Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2733~2741
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.3121

Boosting with crossover for improving imbalanced medical datasets classification

Abeer S. Desuky, Asmaa Hekal Omar, Naglaa M. Mostafa
Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt

Article Info

Article history:

Received Oct 16, 2020
Revised Mar 17, 2021
Accepted Jul 6, 2021

Keywords:

Boosting
Classification
Crossover
Imbalanced datasets
Medical data

ABSTRACT

Due to the common use of electronic health databases in many healthcare services, healthcare data are available for researchers in the classification field to make diseases' diagnosis more efficient. However, healthcare-medical data classification is most challenging because it is often imbalanced data. Most proposed algorithms are susceptible to classify the samples into the majority class, resulting in the insufficient prediction of the minority class. In this paper, a novel preprocessing method is proposed, using boosting and crossover to optimize the ratio of the two classes by progressively rebuilding the training dataset. This approach is shown to give better performance than other state-of-the-art ensemble methods, which is demonstrated by experiments on seven real-world medical datasets with different imbalance ratios and various distributions.

Corresponding Author:

Abeer S. Desuky
Department of Mathematics
Faculty of Science, Al-Azhar University
Nasr city, Cairo, Egypt
Email: abeerdesuky@azhar.edu.eg

1. INTRODUCTION

Imbalanced datasets are referred to the situation where there are much more examples in one class than in the other. Training classifiers with imbalanced datasets is a common problem in the machine learning researchers' community. The trained classifier would become under fitted in categorizing test examples of minority class and over-fitted with massive median examples of the majority class. The classification in class imbalanced datasets has drawn great concern in the medical field because often the classes of instances that are diagnosed as not having a disease are significantly more than the classes of instances that are diagnosed as having a disease. In order to enhance the performance of classification in this field, many efforts have been made and are still being made. Some preprocessing rebalancing methods have been proposed in the past, especially in the aspects of artificially extending the minority class examples (over-sample), resampling down the amount of the majority class examples (under-sample), or the combination of them. Random over-sampling [1] and under-sampling [2] are the simplest methods. The first increase minority amount through copying its examples, and the second randomly delete majority class examples to achieve the balance. Synthetic minority oversampling technique (SMOTE) and its improvements [3]-[6] are the most widespread re-sampling methods that often achieve an efficient performance. In this algorithm, the characteristics of minority class examples' spatial structure are observed and analyzed to fabricate extra minority examples into the dataset.

Another type of proposed algorithm solves the problem of class imbalance during the training phase using cost-sensitive or ensemble-based learning approaches [7]-[9]. In cost-sensitive learning approaches, different weights are assigned to each part of the confusion matrix using the cost matrix to obtain a result

with minimum cost. Since the largest cost is the cost of minority class' misclassified examples, so the classifier will bias to the minority class. Hybridization between cost-sensitive learning techniques and decision tree (DT) [10] or feature selection [11] were proposed for solving the class imbalance problem. Ensemble learning-based mainly on voting and integrating a strong classifier from a collection of weak classifiers produced in several rounds or iterations. Boosting [12], bagging [13], and random forest [14] are the most widely used ensemble-based learning techniques. Most of these techniques were used for solving imbalanced medical data classification problems [14]-[16]. In this paper, we propose a novel approach called boosted crossover (BC) that is derived by the hibernation of the oversampling technique and boosting the classification performance. BC is a two-phase approach that first uses the bio-inspired crossover to rebuild new examples of the minority class, these examples have the same characteristics as their parents. Next, a weighting modification algorithm-the boosting algorithm-gives each classifier a weight-based on its own training error, which provides better performance over other oversampling algorithms, especially with highly imbalanced datasets. While other oversampling algorithms are based on repeating the minority class examples; the main advantage of our algorithm that it is based on using the crossover to build the new minority examples which mean that the resultant examples are new ones but also have the same characteristics as the original examples. The rest of this paper is organized as follows. Section 2 introduces the proposed approach. A description of the datasets and the experimental setup and results are presented in section 3. Finally, section 4 concludes the paper.

2. PROPOSED METHOD

The proposed algorithm is based mainly on two phases that run independently, for fixing imbalance problems in datasets. The idea of the first phase is dividing the training dataset into two groups-one of them contains examples of majority class and the other is the group of minority class-and resemble it back again after adding more examples generated by the bio-inspired crossover operator to the minority group. While the second phase depends mainly on boosting the classification process performance. Figure 1 shows the two phases of the proposed algorithm.

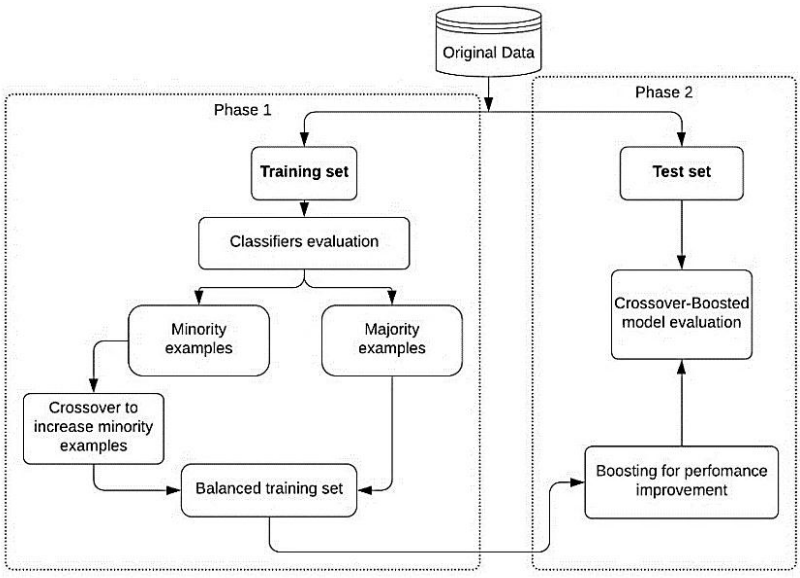

Figure 1. The flowchart of the proposed algorithm

2.1. First phase

The training dataset is divided into two -majority and minority class-groups. The group that contains the examples of minority class is fed to the first phase where the crossover operator is applied to it to generate new instances. Crossover, also called recombination, is a genetic operator used in the bio-inspired genetic algorithms and evolutionary computation, to generate new offspring via combining the genetic

information of two parents. It is a way to randomly generate new solutions from an existing population, and it is similar to the crossover that happens during biological sexual reproduction.

Crossover creates a new offspring by selecting genes from parent chromosomes. There are different types of crossover operators [17] that varies according to the number of crossover points and their locations on each chromosome. The simplest type is Single-point crossover which creates offspring by choosing a crossover point randomly and all genes before or after this point are exchanged between the two parents. This results in two offspring, each carrying some genetic information from both parents. This bio-inspired operator facilitates the inheritance of "characteristics" or "traits" by an offspring from its parents [18], so we choose it to generate new minority examples that carry the same characteristics as the original ones. Single-point crossover is selected among other types of crossover operators to be used in this work, for simplicity and decreasing time consumption of the code.

After applying the crossover operator on the minority class group, the majority class group is recombined with it, resultant in a new balanced training data that is ready to be fed into the second phase. Before applying the second phase, the new balanced training data is tested. Five classifiers [14], [19], [20]; random forest (RF), K-nearest neighbors (KNN), discriminant analysis (DA), naive bayes (NB), and support vector machine (SVM) -are used to test the seven-medical data and a comparison of the results with the existing SMOTE and safe-level SMOTE approaches [3], [4] indicates a significant performance improvement of the proposed method over them. Then, the second phase is applied to them.

2.2. Second phase

After confirming the readiness of the medical data, the second phase is progressed by implementing the adaptive boosting (AdaBoost) algorithm using the new balanced training data and the test data. It is a general approach for improving the classification performance of any given classifier. It converts weak classifiers to strong classifiers by improving the model predictions of the given algorithm. It produces a series of trained classifiers. Each member of the series modifies its training set based on the performance of the prior classifier in the series. All examples that are predicted incorrectly by earlier classifiers in the series are chosen moreover than examples that were predicted correctly. Thus, boosting tries to produce series of improved classifiers that have a better ability to predict examples for which the current classifier's performance is poor. There are many boosting approaches such as AdaBoost (adaptive boosting), gradient tree boosting, LightGBM, and XGBoost. in this paper, adaptive boosting is the selected boosting algorithm. The AdaBoost algorithm is one of the boosting algorithms that were proposed in [21]. The generalized version of the AdaBoost algorithm for binary classification problems is shown in Algorithm 1.

Algorithm 1. A generalized version of the AdaBoost algorithm

Given:

S = {(x_1 , y_1), (x_2 , y_2), . . . , (x_l , y_l)} With $x_i \in \boldsymbol{X}$

And $y_i \in \{-1, +1\}$.

Initialize the distribution:

$$\boldsymbol{D_l^{(i)} = \frac{1}{l}, \quad i = 1, 2, \ldots, l.}$$

For t = 1, 2, …, T:

Train the weak learner using the distribution

$$D_t^{(i)}, \quad i = 1,2,\ldots,l.$$

Get the weak hypothesis $c_t: X \to R$.

Update:

$$D_{t+1}^{(i)} = D_t^{(i)} \, exp(-\alpha_t y_i c_t(x_i))/Z_t, \quad i = 1, 2, \ldots, l,$$

where Z_t is a normalization factor ($D_t^{(i)}$ is still a distribution)

and $\alpha_t = \frac{1}{2} ln \frac{1-\varepsilon_t}{\varepsilon_t}$ with $\varepsilon_t = \sum_{i=1}^{l} D_t^{(i)}[y_i \neq c_t(x_i)]$.

Output the final hypothesis:

$$C(X) = sign(\sum_{t=1}^{T} \alpha_t c_t(X)).$$

AdaBoost combines iteratively the weak classifiers by considering a weight distribution on the training examples such that more weight is attributed to examples misclassified by the previous iterations. The final strong classifier takes the form of a perceptron, a weighted combination of weak classifiers

followed by a threshold. As in Algorithm 1, the algorithm takes as input a training set $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$ where each x_i belongs to some domain or instance X, and each label y_i is in some label set Y. T rounds of AdaBoost training are iterated where T is the number of weak classifiers c_t and ensemble weights α_t are yielded by learning to constitute the final strong classifiers [21], [22]. The weak classifier is the core of an AdaBoost algorithm; in our work, the classification and regression tree (CART) algorithm-proposed by Breiman *et al.* [23] is used as weak classifiers.

3. DATASET DESCRIPTION AND EXPERIMENTAL RESULTS

3.1. Dataset description

To evaluate the different performances, seven medical datasets from the UCI machine learning repository [diabetes (D2), mammographic masses (D3), pima (D4), haberman (D5), diagnostic wisconsin breast cancer (D6), heart disease (D7)] [24] and meander hand parkinson disease (D1) from [25] are used.

The diagnostic wisconsin breast cancer dataset contains information on the diagnostic of wisconsin breast cancer. the Mammographic Masses dataset discriminates the benign and malignant mammographic masses. The heart disease dataset is the part obtained from Cleveland Clinic Foundation and used to detect the presence of the disease in the patient's heart. Diabetes, diagnosis the chronic diabetes disease. Known as AIM-94 diabetes dataset and obtained from two sources: paper records and automatic electronic recording device. The Pima dataset contains two classes to test whether the patient is positive or negative for diabetes. The patients' records are for Pima Indian Women who live near Phoenix Arizona, USA. Haberman data contains records of patients who had undergone surgery for breast cancer at the University of Chicago's Billings Hospital between 1958 and 1970 and collected for a study that was conducted on the survival. Meander Hand Parkinson Disease dataset diagnoses a patient with Parkinson's disease at its early stage utilizing handwriting images acquired during handwriting exams performed by meanders are filled in forms. The numerical information of the data contained in the included imbalanced datasets is summarized in Table 1.

Table 1. Numerical information of imbalanced datasets

Data Sets	Examples	Attributes	MAJ: MIN
D1	368	11	296:72
D2	332	9	223:109
D3	961	6	516:445
D4	768	9	500:268
D5	306	3	255:81
D6	569	32	357:212
D7	209	8	117:92

3.2. Experimental results

To verify the performance, experiments were conducted in the MATLAB R2015a platform. on a computer equipped with 2.20GHZ core i7 processor and 6GB RAM. We performed our experiments in two phases: First, the original dataset is divided into two-Majority and Minority-class groups then, each group is divided equally into two subsets. After that, the equivalent percentage subsets were combined resultant training and testing sets with percent 50% and 50% respectively and have the same percentage of minority and majority classes.

The training set then used in the first phase and its minority class examples oversampled using the crossover operator to have the balance between the two classes. In order to test the validity of the balanced training set the five classifiers mentioned previously were used and the results compared with the results of two widely used oversampling methods (SMOTE and SLSMOTE). Recall, Precision, FScore, and GMean (geometric mean) are the performance measures used in this test besides the accuracy since, it is typically not enough information alone to validate algorithms performance, these measures are defined as [1], [26]:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$\text{Recall (sensitivity)} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (3)$$

Where TP is positive examples truly predicted as positive, FP is negative examples falsely predicted as positive, TN is negative examples truly predicted as negative, and FN is positive examples falsely predicted as negative.

With imbalanced datasets, often increases in recall come at the cost of decreases in precision, since in order to increase the TP for the minority class, also the number of FP is often increased, resulting in reduced precision. FScore provides a way to combine both recall and precision into a single score that achieves both properties and provides a way to express them with a single measure that can give a good indication to the classification of imbalanced data [1], [14].

$$\text{FScore} = 2 \text{ x Precision x Recall}/(\text{Precision} + \text{Recall}) \quad (4)$$

On the other hand, GMean is the square root of the product of class-wise accuracy (sensitivity for positive (minority) examples and specificity for negative (majority) examples). This measure tries to maximize the accuracy of both classes in balance. So, it is often used to evaluate the per-class accuracy of the classifiers. Traditionally if one class is unrecognized well by the classifier, GMean tends to zero [11].

$$\text{GMean} = \sqrt{\text{sensitivity} * \text{specificity}} \quad (5)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (6)$$

Table 2 and Table 3 show the classification performance measures (accuracy and FScore) for the five classifiers-RF, KNN, DA, NB, and SVM- applied on the seven medical datasets. In our experiment, each classifier is applied first on the imbalanced data and the performance measures are calculated for the test dataset (Or.). Then, the first phase is applied to the training data by implementing the bio-inspired crossover operator and the performance measures are recalculated for the same test dataset after train the classifiers using the new balanced data (Cross.). All datasets were rebalanced with SMOTE and SLSMOTE and tested in the same manner and the results were also included to compare the efficiency of our proposal against the other methods.

Table 2. Accuracy of five classifiers applied on the original and oversampled datasets

DATA		RF	KNN	DA	NB	SVM
D1	Or.	89.97	80.15	90.51	53.51	80.71
	Cross	96.03	89.02	83.10	64.78	91.76
	SMOTE	93.10	86.75	93.28	67.56	87.30
	SLSMOTE	93.66	87.12	66.26	66.26	86.75
D2	Or.	78.63	74.11	77.09	76.49	70.19
	Cross	91.67	87.50	81.50	79.02	86.97
	SMOTE	83.51	80.41	81.12	79.69	79.72
	SLSMOTE	84.70	79.22	81.84	80.18	78.77
D3	Or.	80.33	77.93	80.95	80.44	80.12
	Cross	87.62	84.81	84.60	83.46	87.30
	SMOTE	81.27	79.89	80.18	80.87	80.36
	SLSMOTE	81.56	80.29	80.58	80.68	81.25
D4	Or.	77.21	70.05	77.86	76.43	73.42
	Cross	90.27	85.81	82.01	81.25	84.94
	SMOTE	81.42	78.13	80.04	77.28	77.60
	SLSMOTE	81.10	76.32	80.46	78.97	77.07
D5	Or.	68.89	72.16	74.45	74.79	72.23
	Cross	73.51	68.38	61.08	58.32	71.80
	SMOTE	72.31	73.61	65.98	65.70	75.74
	SLSMOTE	73.49	74.20	65.25	63.61	74.47
D6	Or.	96.13	78.39	95.43	93.32	85.23
	Cross	98.08	89.82	97.58	94.86	90.73
	SMOTE	96.75	82.0	96.15	93.20	87.17
	SLSMOTE	97.05	84.35	96.46	93.36	88.35
D7	Or.	79.90	62.21	78.95	77.04	60.19
	Cross	85.36	69.10	78.08	78.76	69.77
	SMOTE	82.41	67.54	79.34	76.26	64.98
	SLSMOTE	82.82	68.87	80.65	78.45	65.41
Win	Or.	-	-	1	1	-
	Cross	7	6	3	4	6
	SMOTE	-	1	1	1	1
	SLSMOTE	-	-	2	1	-

Table 3. FScore of five classifiers applied on the original and oversampled datasets

DATA		RF	KNN	DA	NB	SVM
D1	Or.	83.01	63.93	63.49	63.49	60.27
	Cross	96.03	89.46	82.99	67.05	91.70
	SMOTE	83.71	64.43	84.21	67.32	65.66
	SLSMOTE	85.14	66.58	84.70	66.73	61.37
D2	Or.	75.28	69.09	72.94	73.49	62.70
	Cross	89.72	84.33	76.87	75.30	76.51
	SMOTE	77.47	72.24	73.91	74.04	70.52
	SLSMOTE	79.25	70.82	74.98	75.14	69.07
D3	Or.	80.24	77.80	81.19	80.39	79.98
	Cross	84.08	80.79	80.28	79.77	83.53
	SMOTE	80.98	79.47	80.32	80.68	80.00
	SLSMOTE	81.26	79.87	80.72	80.49	80.89
D4	Or.	74.43	66.60	74.73	73.52	69.16
	Cross	87.32	81.1	75.65	76.00	79.83
	SMOTE	75.91	71.35	73.84	71.31	69.65
	SLSMOTE	75.64	69.77	74.52	73.52	69.12
D5	Or.	57.35	59.88	60.18	56.02	52.89
	Cross	73.70	68.34	61.22	61.65	71.72
	SMOTE	71.09	72.65	63.79	63.65	74.40
	SLSMOTE	73.08	74.26	64.60	63.38	74.11
D6	Or.	95.85	76.34	95.17	92.82	83.97
	Cross	97.92	88.84	97.38	94.47	89.84
	SMOTE	96.74	82.05	96.22	93.23	87.14
	SLSMOTE	97.04	84.44	96.52	93.36	88.38
D7	Or.	79.54	61.29	78.65	76.56	58.79
	Cross	84.45	67.34	77.54	78.57	66.91
	SMOTE	81.63	65.66	78.29	75.17	62.29
	SLSMOTE	82.06	67.19	79.73	77.45	63.07
Win	Or.	-	-	1	-	-
	Cross	7	6	3	5	6
	SMOTE	-	-	-	-	-
	SLSMOTE	-	1	3	2	1

We can notice from the results in Tables 2 and 3 that on 6 out of 7 datasets, the highest performance is achieved by our method. Of course, on the remaining datasets, sometimes the performance of our method is very close to the performances of the other two methods. But in other cases, it shows improvements in performance by more than 10%; as in FScore of D1, D2, and D4. The record named Win in both tables represents the number of datasets with which each method performed the best among the others and it is evident that our proposal has the superiority with all datasets for both accuracy and FScore of the RF classifier. It also gains an excellent performance with KNN and SVM classifiers and good performance with DA and NB classifiers. This can be noticed also in Figures 2-6. Figures 2-6 show the improved distance between the performance measures (Precision, Recall, and GMean) for the five classifiers RF, KNN, DA, NB, and SVM respectively.

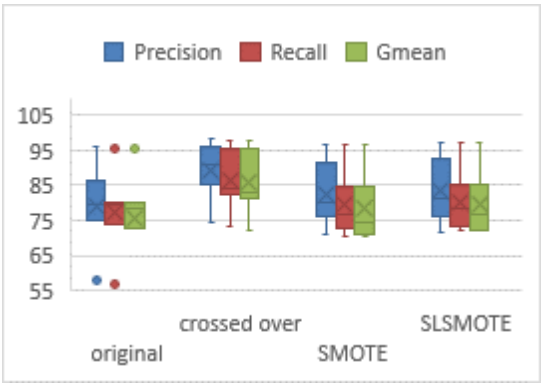

Figure 2. Improved distance for RF classifier

Figure 3. Improved distance for KNN classifier

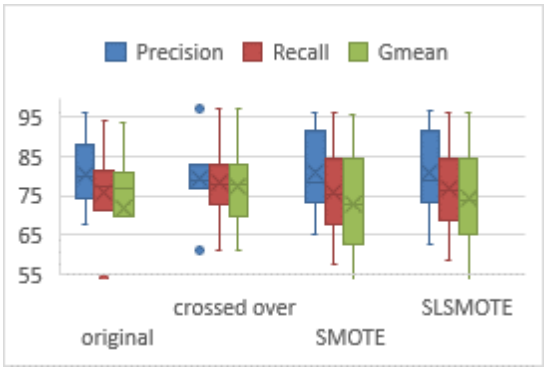

Figure 4. Improved distance for DA classifier

Figure 5. Improved distance for NB classifier

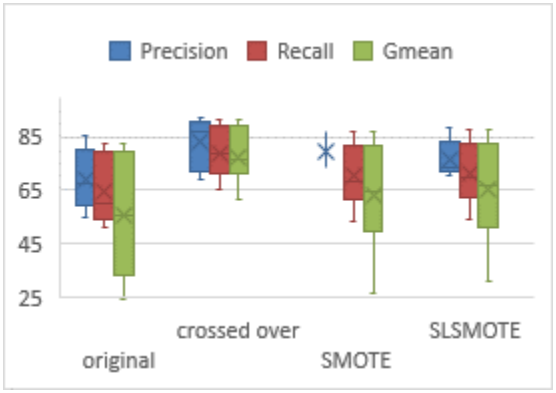

Figure 6. Improved distance for SVM classifier

Table 4 shows all performance measures used after applying boosting–with regression tree (CART) as weak classifier-on the imbalanced training set (Or.) and on the new balanced training set resulted by applying the proposed method (BC) in the second phase. As noticed some data D1, D3, D4, and D7 show a decrease in the precision value due to an increase in the TP for the minority class as mentioned earlier. The remaining datasets show an increase in their precision. Although, All datasets record increase in accuracy, recall, FScore, and GMean measures except the D5 dataset which have a decrease in these measures, except Precession and GMean, This may return to the nature of the data since it has an extremely low number of features.

Table 4. performance measures after applying boosting on original and crossed over datasets

DATA		Acc.	Prec.	Recall	FScore	GMean
D1	Or.	90.21	82.14	63.88	71.87	78.56
	BC	90.76	67.92	100	80.89	94.08
D2	Or.	70.48	56.00	50.90	53.33	63.88
	BC	78.91	63.88	83.63	72.44	80.02
D3	Or.	77.29	75.79	74.77	75.28	77.08
	BC	77.50	70.35	88.73	78.48	77.58
D4	Or.	69.53	56.69	53.73	55.17	64.73
	BC	71.09	55.45	87.31	67.82	73.81
D5	Or.	70.58	78.63	82.14	80.34	56.61
	BC	67.97	89.87	63.39	74.34	71.43
D6	Or.	96.48	100.0	94.38	97.11	97.15
	BC	97.89	100.0	98.28	97.19	98.30
D7	Or.	70.19	68.29	60.86	64.36	68.72
	BC	70.19	61.53	86.95	72.07	70.33

From the conclusions drawn above, the results presented in Tables 2, 3, and 4 also reveal the efficiency of our proposed algorithm where, it behaves excellently on all major performance metrics,

especially for the metrics that can reflect the trade-off between negative and positive classification performance (FScore and GMean) and outperforms SMOTE and SLSMOTE with 3 classifiers out of 5 applied in all used medical datasets.

4. CONCLUSION

Boosted crossover (BC) is an effective method for solving the problem of imbalanced medical data. To best of our knowledge crossover operator with boosting has been utilized for the first time to balance the imbalanced medical datasets. The proposed method rebalances the data by increasing the number of minority class examples which improves the performance of medical diagnosis systems. First, the bio-inspired crossover operator is used to build new examples, then the new balanced data is tested using five different classifiers and compared with two other oversampling methods. Finally, adaptive boosting is used to boost the performance of the system. Experimental results conducted on seven medical datasets prove that Boosted Crossover is very efficient for enhancing the classification performance measures especially with RF, KNN, and SVM classifiers.

REFERENCES

[1] Gongqian Liang, Zhonghui Dong, Baoyu., Tan, and Baosheng Zhang, "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data," *Mathematical Problems in Engineering*, vol. 2019, doi: 10.1155/2019/3526539.

[2] Bin Liua, Grigorios Tsoumakasa, "Dealing with Class Imbalance in Classifier Chains via Random Undersampling," *Knowledge-Based Systems*, vol. 192, 2019, doi: 10.1016/j.knosys.2019.105292.

[3] N.V. Chawla, K.W. Bowyer, "SMOTE: synthetic minority over-sampling technique.," *J. Artif. Intell. Res.*, vol. 16, pp. 341–378, 2002.

[4] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," In *Advances in Knowledge Discovery and Data Mining*, pp. 475–482, 2009, doi: 10.1613/jair.953.

[5] S. Maldonado, J. López and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput.*, vol. 76, pp. 380–389, 2019, doi: 10.1016/j.asoc.2018.12.024.

[6] Paria Soltanzadeh and Mahdi Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences*, vol. 542, pp. 92–111, 2020, doi: 10.1016/j.ins.2020.07.014

[7] G. Rekha, Amit Kumar Tyagi and V. Krishna Reddy, "Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method," International Journal of Hybrid Intelligent Systems, vol. 15, no. 2, pp. 67-76, 2019, doi: 10.3233/HIS-190261

[8] X. Yuan, L. Xie and M. Abouelenien, "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data," *Pattern Recognition*, vol. 77, pp. 160–172, 2018, doi: 10.1016/j.patcog.2017.12.017.

[9] Marcelino Lázaro, Francisco Herrera and Aníbal R. Figueiras-Vidal, "Ensembles of cost-diverse Bayesian neural learners for imbalanced binary classification," *Information Sciences*, vol 520, pp, 31–45, 2020, doi: 10.1016/j.ins.2019.12.050.

[10] M. Aldiki Febriantono, Sholeh Hadi Pramono, Rahmadwati and Golshah Naghdy, "Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0" *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 9, No. 1, pp. 65-72, March 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[11] F. Feng, K. Li, J. Shen, Q. Zhou and X. Yang, "Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification," in IEEE Access, vol. 8, pp. 69979-69996, 2020, doi: 10.1109/ACCESS.2020.2987364.

[12] Sundar R. and Punniyamoorthy M., "Performance enhanced Boosted SVM for Imbalanced datasets," Applied Soft Computing Journal, vol. 83, 2019, 10.1016/j.asoc.2019.105601.

[13] G. Rekha, V. K. Reddy, A. K. Tyagi and M. M. Nair, "Distance-based Bootstrap Sampling in Bagging for Imbalanced Data-Set," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, pp. 1-6, 2020. doi: 10.1109/ic-ETITE47903.2020.345.

[14] Engy El-shafeiy and Amr Abohany, "Medical Imbalanced Data Classification Based on Random Forests," *AICV 2020*, AISC 1153, pp. 81–91, Springer Nature Switzerland AG 2020.

[15] Pattaramon Vuttipittayamongkol and Eyad Elyan, "Overlap-Based Undersampling Method for Classification of Imbalanced Medical Datasets," *AIAI 2020*, IFIP AICT 584, pp. 358–369, 2020.

[16] Ahmed Jameel Mohammed, Masoud Muhammed Hassan, Dler Hussein Kadir, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *International Journal of of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161-3172, May–June 2020, doi: 10.30534/ijatcse/2020/104932020.

[17] A.J. Umbarkar1 and P.D. Sheth., "Crossover Operators in Genetic Algorithms: A Review," *Ictact Journal on Soft Computing*, vol. 06, no. 01, 2015, doi: 10.5120/ijca2017913370.

[18] A. Ghasempour and M. B. Menhaj, "A new genetic based algorithm for channel assignment problem," in *Computational Intelligence, Theory and Applications*, Ed. B. Reusch, Berlin: Springer, 2006, pp. 85-92.
[19] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169-190, 2017, DOI: 10.3233/AIC-170729.
[20] Sen P.C., Hajra M. and Ghosh M., "Supervised Classification Algorithms in Machine Learning: A Survey and Review," *Advances in Intelligent Systems and Computing*, vol 937, Springer, Singapore, 2020, doi: 10.1007/978-981-13-7403-6_11.
[21] Freund, Y. and R. Schapire, "Experiments with a New Boosting Algorithm," ICML, 1996.
[22] Feng D-C, Liu Z-T, Wang X-D, Chen Y, Chang J-Q, Wei D-F and Jiang Z-M., "Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach," *Constr Build Mater*, 230:117000, 2020, doi: 10.1016/j.conbuildmat.2019.117000.
[23] L. Breiman, J. Friedman, R. Olshen, and C. Stone., "Classification and Regression Trees", Chapman and Hall Publisher, New York, USA, 1984.
[24] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[25] S. Dong and J. Jeong., "Onset Classification in Hemodynamic Signals Measured During Three Working Memory Tasks Using Wireless Functional Near-Infrared Spectroscopy," in *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1-11, Jan.-Feb, 2019, doi: 10.1109/JSTQE.2018.2883890.
[26] Abeer S. Desuky, "Severity of Breast Masses Prediction in Mammograms Based on Optimized Naive Bayes Diagnostic System," *Adv. in Systems Science and Appl*., vol. 18, no. 1, 2018, doi: 10.25728/assa.2018.18.1.266.