# Imbalance Data Handling with ESMOTE for Optimal FCM-based Clustering Mechanism

Raid Saleh Ali Shatat
Basic Sciences, Affiliation: Assistant prof, Riyadh ELM University, Riyadh, Kingdom of Saudi Arab.
raid@riyadh.edu.sa

**Abstract:**The imbalanced data classification becomes a serious problem if there is a large difference in sample sizes between classes. Many real-world applications are hampered by this problem, which significantly reduces classifier accuracy in predicting minority classes. Conventional approaches employ undersampling or oversampling strategies to rectify data imbalance. However, these techniques may introduce noise and model bias or result in information loss during clustering operations. Thus, this research suggested an optimal clustering mechanism for solving the issues in the imbalance data clustering. Initially, the Evolutionary Synthetic Minority Over-sampling Technique (ESMOTE) is utilized for handling the imbalanced data. This model removes the noisy sample in the data and determines the minority classes in the data. After that, the clustering is executed via the Optimal Fuzzy C Mean (OFCM), where the parameters are tuned using a Masterpiece Optimization Algorithm (MOA). Finally, the analysis is conducted using the to prove its effectiveness in the clustering process.

**Keywords:** Imbalance Data Handling; Optimal Fuzzy C Mean Clustering; Evolutionary Synthetic Minority Over-Sampling Technique; Masterpiece Optimization Algorithm.

**Received: November 12, 2024 Revised: November 26, 2024 Accepted: January 3, 2025**

**Nomenclature**

| Abbreviation | Expansion |
|---|---|
| MOA | Masterpiece Optimization Algorithm |
| ESMOTE | Evolutionary Synthetic Minority Over-sampling Technique |
| OFCM | Optimal Fuzzy C Mean |
| CBSMOT | Cluster-Based Synthetic Minority Oversampling Technique |
| ML | Machine Learning |
| AKCUS | Adaptive K-means Clustering Undersampling |
| CUTE | Clustering-Based Undersampling Technique |
| auPR | Precision-Recall Curve |
| auROC | area under the Receiver Operator Curve |
| MCC | Mathews Correlation Coefficient |
| NMI | Normalized Mutual Information |
| ARI | Adjusted Rand Index |
| OOA | Osprey optimization algorithm |
| SCO | Single Candidate Optimizer |
| TOA | Teamwork Optimization Algorithm |
| CSO | Cuckoo Search Optimization |

## 1. Introduction

It is extremely challenging to classify samples from an unbalanced dataset [1]. The majority of real-world problem datasets have high dimensions and multiple classes. Since the minority class is frequently misclassified, the existence of an imbalanced class has a substantial impact on classification accuracy. Predicting the minority class is the most crucial scenario[2]. For instance, compared to the non-fraud data, the number of fraudulent cases is extremely low in the case of credit card fraud detection. To identify objects by learning from data, ML is essential. Algorithms and dataset quality determine its performance [3]. However, a major classification difficulty is data imbalance. This is because classifier construction attempts to reduce the classification error rate [4]. Because minority class samples have influenced the overall accuracy, they frequently degrade performance when data imbalance problems occur. Unfortunately, it should be noted that these minority class samples frequently include important information. Inaccurate risk identification can have serious issues in several industries, including fraud detection, natural disaster forecasting, and medical disease diagnosis, where the data imbalance problem is a common problem [5].

The class having major instances in the unbalanced dataset is referred to as the majority class, while the class with the fewest instances is referred to as minority data. The classic classification models typically neglect the minority data to maximize accuracy or error functions. Since their main goal is to increase accuracy, those models are unable to handle imbalanced data well [6]. Accuracy is rarely selected in the majority of scenarios involving imbalanced datasets. Undersampling and oversampling are the most often used techniques for dealing with unbalanced datasets for binary classification [7]. To achieve balance, undersampling involves removing random examples from the majority class. Conversely, oversampling artificially raises minority occurrences. While undersampling runs the danger of losing crucial data, oversampling suffers from overfitting [8]. Cluster-based undersampling, which forms clusters before the random sampling. However, random sampling from the clusters of majority data also degrades cluster-based undersampling [9].

### 1.1. Challenges in Handling Imbalanced Data

In the real-world dataset, the class imbalance is the major issue. The classes of the dataset contain small and large data points which are not handled by the conventional data balancing methods. The conventional approaches are overwhelmed by the minority class because of the poor learning capacity and these methods ignore the minority class in the dataset which provides inaccurate results in the classification process. When executing the clustering operation, the difference between the classes of the dataset is not detected by the conventional model. In addition, the collection of imbalanced datasets is the toughest process. The minority classes of the dataset are distributed in the numerous features space which causes significant issues in the clustering operation. So, in this work, the imbalanced data clustering is performed with the help of the optimal clustering mechanism.

### 1.2. Contributions made for Handling Imbalanced Data

In this research, the optimal clustering model is developed for clustering the imbalanced data which helps to reduce the misclassification error in the detection operation. The ESMOTE is initially proposed to solve the class imbalance issues in the collected imbalanced dataset. The learning process of the classifier is greatly enhanced by this model as it can create a diverse set of minority classes from the collected data based on the synthetic sample and minority algorithm. The MOA is proposed for tuning the fuzziness parameter, epsilon, and iterations count in the OFCM for enhancing the accuracy of the data balancing process. Here, the best exploration and high trapping activity of the MOA provide the optimal solution in the OFCM-based clustering. The OFCM is developed for clustering the imbalanced dataset. The OFCM can enhance the generalization and prevent overfitting issues in the detection and classification.

## 2. Literature Survey

### 2.1. Resampling Techniques for Class Imbalance Problems

In data mining and ML, the distribution of imbalanced data is an unresolved problem. During the classification operation in real-world scenarios, the class imbalance causes severe bias which affects the classification performance. The under-sampling and oversampling approaches are developed by the researcher to solve the class imbalance issues.

***Oversampling***: The minority instance in the dataset is enhanced by the oversampling techniques for solving the class imbalance issues. The imbalanced data oversampling method was proposed in [10] to solve the problem of data imbalance in the field of medical departments. CBSMOT was proposed in [11] to balance the given dataset and this model also solves the issues of redundancy and information loss in the clustering process. The noise-robust oversampling algorithm in [12] was developed for handling the multi-class imbalanced data and this model can handle the noise in the cluster of the data sample while enhancing the class boundaries.

***Undersampling***: The data from the majority instance is selected by the undersampling method for performing the data balancing operation with the help of minority data present in the dataset. In [13], undersampling-based clustering was developed to solve the imbalance issues in the data. In each cluster of data, the undersampling process was carried out. The AKCUS [14] was proposed for performing the clustering over the majority of the class. In this model, the distance factor was considered for executing the undersampling-based clustering. The CUTE was proposed in [15] to reduce the majority of class issues in the collected data.

Here, the representativeness of each member of the cluster was analyzed by this model to attain excellent results in the binary classification process.

## 3. Methodology

In this work, an optimal clustering mechanism is developed for clustering the imbalanced training dataset to enhance the model's performance in the detection task. The optimal clustering model can summarize the majority of the classes in the collected data which helps to develop the balanced training set. The model can learn the features from both classes in the training dataset to enhance the performance of the model under different scenarios. The optimal clustering can prevent overfitting and enhance the generalization capability of the model which provides enhanced classification outcome. The faults diagnosis, pattern recognition, and healthcare department are facilitated by the developed optimal clustering mechanism as it partitions the given data into meaningful groups. At first, the ESMOTE is utilized for handling the imbalance issues in the collected imbalanced dataset. The ESMOTE is the advanced model for solving the imbalance issues by selecting the minority class instances. In addition, the risk of overfitting in the imbalanced data clustering is prevented by the ESMOTE. This model generates the synthetic sample which enhances the performance of the classifier in various types of detection and classification tasks. After solving the class imbalance issues, the imbalanced data is clustered using the OFCM. The OFCM is one of the most widely used clustering approaches. The compactness of the different data clusters is maximized by the OFCM. In between the cluster center and sample points of the given dataset, the squared error is reduced by the OFCM. Here, the parameters like the fuzziness parameter, epsilon, and iterations count are optimized in the OFCM for maximizing the accuracy while addressing the cluster size sensitivity issues. All data points of the cluster have a degree of membership function which is allocated by the OFCM. The degree of membership function in the OFCM handles the overlapping issues in the imbalanced data clustering process. Finally, the experimentation is conducted on the standard dataset to prove the effectiveness of the proposed model in the imbalanced data clustering process. The overall representation of the proposed imbalanced data clustering model is given in Fig. 1.

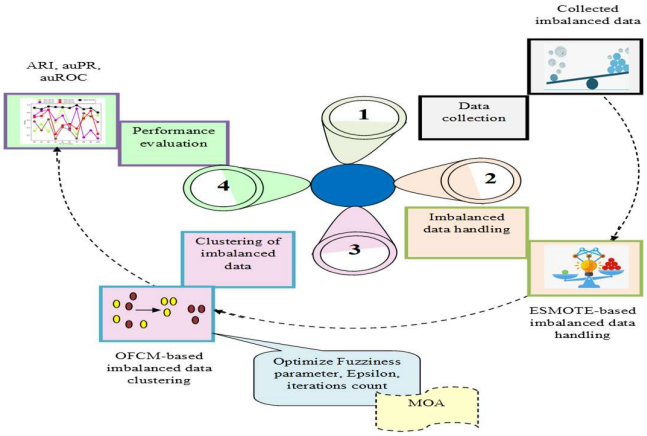


Fig 1. Overall Representation of the Proposed Imbalanced Data Clustering Model

### 3.1. ESMOTE-based Data Handling Process

The ESMOTE [16] is an effective approach for solving the data imbalance issues. Based on the evolutionary strategy, the ESMOTE is developed. The decision vector in the collected dataset was calculated based on the oversampling ratio of the minorities in the collected imbalanced dataset. The mutation steps and their probabilities, minorities of the oversampling ratio, and adjacent neighbors are available in the decision vector and it is represented by the term $Y$ and is expressed in Eq. (1).

$$Y = \left(\{s_k\}, L, \chi, \{q_k\}, \{\phi_k\}\right) \tag{1}$$

The set of new samples is signified as $\chi$ , the vector of the oversampling ratio is denoted as $Y$ . For the minority $k$ , the oversampling ratio is expressed as $s_k$ , the mutation probability is expressed as $q_k$ , and the mutation step is represented as $\phi_k$ , the minority sample's nearest

neighbor is denoted as $L$. With the help of the parameter $s_k$, the synthetic sample $\chi$ and minority neighbors are developed $L$.

### 3.2. Clustering Mechanism

The imbalanced data is effectively clustered by the OFCM [17]. Each data point of the cluster is allocated with the membership function. The OFCM determines the weighted average of each data point for calculating the center of the cluster. The majority class is overlapped by the minority classes in case of imbalanced data. However, the OFCM clustering mechanism considers the degree of membership for solving the overlapping issues. The elimination of overlapping issues helps to differentiate the data within the cluster and provides accurate results in the detection as well as classification tasks. The OFCM approaches determine the distance between the data points in the cluster center for creating the cluster. Further, the OFCM selects the cluster center from the imbalanced data while maintaining the minority classes. This process helps to reduce the majority of classes in the data. The OFCM model helps to develop the balanced dataset which aids to effectively train the model used for the detection and classification. The utilization of the OFCM is greatly increased as it is one of the simple MLmodels. The sample dataset is divided into several groups based on Eq. (2).

$$\begin{cases} K(V,W) = \sum_{j=1}^{d}\sum_{k=1}^{o} v_{jk}^{n} \left\| y_k - w_j \right\|^2 \\ subjected\ to \sum_{k=1}^{o} v_{jk} = 1, v_{jk} \in [0,1] \end{cases} \tag{2}$$

Here, the $j^{th}$ cluster center is denoted as $w_j$, and the $k^{th}$ sample point is expressed as $y_k$, with respect to the cluster centre $w_j$, the membership value of the data point $y_k$ is expressed as $v_{jk}$, the count of the cluster is represented as $d$, the fuzzy index is represented as $n$, the total number of data points are denoted as $o$. The membership function $v_{jk}$ is updated by the Lagrange method and it is given in Eq. (3).

$$v_{jk} = \frac{1}{\Sigma_{l=1}^{d} \left( \frac{\left\| y_k - w_j \right\|}{\left\| y_k - w_l \right\|} \right)^{\frac{2}{n-1}}} \tag{3}$$

The cluster center $w_j$ is upgraded using Eq. (4).

$$w_j = \frac{\Sigma_{k=1}^{o} v_{jk}^{n} y_k}{\Sigma_{k=1}^{o} v_{jk}^{n}} \tag{4}$$

According to the maximum degree of membership, the sample data points are allocated to the corresponding category. The structural representation of the OFCM-based imbalanced data clustering is illustrated in Fig. 2.

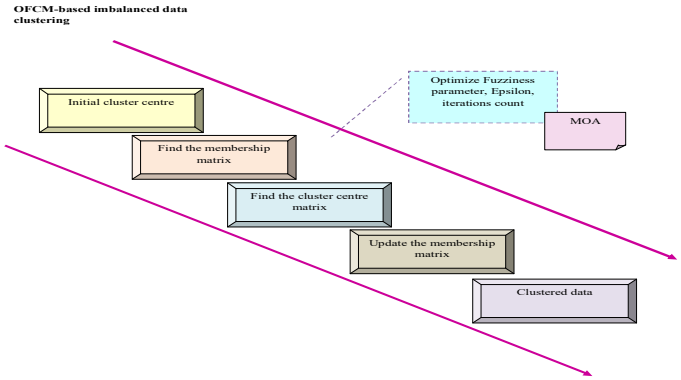


Fig 2. Structural representation of the OFCM-based imbalanced data clustering

### 3.3. MOA-based Parameter Tuning for Achieving Objective Function

The MOA [18] is the nature-inspired algorithm that follows the activity of the Japanese Pufferfish. The MOA is adopted in the FCM for tuning the Fuzziness parameter, Epsilon, and

iterations count for enhancing the accuracy of the imbalanced data clustering process. It is quite easy to form the clusters of the data because of the parameter tuning process. The more accurate center and the optimal answer are determined using the MOA. The parameter tuning helps to overcome the limitation of the FCM in the imbalanced data clustering. The best solution is obtained by the high exploration capability of the MOA. The high trapping activity prevents the entrapment in the local optimal condition. The objective function of the FCM is optimized to improve the performance of the optimal data clustering. The major criteria of the MOA are to optimize the parameters of the FCM so it can exhibit acceptable performance in the clustering of imbalanced datasets. The objective function of the MOA-OFCM is given in Eq. (5).

$$l = \underset{\left\{d_v^{ofcm}, s_h^{ofcm}, k_j^{ofcm}\right\}}{\arg\min} \left( \frac{1}{accuracy} \right) \tag{5}$$

Thus, the objective function is expressed as $l$, the fuzziness parameter is denoted as $d_v^{ofcm}$, the Epsilon is expressed by the term $s_h^{ofcm}$ and the iteration count is denoted as $k_j^{ofcm}$. The accuracy is calculated using Eq. (6).

$$accuracy = \frac{xc + xv}{xc + xv + hk + hl} \tag{6}$$

Here, the true positive and false positive are expressed as $xc, hk$ the true negative and false negative are denoted as $xv, hl$. The pseudocode of the MOA is presented in Algorithm 1.

| **Algorithm 1**: MOA |
|---|
| **Input:** Fuzziness parameter $d_l$, Epsilon $s_k$, iterations count $k_g$ |
| **Output:** Optimized fuzziness parameter $d_v^{ofcm}$, optimized Epsilon $s_h^{ofcm}$ and the optimized iteration count $k_j^{ofcm}$. |
| Initialize the cluster of the optimization problem. |
| Generate the random points in the response area. |
| While termination not satisfied |
| Select the best point within the nest |
| Find the direction of the water flow and determine the velocity value. |
| Develop the new points and search direction. |
| Find the desired point position. |
| Find the optimal solution |
| Return to the best solution |
| Stop |

## 4. Experimental Analysis

### 4.1. Benchmark Datasets for Analysis

The experiment on the clustering undersampling dataset obtained from https://github.com/tanvir097/clustering_undersampling/tree/main/new_data_2 with access date: 2024-11-09 is used for validating the performance of the developed optimal FCM model. This dataset consists of various text files that are imbalanced. The Python codes were used for conducting all experiments on the proposed model. The experimental outcome of the proposed optimal FCM model was correlated with the AKCUS [14], CUTE [15], CBSMOT [11] and the OOA [19], SCO [20], TOA [21], CSO [22].

### 4.2. Validation Metrics

The area under the auPR, auROC, F1-Score, MCC, NMI, and ARI is used for validating the performance of the developed model.

### 4.3. Clustering Comparison with Various Existing Techniques

The clustering result of the proposed MOA-OFCM is presented in Fig. 3. Here, the proposed model is evaluated by adopting the 100 number of clusters. As noticed from Fig. 3, the developed MOA-OFCM provides very effective results in the F1-score metric, and even in the $10^{th}$ cluster also the proposed MOA-OFCM keeps good performance. It can be seen from the graph, that the developed MOA-OFCM desirably attained the highest value of ARI in the $30^{th}$ cluster. The same model attained an ARI value of 0.84 in the $40^{th}$ cluster. Thus, the visualized results showed that the proposed MOA-OFCM provides very effective results in the clustering of imbalanced data. The MOA-OFCM model effectively partitions the given imbalanced

dataset for determining the big difference between each cluster. The high-quality difference of the adjacent cluster is determined using the MOA-OFCM model. The auxiliary analysis of the given input data is performed by executing the clustering operation via the MOA-OFCM. The MOA-OFCM is the combination of MOA and OFCM so it is widely used for dealing with imbalanced problems. The robustness of the imbalanced data clustering is enhanced by the proposed MOA-OFCM. Every sample point in the cluster has a membership function.

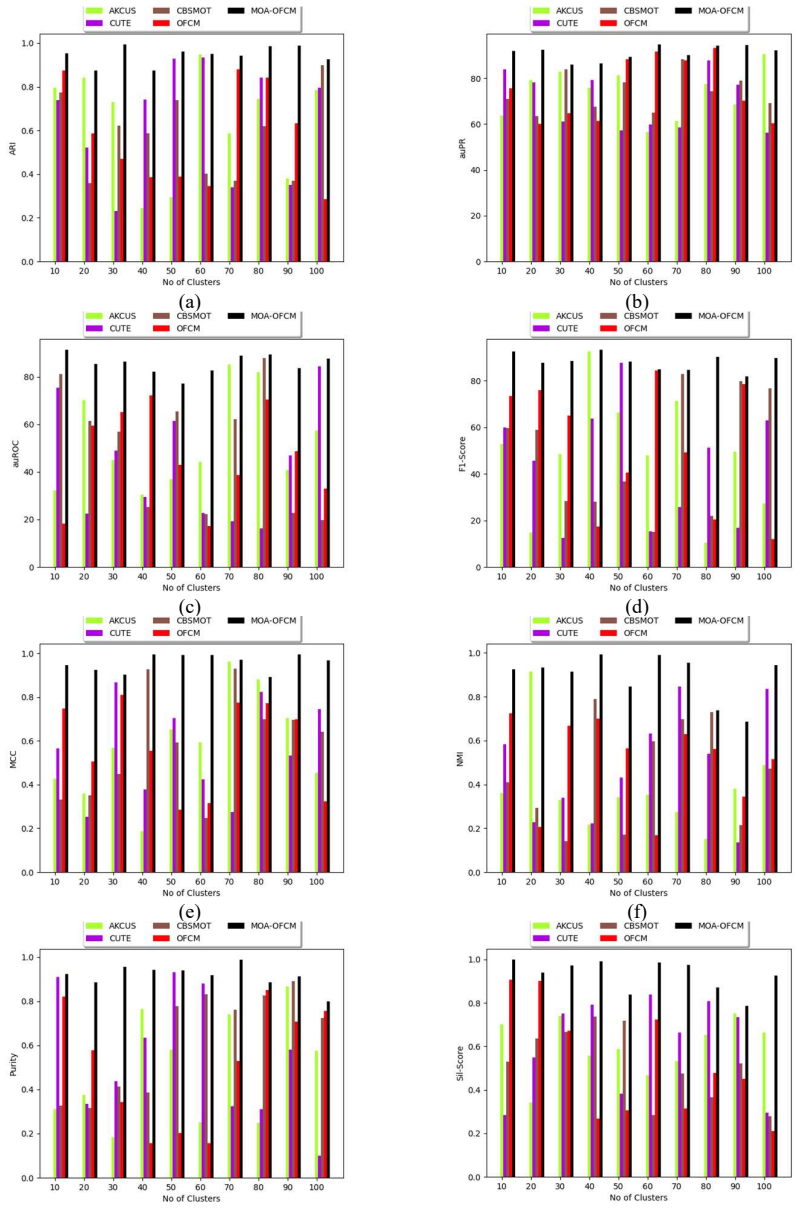


Fig 3. Performance Comparison of the Developed Optimal Clustering Model with existing methods for (a) ARI, (b) auPR, (c) auROC, (d) f1 score, (e) MCC, (f) NMI,(g) Purity, (h) Sil score

### 4.4. Clustering Comparison with Various Existing Algorithms

In the next set of experiments, the proposed MOA-OFCM is compared with a set of algorithms like OOA-OFCM, SCO-OFCM, TOA-OFCM, and CSO-OFCM. It is noticed from Fig. 4 that the proposed MOA-OFCM provides superior results in the auPR metric in the case of all numbers of clusters. The results reported in the graph showed that the proposed MOA-OFCM reaches the purity value of 0.98 in the 70$^{th}$ cluster. Here, the TOA-OFCM gives a very low value of purity in the cluster value of 4. Moreover, the first-best value of the sil score is attained by the TOA-OFCM and the second-best sil score is attained by the SCO-OFCM

model. The conventional OOA-OFCM, SCO-OFCM, TOA-OFCM, and CSO-OFCM provide comparable measurements in the case NMI metric while the proposed MOA-OFCM achieves the best results in the NMI. These results showed that the balanced clustering results are attained by the proposed MOA-OFCM. The conventional OOA-OFCM, SCO-OFCM, and TOA-OFCM fail to provide balanced clustering results on the given imbalanced data. Here, the proposed MOA-OFCM effectively balances the data into several clusters. The proper distribution of classes helps to perform effective learning which enhances the detection and classification results.

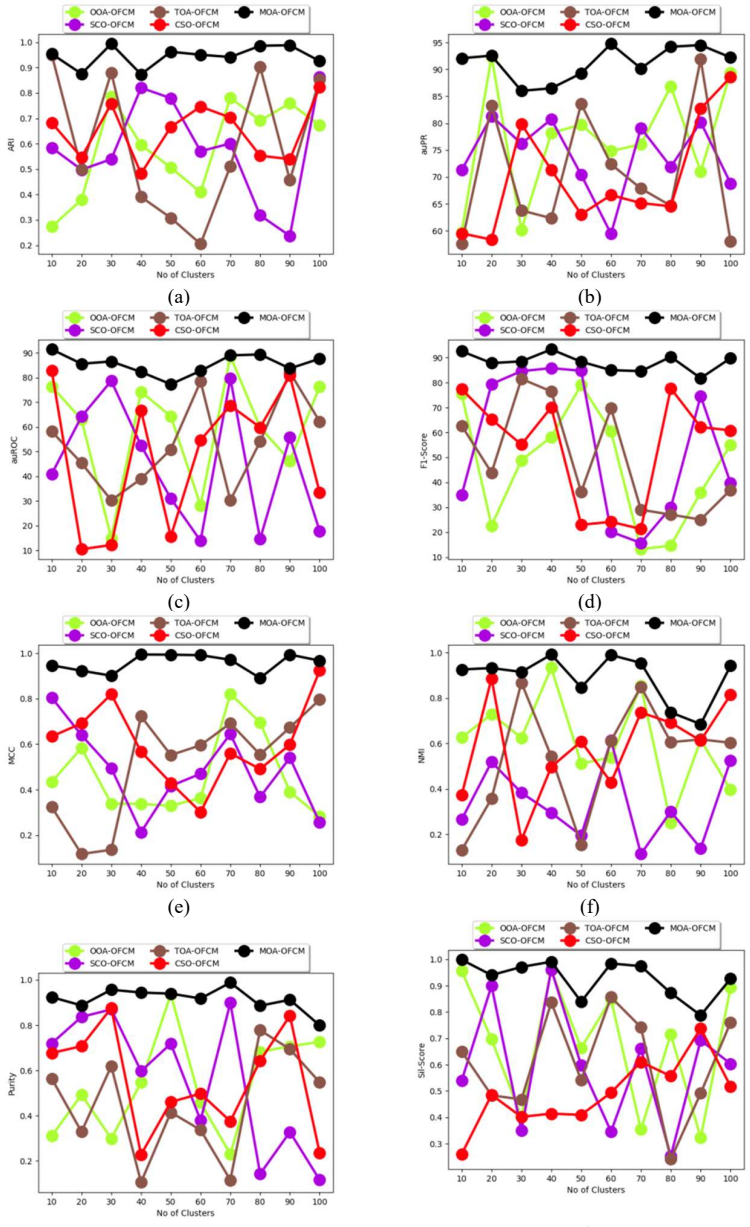


Fig 4. Performance Comparison of the Developed Optimal Clustering Model With Existing Algorithms for (a) ARI, (b) auPR, (c) auROC, (d) f1 score, (e) MCC, (f) NMI,(g) Purity, (h) Sil score

## 5. Conclusion

The purpose of the research work was to develop an optimal clustering model for solving the class imbalance issues in imbalanced data clustering. Here, initially, the ESMOTE was used for handling the imbalance issues in the collected dataset. Further, the MOA-OFCM was applied to the data for clustering the data into separate groups. The robustness of the clustering

and compactness of the group is enhanced by the proposed model. The results reported in the graph showed that the proposed MOA-OFCM reaches the purity value of 0.98 in the 70$^{th}$ cluster. Thus, the performance of the developed model is more efficient than the conventional algorithms. In the future, online learning algorithms will be adopted to enhance the accuracy of the imbalanced data clustering operation.

## 6. References


[1] Yang, W., Pan, C., & Zhang, Y., "An oversampling method for imbalanced data based on spatial distribution of minority samples SD-KMSMOTE," Scientific Reports, vol. 12, Article number: 16820, 2022.

[2] Reza Salehi, A., & Khedmati, M., "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," Scientific Reports, vol. 14, Article number: 5152, 2024.

[3] Liu, Y., Liu, Y., Yu, B.X.B., Zhong, S., Hu, Z., "Noise-robust oversampling for imbalanced data classification," Pattern Recognition, vol. 133, pp. 109008, January 2023.

[4] Nugraha, W., Sony Maulana, M., Sasongko, A., "Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm," Journal of Physics: Conference Series, vol. 1641, pp. 012014, 2020.

[5] Rose Mary, M., and Gunasundari, R., "A Cluster-based Undersampling Technique for Multiclass Skewed Datasets," Engineering, Technology & Applied Science Research, vol. 13, no. 3, pp. 10785-10790, 2023.

[6] Sudhansu Ranjan, L., Kishoro Bisoy, S., Priyadarshini, R., and Nayak, B., "Representative-based cluster undersampling technique for imbalanced credit scoring datasets," In Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021, pp. 119-129. Singapore: Springer Nature Singapore, 2022.

[7] Maram, A., and Ykhlef, M., "Hybrid undersampling and oversampling for handling imbalanced credit card data," IEEE Access, 2024.

[8] Jiaxin, W., Wei, J., Huang, H., Wen, L., Yuan, Y., Chen, H., Wu, R., and Wu, J., "IMWMOTE: A novel oversampling technique for fault diagnosis in heterogeneous imbalanced data," Expert Systems with Applications, vol. 251, pp. 123987, 2024.

[9] Syed As-Sadeq, T., and Chen, Y.,"Comparison of cluster-based sampling approaches for imbalanced data of crashes involving large trucks," Information, vol. 15, no. 3, pp. 145, 2024.

[10] Huajie, H., Liu, B., Xue, X., Cao, J., and Chen, X., "Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique," Applied Soft Computing, vol. 154, pp. 111368, 2024.

[11] Yizhang, W., Qian, J., Hassan, M., Zhang, X., Zhang, T., Yang, C., Zhou, X., and Jia, F., "Density peak clustering algorithms: A review on the decade 2014–2023," Expert Systems with Applications, vol. 238, pp. 121860, 2024.

[12] Haiyan, Y., Xu, X., Li, H., Wu, Y., and Lei, B., "Semi-supervised possibilistic c-means clustering algorithm based on feature weights for imbalanced data," Knowledge-Based Systems,vol. 286, pp. 111388, 2024.

[13] Jingyi, Z., and Liu, X., "An integrated intrusion detection framework based on subspace clustering and ensemble learning," Computers and Electrical Engineering, vol. 115, pp. 109113, 2024.

[14] Mehmet, K., and Düzyel, O., "A new method for GAN-based data augmentation for classes with distinct clusters," Expert Systems with Applications, vol. 235, pp. 121199, 2024.

[15] Zheng, W., Guo, K., Luo, E., Wang, T., Wang, S., Yang, Y., Zhu, X., and Ding, R., "Medical long-tailed learning for imbalanced data: bibliometric analysis," Computer Methods and Programs in Biomedicine, pp. 108106, 2024.

[16] Mohammad Reza, G., Haji Aghajanpour, N., and Ghohani Arab, H., "Masterpiece Optimization Algorithm: A New Method for Solving Engineering Problems," Journal of Soft Computing in Civil Engineering, vol. 1, pp. 21, 2024.

[17] Tao, Y., Wang, K., and Zhang, Y., "Evolutionary synthetic minority oversampling technique with random forest for warfarin dose prediction in Chinese patients," 2019 IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 2019, pp. 2514-2520.

[18] Gao, Y., Yang, C., Lin, K., Pan, J., Li, L., "Conditional semi-fuzzy c-means clustering for imbalanced dataset," IEEE Xplore, Vol. 14, Iss. 11, pp. 2343-2355, 2020.

[19] Mohammad, D., and Trojovský, P., "Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems," Frontiers in Mechanical Engineering, vol. 8, pp. 1126450,2023.

[20] Shami, Tareq M., Grace, D., Burr, A., and Mitchell, P.D. "Single candidate optimizer: a novel optimization algorithm," Evolutionary Intelligence, vol. 17, no. 2, pp. 863-887, 2024.

[21] Tareq, H., Batiha, B., Werner, F., Eguchi, K., Montazeri, Z., and Dehghani, M., "A Completely Different and Innovative Bio-Inspired Metaheuristic Approach for Effectively Solving Complex Optimization Problems Across Various Domains," 2024.

[22] Suresh, K.M., Mansoor Hussain, D., Rohini, M., and Oswalt Manoj, S., "Advanced hybrid optimization algorithms for energy efficient cloud resource allocation," Radio Science, vol. 59, no. 11, pp. e2024RS008012, 2024