Thus, Borderline-SMOTE may decreases the possibility of overlapping between classes. But it is possible to neglect some important minority samples. Third, ADASYN uses the density distribution to determine the numbers of the new sample that are produced from each minority sample, and same as SMOTE and Borderline-SMOTE it uses “Equation 1” to produce new samples. In many of the literature (e.g. [17,18,19]), it is proved that SMOTE may give better performance than ADASYN. Fourth, W-SMOTE it is modification of SMOTE. In W-SMOTE, each minority sample produces different number of new synthetic samples according to a weight matrix that is determined by Euclidean distance. Finally, the produced balanced data from SMOTE, Borderline-SMOTE, and ADASYN does not represent the distribution of the original dataset.

3. **Proposed Method**

To overcome the limitations of the existing approaches, we propose a new oversampling technique based on the region between minority samples and convex combination inside this region to produce the new synthetic samples. Our proposed method consists of two phases. The first phase is how to determine the candidate regions to produce a new synthetic sample. The second phase is how to maintain the distribution of original samples.

3.1. Determining the Candidate Regions

We found the borderline-SMOTE and ADASYN focused on choosing the samples that used to produce the new synthetic samples. Unlike, SMOTE which considers all samples are a candidate to produce new synthetic samples. We follow the borderline-SMOTE approach to split minority samples into three groups: noisy, safe, and danger. Then we exclude the noisy samples. After that, each of the remaining minority samples forms a region with it is KNN samples. The centroid point of each region is then calculated. Then we candidate the regions that have more than or equal K/2 from it is centroid point is majority samples to produce new synthetic samples. Algorithm 1 and Fig. 4 shows how we choose the candidate regions. As shown in algorithm 1, we take the dataset as input and three parameters. The first parameter is K1 that represents the numbers of all nearest samples includes the majority and minority samples from each minority sample, the second parameter is K2 that represents the number of nearest samples that have only minority class label from each minority sample. The third parameter is K3 that represents the number of nearest minority or majority samples from the centroid of minority samples regions. The minority samples Mn which all of its nearest samples K1 are majority samples Mj are considered noisy samples. Each sample from the remaining minority samples makes a minority region with all of each K2 nearest neighbors are minority samples. Then we produce a synthetic sample to represent each region this sample is the centroid point of the region. The minority region that most of its K3 nearest neighbors are majority is selected as candidate region, and its samples are inserted in candidate region list CRL to produce new synthetic samples.

```
Algorithm 1
Inputs:
        Dataset that contains {Mn, MJ} Mn denotes to minority samples, Mj denotes to minority.
Parameters:
            K1 denotes to number of nearest neighbours for each minority samples {contains both
                        minority and majority samples}.
            K2 denotes to number of nearest minority samples for each minority samples.
            K3 denotes to number of nearest neighbours for the centroid of the region {contains both
                        Minority and majority samples}.
Output:
        The Candidate Regions List (CRL)to produce the new synthetic sample.
Steps:
        5- For each Mn samples find its K1 nearest neighbors and exclude the samples which all of
           its K1 nearest neighbors are Mj then select and put the others minority samples into SL
           list denotes to Seed List.
        6- For each s in SL find its K2 nearest neighbors from SL and put them into RL denotes to
           Regions List.
        7- For each r in RL find its centroid point (C) by dividing sum of its samples by K2 and put
           it into CL denotes to Centroid List.
        8- For each c in CL find its K3 nearest neighbors and chose the regions that its k3 nearest
           neighbors contains majority samples greater than or equal to minority samples (k3 / 2 <=
           Mj < K3) and put it into CRL denotes to Candidate Regions List
        9- RETURN: CRL
```

3.2. Maintaining the Distribution of Original Samples.

One of the problems of SMOTE, borderline-SMOTE, and ADASYN is the new synthetic samples do not take the distribution of the original data. The reason for this problem is that the mentioned techniques produce the new samples on the line between the minority samples. We propose to produce the new synthetic samples inside the candidate regions that returned from Algorithm 1 by using the convex combination method. The convex combination is a linear combination of points. Equation (5) show the convex combination of given (n) number of points where αi is the coefficient of each point xi and these coefficients must be non-negative, and their sum is equal to 1.

$$nx = \alpha 1x1 + \alpha 2x2 + \alpha 3x3 \ldots + \alpha nxn. \quad \text{Where} \quad \alpha i \geq 0 \qquad (5)$$

In the convex combination as shown from Equation (5), the secret is lies in the coefficients that are must be non-negative and their sum is equal to 1 This produces samples that lies inside the region between the input samples. Fig. 5 show the convex combination of four points minority samples (X1, X2, X3, and X4) where X2, X3 and X4 are the 3 nearest neighbors of X1. All of the convex combination points as shown in Fig. 5 (c) lies in the convex region between the samples (X1, X2, X3, and X4) which this does not happen when using SMOTE, borderline-SMOTE, and ADASYN where all new samples lie on the line between two samples.

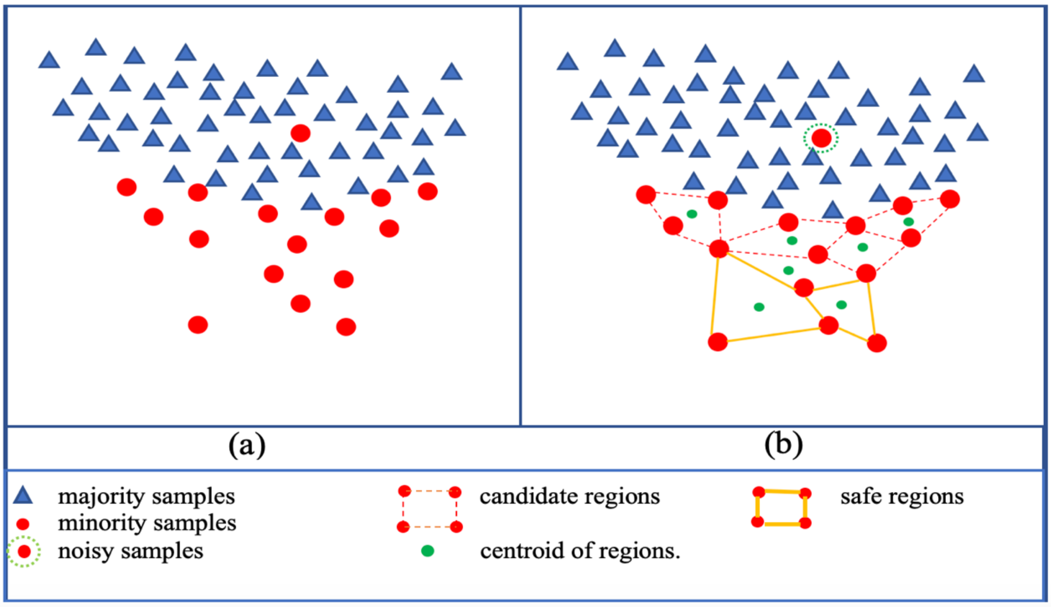

Fig. 4. Determining Candidate Regions; (a) the original data; (b) minority sample regions

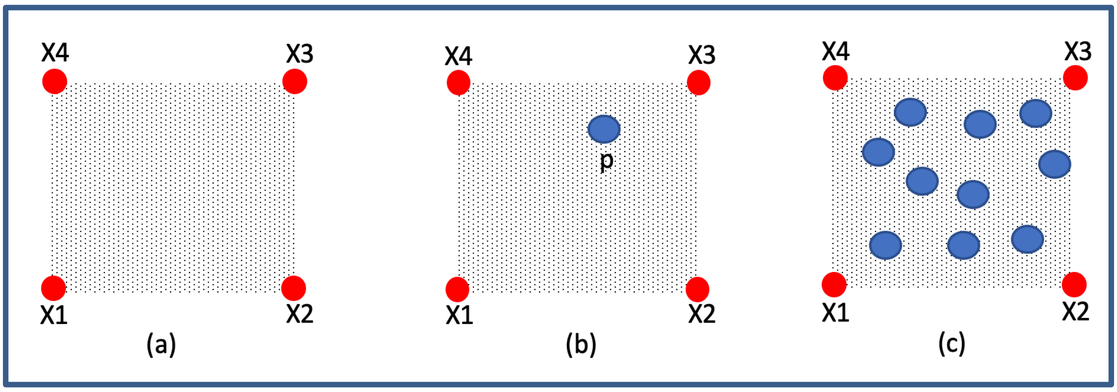

Fig. 5. Convex combination of four minority samples (X1, X2, X3, and X4). (a) distribution of minority samples; (b)the convex combination point "p" of (X1, X2, X3, and X5) ;(c) ten times using convex combination with different coefficients values

4. Experimental Results and Discussion

The overall accuracy is not the most effective metrics to evaluate a model that learns from imbalanced data, because with imbalanced dataset high accuracy not necessary means high performance. Table 2 shows the confusion matrix of binary classification problem and Equation (6) shows how to calculate the overall accuracy by dividing the true positive samples (TP) plus true negative samples (TN) by the size of test set. the overall accuracy is not the suitable metric to evaluate a model that learned from an imbalanced dataset. For example, If we have confusion matrix with TP =0, FN=5, FP=0, and TN=95 the overall accuracy is equal to 95 %. It seems a very high accuracy but in fact this model failed to classify any positive samples. So that overall accuracy cannot be the only metric to evaluate model with imbalanced data. There is alternative metrics prepared to be used with the imbalanced data such as precision, recall, F1-measure, Receiver Operating Characteristic (ROC), Area Under ROC-Curve (AUC).

Table 2. *Confusion Matrix of a binary classification*

Actual / Predicted	Positive	negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

Precision is known as Positive Predicted Value (PPV), it represents the percentage of relevant samples that are identified in prediction. It can be computed by divide true (TP) by (TP) plus false positive samples (FP) as shown in Equation 7:

$$\text{precision} = \frac{TP}{TP+FP} \qquad (7)$$

Recall is known as True Positive Rate (TPR), it represents the percentage of samples in prediction that are relevant. It can be computed by divide true (TP) by (TP) plus false negative samples (FN) as shown in Equation 8:

$$\text{recall} = \frac{TP}{TP+FN} \qquad (8)$$

F1-measure is the harmonic mean between precision and recall. It can be computed by multiplying the result of multiplying precision by recall by 2 then divide the result by precision plus recall as shown in Equation 8. High F1-measure value means that both precision and recall are high.

$$\text{F1} - \text{measure} = \frac{2*precision*recall}{prcision\ +recall} \qquad (9)$$

Receiver Operating Characteristic Curve (ROC-curve) [26], is a plot of recall (TPR) on y-axis with False Positive Rate (FPR) on x-axis for different values of threshold. So, ROC-Curve is representing the trade-offs between benefits (TPR) and costs (FPR) with different thresholds of a classification model. The best model at all is one which gives (TPR=1and FPR=0). For simplicity when comparing the different model by using ROC-Curve, the value that represent the ROC-curve of each model is used, this value is called Area Under ROC-Curve (AUC) [27].

In order to validate our work, we use four available datasets (yeast, Glass Identification, paw, and Wisconsin Prognosis Breast Cancer (WPBC)). All of them are standard imbalanced datasets that widely used in many literatures that deal with imbalanced dataset. The first three ones are available in Knowledge Extraction based on Evolutionary Learning (KEEL) repository [20], and the last one is available in UCI repository [21]. Table 3 contains the number of samples, number of features, and the Imbalanced Ratio (IR) that means the ratio between majority samples and minority samples as shown in Equation 10:

$$\mathrm{IR} = \frac{\#Majority\ samples}{\#Minority\ samples} \tag{10}$$

Table 3. *Datasets description*

Dataset	Number of samples	Number of features	IR
Yeast	528	8	9.35
Glass Identification	172	9	9.12
Paw	600	2	5
WPBC	198	34	3.21

To evaluate our work, we choose two well-known classifiers Random Forest tree (RF) [22], and Support Vector Machine (SVM) [23]. These classifiers are trained using two scenarios: 1) trained using original dataset. 2) the original datasets are oversampled using original, SMOTE, Borderline-SMOTE, ADASYN, W-SMOTE and our proposed method datasets then the resulted datasets are used to train the employed classifiers. To be fair we applied the hold out method {70% for training: 30 % for testing} then the oversampling process will be done on the training set, which means that the test set is the same in (original, SMOTE, Borderline-SMOTE, ADASYN, and our proposed). The results have been recorded as an average of 20 times of experiments.We implement our method using python programming language. Also, we used the scikit-learn library that include implementation of the RF and SVM classifiers. In our experiments, we use these hyperparameters for RF n_estimators =200, criterion = gini, max_depth =default, and for SVM kernel = rbf, gamma= *scale,* decision_function_shape= ovr.

Table 4 and Table 5 show a comparison between our proposed method and the other employed methods with SVM and RF respectively. In these tables the term original means the original dataset without any oversampling and ACC refers the accuracy. The highest value for ACC, AUC, and F1-measure is bold-faced. Fig. 6 and Fig. 7 represents the AUC with RF and SVM respectively when they trained and tested by original, SMOTE, Borderline-SMOTE, ADASYN, W-SMOTE and our proposed method.

Table 4. *Classification performance with RF classifier*

dataset	ORIGNAL			SMOTE			Borderline-SMOTE			ADASYN			W-SMOTE			Our approach		
	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure
Yeast	94.9	0.743	0.79	91.8	0.840	77	90.5	0.795	0.73	90.5	0.834	0.75	89.3	0.827	0.73	**96.2**	**0.864**	**0.86**
Glass	92.3	0.60	0.65	90.3	0.768	0.75	90.3	0.678	0.70	84.6	0.646	0.62	92.3	0.778	0.78	**96.1**	**0.889**	**0.89**
Paw	86.1	0.615	0.64	86.2	0.743	0.73	85.5	0.75	0.73	87.2	0.79	0.77	87.7	0.913	0.82	**91.1**	**0.932**	**0.85**
WPBC	76.2	0.53	0.49	72.8	0.576	0.58	76.2	0.599	0.61	74.5	0.565	0.57	81.3	0.754	0.76	**88.1**	**0.810**	**0.83**

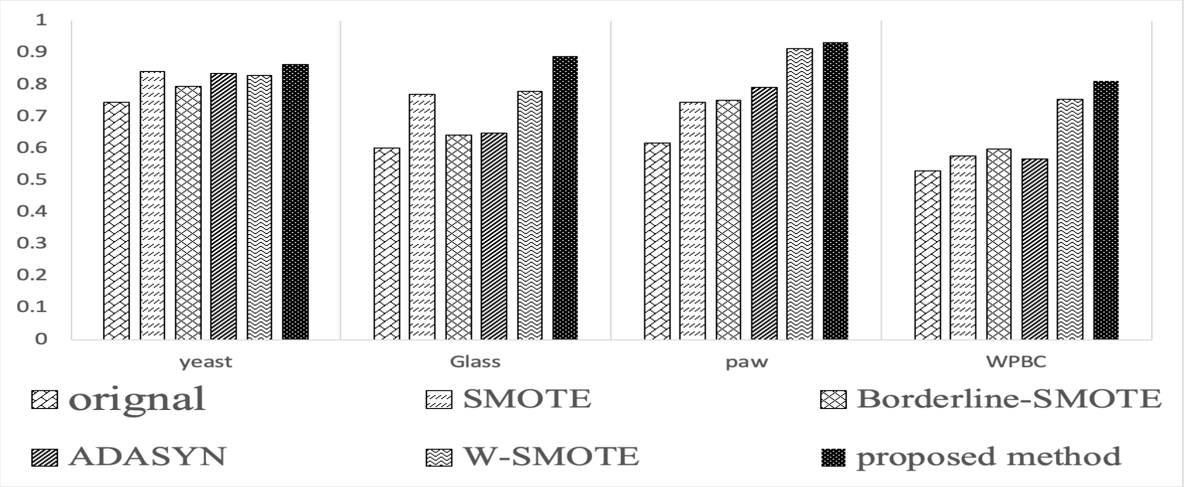

Fig. 6. Comparison between our proposed method and the existing methods in term of AUC with RF classifier

Table 5. *Classification performance with SVM classifier*

dataset	ORIGNAL			SMOTE			Borderline-SMOTE			ADASYN			W-SMOTE			Our approach		
	ACC %	AUC	F1-measure	ACC	AUC	F1-measure	ACC %	AUC	F1-measure	ACC	AUC	F1-measure	ACC %	AUC	F1-measure	ACC %	AUC	F1-measure
Yeast	90.5	0.50	0.48	85.5	0.83	0.71	88.6	0.75	0.71	88.6	0.72	0.70	84.9	0.834	0.67	**94.3**	**0.879**	**0.85**
Glass	90.3	0.50	0.47	88.4	0.757	0.72	86.5	0.757	0.72	86.5	0.657	0.64	90.4	0.873	0.81	**94.2**	**0.968**	**0.87**
Paw	**81.6**	0.619	0.63	78.3	0.831	0.72	81.1	0.764	0.71	81	0.811	0.74	78	0.846	0.73	**81.6**	**0.888**	**0.77**
WPBC	77.9	0.588	0.59	71.2	0.565	0.57	72.8	0.63	0.64	72.8	0.62	0.63	76.3	0.747	0.64	**84.7**	**0.831**	**0.81**

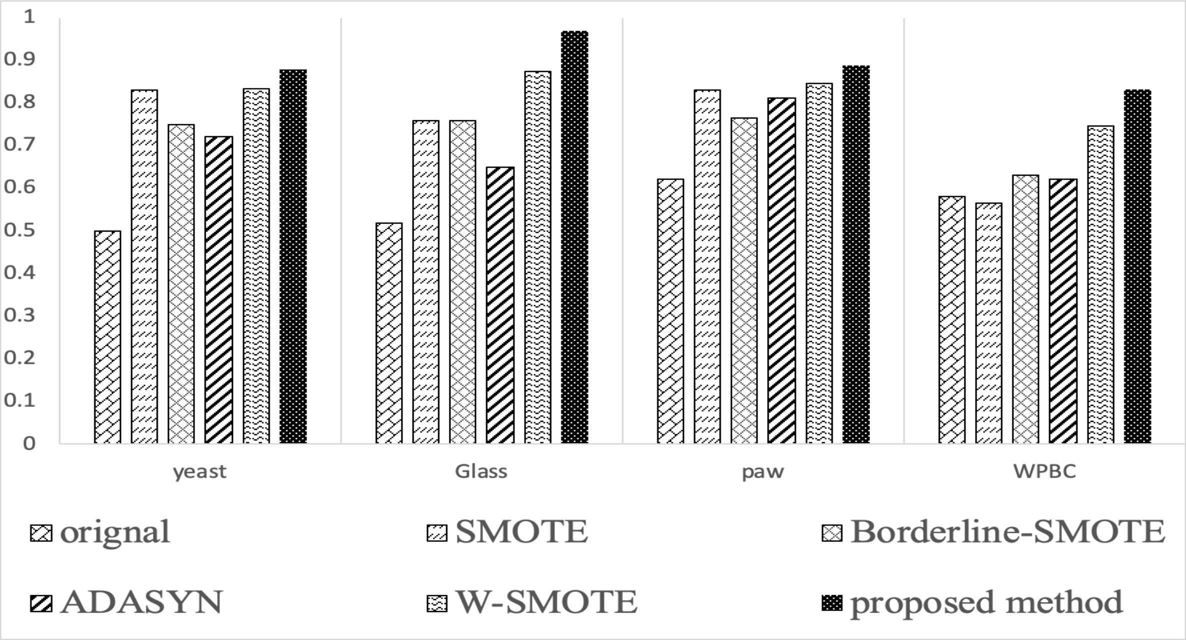

Fig. 7. Comparison between our proposed method and the existing methods in term of AUC with SVM classifier

The results show that, the proposed method achieves the highest classification performance in terms of accuracy, precision, recall. F1-measure and AUC for all datasets. For example, when we test our proposed method with a Random forest tree classifier it gives AUC = 0.864, 0.889, 0.932, and 0.810 with Yeast, Glass Identification, Paw, and WPBC datasets respectively. When we test our proposed method with an SVM classifier it gives AUC = 0.879, 0.968, 0.888, 0.831 with Yeast, Glass Identification, Paw, and WPBC datasets respectively. In Table 5, when we compared our proposed method with other existing methods on the paw dataset, we figured out that the accuracy is very close particularly when the original dataset is used. But, our proposed approach gives a higher AUC and F1-measure compared to the existing method. The improvement in the classification performance came as a result of the new synthetic samples that were generated to balancing the training dataset by using our proposed method.

The main advantage of our proposed approach is the high accuracy, AUC and F1-measure achieved with small and overlapped imbalanced datasets, this is achieved because our proposed approach produces new synthetic samples that represent the distribution of the original sample. However, our proposed approach is similar to most of existing oversampling techniques, where it takes longer time with the bigger datasets.

5. Conclusion

There are a lot of imbalanced data in many real-world applications. However, most machine learning techniques are designed to learn from balanced data. To solve this imbalanced data problem, there are several approaches such as SMOTE, Borderline-SMOTE, and ADASYN. All of these approaches use the line between samples to produce the new samples, which make the produced data does not represent the original data distribution. In this paper, we proposed a technique that determines the best samples that are used to produce new samples by using the convex combination approach. The produced data are then having the same distribution as the original dataset. Our experiments show that our approach gives the best performance in terms of accuracy, precision, recall, AUC compared to the existing approaches in all datasets.

In the future, we will test our proposed method with other datasets that bigger in size and have larger imbalanced ratio than datasets we used in this paper. Also, we will adapt our proposed method to deal with multi-label datasets problems not only binary label datasets.

References

[1] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." *ACM SIGKDD explorations newsletter* 6.1 (2004): 20-29.

[2] Y. Sun, A. C. Wong and M. S. Kamel, "Classification of imbalanced data: A review", Int. J. Pattern Recogn., vol. 23, no. 4, pp. 687-719, 2009.

[3] L. Cao and Y. Zhai, "Imbalanced data classification based on a hybrid resampling svm method", *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom) 2015 IEEE 12th Intl Conf on*, pp. 1533-1536, 2015.

[4] F. Thabtah, S. Hammoud, F. Kamalov and A. Gonsalves, "Data imbalance in classification: Experimental evaluation", Inf. Sci., vol. 513, pp. 429-441, Mar. 2020.

[5] Show-Jane. Yen, and Yue-Shi. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," In Proceedings of the Intelligent Control and Automation, Lecture Notes in Control and Information Sciences (LNCIS), Vol. 344, August 2006, pp. 731-740.

[6] M. Koziarski, "CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification" in Machine Learning, Cornell Univeristy, 2020.

[7] Almutairi, Waleed, and Ryszard Janicki. "On relationships between imbalance and overlapping of datasets." CATA. 2020.

[8] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf. Sci. (Ny)*, vol. 465, pp. 1–20, 2018.

[9] Nagarajan, Vinitha. A critical analysis of Sampling Techniques for imbalanced data classification: An application to Social Media. Diss. Dublin, National College of Ireland, 2017

[10] H. He, Y. Bai, E.A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", Proc. Int’l J. Conf. Neural Networks, pp. 1322-1328, 2008.

[11] I. Mani and I. Zhang, "KNN approach to unbalanced data distributions: A case study involving information extraction", *Proc. WLID*, pp. 1-7, 2003.

[12] R. T. Rockafellar, Convex Analysis, Princeton, NJ, USA: University Press, 1970.

[13] G. Weiss, K. McCarthy and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?", Proc. Int. Conf. Data Mining, pp. 35-41, 2007.

[14] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "*SMOTE: Synthetic Minority Over-Sampling Technique*", J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[15] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of SMOTE for mining imbalanced data", Proc. CIDM, pp. 104-111, Apr. 2011.

[16] H. Han, W.Y. Wang and B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", Proc. Int’l Conf. Intelligent Computing, pp. 878-887, 2005.

[17] S. Taneja, B. Suri and C. Kothari, "Application of Balancing Techniques with Ensemble Approach for Credit Card Fraud Detection", International Conference on Computing Power and Communication Technologies (GUCON), 2019.

[18] T. M. Barros, P. A. SouzaNeto, I. Silva, and L. A. Guedes, “Predictive models for imbalanced data: A school dropout perspective,” *Educ. Sci. (Basel)*, vol. 9, no. 4, p. 275, 2019

[19] K. Davagdorj, J. S. Lee, V. H. Pham, and K. H. Ryu, “A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention,” *Appl. Sci. (Basel)*, vol. 10, no. 9, p. 3307, 2020.

[20] J. Alcal-Fdez, A. Fernndez, J. Luengo, J. Derrac, S. Garca, L. Snchez, et al., "KEEL Data-Mining Software Tool: Data Set Repository Integration of Algorithms and Experimental Analysis Framework", *J. Multiple-Valued Logic and Soft Computing*, pp. 255-287, 2011.

[21] UCI Machine Learning Repository: Breast Cancer Wisconsin (Prognostic) Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic). [Accessed: 13-Mar-2021].

[22] G. Louppe, “Understanding random forests: From theory to practice,” Ph.D. dissertation, University of Liege, Belgium, 2014.

[23] L. Wang,Support vector machines: theory and applications. Berlin, Springer, 2005

[24] I. Tomek, "An Experiment with the Edited Nearest-Neighbor Rule", *IEEE Trans. Systems Man and Cybernetics*, vol. 6, no. 6, pp. 448-452, June 1976.

[25] Q. Wang, “A hybrid sampling SVM approach to imbalanced data classification,” *Abstr. Appl. Anal.*, vol. 2014, pp. 1–7, 2014.

[26] J. Fan, S. Upadhye and A. Worster, "Understanding receiver operating characteristic (ROC) curves", *Cjem*, vol. 8, pp. 19-20, 2006.

[27] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiol.*, vol. 143, pp. 29-36, 1982.

[28] MR. Prusty, T. Jayanthi and K. Velusamy, "Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors", *Progress in Nuclear Energy 2017*, vol. 100, pp. 355-364, 2017.

[29] O. M. Elzeki, M. Abd Elfattah, H. Salem, A. E. Hassanien, and M. Shams, “A novel perceptual two layer image fusion using deep learning for imbalanced COVID-19 dataset,” *PeerJ Comput. Sci.*, vol. 7,p. e364, 2021.