Arabian Journal for Science and Engineering (2021) 46:3853–3864
https://doi.org/10.1007/s13369-021-05347-7

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

An Improved Hybrid Approach for Handling Class Imbalance Problem

Abeer S. Desuky[1] · **Sadiq Hussain**[2]

Received: 23 September 2020 / Accepted: 12 January 2021 / Published online: 28 January 2021

Abstract
Class imbalance issue that presents in many real-world datasets exhibit favouritism toward the majority class and showcases poor performance for the minority class. Such misclassifications may incur dubious outcome in case of disease diagnosis and other critical applications. Hence, it is a hot topic for the researchers to tackle the class imbalance issue. We present a novel hybrid approach for handling such datasets. We utilize simulated annealing algorithm for undersampling and apply support vector machine, decision tree, *k*-nearest neighbor and discriminant analysis for the classification task. We validate our technique in 51 real-world datasets and compare it with other recent works. Our technique yields better efficacy than the existing techniques and hence it can be applied in imbalance datasets to mitigate the misclassification.

Keywords Imbalance datasets · Simulated annealing · Undersampling · Oversampling · Classification

1 Introduction

Machine learning is a well-known research domain in computers science to use a variety of algorithms to extract useful information among huge raw data. These algorithms have widely been applied to different subjects such as medical data analysis [1–4], class noise detection [5], image processing [6], sentiment analysis [7, 8], signal processing [9, 10], road accident analysis [11], social data mining [12] and many more. However, in most cases, we may not have a balanced dataset. Having a balanced dataset can benefit different machine learning algorithms to learn various circumstances of different classes. For this reason, dealing with imbalanced datasets is a challenging task in machine learning. In the case of imbalanced data, an uneven distribution of samples occurs among a variety of classes as the minority class has notably fewer samples than the majority class. This imbalanced data case is very widespread in real applications such as disease diagnosis, network intrusion recognition, and software flaw prediction. This biasness arises due to the favouritism of the learned classifier toward the majority population while ignoring the minority samples. Nevertheless, the recognition of the minority records with special property is of utmost importance while dealing with imbalance domains. For instance, in the medical data analytics area, the wrong classification of COVID-19 patient (a minority sample) as a non-COVID-19 subject will incur a high or even unacceptable cost. Software security, financial fraud prediction and helicopter fault monitoring are similar examples of this sort. With great attention or afflux devoted to combating class imbalance issue, various solutions have been devised. These solutions may be categorised as two forms: data-level and algorithm-level methods. The data-level method mitigates the majority records (undersampling) and the number of minority records is enhanced (oversampling) or integrate both of them to correct imbalance scenario. The inductive bias toward the majority samples is lessened by adjusting the prevailing learning approaches in the algorithm-level techniques. The data-level approaches are more frequently utilized in comparison to algorithm-level strategies as they can be integrated with other techniques such as ensemble as well as active learning to devise intricate hybrid methods and this approach does not depend on any specific classifier.

In our study, we have utilized an undersampling method for the instances of the majority samples. Majority class samples from a definite number of clusters are eradicated by applying cluster-oriented undersampling to balance the training dataset [13]. Underlying data distribution affects

✉ Sadiq Hussain
sadiq@dibru.ac.in

Abeer S. Desuky
abeerdesuky@azhar.edu.eg

[1] Faculty of Science, Al-Azhar University, Cairo, Egypt

[2] Examination Branch, Dibrugarh University, Dibrugarh, India

the distance-based abolition of occurrences. Weighted learning principle of infrequent instances makes ensemble learning methods an effective solution. Devi et al. [13] utilized the AdaBoost ensemble method to eliminate irrelevant majority class records from the clusters. Mohammed et al. [14] in their study empirically examined two resampling approaches-undersampling and oversampling. They exploited several machine learning techniques with various hyperparameters that yielded superior outcomes for both the resampling techniques. Liu et al. [15] coupled Ensemble of Classifier Chains (ECC) with random undersampling to make ECC flexible to class imbalance. Chains of numerous sizes were constructed to increase the exploitation of majority class records and binary models per label were also devised. If there is an overlap of records along with imbalance, then it makes the learning task trickier [4]. Overlapped data points were eliminated in binary datasets by introducing an undersampling approach by Vuttipittayamongkol et al. [16]. Their techniques were devised to recognize and remove majority class records from the region of overlapping. Possible overlapped examples were identified by employing four techniques that exploited neighbourhood searching with numerous criteria. The authors [17] exploited the elimination threshold and soft clustering techniques to take out negative records in the overlapping area. Sarkar et al. [18] devised an ensemble learning-based undersampling method utilizing Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM). They validated their method on a steel plant accident dataset. The outcome showed that their novel method effectively resolved the class imbalance problem.

In imbalance settings, there is lots of scientific literature that utilized discriminant analysis (DA), decision tree (DT), SVM and *k*-nearest neighbor (*k*-NN). Bejaoui et al. [19] presented an improved regularized-quadratic discriminant analysis (R-QDA) that utilized a modified bias and two regularization parameters, properly selected to evade improper characteristics of R-QDA in the imbalanced scenario and hence ensured enhanced the result of the classifier in best possible way. The presented classifier used a random matrix theory-based scrutiny of its presentation when the number of features and that of samples grew huge simultaneously. Jian et al. [20] proposed a novel contribution sampling method based on the contributions of the non-support and support vectors to classification. Dubey et al. [21] devised a modified kNN technique so that it could handle the class allocation in a broader region around the query example. They empirically validated their method on several real-world datasets. Liu et al. [22] introduced a novel decision tree method which generated rules and was statistically significant and also insensitive and robust to the size of classes. They employed the metric applied in C4.5, Information Gain concerning confidence of a rule to make decision trees robust.

Simulated annealing (SA) is also used in machine learning paradigm in various applications. Tóth et al. [23] utilized SA for speedy optimization of parameters of an object recognizer ensemble over huge image databases. Yang et al. [24] proposed a novel edition of monarch butterfly optimization (MBO) with SA method termed as SAMBO. The migration operator and butterfly adjusting operator was exploited by the SA method. The experiments were carried on 14 continuous nonlinear functions. Camelo et al. [25] demonstrated empirically utilizing Fast Simulated Annealing and metaheuristics Simulated Annealing for optimization of Multilayer Perceptron (MLP) to optimize the hyper-parameters. The model was applied to optimize two parameters: the configuration of the neural network layers and its neuron weights.

We devise an improved hybrid method to handle the unbalanced settings and hence improve the overall performance. To best of our knowledge, simulated annealing strategy with machine learning classifiers has been utilized for the first time to balance imbalanced datasets. Optimization is the process of achieving the best solution for a problem (selecting an optimal subset of majority class instances); in this work, using simulated annealing optimization technique helps in improving the objective function (classification performance) value. We adopt undersampling using simulated annealing, different classifiers viz. discriminant analysis, SVM, decision tree, kNN. Unlike most other optimization algorithms, SA applies a cooling strategy to locate optimal solutions ignoring local optima while searching the solution space and converging to the global optimum. Moreover, using the *F*-score metric as an objective function helping in selecting the examples that improve the overall accuracy for both majority and minority class. We evaluate our technique in several real-world datasets including UCI and KEEL datasets. The performance of our technique is comparable to many existing methods in this domain. It also outperformed different techniques in terms of *F*-score, accuracy, AUC, and *G*-mean. We also perform the Wilcoxon's Signed Rank test and the *p*-values indicate that there is a significant difference between the proposed method and state-of-the-art pre-processing techniques. Our hybrid technique is simple, efficient and easy to implement. Hence, it can be applied to balance many real-world datasets as it has been tested in 51 benchmark datasets.

2 Literature Review

Sampling methods in combination with ensemble classification techniques have demonstrated its efficacy in real-world problems, especially to resolve class imbalance issue. Tsai et al. [26] devised a novel undersampling technique that integrated instance selection and clustering analysis. Analogous

data records of the majority cohort were assembled into subgroups by the clustering technique. Misleading data samples were sorted out from the subclasses by the instance selection method.

The classification problem with class imbalanced data in the medical domain has attracted many researchers. Most of the prevailing techniques categorize samples into the majority class that resulted from bias and inadequate recognition of minority class. Zhu et al. [27] proposed a new method called class weights random forest to tackle this issue. Their technique could detect both minority and majority class with high accuracy and hence improved the overall performance of the classification algorithm.

Li et al. [28] presented a unified pre-processing method utilizing stochastic swarm heuristics to jointly optimize the mixtures from the two classes by gradually reconstructing the training dataset. Their method exhibited competitive performance in comparison with popular techniques.

Li et al. [29] devised a new hybrid approach dubbed as ant colony optimization resampling (ACOR) to tackle class imbalance issue. ACOR consists of two stages: at first, a particular oversampling method was employed to rebalance an imbalanced dataset; in next stage, it applied ant colony optimization to detect an (sub) optimal subset from the balanced dataset. The benefit of using this approach was that a perfect training set could be achieved by the optimization technique and prevailing oversampling methods could be fully applied. The evaluation metrics confirmed that enhanced performance was recorded by ACOR and yielded a better outcome than four popular oversampling methods.

The analysis of medical data from electronic health records (EHRs) poses a great challenge due to its imbalanced and heterogeneous characteristics. Huda et al. [30] reviewed the challenges by utilizing brain tumor images. They integrated ensemble-based classification and feature selection methods to demonstrate an affordable and fast detection of the genetic variant of a brain tumor. To mitigate the effect of imbalanced characteristics of medical data, they hybridized ensemble classification with feature selection.

Febriantono et al. [31] applied decision tree C5.0 of cost-sensitive type to work out imbalanced data issue of multi-class nature. At the first step, C5.0 algorithm was utilized by the decision tree model. Afterwards, the minimum cost model was obtained by using the cost-sensitive learning. The results performed on testing dataset asserted that C5.0 demonstrated better performance than its counterparts ID3 and C4.5 algorithms.

Babu et al. [32] proposed a genetic algorithm (GA)-based error classification for imbalanced dataset. For error identification and dataset processing, principle component analysis (PCA) was utilized. The errors presented in a dataset exhibited in a binary form by their approach. Error location identification was achieved through GA. The GA-based approach had successfully recognized the error location and enhanced the processing time of the imbalanced dataset.

The classical extreme learning machine (ELM) algorithm is unable to generate better performance in case of the imbalanced dataset. Ri et al. [33] defined a novel cost function based on *G*-mean for ELM optimization problem in imbalanced data learning. They tested their methodology on 11 multi-class and 58 binary repositories having diverse gradation of imbalance ratio. Their approach outperformed the classical ELM and yielded competitive performance in comparison to prevailing methods.

Susan et al. [34] applied a new hybrid technique of learning from imbalanced datasets by undersampling and oversampling of the majority and the minority cohorts' samples, respectively. They utilized different and intelligent versions of oversampling methods. The decision tree method was fed to the datasets after balancing it. Empirical experiments proved the efficiency of their technique as higher accuracies were achieved compared to the baseline techniques.

El-Shafeiy et al. [35] carried out a study of class imbalance in the domain of medicine. They applied random forests (RF) for oversampling and undersampling strategies by integrating decision trees to subgroups of the dataset. Their RF-based techniques yielded enhancement in the area of imbalance medical dataset.

Yang et al. [36] devised an integrated scheme by combining weight functions and weight constant of cost-sensitive learning techniques into the regularized risk minimization approach. Their results showcased that their methods could mitigate the misclassification cost efficiently while taken care of the privacy requirement. Their empirical evidence revealed that the selection of weight functions and weight constant did not have an impact on the Fisher-consistent property but the performance of the classifiers were influenced highly by interacting with privacy-preserving levels.

Abnormal state detection and feature extraction are the key issues in class imbalance thermal signals. Wang et al. [37] designed an improved framework that incorporated hidden information and prior knowledge in the class imbalance condition for sintering state recognition. They fused hidden information and prior knowledge to devise a cascaded stack Autoencoder model for distinguished feature extraction of imbalance records. They also presented a data-dependent kernel modification optimal margin distribution machine (ddKMODM) as a sintering state recognition model.

3 Methods

3.1 Simulated Annealing (SA)

SA is a simple and well-known metaheuristic technique utilized in global optimization issues, whose objective function

can be examined via computer simulation [36]. Real-world issues are tackled by it. Annealing in the early 1980s was proposed by Kirkpatrick et al. [37] in the optimization of combinatorial nature. The process involves increasing the temperature of a solid and then make the state of energy lower. The two stages are depicted as follows:

- Bring the solid to a very elevated temperature until "melting" of the structure;
- Cool the solid consistent with a very specific temperature declining scheme to attain a solid-state of minimum energy.

Arbitrary allocation of particles is done at the liquid stage. With long cooling time and high temperature at the initial state facilitates the least energy phase. A metastable position with the energy of non-minimal can be achieved by the solid due to non-occurrence of the case. Abrupt cooling of the solid is termed as hardening.

In the state space S, a categorization of the solution is produced by the Metropolis algorithm which is utilized in SA method. An equivalence is set between a multiple particle system and the optimization issue to implement it as follows:

- The potential solid states are described by the solutions
- The solid's energy is denoted by the minimized function

After that initialization of control parameter occurs. The objective and the parameter conveyed with units of the same type. A neighborhood, a solution generated by the system in the neighborhood and points in the state space are assumed to be provided by the user. The principle of acceptance is described as below:

Definition 1 Let two points of state space be p,q and (R,fn) be an instance of combinatorial minimization issue. The probability of accepting solution q from the present solution p is described by the condition of acceptance as below:

$$\text{Prob}\{\text{accept } q\} = \begin{cases} 1 & \text{if fn}(q) < \text{fn}(p) \\ e^{\left(\frac{\text{fn}(p)-\text{fn}(q)}{g_y}\right)} & \text{else} \end{cases} \quad (1)$$

By comparison, the perturbation method of the Metropolis technique is agreed upon by the principle of creating a neighbour and Metropolis condition is governed by the principle of acceptance.

Definition 2 The alternation of the existing solution by a neighboring solution is dubbed as a transition. The transition is executed in acceptance and generation phases.

Let g_y be the value of the temperature parameter and X_y be the total transitions produced with some iteration y in the sequel. The norm of SA is denoted as below:

1. Start with setting (p:=p_{start}, y:=0,g_y=g_0, X_y:=X_0);
2. do
3. For x=0 to X_y do
   - From the present solution p, a solution q is generated from the neighbourhood $R_{p;}$
   - p:=q if fn(q)<fn(p) and the present solution is then q;
   - Otherwise, the present solution with probability $e^{\left(\frac{fn(p)-fn(q)}{g_y}\right)}$ is q;
4. y:=y+1;
5. Calculate (X_y, g_y);
6. while $g_y \simeq 0$.

The capacity to recognize changeovers that demean the objective function is one of the prime characteristics of SA.

3.2 Datasets

Public datasets that are employed in our experiments are listed in Table 1. UCI and KEEL are the repositories from where these datasets belong. The class imbalance degree is defined as follows:

$$\text{Imbalance degree (imb)} = \frac{N}{P} \quad (2)$$

where positive and negative records are marked as P and N, respectively [16]. One can differentiate these datasets with respect to imbalance degree, number of features and instances. The model was chosen in the training phase with tenfold cross-validation.

3.3 Proposed Method

The novel hybrid technique is described as below:

Our undersampling method is using SA with each classifier by enhancing the F-score (Eq. 7) to derive the best

Table 1 List of datasets

Dataset	Records	Number of features	Imbalance ratio
abalone19	4174	8	129.44
Page-blocks	5473	10	61.19
Yeast6	1484	8	41.4
Ecoli0137vs26	281	7	39.14
Abalone	4177	8	35.32
yeast5	1484	8	32.73
Yeast4	1484	8	28.1
Glass5	214	9	22.78
Abalone09-18	731	8	16.4
Page-blocks13vs2	472	10	15.86
Ecoli4	336	7	15.8
Glass4	214	9	15.46
Yeast1vs7	459	7	14.3
shuttle-c0-vs-c4	1829	9	13.87
ecoli-0–1-4-6_vs_5	280	6	13
Glass2	214	9	11.59
Vowel	990	10	10
Vowel0	988	13	9.98
Yeast2vs4	514	8	9.08
Ecoli3	336	7	8.6
Yeast3	1484	8	8.1
Cleveland	297	13	7.48
Segmemt0	2308	19	6.02
Ecoli2	336	7	5.46
Hepatitis	80	19	5.15
New-thyroid1	215	5	5.14
New-thyroid2	215	5	5.14
Newthyroid	215	5	5.14
Libra	360	90	4
Contraceptive	1473	9	3.42
Ecoli1	336	7	3.36
Vehicle0	846	18	3.25
Transfusion	748	4	3.2
Parkinsons	195	22	3.06
Vehicle1	846	18	2.9
vehicle2	846	18	2.88
Haberman	306	3	2.77
ILPD	579	10	2.49
Breast	277	9	2.41
Glass0	214	9	2.06
Iris	150	4	2
Breast_tissue	106	9	1.94
Tic-tac-toe	958	9	1.89
Pima	768	8	1.87
Wisconsin	683	9	1.86
Ionosphere	351	34	1.79
BreastEW	569	30	1.68
Wine	178	13	1.51
Bupa	345	6	1.37
Liver_disorders	345	6	1.38
Heart	270	13	1.25

possible subset of the majority class records from the training set. The steps are as follows:

1. Split the data into training (50%), validation (25%), and testing (25%).
2. Send the training and validating to the SA algorithm that uses the *F*-score of the classifier as an objective function.
3. The population used is vectors of zeros and ones; each vector is the same size as the majority examples in the set meant for training purpose. One denotes the corresponding example stay in the training set, zero implies removing the corresponding example from the set.
4. Each classifier is trained and tested using the validation set and SA uses *F*-score as its objective function.
5. Train each classifier using the undersampled training subset resulted from the SA, and test the model using the testing subset. The evaluation metrics used are accuracy (Eq. 3), *G*-mean (Eq. 8), AUC, and *F*-score.

3.3.1 Details of the Selection of Optimal Subset of Majority Instances by Using SA

The optimal subset of majority instances selection problem is defined as follows:

Definition 1 The optimal subset of majority instances selection problem.

Given a set of majority instances $G = \{G_1, G_2, G_3, \ldots, G_m\}$ and a cost function $C{:}G{->}s$ $(0 \leq s)$, find the subset such that the value of the cost function is minimized.

An initial solution is needed by most of the optimization problem including SA [38]. A feasible solution is randomly selected and marked as an initial solution. The one-bit difference in the binary vectors from the projected solution is exploited as the neighboring solutions. The cost function is one of the significant factors for examining individual solutions and hence critical in heuristic optimization method like SA. The basic concept used in this paper for the cost function is to exploit the *F*-score of classification by applying majority instances represented by the given solution.

While exploring the solution space by evading the local optima to locate the best possible solutions, a cooling strategy is adopted by SA. The cooling strategy indicates a

scheme for how to search. Initial temperature, termination condition and temperature declining functions are applied as parameters. Adequate transitions can be achieved by allocating the starting temperature enormous. The product of temperature and a constant x is used as temperature declining function. If the temperature is less than a particular value 0.0001, then the method terminates. That particular value is obtained by executing many trails.

3.3.2 The Algorithm: Optimal Subset of Majority instances selection method utilizing SA

At first, randomly an initial solution is chosen and it is considered as the optimal solution. The cost function is utilized to compute the cost of the initial solution. As long as temperature Temp does not satisfy the terminating criteria, a neighbouring solution of the current optimal solution is chosen and its cost is also determined. If the current optimal solution's cost is equal to or less than the newly chosen neighboring solution, the newly chosen optimal solution replaces the current optimal solution. If the cost of the neighboring solution is higher than the current optimal solution, a random value s is chosen in the range of (0,1). Following 6, temperature T is reduced and the whole strategy continues until Temp satisfies the terminating condition.

3.4 Evaluation Metrics [39]

Let us say TP, TN denotes true positives, true negatives, while FP, FN depicts False positives and False negatives, respectively. N_{TP} denotes the number of true positives and so on. Accuracy, precision, TNR, F-measure, G-mean are calculated as follows:

$$\text{accuracy} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{P}} + N_{\text{N}}} \tag{3}$$

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{4}$$

$$\text{recall} = \text{sensitivity} = \text{TPR} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{5}$$

$$\text{Specificity} = \text{TNR} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}} \tag{6}$$

$$F\text{ - score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{7}$$

$$G\text{ - mean} = \sqrt{\text{TPR} \times \text{TNR}} \tag{8}$$

Input: The dataset meant for training

Output: Optimal Subset of Majority instances: O_m

1. $O_m \leftarrow$ Empty; //terminal solution
2. Temp $\leftarrow$ 80000;
3. x $\leftarrow$ 0.8;
4. Produce the starting solution O_i;
5. $O_m \leftarrow O_i$;
6. Cost (O_m) , starting solution, is computed;
7. While loop begins with condition (Temp > 0.0001)
8. start:
9. The neighbor solution, O_n, is chosen randomly with a one-bit difference from starting solution O_m;
10. If the cost function of the starting solution is equal to the neighbor solution:
11. $O_m \leftarrow O_n$;
12. otherwise
13. In (0,1) range, a random number s is selected;
14. $if\ \ (s < e^{\frac{-(\text{Cost}(\vartheta n) - \text{Cost}(\vartheta b))}{\text{Temp}}})$
15. $O_m \leftarrow O_n$;
16. Temp $\leftarrow$ x multiplied by Temp
17. Stop // while loop ends

Table 2 *G*-mean values of the proposed method and the best values obtained by methods in [16] for each dataset

Dataset	Proposed				Best in [16]							
	DA	DT	SVM	KNN	NB-SVM	NB-RF	SMOTE	BLSMOTE	ENN	kmUnder	OBU	Baseline
Wisconsin	96.35	94.9	95.85	*96.76*	**97.12**	**97.12**	96.66	96.66	*96.77*	96.66	51.65	*96.77*
Pima	**74.71**	*70.77*	*74.02*	58.4	55.24	*74.02*	66.67	66.67	66.04	67.8	50.26	64.08
Glass0	67.08	*77.78*	64.66	70.01	73.19	77.78	*78.26*	75	*78.26*	**80.34**	68.29	66.67
Vehicle1	*79.41*	*76.28*	**81.03**	68.97	55.92	**81.03**	61.54	57.45	60.76	70.4	42.16	55.7
Vehicle0	**97.15**	*93.59*	*96.42*	87.82	89.94	*96.42*	91.76	89.66	90.7	85.71	92.86	92.86
Ecoli1	*92.24*	89.24	88.75	88.68	*91.82*	*91.82*	80	80.85	85.71	85.71	**93.93**	85.71
New-thyroid1	90.56	*93.92*	*98.88*	92.84	95.74	*98.88*	92.31	92.31	92.31	**100**	89.75	92.31
New-thyroid2	*99.26*	*98.88*	**100**	96.75	95.74	**100**	**100**	**100**	**100**	95.74	93.33	**100**
Ecoli2	91.49	86.28	89.79	**97.38**	*94.02*	**97.38**	88.89	89.44	*94.87*	92.29	81.82	89.44
Segmemt0	98.28	98.25	**99.33**	98.54	94	**99.33**	98.46	*99.24*	*99.22*	97.95	84.64	*99.22*
Yeast3	87.68	*88.25*	81.57	87.63	**93.74**	**93.74**	70.97	80.6	74.14	*90.11*	78.99	71.43
Ecoli3	*87.6*	81.14	86.74	85.67	**94.87**	**94.87**	63.16	47.06	74.32	*89.44*	92.2	60
Yeast2vs4	81.13	87.43	79.56	*88.15*	86.98	*88.15*	77.78	70.59	70.71	**88.47**	84.21	66.67
Vowel0	92.22	96.41	93.01	**100**	*99.72*	**100**	*97.14*	*97.14*	*97.14*	94.28	*99.72*	*97.14*
Glass2	**80.7**	*68.05*	53.52	64.62	*76.24*	*76.24*	57.74	57.74	0	66.67	55.47	0
Yeast1vs7	**78.45**	*77.81*	70.72	*72.76*	61.83	*77.81*	30.77	40	66.67	64.17	66.67	66.67
Glass4	*96.25*	84.4	**97.64**	*97.29*	82.16	**97.64**	80	70.71	70.71	82.16	70.71	70.71
Ecoli4	76.97	66.68	67.98	*81.1*	**100**	**100**	**100**	*86.6*	**100**	**100**	**100**	**100**
Page-blocks13vs2	81.42	**100**	*97.25*	93.18	97.7	**100**	**100**	**100**	*99.43*	97.12	**100**	*99.43*
Abalone09-18	**68.02**	*66.26*	42.81	56.2	64.5	*66.26*	35.29	40	54.55	*66.52*	67	54.55
Glass5	**96.29**	*61.57*	52.53	*78.31*	0	*78.31*	0	0	0	28.57	40	0
Yeast4	*76.52*	49.83	**88.27**	64.3	84.29	**88.27**	54.29	54.29	33.33	*72.13*	66.64	31.62
Ecoli0137vs26	*70.71*	*70.71*	70.54	68.78	**100**	**100**	**100**	**100**	**100**	*96.23*	**100**	**100**
Yeast6	59.41	50.34	56.5	59.98	**91.13**	**91.13**	*75.46*	46.15	54.55	74.8	*90.64*	54.55
# of best value(win)	14				14		4	3	3	4	4	2
Top 3	19				11		3	3	6	5	2	4

AUC: In signal detection theory related to radio signals, the Receiver Operating Characteristics (ROC) curves were initially devised. For model evaluation strategies, ROC has been used recently by data mining and machine learning communities. For a binary classification problem, the ROC curve plots the true positive rate as a function of the false-positive rate. The AUC is denoted as the area under the ROC curve and is closely associated with the ranking quality of the classification.

4 Experimental Results

To assess the performance of the proposed method, the experiments were conducted in MATLAB R2020a platform on a laptop equipped with 2.20GHZ core i7 processor and 6 GB RAM. Our experiments were performed on 51 real-world datasets. First, original dataset is divided into three subsets—50%, 25% and 25%—for training, validation and test, respectively; each subset has an equivalent percentage of majority class and minority class as the other two sets. The training and validation sets are fed to the undersampling phase where the training data is divided into two—majority and minority-class groups. The best examples are selected from the majority class examples based on the *F*-score fitness function.

To emphasize the efficiency of our method, we include the best results obtained in [16], as a baseline for comparison. In [16], 24 out of the 51 datasets were used to assess different classification methods. In [16], four undersampling techniques based on neighbourhood searching (NB-based) by utilizing the *k*-NN rule to select and remove majority class examples from the potential region of overlapping. RF (Random Forest) and SVM (Support Vector Machine) were used in [16] for learning and their results compared with several pre-processing state-of-the-art techniques to rebalance datasets before applying the learning algorithm, like the SMOTE (Synthetic Minority Over-Sampling Technique) [40], kmUnder (k -means undersampling) [29], OBU [41], BLSMOTE [42] and ENN [43]. Columns "NB-SVM"

Table 3 *F*-score values of the proposed method and the best values obtained by methods in [16] for each dataset

Dataset	Proposed				Best in [16]							
	DA	DT	SVM	KNN	NB-SVM	NB-RF	SMOTE	BLSMOTE	ENN	kmUnder	OBU	Baseline
Wisconsin	94.74	92.82	93.51	95.05	94.95	**98.95**	94.85	95.74	*96.77*	*95.74*	51.65	*96.77*
Pima	**67.4**	64.24	*66.67*	54.27	58.29	59.52	*66.67*	66.67	*66.04*	*66.67*	52.31	64.08
Glass0	61.43	70.03	59.87	65	64.86	*70.97*	**78.26**	75	**78.26**	*75.86*	68.29	66.67
Vehicle1	*65.02*	*61.44*	**66.82**	55.17	48.81	50	61.54	57.45	60.76	60.38	42.16	55.7
Vehicle0	*92.28*	86.59	**93.77**	75.14	77.89	77.23	91.76	89.66	90.7	85.71	*92.86*	*92.86*
Ecoli1	82.29	*82.88*	77.45	79.11	78.95	*83.33*	80	77.42	**85.71**	**85.71**	*83.33*	**85.71**
New-thyroid1	86.94	86.47	*94.12*	89.44	82.35	**100**	*92.31*	92.31	92.31	**100**	87.5	*92.31*
New-thyroid2	*96.08*	94.12	**100**	91.91	82.35	**100**	**100**	**100**	**100**	87.5	*93.33*	**100**
Ecoli2	71.06	65.36	76.48	87.92	*90*	*88.89*	*88.89*	*88.89*	**94.74**	88.89	81.82	88.89
Segmemt0	93.2	93.88	97.08	92.16	73.86	*99.22*	*98.46*	**99.24**	*99.22*	96.97	53.72	*99.22*
Yeast3	70.4	*77.96*	75.73	64.2	66.67	*77.92*	70.97	**80.6**	70.18	70.89	63.64	71.43
Ecoli3	63.88	*64.05*	60.71	62.01	**70**	58.82	63.16	47.06	*66.67*	54.55	60.87	60
Yeast2vs4	65.2	74.67	73.48	72	70	**90**	77.78	70.59	66.67	*80*	*84.21*	66.67
Vowel0	74.6	86.08	84.68	**100**	*97.3*	82.93	*97.14*	*97.14*	*97.14*	94.12	*97.3*	*97.14*
Glass2	*39.22*	23.33	16.57	32.96	*40*	37.5	**50**	**50**	0	22.22	20.69	0
Yeast1vs7	*47.54*	47.53	*49.25*	34.46	17.54	54.55	30.77	40	**66.67**	25	**66.67**	**66.67**
Glass4	63.89	51.52	*72.22*	*70*	66.67	50	**80**	66.67	66.67	40	66.67	66.67
Ecoli4	66.67	45.12	63.1	*73.89*	**100**	66.67	**100**	*85.71*	**100**	**100**	**100**	**100**
Page-blocks13vs2	53.26	**100**	71.46	*72.59*	71.43	66.67	**100**	**100**	*90.91*	*90.91*	**100**	*90.91*
Abalone09-18	**59.44**	28.97	20.72	30.88	22.86	28.57	*35.29*	40	*54.55*	18.87	21.05	*54.55*
Glass5	**62.12**	25.66	0	*43.49*	0	40	0	0	0	28.57	*40*	0
Yeast4	*41.93*	25.81	**42.05**	31.23	32	*38.89*	33.33	37.5	33.33	22.22	34.04	18.18
Ecoli0137vs26	*66.67*	*66.67*	*61.11*	30.56	**100**	*66.67*	**100**	**100**	**100**	33.33	**100**	**100**
Yeast6	31.03	29.84	22.13	31.79	54.55	46.15	**66.67**	46.15	*54.55*	47.06	*48*	*54.55*
# of best value(win)	9				7		8	5	7	3	4	5
Top 3	15				10		6	3	8	4	6	8

and "NB-RF" in Tables 2 and 3 present the best *G*-mean and *F*-score values, respectively, selected for each classifier (SVM and RF) among the four (NB-based) methods, while posterior columns present the best value selected from the classifiers (SVM and RF) after applying the state-of-the-art pre-processing techniques.Bold values in Tables 2 and 3 represents the best value in that row while Italic represnts the second and third best values in the same row.

As can be noticed from Table 2, our SA-based classifiers have achieved an overall superior performance in *G*-mean over other methods. While the best pre-processing technique achieved the best *G*-mean for only 4 datasets out of the 24 datasets used; our method achieved the best *G*-mean for 14 datasets like the number achieved by methods proposed in [16]. Our method also shows better performance with the dataset (Glass5) that recorded zero *G*-mean with most pre-processing methods. These enhancements in *G*-mean indicate that our proposed method has achieved a better balance in the classification accuracy between both (majority and minority) classes and it has not affected by the class distribution that affected on other state-of-the-art methods.

Table 3 presents another comparison with [16] based on the *F*-score measure. *F*-score provides a good measure to assess the trade-offs between the accuracy of positive class and the negative class' errors. This measure is very useful to appreciate classifiers performance, especially when used to give more insights on performance if *G*-mean metric is competitive for two different methods; as in our case where the number of best *G*-mean values is equivalent to [16].

It can be noticed from Table 3 that our proposed method ranks top in *F*-score; where it provides a significantly higher number of best *F*-score values than [16] and in all pre-processing methods. This superiority in *F*-score proves that our method has enhanced the trade-offs between specificity and sensitivity, which means the reduction obtained by our method for false positives and negatives, over state-of-the-art methods.

Accuracy is also an important measure for any classifier performance so, we cannot ignore it. As accuracy has not been used in [16], we have compared the classification accuracy of our proposed method with another recent work [44]. Researchers in this work introduced a hybrid approach

Table 4 Accuracy values of the proposed method and the best values obtained by methods in [44] for binary datasets

Dataset	Proposed				Best in [44]			
					Reduced datasets		Non-reduced datasets	
	DA	DT	SVM	KNN	AOUSID	AISAID	C4.5	10NN
abalone19	95.78	90.8	**99.23**	94.47	82.04	81.42	82.02	48.05
shuttle-c0-vs-c4	99.85	**100**	**100**	99.71	97.62	98.01	97.17	90
vowel0	94.33	97.03	97.03	**100**	93.72	91.05	94.94	**100**
yeast5	96.41	**98.38**	96.5	97.12	88.4	89.12	87.50	79.42
glass2	79.87	59.75	42.77	**84.91**	71.69	71.2	60.08	33.4
ecoli-0-1-4-6_vs_5	95.24	92.38	91.43	**95.71**	80.21	77.13	81.36	83.9
glass0	62.89	73.58	60.38	65.41	77.45	**79.24**	78.13	70.57
yeast2 (Yeast2vs4)	92.45	94.53	**95.31**	93.49	79.81	68.49	62.82	81.63
vehicle2	**96.21**	95.1	93.52	81.52	94.06	93.67	94.85	88.31
# of best value (Win)	8				0	1	0	1

to handle the problem of imbalanced data using oversampling and the instance selection undersampling algorithms. They relied on clustering to select instances from majority class using an agent-based population learning algorithm. Their experiment was based mainly on proving that their proposal performed better than the methods of traditional learning where machine learning techniques were applied for learning on original imbalanced data so, they compared the performance accuracy of their proposed technique—Agent-based Over and Undersampling for the Imbalanced Data (AOUSID)—with the classification accuracy obtained by another 7 techniques. Three of these techniques introduced by other researchers for undersampling, while the rest of the 7 techniques were traditional machine learning algorithms.

Table 4 presents the results obtained by our proposal and the best results presented in [44] in a comparison based on the classification accuracy.

Only 3 from the 7 techniques that have been used in [44] besides their proposal (AOUSID) are listed in Table 4 because only these 4 techniques have gained best results with the imbalanced data used. These three techniques are:

- AISAID—an algorithm introduced by [45] for solving the imbalance problem by applying the instance selection procedure to resample the majority class.
- Traditional ML algorithms: C4.5 algorithm [44], and *K*-nearest neighbor (*k*-NN) [46].

The experiment conducted by [44] used 11 datasets, from the dataset repository of KEEL [47]. We used the binary class datasets—9 datasets—in our comparison. Data descriptions of these 9 datasets are also listed in Table 1.

It can be noticed from the results shown in Table 4, that in comparison to other algorithms, our proposed method asserts competitive results. The SA algorithm with the 4 classifiers performs best with almost all the imbalanced

Table 5 AUC and G-mean values of the proposed method and RCSMOTE in [48]

Data	AUC		*G*-mean	
	Proposed	RCSMOTE	Proposed	RCSMOTE
Bupa	**68.6**	**68.60**	**66.84**	64.38
Pima	76.46	**78.38**	74.71	**76.22**
Breast	**68.57**	67.82	**67.53**	55.87
Haberman	60.06	**71.51**	**59.35**	58.24
Newthyroid	**98.15**	97.74	**98.88**	94.47
Hepatitis	82.74	**89.4**	84.38	**87.06**
Cleveland	72.22	**76.37**	**71.88**	66.96
Ecoli (Ecoli3)	88.96	**92.12**	**94.87**	86.61
Breast_tissue	81.5	**86.75**	82.57	**83.81**
Glass (Glass0)	81.48	**91.23**	77.78	**90.86**
Heart	73.52	**83.35**	73.99	**83.39**
Iris	**100**	94.42	**100**	94.39
Libra	**99.07**	89.69	**99.07**	89.68
Liver_disorders	**66**	65.83	**63.38**	55.83
Segment (Segment0)	**99.29**	97.86	**99.33**	97.03
Vehicle (Vehicle0)	**96.95**	94.43	**97.15**	94.43
Wine	**97.51**	95.83	**97.47**	95.3
Contraceptive	65.98	**66.46**	**65.98**	59.93
Ionosphere	**89.35**	89.19	**89.44**	88.99
Parkinsons	88.43	**88.95**	**87.64**	85.98
Tic-tac-toe	**99.15**	94.46	**99.14**	93.65
Transfusion	66.9	**68.03**	**66.18**	63.12
ILPD	**70.03**	66.94	**66.03**	59.63
BreastEW	**93.71**	93.62	**93.81**	92.46
Abalone	**85.02**	74.11	**84.61**	69.29
Yeast (Yeast4)	**88.26**	86.05	**88.27**	73.21
Vowel	**99.93**	99.77	**100**	97.83
Page-blocks	**96.67**	96.52	**96.63**	92.14
# of best value (Win)	17	12	23	5

Table 6 *P*-values of the Wilcoxon's signed rank tests for the proposed method

	Best in [46]	SMOTE	BLSMOTE	ENN	kmUnder	OBU	Baseline
Pre-processing techniques in [16]							
G-mean	1.4E−01	**1.0E−03**	**0.0**	**0.0**	**2.0E−03**	**1.0E−03**	**0.0**
F-score	9.27E−01	8.84E−01	3.55E−01	6.27E−01	**2.0E−02**	7.8E−02	1.62E−01
		AOUSID		AISAID		C4.5	10NN
Pre-processing techniques in [46]							
ACC		**2.1E−02**		**2.1E−02**		**2.1E−02**	**1.2E−02**
		RCSMOTE					
Pre-processing technique in [50]							
AUC		9.62E−01					
G-mean		**1.0E−03**					

datasets compared to the best results obtained by other undersampling techniques and the traditional machine learning algorithms.

For more validation for our proposal, another comparison has been conducted; where 28 datasets with different overlapping degrees, some features with outliers, noisy samples, and multiclass have been used in this last comparison (more details about datasets characteristics are available in [48]). These datasets used in [48] to compare between its authors' proposal named RCSMOTE and state-of-the-art over-sampling SMOTE-based techniques. RCSMOTE (Range-Controlled SMOTE) is an improved SMOTE method based on over-sampling the borderline samples (considering a safe range) after identifying them from noisy ones in minority class samples [49]. Table 5 shows the comparison between our proposed method and RCSMOTE method based on AUC and *G*-mean values.Results in Table 5 demonstrate that the proposed method outperforms RCSMOTE for both AUC and *G*-mean in 17 datasets out of 28 compared to the superiority of RCSMOTE in just 5 datasets.

Wilcoxon's signed ranks test [51] is also applied as a statistical test to compare the performance of the proposed method with all other pre-processing techniques involved in the performed comparisons in our experiments. The *p*-values associated with these comparisons has been obtained to indicate the degree of difference between the methods. The difference considered to be significant if the *p*-value is lower than 0.05.

Table 6 shows the results of Wilcoxon's test for the three performed comparisons. The results present the decrement in the *p*-values, which indicates a great significance in the differences between the proposed method and almost all other pre-processing methods. Particularly, in the *G*-mean values that clearly shows an improvement in the performance obtained with the proposed method. Since *G*-mean is the square root of the product of class-wise sensitivity (sensitivity for positive examples and specificity for negative examples) and this measure tries to maximize the accuracy of both classes in balance.

All the preceding results indicate the efficiency of the proposed method in improving the classification performance for the used datasets. Using *F*-score measure as an objective function in SA technique, helped in improving the classification accuracy for both minority and majority classes; since the *F*-score 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. Also, using SA optimization itself helps avoid falling into a local optimum solution trap and converges to the global optimum solution. Finally, applying SA on different classifiers allows the proposal to deal with the diverse and variation in datasets.

5 Discussion

The real-world imbalance datasets exhibited erroneous classification results and showed a bias toward majority class. To tackle the imbalance issue, various techniques were devised by the researchers. We introduce a modified hybrid strategy to take care of this problem. We use simulated annealing to pick the best possible subset of major class records (rows of data). Afterwards, KNN, DA, SVM and DT classifiers are utilized to assess the efficiency of our technique. We evaluate our empirical results with the two recent works [16, 44]. We explore 51 real datasets from different data repositories for the experiments. In [16], 24 datasets were used. Out of 24 datasets, our method outperforms method proposed in [16] in 14 datasets and yields comparable performance with the rest of the datasets. The evaluation metrics used are *G*-mean and *F*-score for this comparison. Accuracy was not considered in [16]. To evaluate our findings using accuracy, the

comparison is done with [44]. In both cases, our approach proves its efficacy and hence can be applied in real-world settings where the dataset is an imbalance. The proposed technique is further validated with the presented method [49] in terms of AUC and *G*-mean. Our technique showcased superiority 17 datasets whereas RCSMOTE method yielded better results only in 5 datasets out of 28 datasets. We also perform Wilcoxon's signed ranked test and results demonstrate that there is a great difference between the proposed method and the other pre-processing techniques.

References

1. Abdar, M.; Acharya, U.R.; Sarrafzadegan, N.; Makarenkov, V.: NE-nu-SVC: a new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease. IEEE Access **7**, 167605–167620 (2019)
2. Abdar, M.; Nasarian, E.; Zhou, X.; Bargshady, G.; Wijayaningrum, V.N.; Hussain, S.: Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach. In: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 26–30). IEEE (2019)
3. Zomorodi-moghadam, M.; Abdar, M.; Davarzani, Z.; Zhou, X.; Pławiak, P.; Acharya, U.R.: Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease. Expert Syst. **38**, e12485 (2019)
4. Nasarian, E.; Abdar, M.; Fahami, M.A.; Alizadehsani, R.; Hussain, S.; Basiri, M.E.; Zomorodi-Moghadam, M.; Zhou, X.; Pławiak, P.; Acharya, U.R.; Tan, R.S.: Association between work-related features and coronary artery disease: a heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recogn. Lett. **133**, 33–40 (2020)
5. Samami, M.; Akbari, E.; Abdar, M.; Plawiak, P.; Nematzadeh, H.; Basiri, M.E.; Makarenkov, V.: A mixed solution-based high agreement filtering method for class noise detection in binary classification. Phys. A Stat. Mech. Appl. **553**, 124219 (2020)
6. Tuncer, T.; Dogan, S.; Abdar, M.; Ehsan Basiri, M.; Pławiak, P.: Face recognition with triangular fuzzy set-based local cross patterns in wavelet domain. Symmetry **11**(6), 787 (2019)
7. Abdar, M.; Basiri, M.E.; Yin, J.; Habibnezhad, M.; Chi, G.; Nemati, S.; Asadi, S.: Energy choices in Alaska: mining people's perception and attitudes from geotagged tweets. Renew. Sustain. Energy Rev. **124**, 109781 (2020)
8. Basiri, M.E.; Abdar, M.; Cifci, M.A.; Nemati, S.; Acharya, U.R.: A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. Knowl.-Based Syst. **198**, 105949 (2020)
9. Pławiak, P.; Abdar, M.: Novel methodology for cardiac arrhythmias classification based on long-duration ECG signal fragments analysis. In: Biomedical Signal Processing (pp. 225–272). Springer, Singapore (2020)
10. Stoean, C.; Stoean, R.; Atencia, M.; Abdar, M.; Velázquez-Pérez, L.; Khosravi, A.; Nahavandi, S.; Acharya, U.R.; Joya, G.: Automated detection of presymptomatic conditions in Spinocerebellar Ataxia type 2 using Monte Carlo dropout and deep neural network techniques with electrooculogram signals. Sensors **20**(11), 3032 (2020)
11. Hussain, S.; Muhammad, L.J.; Ishaq, F.S.; Yakubu, A.; Mohammed, I.A.: Performance evaluation of various data mining algorithms on road traffic accident dataset. In: Information and Communication Technology for Intelligent Systems (pp. 67–78). Springer, Singapore (2019)
12. Hussain, S.; Muhammad, L.J.; Yakubu, A.: Mining social media and DBpedia data using gephi and R. J. Appl. Comput. Sci. Math. **12**(1), 14–20 (2018)
13. Devi, D.; Namasudra, S.; Kadry, S.: A boosting-aided adaptive cluster-based undersampling approach for treatment of class imbalance problem. Int. J. Data Warehousing Min. (IJDWM) **16**(3), 60–86 (2020)
14. Mohammed, R.; Rawashdeh, J.; Abdullah, M.: Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 243–248). IEEE (2020)
15. Liu, B.; Tsoumakas, G.: Dealing with class imbalance in classifier chains via random undersampling. Knowl.-Based Syst. **192**, 105292 (2020)
16. Vuttipittayamongkol, P.; Elyan, E.: Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Inf. Sci. **509**, 47–70 (2020)
17. Vuttipittayamongkol, P.; Elyan, E.: Improved overlap-based undersampling for imbalanced dataset classification with application to Epilepsy and Parkinson's disease. Int. J. Neural Syst. **30**, 2050043 (2020)
18. Sarkar, S.; Khatedi, N.; Pramanik, A.; Maiti, J.: An ensemble learning-based undersampling technique for handling class-imbalance problem. In: Proceedings of ICETIT 2019 (pp. 586–595). Springer, Cham (2020)
19. Bejaoui, A.; Elkhalil, K.; Kammoun, A.; Alouni, M.S.; Alnaffouri, T.: Improved design of quadratic discriminant analysis classifier in unbalanced settings. arXiv preprint arXiv:2006.06355 (2020)
20. Jian, C.; Gao, J.; Ao, Y.: A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing **193**, 115–122 (2016)
21. Dubey, H.; Pudi, V.: Class based weighted k-nearest neighbor over imbalance dataset. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 305–316). Springer, Berlin, Heidelberg (2013)
22. Liu, W.; Chawla, S.; Cieslak, D.A.; Chawla, N.V.: A robust decision tree algorithm for imbalanced data sets. In Proceedings of the 2010 SIAM International Conference on Data Mining (pp. 766–777). Society for Industrial and Applied Mathematics (2010)
23. Tóth, J.; Tomán, H.; Hajdu, A.: Efficient sampling-based energy function evaluation for ensemble optimization using simulated annealing. Pattern Recognit. **107**, 107510 (2020)
24. Yang, D.; Wang, X.; Tian, X.; Zhang, Y.: Improving monarch butterfly optimization through simulated annealing strategy. J. Ambient Intell. Hum. Comput., 1–1, 2020
25. Camelo, P.H.C.; de Carvalho, R.L.: Multilayer perceptron optimization through simulated annealing and fast simulated annealing. Acad. J. Comput., Eng. Appl. Math. **1**(2), 28–31 (2020)
26. Tsai, C.F.; Lin, W.C.; Hu, Y.H.; Yao, G.T.: Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf. Sci. **477**, 47–54 (2019)
27. Zhu, M.; Xia, J.; Jin, X.; Yan, M.; Cai, G.; Yan, J.; Ning, G.: Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access **6**, 4641–4652 (2018)
28. Li, J.; Fong, S.; Yuan, M.; Wong, R.K.: Adaptive multi-objective swarm crossover optimization for imbalanced data classification. In: International Conference on Advanced Data Mining and Applications (pp. 374–390). Springer, Cham (2016)
29. Li, M.; Xiong, A.; Wang, L.; Deng, S.; Ye, J.: Aco resampling: enhancing the performance of oversampling methods for class imbalance classification. Knowl.-Based Syst. **196**, 105818 (2020)
30. Huda, S.; Yearwood, J.; Jelinek, H.F.; Hassan, M.M.; Fortino, G.; Buckland, M.: A hybrid feature selection with ensemble

classification for imbalanced healthcare data: a case study for brain tumor diagnosis. IEEE Access **4**, 9145–9154 (2016)
31. Febriantono, M.A.; Pramono, S.H.; Rahmadwati, R.; Naghdy, G.: Classification of multiclass imbalanced data using cost-sensitive decision tree C50. IAES Int. J. Artif. Intell. **9**(1), 65 (2020)
32. Babu, M.C.; Pushpa, S.: Genetic algorithm-based PCA classification for imbalanced dataset. In: Intelligent Computing in Engineering (pp. 541–552). Springer, Singapore (2020)
33. Ri, J.; Kim, H.: G-mean based extreme learning machine for imbalance learning. Dig. Signal Process. **98**, 102637 (2020)
34. Susan, S.; Kumar, A.: Hybrid of intelligent minority oversampling and PSO-based intelligent majority undersampling for learning from imbalanced datasets. In: International Conference on Intelligent Systems Design and Applications (pp. 760–769). Springer, Cham (2018)
35. El-Shafeiy, E.; Abohany, A.: Medical imbalanced data classification based on random forests. In: Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 81–91). Springer, Cham (2020)
36. Yang, Y.; Huang, S.; Huang, W.; Chang, X.: Privacy-preserving cost-sensitive learning. In: IEEE Transactions on Neural Networks and Learning Systems (2020)
37. Wang, D.; Zhang, X.; Chen, H.; Zhou, Y.: A sintering state recognition framework to integrate prior knowledge and hidden information considering class imbalance. In: IEEE Transactions on Industrial Electronics (2020)
38. Delahaye, D.; Chaimatanan, S.; Mongeau, M.: Simulated annealing: from basics to applications. In Handbook of Metaheuristics (pp. 1–35). Springer, Cham (2019)
39. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
40. Jeong, I.S.; Kim, H.K.; Kim, T.H.; Lee, D.H.; Kim, K.J.; Kang, S.H.: A feature selection approach based on simulated annealing for detecting various denial of service attacks. Softw. Netw. **2018**(1), 173–190 (2018)
41. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
42. Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409**, 17–26 (2017)
43. Vuttipittayamongkol, P.; Elyan, E.; Petrovski, A.; Jayne, C.: Overlap-based undersampling for improving imbalanced data classification. In: International Conference on Intelligent Data Engineering and Automated Learning (pp. 689–697). Springer, Cham (2018)
44. Czarnowski, I.; Jędrzejowicz, P.: An approach to imbalanced data classification based on instance selection and over-sampling. In: International Conference on Computational Collective Intelligence (pp. 601–610). Springer, Cham (2019)
45. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst., Man, Cybern. **3**, 408–421 (1972)
46. Czarnowski, I.; Jędrzejowicz, P.: Cluster-based instance selection for the imbalanced data classification. In: International Conference on Computational Collective Intelligence (pp. 191–200). Springer, Cham (2018)
47. Quinlan, J.: C4. 5: Programs for Machine Learning. Elsevier (2014)
48. Alcalá-Fdez, J.; Fernández, A.; Luengo, J.; Derrac, J.; García, S.; Sánchez, L.; Herrera, F.: Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Log. Soft Comput. **17**, 255–287 (2011)
49. Soltanzadeh, P.; Hashemzadeh, M.: RCSMOTE: range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf. Sci. **542**(2021), 92–111 (2021)
50. Han, H.; Wang, W.Y.; Mao, B.H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing (pp. 878–887). Springer, Berlin, Heidelberg (2005)
51. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)