Novel Oversampling Algorithm for Handling Imbalanced Data Classification

Novel Oversampling Algorithm

Anjali S. More, Dipti P. Rana
Department of Computer Science and Engineering
Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India

***Abstract*—In the current age, the attention of researchers is immersed by numerous imbalanced data applications. These application areas are intrusion detection in security, fraud recognition in finance, medical applications dealing with disease diagnosis pilfering in electricity, and many more. Imbalanced data applications are categorized into two types: binary and multiclass data imbalance. Unequal data distribution among data diverts classification performance metrics towards the majority data instance class and ignores the minority data, instance class. Data imbalance leads to an increase in the classification error rate. Random Forest Classification (RFC) is best suitable technique to deal with imbalanced datasets. This paper proposes the novel oversampling rate calculation algorithm as Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate). Experimentation analysis of the proposed novel approach IDBMORate on Page-block (Binary) dataset shows that instances of positive class is increased from 559 to 1118 whereas negative instance class remains same as 4913. In case of referred multiclass dataset (Ecoli), IDBMORate produces the consistent result as minority classes (om, omL, imS, imL) instances are oversampled majority class instances remains unchanged. IDBMORate algorithm reduces the ignorance of minority class and oversamples its data without disturbing the size of the majority instance class. Thus, it reduces the overall computation cost and leads towards the improvisation of classification performance.**

Keywords*—Binary imbalance; multiclass imbalance; oversampling; random forest classification; classification

I. Introduction

Numerous ranges of applications in today's real-world deal with imbalanced data applications. Numerous domains specifically medical diagnosis text mining, tracking of financial transactions, telecommunication, and industrial and engineering applications [1,2,3]. Dealing with these applications attracts researchers to resolve the data imbalance challenge. For the rapid development of real-world applications, information management with imbalanced classification is a decisive task. The upcoming needs of this digitized world comprise the utilization of technologies that can handle complex unevenness within the data sample distribution within data. There are a variety of functional application areas which need to reshape unbalanced, complex, and huge volumes of data by incorporating sampling techniques [4, 5, 6].

Data sampling methods are trendy in addressing class inequality at the data level and generally show improvement in classification results. The existing sampling approaches show that there is performance inconsistency if it is applied on both binaries as well as multiclass imbalance data application. The existing imbalanced data applications and work depict that there is an excessive sample generation in the existing oversampling methods which diverse the classification accuracy towards the majority data sample class [7,8]. It also increases the computation cost due to excessive sample generation. Present scenarios also have a diversion in data size of majority data sample in oversampling process and ignorance of minority data sample class. Data sample ignorance in the minority class leads to missing important information and overfitting in the majority class due to excessive data generation in the oversampling process. These challenges motivated this research work to derive a novel oversampling algorithm.

Imbalanced data classification biases performance towards majority numbered class in case of a binary application or majority classes in case of multiclass applications [9]. Traditional approaches lean towards abridged accuracy due to the massive amount of biased data towards the majority [10]. The proposed research work deals with a novel oversampling rate algorithm. In the existing study, the sampling methods which are suitable for the binary imbalance category are not suitable for multiclass imbalance application domains. The proposed IDBMORate algorithm is targeted to calculate oversampling rate which is dynamically applicable to binary as well as multiclass data imbalance and get enhanced classification performance.

In the first attempt, the proposed novel oversampling algorithm deals with the dynamicity of data oversampling which applies to both categories. The second advantage of the proposed algorithm is it will not disturb the majority data instance class and only focus on oversampling the minority data sample class. These two advantages indicate the strengths of the proposed algorithm in terms of less computation time and enhanced classification performance. The main objective of the paper is to identify imbalanced application areas and study existing sampling techniques. The subsequent objective of this research study is to propose a novel oversampling algorithm that leads to performance improvement. Experimental analysis of proposed IDBMORate on selected binary and multiclass datasets shows improved performance metrics.

A. *Organization of the Paper*

The research study in this paper is organized as follows. The next section deals with a brief review of the related literature study of binary and multiclass imbalanced application domains and suitability of classifier. The third section emphasis on existing sampling approaches. Subsequent fourth section deals with the study of proposed Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) algorithm and experimentation. Experiment analysis is carried on both binary (Page-block Dataset) as well as multiclass imbalanced (Ecoli Dataset) for verifying the dynamicity of proposed algorithm. Subsequent section deals with computational results of proposed IDBMORate. The final section outlines the major advantages and dynamicity of the proposed research work.

B. *Research Gap*

Excess time and computation cost required for generating new data samples for balancing the data. Proposed IDBMORate overcomes this research gap by oversampling minority class without disturbing majority data class and improvises classification performance.

II. Imbalanced Application Domains

This section of the paper focuses on imbalanced application domains and the suitability of the classifier for binary and multiclass imbalanced application domains [11,12]. It also highlights the issues raised due to data imbalance [13,14].

A. *Imbalance Application and Suitability of Classifier*

Classification with Imbalanced Dataset (ID) deals with heterogeneous and other imbalances with a massive amount of data.

Fig. 1 depicts the compatibility flow of classifiers depending upon the type of massive and streamed data. It shows that traditional classifiers are best suitable for balanced datasets [15,16] and Random Forest Classifier is best suitable for imbalanced data applications [24].

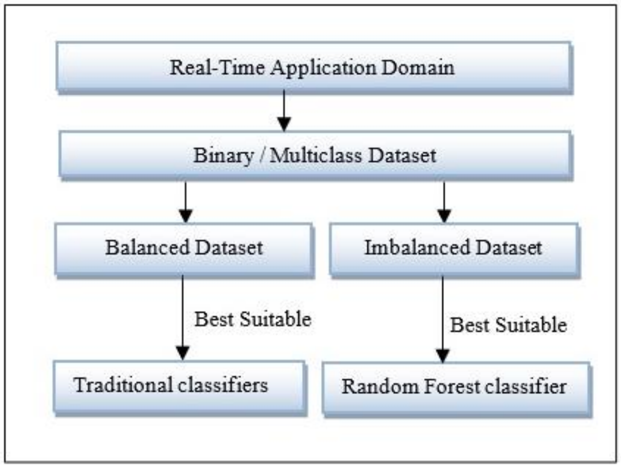

Fig. 1. Data Types and Suitability of Classifier.

B. *Binary and Multiclass Imbalanced Application Domains*

There is a list of numerous numbers of imbalanced applications which belong to class types as either binary imbalance or multiclass imbalance [17]. Table I nominates a list of selected applications with data domain analysis and categorization as binary, multiclass, or of both binary and multiclass imbalance. Binary classification techniques are the most progressive technique to deal with several applications such as medical diagnosis, and fault-finding activities in various business domains which always put forth the statistical results either belonging to one category of data or belonging to a second category [18,19,20].

TABLE I. Imbalance Applications with Domain and category [11-22]

Sample Application		**Imbalance Category**
Diagnosis of cancer-infected patients and patient categorizing	Medical	Binary and Multiclass
Detection of an error occurring in code blocks in software projects	Software development	Binary class
Analyzing the count of faulty machines in industries	Industrial monitoring	Binary class
Multi-dimensional image categorization in various smart city applications	Hyperspectral image processing	Multiclass
Recognition of actions sequences and objects in videos	Mining of video	Binary and multiclass
Analyzing normal and dangerous actions	Action analysis	Binary class
Target specified classification with defined and varied frequency	Targeted classification domain	Multiclass
Analysis of literature relations in text	Mining of text	Binary class
Occurrence of frequent and rare activities in various domains	Activity recognition	Imbalance Multiclass
Recognition of annoyance and sentiment in text	Sentiment analysis	Binary and multiclass
Detection of normal and fraudulent transactions	Finance	Binary class
Categorization of deceptive and ordinary calls	Telecommunication	Binary class

To deal with the classification analysis of these binary and multiclass imbalance data applications, numerous approaches are discussed in the upcoming sections. Data imbalance approaches works at different data level or algorithmic level. At the data level based on the nature of the data, the approaches are categorized [23]. Table I summarizes selected applications, related application domains, and class categories.

III. Related Work

A. *Existing Sampling Approaches*

Sampling techniques are used to balance distorted data distribution Fig. 2 depicts categories of probability and non-probability sampling techniques [24].

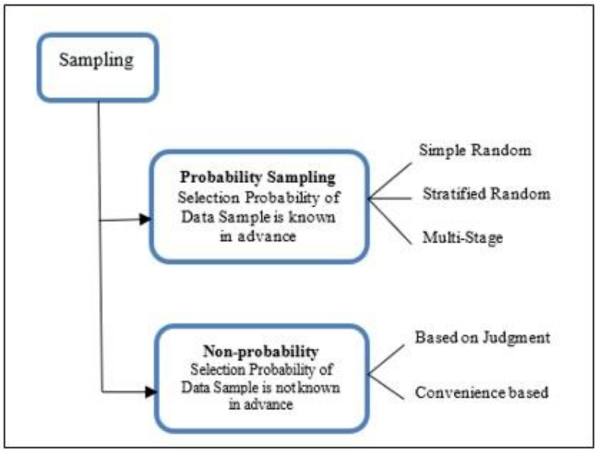

Fig. 2. Probability-based Sampling Strategies.

Both strategies have different sampling approaches to balance the dataset. Table II indicates the simple random sampling techniques steps [22], [23].

TABLE II. SIMPLE RANDOM SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalanced data set.
Steps:2	Distribution of dataset into x number of subsets with equal selection probability.

Table III indicates the stratified random sampling techniques steps [24].

TABLE III. STRATIFIED RANDOM SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalanced Data Set.
Step:2	Dataset distribution into “Strata”.
Step:3	From each stratum select x as any random data sample.
Step:4	Merge the stratum x into the overall data sample.

Sample Case:

Game X has a team of 600 girl participants and 400 boy participant members. For applying a 30-number stratified random sample there is a need to select 12 boy participants from 400 and 18 girl participants from the overall count of 600 participants [25].

Table IV indicates the multistage sampling techniques steps.

TABLE IV. MULTI-STAGE SAMPLING ALGORITHMIC STEPS

Input: Imbalance data of sample size X provided	
Step:1	Take input as an imbalance data Set.
Step:2	Stage I sampling is based on one data attribute as selection criteria for all data samples provided in the data set.
Step:3	Stage II sampling is based on another data attribute as selection criteria for all data samples.

Sample case: Compilation of region-wise voters list based on numerous attributes like city, gender, etc. [26,27].

IV. PROPOSED ALGORITHM AND EXPERIMENTATION

This section of the paper deals with the evolution of the proposed algorithm Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) to balance the imbalance ratio for both the category that is binary as well as multiple classes. The proposed algorithm targets the aim of oversampling minority data sample classes.

TABLE V. PROPOSED ALGORITHM

```
Algorithm: IDBMIORate
         Input: Total # of Classes C, Distribution D Original Imbalanced
         Dataset S
         Output: Oversampling Rate of minority Class
Step 1   Calculate n_min through D
Step 2    n_max = len(D)
Step 3   Assign N= S
Step 4   max = math. Ceil((n_max*(len(c)-1))/n_min)
Step 5   Declare D, i=0, Declare N
Step 6   while i < max
Step 7    Total samples = len(_N)
Step 8    Samples in min Sample Class=min(_D)
          pmin =Calculate current ratio of minimum samples
          if (pmin < (2/(3*lenI-1))) then
          current_min_class = sort_ values_D
Step 9    End if
          Update Values of _D and _N
          S[Class] = = current_min_class

Step 10  ._N = append(samples_in_original_data)
Step 11  _D = sort_values(_N['Class']).
Step 12   Update value of i =i+1
Step 13   Compute IOrate = _D – D
Step 14   data = IOrate Data (IOrate, _N)
Step 15  return IOrate
```

IDBMORate is successfully targeting data rescaling, selection of data, the invention of extra data, and transformation of data. The proposed algorithm deals with the dynamic approach of oversampling rate calculation.

Table V deals with the proposed IDBMIORate algorithm and Table VI deals with the related terminology.

Table VII deals with the data distribution of referred dataset.

TABLE VI. TERMINOLOGY USED FOR PROPOSED ALGORITHM

Key Term	Specifications
S	Original Imbalanced Dataset
C	Total number of Classes
N	Number of Total Data Sample Dataset
D	Data Distribution
n_{Min}	Number of Minority instance data Sample
n_{Max}	Count of Minority instance data Sample
RFC	Random Forest Classification

TABLE VII. DATA DISTRIBUTION SUMMARY

Dataset [28]	Total # of Instances	Imbalanced Category	# Data distribution according to classes
Page-blocks	5472	Binary	Positive 4913 Negative 559
Ecoli	336	Multiclass	cp 143, im 77, pp 52, imU 35, om 20, omL 5, imS 2, imL 2

A. Experimental Analysis of IDBMORate for Binary Datasets

For the Proposed Algorithm IDBMORate the experimentation has been carried out in Python Programming Platform for binary Dataset. Execution on Binary Imbalanced Dataset -1 is set Page-blocks with Random Forest Classification Model Total Data size: 5472.

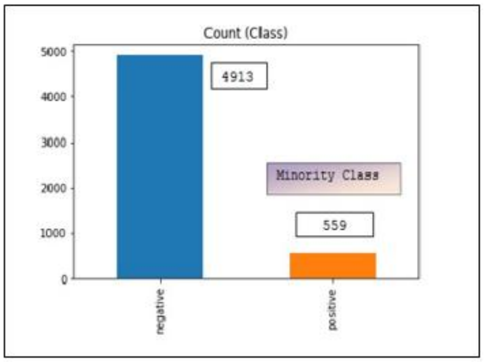

Fig. 3. Result before Sampling – Page-Block Dataset.

Fig. 3 specifies the distribution of class labels before sampling for class negative is 4913 and for class for the positive class are 559. Result after Sampling through IDBMIORate is as depicted in Fig. 4.

Fig. 5 deals with classification result of Page-block dataset with the proposed algorithm

Table VIII deals with the classification performance metrics of the proposed algorithm with RFC for the page block binary dataset.

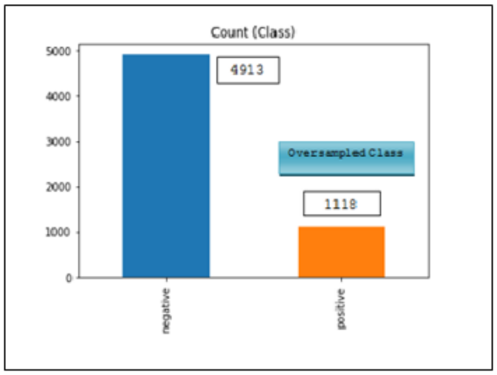

Fig. 4. Result with a Proposed Algorithm – Page-Block Dataset.

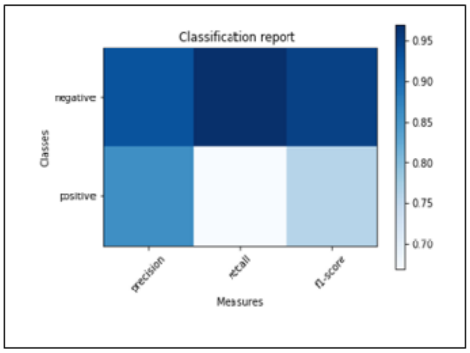

Fig. 5. Page-Block Dataset Classification Performance Graph.

TABLE VIII. PERFORMANCE METRICS FOR PAGE BLOCK DATASET

Performance Parameters	**Precision**	**Recall**	**F1-Score**	**Support**
Class				
negative	0.93	0.97	0.95	1463
positive	0.86	0.67	0.76	347
Avg / Total	0.91	0.92	0.91	1810

B. Experimental Analysis of IDBMORate for Multiclass Datasets

The proposed algorithm also outperforms in the case of multiclass dataset. For performance evaluation of the multiclass dataset, this research study has used Ecoli dataset which contains multiple classes. The total sample size of the Ecoli dataset is 336.

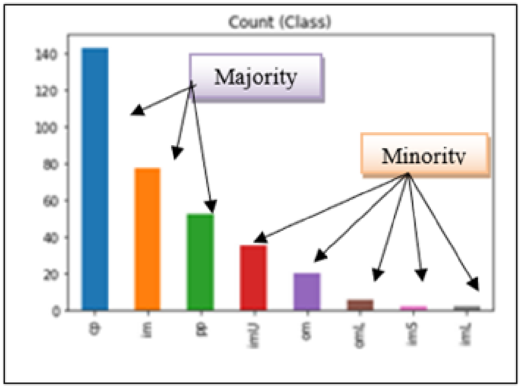

Fig. 6. Result before Sampling – Ecoli Dataset.

Fig. 6 indicates results before Sampling.

Distribution of class labels before Sampling for class cp 143

Distribution of class labels before Sampling for class im 77

Distribution of class labels before Sampling for class pp 52

Distribution of class labels before Sampling for class imU 35

Distribution of class labels before Sampling for class om 20

Distribution of class labels before Sampling for class omL 5

Distribution of class labels before Sampling for class imS 2

Distribution of class labels before Sampling for class imL 2

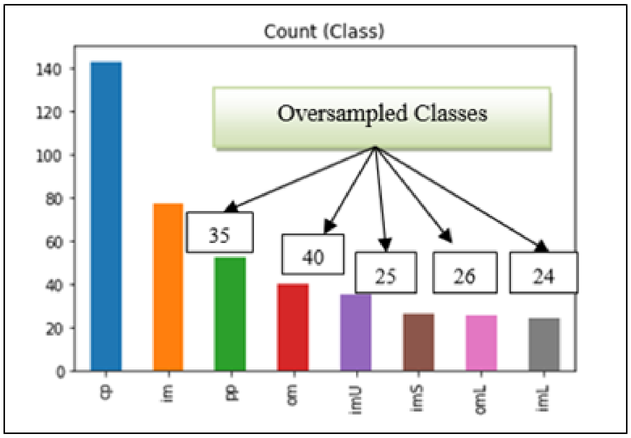

Fig. 7. Result with a Proposed Algorithm – Ecoli Dataset.

Fig. 7 indicates results after Sampling.

Distribution of class labels after Sampling for class cp 143

Distribution of class labels after Sampling for class im 77

Distribution of class labels after Sampling for class pp 52

Distribution of class labels after Sampling for class imU 35

Distribution of class labels after Sampling for class om 40

Distribution of class labels after Sampling for class omL 25

Distribution of class labels before Sampling for class imS 26

Distribution of class labels before Sampling for class imL 24

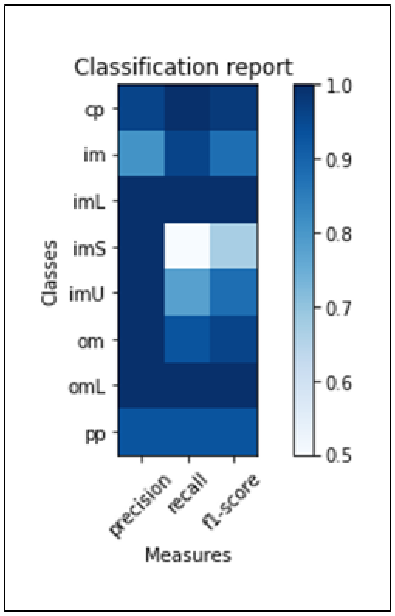

Fig. 8. Ecoli Dataset Classification Performance Graph.

Fig. 8 depicts the multiclass classification result with the proposed novel sampling approach. Precision, Recall, F1 Score, and support parameters are used for measuring the classification performance for both page block (binary) and Ecoli (Multiclass) dataset.

TABLE IX. PERFORMANCE METRICS WITH A PROPOSED ALGORITHM FOR MULTICLASS-ECOLI DATASET

Performance Parameters	**Precision**	**Recall**	**F1- Score**	**Support**
Class				
cp	0.96	1.00	0.98	44
im	0.81	0.96	0.88	26
imL	1.00	1.00	1.00	4
imS	1.00	0.50	0.67	8
imU	1.00	0.78	0.88	9
om	1.00	0.93	0.96	14
omL	1.00	1.00	1.00	7

Table IX shows the RFC classification result with the proposed oversampling rate algorithm to compute the effectiveness of the proposed algorithm.

V. CONCLUSION AND FUTURE WORK

This research work addressed binary and multiclass imbalanced application domains, associated problems, and approaches to dissolve data imbalance dynamically. The proposed algorithm Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate (IDBMORate) balances the minority classes without affecting the majority class which minimizes the cost of computation. Experimentation analysis on dataset page block and Ecoli has been carried out. IDBMORate algorithm overcomes the problem of the generation of extreme synthetic data samples for the minority classes, which leads to improved classification accuracy with the Random Forest Classification Model. Experimental analysis shows that IDBMORate efficiently outperforms the existing oversampling techniques for both binary as well as Multiclass imbalanced real-life scenarios. (IDBMORate) balances the minority classes without affecting the majority class which minimizes the cost of computation. The Proposed algorithms Improvised Dynamic Binary-Multiclass Imbalanced Oversampling Rate proposed algorithm which shows improvised results for both binary as well as multiclass DATA. THE hybrid sampling method will be focused in the future to upgrade the performance. The more dynamic method can be focused to work in a distributed environment.

REFERENCES

[1] Ahmed M. Sayed,"Machine Learning Augmented Breast Tumors Classification using Magnetic Resonance Imaging Histograms" International Journal of Advanced Computer Science and Applications, vol.12., no.12, pp.1-9, 2021.

[2] Prakruthi M K, Komarasamy G, "Novel Framework for Enhanced Learning-based Classification of Lesion in Diabetic Retinopathy", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.37-44, 2022.

[3] Angelo, P., Resende, A.and Drummond, A. C. "A Survey of Random Forest Based Methods for Intrusion Detection Systems", ACM Comput. Surv. 51(3), 48-48.36,2018.

[4] Kamlesh Upadhyay, Prabhjot Kaur, Ritu Sachdeva," Fast and Robust Fuzzy-based Hybrid Data-level Method to Handle Class Imbalance", International Journal of Advanced Computer Science and Applications vol.13., no.6, pp.65-74, 2022.

[5] Delplace, A., Hermoso, S., and Anandita, K. ,"Cyber Attack Detection thanks to Machine Learning Algorithms", COMS7507: Advanced Security, 1-46, 2019.

[6] Jyoti Islam, Yanqing Zhang, "Brain MRI Analysis for Alzheimer's Disease Diagnosis Using an Ensemble System of Deep Convolutional Neural Networks", International Journal of Springer, 2018.

[7] Elyan, E., Francisco, C., Garcia, M. and Jayne, C, CDSMOTE: Class Decomposition and Synthetic Minority Class Oversampling Technique for Imbalanced Data Classification, Neural Computing & Applications, 33, 2839–2851,2020.

[8] Hamad, R. A., Kimura, M., and Lundström, J. Efficacy of Imbalanced Data Handling Methods on Deep Learning for Smart Homes Environments. SN COMPUT. SCI. 1, 204,2020.

[9] SMahmoud B. Rokaya, "Shallow Net for COVID-19 Classification based on Biomarkers", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.97-103, 2022.

[10] Mohd Hakim Abdul Hamid, Marina Yusoff, Azlinah Mohamed "Survey on Highly Imbalanced Multi-class Data" International Journal of Advanced Computer Science and Applications, vol.13., no.16, pp.211-229, 2022.

[11] Evangeline D, Amy S Vadakkan, Sachin R S, Aakifha Khateeb, Bhaskar C5 "Cyberbullying Detection in Textual Modality" International Journal of Advanced Computer Science and Applications, vol.12., no.12, pp.217-229, 2021.

[12] Pramod Sunagar, Anita Kanavalli,"A Hybrid RNN based Deep Learning Approach for Text Classification", International Journal of Advanced Computer Science and Applications, vol.13., no.6, pp.289-295, 2022

[13] Ainul Yaqin, Majid Rahardi, Ferrian Fauzi Abdullah, "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University," International Journal of Advanced Computer Science and Applications, vol.13. ,no.6, pp. 418-422, 2022.

[14] More, A. S., Rana, D. P., and Agarwal, IRandom Forest Classifier Approach for Imbalanced Big Data Classification for Smart City Application Domains. International Journal of Computational Intelligence & IoT, 1(2), 261-266,2018.

[15] Cao, L., and Shen, H. , Imbalanced Data Classification Using ImprovedClustering Algorithm and Under-sampling Method, In Proceedings of 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, pp.361-366, 2019.

[16] Chee Keong Chan., Alexander W., "Development of a platform to explore network intrusion detection system for cyber security." Journal of Computer and Communications, Vol. 6, pp.1-11,2018.

[17] E.Abrahim., A.Saleem, T. Dao, Zhaoheng Liu, "Multiple-objective optimization and design of series-parallel systems using novel hybrid genetic algorithm meta-heuristic approach," World Journal of Engineering and Technology, Vol. 6, No.1 pp. 532-555,2018.

[18] Holewik, J., Schaefer, G., Korovin, I., "Imbalanced ensemble learning for enhanced pulsar identification," Proceedings of International Conference , pp.515-524, 2020.

[19] Jegierski H., Saganowski, S, "An outside the box'' solution for imbalanced data classification," IEEE Access, Vol. 8, pp. 125191-125209,2020.

[20] Khaja Mohammad Shahzad., Jong Sou Park, "Optimization of intrusion detection through fast hybrid feature selection," Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, pp.1-5, 2019.

[21] Kim, J., Jeong, J., and Shin, J, "M2m: Imbalanced classification via major-to-minor translation," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", pp.13893-13902, 2020.

[22] Q. Wang, "Imbalanced Classification Based on Over-sampling and Feature Selection," IEEE 5th International Conference on Cloud Computing and Big Data Analytics), pp. 325-330, 2020.

[23] Du, G.; Zhang, J.; Li, S.; Li, C. Learning from class-imbalance and heterogeneous data for 30-day hospital readmission. Neurocomputing 420, 27–35, 2021.

[24] Anjali S. More and Dipti P. Rana," Performance enrichment through parameter tuning of random forest classification for imbalanced data applications", Materials Today: Proceedings, Vol. 56, No. 6, pp. 3585-3593, 2022.

[25] Srinivasan, R.; Subalalitha, C. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distrib. Parallel Databases, Vol. 39, pp.1–16, 2021.

[26] H.X. Guo, Y.J. Li, J. Shang, M.Y. Gu, and Y.Y. Huang, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, 2017.

[27] M. Bach, A.Werner, J. Zywiec, and W. Pluskiewicz, "The study of under- and oversampling methods' utility in the analysis of highly imbalanced data on osteoporosis," Inf. Sci., vol. 384, pp. 174–190, 2017.

[28] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.