Received: May 19, 2021. Revised: June 20, 2021. 10

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hybrid Chicken Swarm Optimization (CSO) and Fuzzy Logic (FL) Model for Handling Imbalanced Datasets

Farid Ali Mousa[1]* **Ibrahim Eldesouky Fattoh[2]**

[1]*Information Technology Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University Beni Suef, Egypt*
[2]*Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University Beni Suef, Egypt*
* Corresponding author's Email: fared.ali@fcis.bsu.edu.eg

Abstract: The unequal representation of the various divisions existing in the data is one of the main data inconsistency issues. Data with an imbalanced distribution negatively influence the efficiency of most conventional classifiers. This paper introduces a new method for over-sampling the handling of imbalanced data sets. The method hybridize chicken swarm optimization and fuzzy logic (CSO-FL). The proposed model ensures that the synthetic samples generated reside in minority regions. The proposed hybrid CSO-FL applied on three datasets with different imbalanced ratios between 1.78 and 129.44. It demonstrated significant improvements in the efficiency of various classification techniques. During the classification process, we used KNN, DT, SVM and Naïve classifiers. The obtained results were very promising; the precision, sensitivity, and F_score values are enhanced in all classifiers. The values in one dataset improved with ratios >90 % in many classifiers because of the high imbalanced ratio in this dataset, while in the other two datasets, the measurement values enhanced with ratios from nearly 10% to 30%. The CSO-FL approach compared with three different approaches on the same datasets. The approaches are SMOTE algorithm, modified SMOTE, and TGT algorithm and proved to outperform their results.

Keywords: Chicken swarm optimization, Imbalanced dataset, Fuzzy logic, Decision tree, Naïve, Support vector machine, K-nearest neighbour.

1. Introduction

Data quality is one of the main data processing problems, since dirty data often lead to inaccurate analysis outcomes and weak business decision making. Companies also collected large volumes of data from various sources in to create their own data lakes to enhance data and enhance analytics. The treatment and acquisition of data also leads to inconsistencies in the data, for instance missed values, spelling errors, mixed types, double entries and breaches of company rules. It is not shocking that the implementation of effective and successful processes for data cleaning is challenging [1].

If the dataset is imbalanced, the dilemma becomes more complicated. If the distribution of its groups is not identical, a dataset is called imbalanced. In this case, a few samples (minority class) represent at least one class, while the other class is represented by other samples (majority class). Recent experiments on machine learning have shown that the sluggish allocation of classes is causing a performance gap. This suggests that classifiers appear to give the class of the majority high precision while giving low precision to the class of the minorities. The minority classes are more significant in many real-world applications than cancer diagnosis in the medical community. This is why both academia and industry have increased interest in the classification of imbalanced data sets [2].

Every day, massive quantities of data are produced in today's internet environment. Thus, a deep understanding of information discovery and interpretation of raw data should be advanced in order

to facilitate decision-making in companies. There has been an evolution of data classification through the learning process. If the dataset is imbalanced, the dilemma becomes more complicated. If the class distribution is not uniform, a dataset is considered to be imbalanced. There are examples from one class in this case that are greater than the other. The class with a larger number of samples is called 'mainstream class' and 'minority class' is called the class with a comparatively lower range of instances [3].

Recent findings of machine learning have shown that an uneven class distribution may result in a bias in model output. The explanation is that the classification offers the majority class with high accuracy and the minority class with low accuracy. That is because the vast number of big classes are inclined to conventional training behavior, such as overall performance. The minority groups are more significant in many real-life applications than in medical applications when diagnosing cancer. Therefore, both academia and industry are deeply interested in classifying imbalanced data sets. Many academic studies have previously suggested some methods for misclassifying the issue of imbalanced data sets [4]. For the previous reasons we advise of using proposed hybrid technique that it should be pointed out that ensemble techniques also have the challenge of ensuring that the variations in each approach complement each other and achieve greater efficiency together in comparison with each individual approach on its own that leads to high performance in all data sets. The proposed model is important because it prepares the data in the most meaningful way for the subsequent detailed analysis that ensure that the generated samples in the minority class are farthest from the majority class by two tests; the first by using fuzzy membership function that will give each sample a fuzzy number and in the second test by using CSO that will added the sample s to the minority class if and only if satisfies the objective function.

1.1 Chicken swarm optimization (CSO)

Organic meta-heuristic algorithms have demonstrated a great number of optimization implementations to be solved [5]. It utilizes the tolerance for inaccuracy or complexity of problems with optimization and can reach suitable solutions at low calculation costs. One of the potential research in coping with optimization applications with algorithms of mix-heuristic like Particle Swarm Optimization (PSO) [6], Bat Algorithm [1], Differential Evolution (DE) [7] and Chicken Swarm Optimization (CSO) [8]. Chicken Swarm Optimization (CSO) imitates the hierarchy of the swarm and of the chicken swarm behavior. The swarm of chicken may be split into multiple sections, each comprising one rooster and a number of hens and chicks. Various chickens obey various motions rules. There are competing chickens in a particular hierarchical order. [9]

The following rules explain the actions of chicken's behaviours; Number of classes in the chicken swarm. The dominant rooster, chicks and a couple of hens are in each group.

Chickens imitate their rooster group mates to look for food to discourage them from consuming their own food. Suppose chickens will randomly rob other people's healthy food. The chicks look to their mother for food (hen). In competition for food the dominant people have advantages. The numbers of roosters (RN) and hens (HN), chicks (CN) and mother’s hens (MN) are assumed. The better RN chickens were supposed to be roosters, while the worst RN chicks were considered chicks. The remaining ones are handled like hens. [8].

1.2 Fuzzy logic (FL)

At the end of the 1980s the fuzzy logic has been seen as an emerging technology, mostly as a controversy technology for two decades. Any of this is attributed to a wide range of successful applications from consumer electronics to industrial process controls to automobiles. We ought, to put this paradigm first in context before undertaking a profound discussion of technical questions related to fuzzy logic. In this regard, two meanings of 'Fuzzy Logic' are first clarified. In two opposite directions, the expression "Fuzzy logic" was used. In a narrow sense, Fuzzy logic is a logical way of generalizing the classical two-value logic for complexity reasoning. In the broadest context, fluid logic extends to all the philosophies and technology using fuzzy sets, classes of sharp limits[10].

Lotfi A. Zadeh [11], initiated the Fuzzy Logic in 1965. In essence, it is a multi-value logic that permits the definition of intermediate values between standard assessments such as yes | no, true | false, high | low and others. Computers may formulate and process notions such as large or very quick mathematically and use human reasoning in computer programming [12].

A member function (MF) is a curve which specifies whether a member value (membership degree) has been mapped to each input area (discourse universe) point from zero to one.

The rules use the weighting factors for the input membership values to decide how they affect the

fuzzy performance of the output sets. If the functions are inferred, merged, scaled and combined and de-fuzzified the output that drives the system into a crisp [13].

The membership grade of μA(x) quantifies to the blurry set the membership degree of x. 0 implies that x does not belong to the fuzzy group; the value of 1 means that x is a whole member of the fuzzy set. Fuzzy members, which are a part of the Fuzzy set, are characterized by values from 0 to 1. [13]. In designing the membership function specification is a sensitive point, since the only limitation a membership feature has to meet is that its values must be [0, 1]. Therefore, as opposed to a narrow-minded set, an infinite number of member functions can be defined. The simplest membership features are created by direct lines. In real time applications both Triangular Member function and Trapezoidal Member Functions were commonly used due to their simplicity of calculations and computational efficiencies [14]. Unique implementations can also have other advanced MFs if needed. Specifically, any kind of continuous probability function may be used as MF, if the relevant definitions of the MF are specified with a set of parameters [15].

This paper is prepared as follow: section 2 will cover the area of related work, section 3 will discuss the proposed model phases, section 4 will cover the used dataset, applied experiments and discusses the obtained results, and section 5 will provides the conclusion and future work.

2. Related work

There are many methods for solving unbalanced data; The aim in [16] is to change the data collection to make the standard learning algorithm more suited for data level approaches. To alter, under-sample, and over-sample datasets, two sub-approaches are used. Samples from the main class are to be removed when over samples create new minority class artefacts. The selection of the samples is done using random techniques of traditional approaches. But this also leads to the elimination of fresh, irrelevant samples of relevant samples or appearances.

Methods of under-sampling eliminate the majority class samples. This decrease can be performed by random under-sampling or by educated under-sampling using certain statistical information. For certain class examples, certain educated methods of under-sampling are based on data cleaning techniques [17].

In [18], the authors reported that the study of under-sampling approaches is deficient in contrast with over-sampling approaches. In addition, existing under-sampling methods are affected by output instability.

A new SMOTE approach for tackling the issue of imbalanced data was presented in [19]. By refusing the synthetic samples, the SMOTE process was updated. They showed that it does not interfere as much as conventional approaches when calculating the worth of a closest neighbour. Eight databases were experimented; the new approach achieved greater efficiency. This is because each new instance is created with its place in the distribution boundary in mind.

In [20], they said SMOTE is an intelligent over-sampling method. Over-sampling techniques may lead the learner to over fit and to rise training dataset size. The authors submitted that over fitting is not a significant problem for SMOTE, as it generates new instances in synthesis compared with replication of current instances.

Researchers in [21] proposed the Borderline-SMOTE this algorithm is supposed to make a small contribution to the success of the classification by instances beyond the boarder rows. The approach thus identifies these borderline instances by over-sampling the proportion of the majorities and minorities in each instance. The mostly neighbouring noisy examples are not taken into account. The so-called dangerous instances shall be over-sampled accordingly.

Significant over-sampling disadvantages refer to the reality that it can result in overfitting, boost the time needed to create the classifier, or even hurt the learning process. Under-sampling do rebalancing by deleting instances from the majority class. While this enables to identify the specific space, it can trigger information loss by decreasing the size of the dataset. Another significant factor which affects sampling is the noise that may exist in the dataset which negatively affects the minority classes more than majority.

Researcher should look at the bigger image while thinking about sampling. In other words, one should think of the nature of the problem being addressed, and the suitable classifier for the problem under consideration. Various classifiers achieve higher performance when accompanied with sampling approaches.

In [22], researchers proposed the AHC. It was the first attempt to construct synthetic instances through the application of clusters to balance knowledge. The K-means algorithm was used for the most cases and agglomerated hierarchical clusters were used for exaggeration of the minorities' example. Clusters are obtained here from all classes of dendro-grams and

their centers are calculated using the original minority class instances.

Safe-Level-SMOTE proposed in [23], it gives each minority instance a safe level prior to the generation of synthesized instances. One of the newest multiclass methods for Mahalanobis inspired by space is MDO [24]. For the Mahalanobis from each class studied, MDO builds synthetic instances within the same scope as for the other minority cases.

The authors modified the smote algorithm in [25] by generating a new synthetic sample based on a randomly chosen minority neighbour and adding the distance between the closest majority and minority neighbours.

The authors of [26] developed a new technique that uses two classifiers to extract and model minority class data information. On the given data, a decision tree is trained to model the minority class data as a collection of classification rules, which are then used to create new minority class samples. Then, using the provided data, a neural network is trained to verify that all of the generated samples belong to the minority class data distribution.

The major problem of the modified Smote proposed in the literature is to adjust current algorithms according to the new circumstances is to select the best way to achieve the new goal. This could be accomplished by analysing the particular characteristics and conditions of the dataset and the problem itself. The proposed model uses CSO to solve the mentioned problem by iterative the algorithm steps until reach to the best generated samples that fir the objective function.

Most of the classical methods of machine learning have demonstrated shortcomings when used in the field of imbalanced data. Conventional machine learning algorithms do not work well for imbalanced data classification as they assume equal costs for each class. As a result, the traditional machine learning algorithms become biased towards the majority group. Therefore, intelligent systems must be designed to overcome such problem especially that learning from imbalanced data is still a focus of intense research

As discussed above, most of the methods used to generate a new samples in the minority class used the data in the minority class only that is may cause an over fitting if the scope of the class is small and we need to generate a lot of samples; this drawback is solved in the proposed model as we used the majority and minority classes when generating the samples; also we used an heuristic method for checking the generated samples to fit in the correct scope.

3. Proposed model

The hybrid CSO-FL technique main focus is to produce new synthetic samples of the minority class that lessen the space between majority and minority data based on the fuzzy membership function that give the nearest sample higher value and lower values to farthest samples. Towards this goal, majority data samples are considered while generating the new minority data samples. Algorithms 1 and 2 provides detailed steps of the proposed methodology.

We argue that using an oversampling technique that simulates the adversarial architecture can yield better results during the oversampling process and consequently to handle binary classification of imbalanced dataset in a better way. In more specific words, we argue that generation process of oversampling can be guided by two steps, the first step is by getting the boundary of the generated samples and then we used CSO algorithm to generate a sample that will be tested using the objective function.

The algorithm starts by considering samples in minority class. For each sample, we get the k-nearest neighbours of it from the majority class and the k-nearest neighbours from minority class by using Euclidean distance as shown in Eq. (1).

$$d(S, C_i) = \sqrt[2]{\sum_{p=1}^{n}(S_p - C_p)} \tag{1}$$

Where; S represent selected sample, C is the class samples and p is the features for each sample.

We select randomly one of the minority neighbours and calculate the distance between it and both the nearest and farthest majority neighbour and the nearest and farthest minority neighbour. We calculate the membership function to minimum and maximum sample of the minority regarding the minority and majority class using Eq. 2 Then we calculate the minority membership and majority membership as shown in Algorithm 1; these values will be used in CSO algorithm to check the new generated sample exist in the correct space.

$$\mu_i(x) = \frac{\sum_{j=1}^{k} \mu_{ij}(1/(\|x - x_j\|^{2/(m-1)})}{\sum_{j=1}^{k}(1/(\|x - x_j\|^{2/(m-1)})} \tag{2}$$

When determining each neighbour's contribution to the membership value, variable m defines how heavily the distance is weighted and it can ranged from 0 to 1 as a fuzzification parameter. The inverse of the distances from the nearest neighbours, as well as their class memberships, influence the assigned memberships of x, as shown in Eq. (2). The inverse

distance is used to weight a vector's membership more if it is closer to the vector under consideration and less if it is far away.

CSO algorithm will start with upper and lower values calculated from minimum and maximum neighbours in minority and majority classes and then will generate a new samples, if this generated sample satisfy the objective function that depends on fuzzy logic then the generated sample will added to the data set. The objective function check if the member ship value of the generated sample fall in scope of the minority samples and farthest from the majority samples. It should be greater than the minority member and smaller than the majority member. Through using this algorithm, the generated minority class sample by objective function is validated by an entirely different and unexplained process.

The chicken of the best fitness values of the next generation are picked of flocks.

$$X_{i,j}^{t+1} = X_{i,j}^{t} * (1 + Randn(0, \sigma^2)) \quad (3)$$

Where; $\sigma^2 =$

$$\begin{cases} 1, & if\ f_i \le f_k \\ \exp(\frac{f_k - f_i}{|f_i| + \varepsilon}), & otherwise \end{cases} \qquad k \in [1, N], k \neq i \quad (4)$$

$$X_{i,j}^{t+1} = X_{i,j}^{t} + S1 * Rand * \left(X_{r1,j}^{t} - X_{i,j}^{t}\right) + S2 * Rand * \left(\mathrm{X}_{\mathrm{r2,j}}^{\mathrm{t}} - \mathrm{X}_{\mathrm{i,j}}^{\mathrm{t}}\right) \quad (5)$$

$$S1 = \exp((f_i - f_{r1})/abs(f_i) + \varepsilon)) \quad (6)$$

$$S2 = \exp\left((f_{r2} - f_i)\right) \quad (7)$$

$$X_{i,j}^{t+1} = X_{i,j}^{t} + FL * (X_{m,j}^{t} - X_{i,j}^{t}) \quad (8)$$

At time t, The N number of chickens, are referred as $\mathrm{X}_{\mathrm{i,j}}^{\mathrm{t+1}}$, where i ϵ [1, 2, …, N], j ϵ [1, 2, …, D] in D-dimensional space as shown in Eqs. (3), (5) and (8). The optimization problem is actually the problem of finding the minimum value of nonlinear equations. Therefore, the best Par corresponds to the minimum fitness value. Fit, is the corresponding fitness value. The steps of the hybrid CSO-FL approach are described in algorithms 1 and 2.

Algorithm 2 shows phase 2 of the proposed hybrid CSO-fuzzy logic that take upper, lower values, membership function of majority and minority classes from Imblanced_Fuzzy function and then check the objective function of the generated samples, and if its satisfy the objective function then this sample will be added to the minority class.

ALGORITHM 1: Imblanced_Fuzzy

Function Imblanced_Fuzzy (K,T,M1,M2,R)
Input: K,T,M1,M2,R where
K:#neighbors,
T:Number of required Samples,
M1: Minority class samples,
M2: Majority class samples,
R:#iterations
Output: Original Data + T * Minority class samples
For i=1 to R **do**
S← M1(i)
Get k-nearest neighbors of S from M1 along with their distances
minA← the nearest neighbor of M1 to S
maxA← the farthest neighbor of M1 to S
Get k-nearest neighbors of S from M2 along with their distances
minB← the nearest neighbor of M2 to S
maxB← the farthest neighbor of M2 to S
x← Randomly select one of the nearest neighbors of S from M1
μ(minA_M1)← calculate membership value to minA with respect to M1
μ(maxA_M1)← calculate membership value to maxA with respect to M1
μ(minA_M2)← calculate membership value to minA with respect to M2
μ(maxA_M2)← calculate membership value to maxA with respect to M2
μ(minority)=Min(μ (minA_M1), μ (maxA_M1))
μ(majority)=Max(μ (minA_M2), μ (maxA_M2))
For j=1 to N **do** //loop to generate samples
For p=1 **to** P **do** //loop all attributes
lower[p]← maxA - minA // an array of lower limits of all attributes.
upper[p]← maxB - minB // an array of upper limits of all attributes.
End for
End for
End

4. Datasets, experiments, and results

Datasets and experimental results are discussed in this section.

4.1 Datasets

In this section, we describe the basic properties of the chosen datasets to apply the proposed hybrid model on it [27].

We choose 3 different imbalanced datasets (abalone 19, page-blocks 0, and Pima). They are usually comprised of two classes: the negative majority and the positive minority. Both Imbalanced data sets are divided by five folds stratified cross validation. Notice that it is considered to divide the dataset into five folds in order to obtain enough

ALGORITHM 2: CSO_Based_Fuzzy

```
Function CSO_Based_Fuzzy (lower,
upper, μ(minority), μ(majority))
// Input to CSO
// Fit_Func,
M: Number of (iterations),
population size,
rPercent: population size of roosters, percent of the total population size, hPercent: Population size of hens  mPercent: The mother hens accounts
rNum = round( pop * rPercent );
//The population size of roosters
hNum = round( pop * hPercent );
//The population size of hens
cNum = pop - rNum - hNum;
/The population size of chicks
mNum = round( hNum * mPercent );
//The population size of mother hens
lb= lower[p];% Lower bounds
ub= upper[p];
%Initialization
for i = 1 to pop do
x( i, : ) = lb + (ub - lb) .* rand( 1, dim );
fit( i ) = FitFunc( x( i, : ) );
End
If Member(x)> minority_memb && Member(x)< majority__memb
 objective= Member(x);
else
objective=9999;
End
Optimal_solution =  bestX;
The_objective_value=  fMin;
Sample[p]=PSO(lower[i], upper[i])
End
```

minority class examples from the test partitions. Thus, test partition examples show the fundamental knowledge more clearly.

Abolone 19 dataset: the abalone 19 dataset also has 8 attributes, and composed of 4174 instances with the imbalanced ration 129.44. The number of positive instances is 32 and number of negative instances is 4142.

Page-blocks 0 dataset: the page-blocks 0 dataset has 10 attributes and composed of 5472 instances with imbalanced ratio 8.79. The number of positive instances is 559 and number of negative instances is 4913.

Pima dataset: Pima dataset has 8 attributes and composed of 768 instances with imbalanced ratio 1.87, the number of positive instances is 268 and negative instances is 500.

The reason for choosing these three datasets is (the difference between the imbalanced ratios between them, to clarify the effect of the proposed model. [27]

Table 1. Classification results on the original Abalone 19 dataset

abolone 19 (4142,32)	**KNN**	**SVM**	**DT**	**Naïve**
Accuracy	0.9923	0.9923	0.9863	0.9463
Sensitivity	0	0	0	0.02
Precision	0	0	0	0.125
F_score	0	0	0	0.0345

Table 2. Proposed hybrid CSO-FL model with abalone 19 dataset

abalone 19	**KNN**	**SVM)**	**DT**	**Naïve**
Accuracy	0.9706	0.9704	0.9672	0.8395
Sensitivity	0.9477	0.9341	0.9299	0.9519
Precision	0.9006	0.9103	0.8989	0.541
F_score	0.9235	0.922	0.9141	0.6899

4.2 Experiments and Results

In this section, we will describe the output of the proposed model for each dataset, and then we will show a comparison for our model with other three different models.

At the beginning, we run four classification algorithms K-nearest neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Naïve [28] on the three datasets to show the bad effect of the imbalanced data on the accuracy, sensitivity, precision, and F_score measuring values. Here we will describe the results obtained for each dataset lonely.

For the abalone 19 dataset, Table 1 shows the results obtained from the four classifiers on the abalone 19 (4142, 32) dataset.

It shown that the accuracy value is > 90 % for all classifiers. The high accuracy values don't reflect the actual accuracy resulted because there is big difference between the numbers of instances in the two classes. The sensitivity, precision, and F_score values are 0 for all algorithms except Naïve, they are smaller than 0.2%. The main cause of this is the high imbalanced ratio in this dataset that reached 129.44. We applied the proposed hybrid CSO-FL model on the abolone 19 dataset to increase the number of positive instances and the obtained results shown in Table 2.

As shown in Table 2, the number of positive instances reached to 956 instances, this cause the sensitivity, precision, and F_score values to be enhanced for all classifiers. The sensitivity value reached > 95 %, precision >91 %, and F_score > 92 % in different classifiers. This reflects the effect of the

Table 3. Classification results on the original page- blocks 0 dataset

page-blocks0 (4913,559)	KNN	SVM	DT	Naïve
Accuracy	0.9609	0.966	0.9682	0.9379
Sensitivity	0.8694	0.8894	0.8279	0.6588
Precision	0.7263	0.7621	0.8694	0.8058
F_score	0.7914	0.8208	0.8482	0.7249

Table 4. The proposed hybrid CSO-FL model with page-blocks0 dataset

page-blocks0 (4913,4899)	KNN	SVM	DT	Naïve
Accuracy	0.9823	0.9809	0.9888	0.9367
Sensitivity	0.9965	0.9888	0.9916	0.9481
Precision	0.9688	0.9735	0.9858	0.9226
F_score	0.9825	0.9811	0.9887	0.9351

added instances to the dataset.

For the page-blocks 0 dataset, Table 3 shows the results of the four classifiers on that dataset.

It is shown that the accuracy values for all algorithms are > 90 %, while sensitivity, precision and F_score values are still affected by the imbalanced ratio in this dataset which is 8.79. To increase the positive instances number in this dataset in order to enhance the sensitivity, precision and F_score values, we applied the proposed hybrid CSO-FL model, and Table 4 shows the results.

The number of positive instances reached to 4899, all measurements values are enhanced with the new instances. The sensitivity values reached to 99.65% in KNN, while the smallest sensitivity was achieved by Naïve reached to 94.81% after it was 65.88%. The precision value reached to 98.58% in DT, and the smallest precision was in KNN 72.63% reached to 96.88%. Also the F_score value enhanced and the highest value reached to 98.87% in DT. The smallest one was in Naïve 72.49% reached to 93.51%. These values reflects the positive effects of the added positive instances by the proposed model.

The last experiment on the Pima dataset, the results of the four classifiers on that dataset appear in Table 5.

The imbalanced ratio of the Pima dataset was 1.87 affected all the measurements values, we can notice also that the accuracy value is affected since highest accuracy resulted is 73.96%. That is beside the values of sensitivity, precision, and F_score also. After applying the proposed model on the Pima dataset the number of positive instances reached to

Table 5. Classification results on the original Pima dataset

Pima (500,268)	KNN	SVM	DT	Naïve
Accuracy	0.7396	0.7331	0.7148	0.7396
Sensitivity	0.6405	0.6438	0.5904	0.7742
Precision	0.5784	0.5261	0.597	0.3582
F_score	0.6078	0.5791	0.5937	0.4898

Table 6. Proposed model with Pima dataset

Pima (500, 436)	KNN	SVM	DT	Naïve
Accuracy	0.8611	0.86	0.8162	0.861
Sensitivity	0.9083	0.9312	0.8647	0.8828
Precision	0.8148	0.8008	0.7694	0.8433
F_score	0.859	0.8611	0.8143	0.8626

We can notice from table 6 that all measurements values have been increased and lowest accuracy become 81.62% and highest reached to 86.11% in KNN. In addition, sensitivity reached to 93.12, precision 84.33%, and F_score 86.26%. The resulted enhancements show that the added positive instances have appositive effect on the different measurements which reflects the value of these added instances.

To evaluate the proposed model, we run the smote algorithm, modified smote [25], and TGT [26] on the three datasets . Figures 1, 2, and 3 show the obtained results for all datasets abalone 19, page-blocks0, and pima respectively for all experiments applied in this research. Before any enhancements, with smote, modified smote [25], TGT [26], and the proposed CSO-FL.

From Fig. 1 to 3 we can see that smote and modified smote yield positive instances more than TGT and Proposed CSO-FL model, while the measurements values of smote and modified smote are less than the obtained in both TGT and the proposed model. Also we can observe that the results obtained from the proposed CSO-FL model outperforms the TGT [26] model in most measurements values for all classifiers.

We can observe that the hybrid CSO -FL over performed for separate classifiers in all three datasets. We note that the classifiers display high precision metrics in Fig. 1, Fig. 2, and Fig. 3 before over-sampling. Generally speaking, the precision metric calculates the proportion of all properly classified cases. But here, if used alone for other evaluations [26], this metric value is false. As a result of data inequality, all large samples are usually accurate and

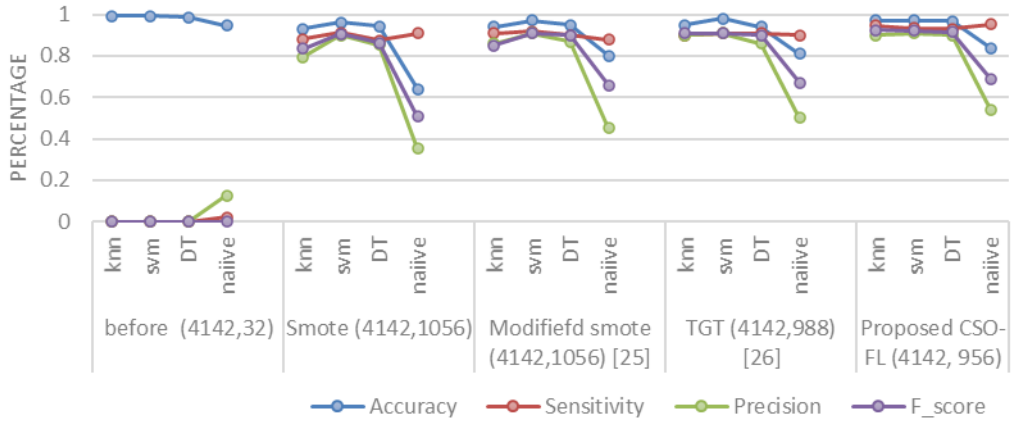

Figure. 1 Comparison on the abalone 19 dataset (before enhancements, Smote, modified smote [25], TGT [26], and Hybrid CSO-FL model)

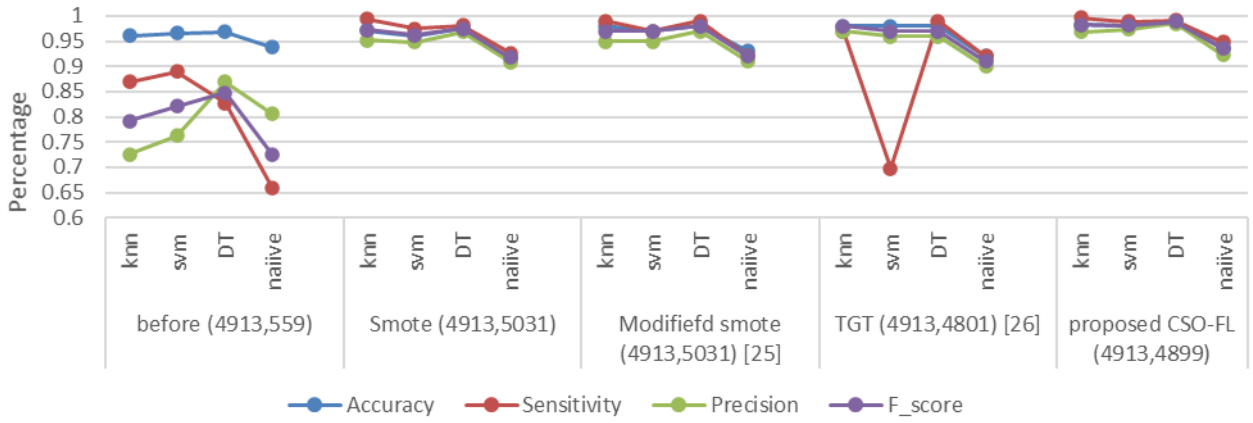

Figure. 2 Comparison on the page-blocks0 dataset (before enhancements, Smote, modified smote [25], TGT [26], and Hybrid CSO-FL model)

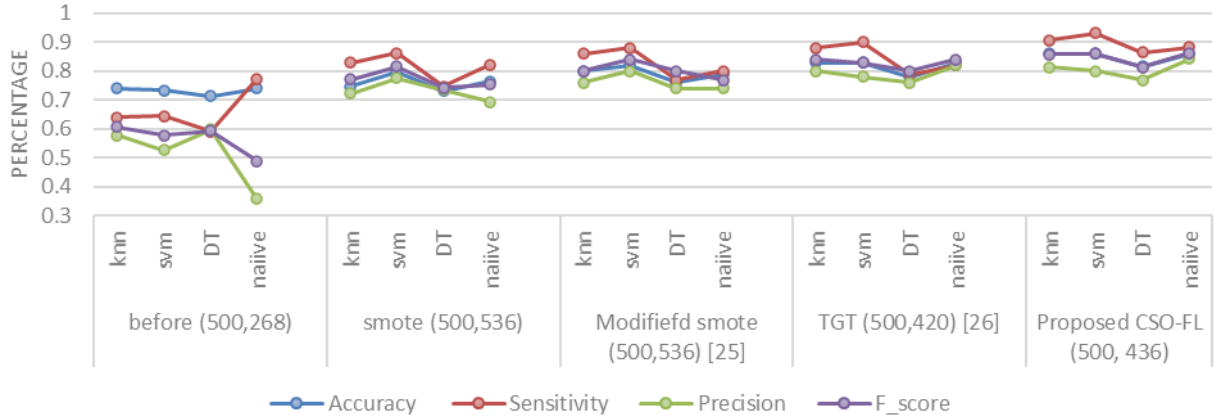

Figure. 3 Comparison on the pima dataset (before enhancements, Smote, modified smote [25], TGT [26], and Hybrid CSO-FL model)

all minority samples are inaccurate. Moreover we can observe that when the ratio of imbalanced data is high; this lead to a higher improvement in the evaluation metrics because the use of the CSO can help us of generating a lot of samples between the scope of the minority and majority samples based on fuzzy logic system.

The performance of the methodology suggested against the SMOTE algorithm may be linked to the discrepancy between the method's workings. In a minority data space the traditional SMOTE algorithm generates synthetic samples; the proposed approach generates synthesis samples that are governed by the law of the minority and majority groups, derived

from fuzzy logic. These produced samples are verified with two controls after the led generation; in the first check we ensured that it is in the minority scope using lower and upper bounds and second check using objective function in CSO in order to ensure that the minority class includes all synthetic samples. If not, they will be thrown out. This new double-checked samples generated with the p hybrid CSO-FL sampling methodology resulted in better rankings, which prove our original claim. In addition, the interaction between fuzzy logic and CSO has shown that it is efficient enough to produce new synthetic samples better than those produced by the sample SMOTE. The fuzzy logic addresses reasoning which is approximate instead of accurate in a manner that is much like human logic. Furthermore, fuzziness decreases away from locations with a higher possibility or existence. This aspect satisfies the condition that the generated samples cannot be placed near of majority class area to protect of generating a misclassified sample.

The generated samples are Uncertainty samples in the minority class so the use of fuzzy logic in the proposed system will ensure that the generated samples are correct before adding them to the data set and authors in literature advice of using fuzzy logic in solving problems of uncertainty.

5. Conclusion and future work

The paper explored the nature of the imbalanced data and its current real-life applications. We provided a taxonomy for the solutions found in the literature. Then, we presented a comparative study for the efforts done with the aim of addressing the challenge of the classification of imbalanced data. At last, we introduced our proposed solution based on hybridization between CSO and FL for handling the imbalanced data problem along with our experiment, which showed a noticeable higher performance results in the three datasets using different classifiers. The Proposed CSO-FL approach gave us a higher evaluation metrics in case of the ratio between the minority and majority class is high. The precision, sensitivity, and F_score values enhanced in all classifiers. The values in the abalone 19 dataset improved with ratios >90 % in many classifiers because of the high imbalanced ratio in this dataset, while in pima and page-blocks datasets, the measurement values enhanced with ratios from nearly 10% to 30%. In addition, the proposed CSO-FL approach compared with SMOTE, modified SMOTE, and TGT algorithms on the same datasets and proved to outperform their results. There are many directions available for future work. The three used datasets in our experiments were numerical datasets. The proposed methods can be augmented with datasets having categorical and mixed attributes. One may think to examine applying the proposed methods on real-life datasets which we expect to be very helpful if used in medical field.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources and supervision have been done by 1st author. The data curation, writing—original draft preparation, writing—review and editing, visualization and project administration have been done by 2nd author.

References

[1] M. Jesmeen, J. Hossen, S. Sayeed, C. Ho, K. Tawsif, A. Rahman, and M. Arif, “A survey on cleaning dirty data using machine learning paradigm for big data analytics”, *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 10, No. 3, pp.1234-1243, 2018.

[2] S. Tyagi and S. Mittal, “Sampling approaches for imbalanced data classification problem in machine learning”, In: *Proc. of ICRIC 2019*, pp. 209-221, 2020.

[3] Z. A. Elassad, H. Mousannif, and H. A. Moatassime, “Class-imbalanced crash prediction based on real-time traffic and weather data: a driving simulator study”, *Traffic Injury Prevention*, Vol. 21, No. 3, pp. 201-208, 2020.

[4] J. Zhao, J. Jin, S. Chen, R. Zhang, B. Yu, and Q. Liu, “A weighted hybrid ensemble method for classifying imbalanced data”, *Knowledge-Based Systems*, Vol. 203, p. 106087, 2020.

[5] S. Li, W. Gong, and Q. Gu, “A comprehensive survey on meta-heuristic algorithms for parameter extraction of photovoltaic models”, *Renewable and Sustainable Energy Reviews*, Vol. 141, p. 110828, 2021.

[6] H. Chen, D. Fan, L. Fang, W. Huang, J. Huang, C. Cao, L. Zeng, L. Yang, and Y. He, “Particle swarm optimization algorithm with mutation operator for particle filter noise reduction in mechanical fault diagnosis”, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 34, No. 10, p. 2058012, 2020.

[7] J. Luo, F. He, and J. Yong, “An efficient and robust bat algorithm with fusion of opposition-

based learning and whale optimization algorithm", *Intelligent Data Analysis*, Vol. 24, No. 3, pp. 581-606, 2020.

[8] X. Meng, Y. Liu, X. Gao, and H. Zhang, "A new bio-inspired algorithm: chicken swarm optimization", In: *Proc. of International Conference in Swarm Intelligence*, pp. 86-94, 2014.

[9] S. Deb, X. Gao, K. Tammi, K. Kalita, and P. Mahanta, "A new teaching–learning-based chicken swarm optimization algorithm", *Soft Computing*, Vol. 24, No. 7, pp. 5313-5331, 2020.

[10] M. E. Alaoui, "Fuzzy TOPSIS: Logic, Approaches, and Case Studies", *CRC Press*, 2021.

[11] L. Zadeh, "Fuzzy logic", *Computer*, Vol. 21, No. 4, pp. 83-93, 1988.

[12] M. Kumar, L. Misra, and G. Shekhar, "Survey in fuzzy logic: an introduction", *International Journal for Scientific Research & Development*, Vol. 3, No. 6, pp. 822-824, 2015.

[13] O. A. M. Ali, A. Y. Ali, and B. S. Sumait, "Comparison between the effects of different types of membership functions on fuzzy logic controller performance", *International Journal of Emerging Engineering Research and Technology*, Vol. 3, No. 3, pp. 76-83, 2015.

[14] M. Masdari and H. Khezri, "A survey and taxonomy of the fuzzy signature-based intrusion detection systems", *Applied Soft Computing*, Vol. 92, pp. 106301, 2020.

[15] R. Das, S. Sen, and U. Maulik, "A survey on fuzzy deep neural networks", *ACM Computing Surveys (CSUR)*, Vol. 53, No. 3, pp. 1-25, 2020.

[16] H. Aydadenta and A. Adiwijaya, "A clustering approach for feature selection in microarray data classification using random forest", *Journal of Information Processing Systems*, Vol. 14, No. 5, pp. 1167-1175, 2018.

[17] Y. Hou, B. Li, L. Li, and J. Liu, "A density-based under-sampling algorithm for imbalance classification", *Journal of Physics: Conference Series*, Vol. 1302, No 2, p. 022064, 2019.

[18] J. Ha and J. Lee, "A new under-sampling method using genetic algorithm for imbalanced data classification", In: *Proc. of the 10th International Conference on Ubiquitous Information Management and Communication*, pp. 1-6, 2016.

[19] J. Lee, N. Kim, and J. H. Lee, "An over-sampling technique with rejection for imbalanced class learning", In: *Proc. of the 9th International Conference on Ubiquitous Information Management and Communication*, pp. 1-6, 2015.

[20] A. Fernández, S. del Río, N. Chawla, and F. Herrera, "An insight into imbalanced big data classification: outcomes and challenges", *Complex & Intelligent Systems*, Vol. 3, No.2, pp. 105-120, 2017.

[21] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning", In: *Proc. of International Conference on Intelligent Computing*, pp. 878-887, 2005.

[22] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection", *Artificial Intelligence in Medicine*, Vol. 37, No. 1, pp. 7-18, 2006.

[23] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem", In: *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 475-482, 2009.

[24] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 1, pp. 238-251, 2015.

[25] A. Mahmoud, F. Ali, A. El-Kilany, and S. Mazen, "A Novel Oversampling Technique To Handle Imbalanced Datasets", In: *Proc. of 34th International European Council Conference on Modelling and Simulation*, 2020.

[26] A. Mahmoud, A. El-Kilany, F. Ali, and S. Mazen, "TGT: A Novel Adversarial Guided Oversampling Technique for Handling Imbalanced Datasets", *Egyptian Informatics Journal*, 2021.

[27] I. Triguero, S. González, J. Moyano, S. García, J. Alcalá-Fdez, J. Luengo, and F. Herrera, "KEEL 3.0: an open source software for multi-stage analysis in data mining", *International Journal of Computational Intelligence Systems*, Vol. 10, No. 1, pp. 1238-1249, 2017.

[28] V. Geetha and S. Gowsalya, "An Efficient Classification Techniques in Data Mining", *Adalya Journal*, Vol. 9, 2020.