International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *01 October 2023* ***Accepted:*** *20 October 2023*

A Hybrid Resampling Approach for Multiclass Skewed Datasets and Experimental Analysis with Diverse Classifier Models

Rose Mary Mathew[1], Dr. R. Gunasundari[2]
[1]Research Scholar, Department of Computer Science,
Karpagam Academy of Higher Education ,
Coimbatore, India
rosem.mathew@gmail.com
https://orcid.org/0000-0003-0555-4873
[2]Professor, Department of Computer Applications,
Karpagam Academy of Higher Education,
Coimbatore, India
gunasoundar04@gmail.com
https://orcid.org/0000-0003-4157-285X

Abstract— In real-life scenarios, imbalanced datasets pose a prevalent challenge for classification tasks, where certain classes are heavily underrepresented compared to others. To combat this issue, this article introduces DOSAKU, a novel hybrid resampling technique that combines the strengths of DOSMOTE and AKCUS algorithms. By integrating both oversampling and undersampling methods, DOSAKU significantly reduces the imbalance ratio of datasets, enhancing the performance of classifiers. The proposed approach is evaluated on multiple models employing different classifiers, and the results demonstrate its superiority over existing resampling measures, making it an effective solution for handling class imbalance challenges. DOSAKU's promising performance is a substantial contribution to the field of imbalanced data classification, as it offers a robust and innovative solution for improving predictive model accuracy and fairness in real-world applications where imbalanced datasets are common.

Keywords- hybrid, multiclass, oversampling, skewed, undersampling

I. INTRODUCTION

Multiclass imbalanced data classification plays a critical role in various domains, including healthcare, finance, and customer analytics [1]. The prevalence of imbalanced datasets in real-world scenarios poses challenges for developing accurate learning models. Unique complexities arise when dealing with multiclass imbalanced data, with the main objective being the development of models that effectively classify all classes, including the minority ones. Techniques for multiclass imbalanced data classification are crucial in addressing fairness and bias concerns. Existing methods primarily focus on balancing the representation of minority classes, thereby enhancing the performance of classification models [2].

Balancing data in multiclass classification is essential to promote model generalization and reduce bias towards majority classes. As new applications emerge in domains like cybersecurity, social media analysis, and healthcare, the significance of multiclass imbalanced data classification remains evident [3]. Addressing imbalances through appropriate techniques ensures the development of unbiased models and contributes to fairness, transparency, and accountability. By achieving a balanced dataset, these methods empower models to make reliable and unbiased predictions for all classes. The ultimate objective is to construct fair, robust models capable of providing accurate predictions across all classes, thereby promoting fairness, trustworthiness, and improved decision-making while adhering to ethical principles.

Hybrid resampling is an approach that combines various techniques, such as oversampling and undersampling, to create a balanced dataset that captures the characteristics of all classes [4]. This approach may involve applying oversampling first and then undersampling or vice versa, effectively leveraging both techniques to achieve a balanced dataset that encompasses the essential aspects of both majority and minority classes.

This article introduces a novel hybrid resampling approach called DOSAKU, designed to address the issue of class imbalance in multiclass skewed datasets. DOSAKU combines the strengths of the DOSMOTE algorithm [5] and AKCUS algorithm [6] to achieve a balanced dataset by minimizing the imbalance ratio. To assess the effectiveness of the proposed algorithm, an experimental study is conducted using benchmark datasets. The structure of the article is as follows: Section 2 discusses related works, Section 3 describes the proposed algorithm, Section 4 presents the experimental study and the results are discussed in Section 5. Finally, Section 6 concludes the article by summarizing the major findings and suggesting future directions.

II. RELATED WORKS

Researchers have extensively explored various resampling methods to address imbalanced data in classification tasks. While oversampling and undersampling techniques have been extensively studied, the attention given to hybrid approaches that combine these methods with classifiers has been limited. Most of these hybrid techniques have been focused on binary class imbalances, leaving only a few studies dedicated to tackling challenges in multiclass scenarios.

One such hybrid approach is RHSBoost, proposed by Gong, which combines random undersampling and ROSE (Random Over-Sampling Examples) sampling within a boosting framework [7]. Experimental results demonstrate RHSBoost's promise in handling imbalanced datasets. To address both class imbalance and overlapping in multiclass problems, R. Alejo introduced MBP + GGE, a hybrid approach that combines a modified back-propagation technique with Gabriel graph editing [8]. The method shows improved performance through a new cost function based on mean squared error (MSE).

Another novel hybrid algorithm, EUSBoost, was devised by Galar, building upon the effectiveness of RUSBoost by incorporating random undersampling with a Boosting algorithm [9]. Zhou et al. presented SMOTE-CUT (SMOTE-Clustered Undersampling Technique), a method that combines oversampling, clustering, and undersampling [10]. This approach uses SMOTE for oversampling, performs clustering on the original and SMOTE-generated samples, and then removes majority class samples from the clusters. In the context of boosting, SMOTEBoost, proposed by Zhang et al., integrates the SMOTE algorithm to enhance the prediction of the minority class [11]. To address noise in the minority class, Zhang et al. introduced Borderline-SMOTE-NC (Borderline-SMOTE with Noise Correction), a variation of Borderline SMOTE [12]. M. Mukherjee et al. proposed SMOTENC (SMOTE with Noise Filtering), which is a variation of SMOTE that filters out noisy instances from the minority class before generating synthetic instances [13].

Researchers employ hybrid approaches to address imbalanced data classification in binary and multiclass scenarios. These methods combine resampling techniques with classifiers or boosting algorithms to enhance performance. Hybrid ensemble approaches create classifier ensembles trained on diverse resampled datasets, leveraging various techniques to achieve improved predictive capabilities and robustness. By exploiting ensemble diversity, these methods handle class imbalances and capture complex data relationships. They generalize better, avoid overfitting, and provide accurate predictions for different classes. These are valuable tools in real-world applications, and these approaches offer promising directions for addressing imbalanced data challenges.

III. PROPOSED WORK

This article introduces a novel hybrid resampling procedure, DOSAKU (Darwinian Optimized SMOTE with Adaptive K-Means Cluster Undersampling), designed to effectively handle imbalanced multiclass datasets with varying degrees of skewness. The method combines the strengths of DOSMOTE for oversampling the minority classes and AKCUS for undersampling the majority classes. The proposed approach aims to address the class imbalance problem in multiclass skewness scenarios, providing a robust solution for improving the performance of classification models. DOSAKU offers a promising direction for handling the challenges posed by imbalanced data in multiclass settings, enhancing the accuracy and fairness of predictive models across all classes.

TABLE 1. ALGORITHM FOR HYBRID RESAMPLING

```
Algorithm
Input: Imbalanced dataset
Output: Balanced Dataset

1. Read the multiclass imbalanced dataset and identify the frequency of data available in each class.
2. Store the highest and lowest frequency in max variable and min variable respectively.
3. Assign variable $difference = max - min$
4. Set variable $limit = difference/2$.
5. Set variable $threshold = min + limit$
6. For every class do the following

    6.1. Check whether the class frequency is less than threshold then do step 6.2 otherwise go to step 6.3.
    6.2. Call DOSMOTE () for the class and generate synthetic samples to get a maximum class frequency up to threshold value.
```

```
6.3. Call AKCUS () for the class to remove the samples from the class to get the class frequency same as of threshold value.

7. Combine the data in all the classes to get a balanced multiclass dataset for creating models.
8. Stop
```

DOSAKU stands as a versatile and effective solution for handling various types of skewness in multiclass datasets by integrating the strengths of both DOSMOTE and AKCUS techniques. Through selective application of oversampling to minority classes and undersampling to majority classes, the procedure aims to achieve a well-balanced dataset, ensuring a fair distribution of instances across all classes. This hybrid strategy is particularly valuable when faced with scenarios lacking clear distinctions between multimajority and multiminority classes, necessitating a more nuanced balancing technique. The ultimate objective is to strike the optimal balance between oversampling and undersampling to enhance classifier performance and improve prediction accuracy on imbalanced datasets. A detailed algorithmic description is provided in TABLE 1.

The proposed method commences by taking the original imbalanced dataset as input and conducts an analysis to determine the number of classes and instances within each class. The highest and lowest frequencies are stored in the *max* and *min* variables, respectively. The difference between the *max* and *min* values is calculated and stored in the *difference* variable. Additionally, the *limit* variable is set to half of the value in the *difference* variable. A fixed frequency for the classes is determined by setting the *threshold* variable as the sum of the *min* and *limit* variables. Proceeding class by class, each class's frequency is compared with the *threshold* value. If the class frequency is less than the *threshold* data, the DOSMOTE method is called; otherwise, the AKCUS method is invoked.

The class employing the DOSMOTE method generates synthetic samples using a combination of SMOTE (Synthetic Minority Over-sampling Technique) and Darwinian Particle Swarm Optimization [14]. To create synthetic samples, the method identifies the nearest neighbours of existing instances within this class and generates artificial samples using a formula involving the original sample and a random value.

$$X_{syn} = x_i + rand\ (0,1) * |x_i - x_{near}|$$

Where x_i *-> Minority sample*
x_{near} *-> Nearest neighbor of* x_i *in the minority data*

The generated samples are integrated into the DPSO (Darwinian Particle Swarm Optimization) structure, which utilizes an evolutionary process to globally optimize the particles (synthetic samples) by considering information from neighboring data in other classes. Each class is treated as a separate swarm, and for each swarm, the fitness values of the newly generated samples are evaluated to identify the best data. Subsequently, the position and velocity of the particle (synthetic sample) are updated accordingly, and any obsolete particles or samples are discarded during this iterative process. The DPSO algorithm continues until the specified number of iterations is reached, and only the best particles are selected and added to the actual data. This iterative optimization process helps achieve the desired class frequency up to the defined threshold. Ultimately, the method produces optimized synthetic samples that effectively contribute to balancing the class distribution in the imbalanced dataset.

After generating synthetic samples, the subsequent class data is analyzed to determine whether to call the AKCUS (Adaptive K-Means Clustering Undersampling) method, and if so, the AKCUS algorithm is executed for that class separately [6]. In this method, K elements from the class data are selected as seed elements to define the properties of corresponding clusters. Euclidean distance is utilized to compute distances, including the distance between each element and a cluster, the distance between two elements, and the distance between two clusters.

The distances between clusters are computed and stored in an array as a triangular matrix. The shortest distance between clusters (D_{mini}) and the nearest clusters associated with that shortest distance are identified. For non-clustered elements distances from each cluster are computed, and the element is assigned to that cluster which is having the shortest distance. The algorithm then proceeds with other non-clustered elements. Instances assigned to the closest cluster are added to the cluster, if their distance from the cluster centroid is smaller than D_{mini}. The centroid value is updated as the mean of all instances in the cluster, and distances from the updated cluster to other clusters, the shortest distance between clusters, and the nearest clusters are identified. If the value of D_{mini} is lower than the distance of an instance from the closest cluster, the two closest clusters (C1 and C2) are selected for merging. Cluster C2 is eliminated by removing all instances from the cluster and deleting its representation. New instances are added to the now empty cluster, creating a new cluster. The distances among all clusters are recomputed, and the nearest two clusters are identified.

The AKCUS method effectively employs the Adaptive K-means algorithm for cluster center calculation and performs clustering, instance assignment, centroid computation, cluster merging, and cluster creation based on distance measurements and element properties [15]. After completing the clustering and allocating centroid points for all data values, undersampling is applied to the class. Data points with a greater distance from the centroid in each cluster are removed in a fixed proportion until the class frequency matches the threshold value.

Once the threshold value is achieved, the class data is merged with the previously evaluated data. The class analysis continues for other classes, ensuring that all classes in the dataset undergo either the DOSMOTE method or the AKCUS method. This results in a balanced dataset where each class has a frequency equal to the threshold, facilitating better model training and ensuring equal representation of all classes in the dataset.

IV. EXPERIMENTAL STUDY

An empirical investigation was conducted to evaluate the effectiveness of the proposed hybrid resampling technique, known as DOSAKU (Darwinian Optimized SMOTE with Adaptive K-Means Undersampling). Multiclass skewed datasets were obtained from various data repositories, and data pre-processing was performed, which involved handling categorical values, missing values, and noise data. Subsequently, the dataset was split into an 80:20 ratio for training and testing purposes. The training set underwent the DOSAKU hybrid resampling technique to mitigate the imbalance ratio, resulting in a transformed training set with reduced imbalance. Models were then constructed using different classifiers, such as K-NN, SVM, and Random Forest, based on the transformed training data. The performance of each model was evaluated using the test data, and performance metrics were calculated for further analysis. Figure 1 illustrates the various stages in the experimental study, which are explained below in different steps.

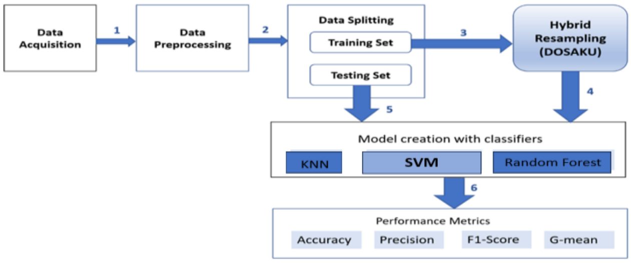

Figure 1. Describing the various stages in the experimental study.

Step 1. Getting the data

The datasets utilized in this study were sourced from the KEEL data repository [16][17]. Two datasets were selected, each exhibiting varying degrees of class imbalance. Specifically, the Thyroid and Glass datasets were chosen for this experiment. A comprehensive description of the characteristics and structure of these datasets is presented in the following TABLE 2.

TABLE 2. DETAILED DESCRIPTION OF THE DATASET

Name of the dataset	**No. of Instances**	**Imbalance Ratio**	**No. of Attributes**	**Class Label**	**No. of records**
Glass	214	8.44	9	1	70
				2	76
				3	17
				5	13
				6	9
				7	29
Thyroid	720	40.16	22	1	17
				2	33
				3	666

Step 2. Data Preprocessing

This data preparation stage involves transforming the raw data for analysis. Processing eliminates noise, outliers, duplicate records, and handles missing values. Data integration techniques ensure consistency when merging from multiple sources. Transformation techniques, such as converting categorical data to numerical format, normalizing, and scaling it, are applied. Feature selection methods identify the most relevant features for model training and prediction. This comprehensive data preparation phase establishes a reliable foundation for subsequent analysis, guaranteeing accuracy and efficiency in model development and evaluation.

Step 3. Model Training

Following the data pre-processing stage, a clean and transformed dataset is obtained, poised for further analysis. This dataset is divided into two portions, with 80% allocated for model training and the remaining 20% for testing. The study centers on the training dataset, revealing the initial imbalance ratio across different classes. It is vital to acknowledge that the selected datasets inherently exhibit class imbalance. To mitigate this imbalance, the proposed hybrid resampling technique, DOSAKU, is applied to the training data, leading to a reduction in the imbalance ratio and the creation of a more balanced dataset. This balanced data is then employed for model creation. Several classifiers, including K-Nearest Neighbor, Support Vector Machine (SVM), and Random Forest (RF), are chosen for training on the resampled data, generating diverse models for subsequent classification tasks. These models are trained on the resampled data and prepared for the subsequent testing phase. The evaluation of the models performance on the test data allows for a comprehensive assessment of the efficacy of the DOSAKU resampling technique in addressing class imbalance challenges in multiclass datasets.

Step 4. Model Testing

The testing dataset comprises approximately 20% of the original dataset, carefully selected to include instances from different classes for a comprehensive assessment of model performance. The effectiveness of the models is evaluated during the testing phase, utilizing confusion matrices to compare actual and predicted data. These matrices provide insights into True-Positive (TP), False-Positive (FP), True-Negative (TN), and False-Negative (FN) values. Performance evaluation metrics, including Accuracy, Error, Precision, F1-Score, and G-mean, are then calculated using these values to gauge the models' effectiveness in classifying imbalanced multiclass datasets.

V. RESULTS AND DISCUSSION

This research primarily centers on multiclass imbalanced datasets and aims to assess the efficacy of a hybrid resampling technique. The evaluation encompasses the classification performance of the original highly imbalanced dataset, employing a variety of classifiers. Furthermore, it analyses the influence of applying the suggested resampling method to the same dataset with different classifiers. The results are obtained through rigorous dataset testing, computing various performance metrics, including accuracy, error rate, precision, F-Score, and G-mean, for the diverse datasets.

The glass dataset, a seven-class dataset comprising 9 features and a class label, serves as the subject of this study. The proposed hybrid resampling technique is applied to the training subset of the dataset. To explore the technique's effectiveness, four models are developed: one trained on the original dataset and the other three trained on the resampled data using SMOTE, ENN, and DOSAKU, respectively. The model creation process involves the utilization of three distinct classifiers: K-NN, SVM, and Random Forest. To address the dataset's imbalance, the results are presented with a focus on the G-mean metrics over accuracy. Notably, across all three classifiers, the models trained on the DOSAKU resampled data consistently demonstrate higher G-mean results when compared to the models trained on the original dataset. TABLE 3 and Figure 2 showcase the comprehensive results, encompassing various metrics for the different models trained with their respective resampled datasets.

TABLE 3: PERFORMANCE METRICS OF THE GLASS DATASET

Dataset: Glass					
Classifiers	**Metrics**	**Original dataset**	**SMOTE**	**ENN**	**DOSAKU**
K-NN	Accuracy	0.6046	0.6744	0.5581	0.5412
	Error Rate	0.3954	0.3256	0.4419	0.4588
	Precision	0.6726	0.7743	0.6684	0.5357
	F-Score	0.5771	0.7645	0.5579	0.5495
	G-Mean	0.6133	0.7742	0.6018	**0.7811**
SVM	Accuracy	0.6279	0.6744	0.6046	0.6471
	Error Rate	0.3721	0.3256	0.3954	0.3529
	Precision	0.6781	0.7673	0.5948	0.4894
	F-Score	0.5874	0.7245	0.5445	0.4607
	G-Mean	0.6207	0.7454	0.5669	**0.7731**
Random Forest	Accuracy	0.6279	0.6976	0.5348	0.5882
	Error Rate	0.3721	0.3024	0.4652	0.4118
	Precision	0.7373	0.7912	0.6303	0.6010
	F-Score	0.7161	0.7929	0.5777	0.6257
	G-Mean	0.7346	0.7993	0.6090	**0.8205**

The Thyroid dataset represents a three-class dataset comprising 21 features and a class label. Within this study, the proposed hybrid resampling technique is applied to the training subset of the dataset. To evaluate the performance of proposed algorithm, four models are developed: one trained on the original dataset, and the other three trained on the resampled data using SMOTE, ENN, and DOSAKU methods. The model creation process involves the application of three distinct classifiers: K-NN, SVM, and Random Forest. Given the dataset's inherent imbalance, the evaluation prioritizes the G-mean and F-Score metrics over accuracy. Across all three classifiers, the models trained on the DOSAKU resampled data consistently exhibit superior F-Score and G-mean results when compared to the models trained on the original dataset. The results, presented in TABLE 4, offer an in-depth view of the various metrics for the different models trained with their respective resampled datasets. Furthermore, Figure 3 visually displays the results of these metrics, providing a comprehensive overview of the findings.

Article Received: *15 August 2023* ***Revised:*** *01 October 2023* ***Accepted:*** *20 October 2023*

TABLE 4: PERFORMANCE METRICS OF THE THYROID DATASET

Dataset: Thyroid					
Classifiers	**Metrics**	**Original dataset**	**SMOTE**	**ENN**	**DOSAKU**
K-NN	Accuracy	0.9954	0.84	0.9111	0.9954
	Error Rate	0.0046	0.16	0.0889	0.0046
	Precision	0.949	0.5229	0.4198	0.9983
	F-Score	0.9204	0.5108	0.3972	0.9735
	G-Mean	0.952	0.530	0.4019	**0.9673**
SVM	Accuracy	0.9861	0.909	0.9277	0.9907
	Error Rate	0.0139	0.091	0.0723	0.0093
	Precision	0.949	0.6229	0.6423	0.9507
	F-Score	0.9204	0.643	0.4873	0.937
	G-Mean	0.9523	0.6507	0.5342	**0.9673**
Random Forest	Accuracy	0.9861	0.9722	0.9333	0.9861
	Error Rate	0.0139	0.0278	0.0667	0.0139
	Precision	0.915	0.8222	0.8957	0.915
	F-Score	0.8945	0.8628	0.5882	0.9022
	G-Mean	0.9071	0.8724	0.6922	**0.9371**

The proposed hybrid resampling algorithm, DOSAKU, has exhibited promising outcomes in enhancing the G-mean metric for both the Glass dataset and the Thyroid dataset, utilizing various classifiers, namely K-NN, SVM, and Random Forest. Notably, in the case of the Glass dataset, the models trained using DOSAKU demonstrated significantly higher G-mean scores in comparison to those trained on the original imbalanced dataset. This noteworthy improvement signifies that DOSAKU effectively addressed the class imbalance issue, resulting in more balanced and accurate predictions for all classes. Likewise, the application of DOSAKU to the Thyroid dataset showcased its capability in elevating the G-mean metric for all three classifiers. The resampling technique successfully mitigated the detrimental effects of class imbalance, leading to models with overall improved performance.

The consistent and remarkable enhancement in G-mean scores across both datasets and all classifiers underscores the effectiveness of the proposed DOSAKU algorithm. By creating a more balanced training dataset, DOSAKU equips the classifiers with a better understanding of the minority classes, ultimately leading to more robust and reliable predictions. These findings establish DOSAKU as a valuable tool in addressing class imbalance challenges and enhancing the performance of classifiers in real-world applications.

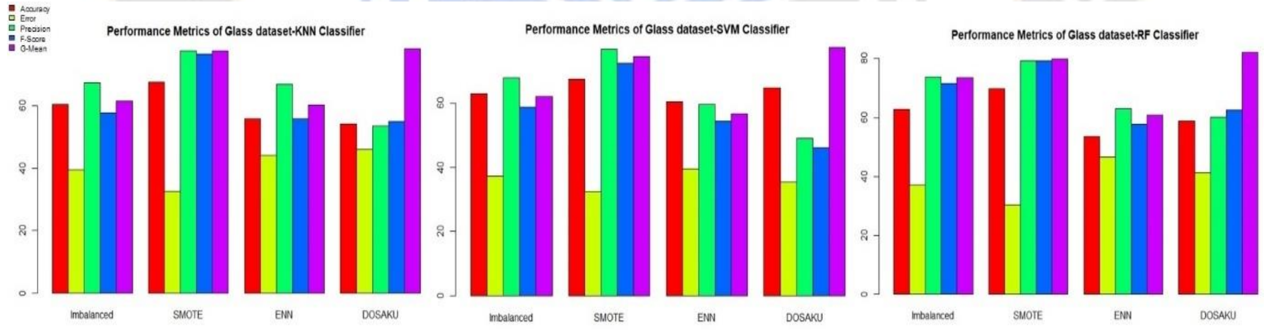

Figure 2. Performance results of various models with different classifiers and resampling techniques using Glass dataset

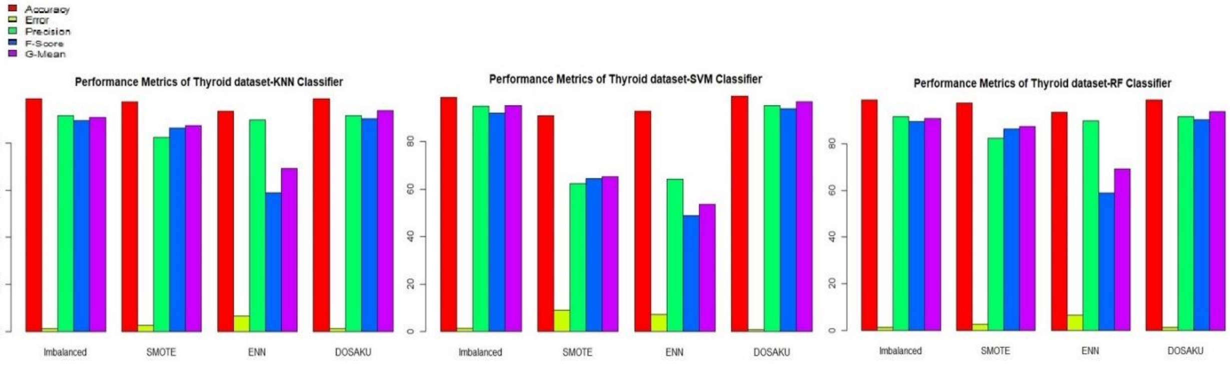

Figure 3. Performance results of various models with different classifiers and resampling techniques using Thyroid dataset

VI. CONCLUSION

The remarkable performance of DOSAKU, especially in terms of the G-mean metric, underscores its effectiveness in addressing class imbalance in multiclass datasets. These findings carry significant relevance in real-world applications, where accurate predictions for all classes are vital for informed decision-making and effective problem-solving. Overall, the positive outcomes of this study strongly advocate for the adoption of DOSAKU as a valuable tool in handling imbalanced data across various domains. DOSAKU's ability to enhance not only the G-mean metric but also other performance metrics opens avenues for constructing fair, reliable, and precise models. As such, it presents a promising choice for both practitioners and researchers dealing with multiclass skewed datasets. However, to bolster its validity and generalizability, it is advisable to conduct further investigations and experiments on larger and more diverse datasets. In conclusion, the study demonstrates DOSAKU's potential as an effective solution for class imbalance, paving the way for more robust and trustworthy models in practical applications. The positive impact it brings to various performance metrics positions it as an asset in the toolkit of data scientists and researchers working in the realm of multiclass imbalanced datasets.

References

[1] R. Singh and R. Raut, “Review on Class Imbalance Learning: Binary and Multiclass,” *Int. J. Comput. Appl.*, vol. 131, no. 16, pp. 4–8, 2015, doi: 10.5120/ijca2015907573.

[2] H. Patel and G. S. Thakur, “Classification of imbalanced data using a modified fuzzy-neighbor weighted approach,” *Int. J. Intell. Eng. Syst.*, vol. 10, no. 1, pp. 56–64, 2017, doi: 10.22266/ijies2017.0228.07.

[3] H. Kaur, H. S. Pannu, and A. K. Malhi, “A systematic review on imbalanced data challenges in machine learning: Applications and solutions,” *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, Aug. 01, 2019, doi: 10.1145/3343440.

[4] Y. Pristyanto, N. A. Setiawan, and I. Ardiyanto, “Hybrid resampling to handle imbalanced class on classification of student performance in classroom,” *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, pp. 207–212, 2017, doi: 10.1109/ICICOS.2017.8276363.

[5] R. M. Mathew and R. Gunasundari, “An Oversampling Mechanism for Multimajority Datasets using SMOTE and Darwinian Particle Swarm Optimisation,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 2, pp. 143–153, 2023, doi: 10.17762/ijritcc.v11i2.6139.

[6] R. M. Mathew, “A Cluster-based Undersampling Technique for Multiclass Skewed Datasets,” *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 10785–10790, 2023.

[7] H. K. Joonho Gong, “RHSBoost: Improving classification performance in imbalance data,” *Comput. Stat. Data Anal.*, vol. 111, pp. 1–13, 2017.

[8] V. A. Alejo R. et.al, “A hybrid approach for max-sat: MBP + GGE,” *J. Artif. Intell. Res.*, vol. 49, pp. 619–652, 2014.

[9] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, “EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling,” *Pattern Recognit.*, vol. 46, no. 12, pp. 3460–3471, 2013, doi: 10.1016/j.patcog.2013.05.006.

[10] C. Zhou, Z., Zhang, J., Wu, G., and Zhang, “SMOTE-CUT: A Hybrid Approach to Imbalanced Data Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1467–1481, 2013.

[11] K. W. Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, “SMOTEBoost: Improving prediction of the minority class in boosting,” in *European Conference on Principles of Data Mining and Knowledge Discovery, Cavtat/Dubrovnik, Croatia*, 2003, pp. 107–119.

[12] J. Han, H., Wang, W., and Tang, “Borderline-SMOTE-NC: A hybrid approach for imbalanced classification,” *Pattern Recognit.*, vol. 48, no. 10, pp. 3392-3408., 2015.

[13] M. Mukherjee, M.; Khushi, “SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features.,” *Appl. Syst. Innov.*, vol. 4, no. 18, 2021.

[14] J. Tillett, T. M. Rao, F. Sahin, and R. Rao, “Darwinian particle swarm optimization,” *Proc. 2nd Indian Int. Conf. Artif. Intell. IICAI 2005*, pp. 1474–1487, 2005.

[15] S. K. Bhatia, “Adaptive K-Means Clustering . Adaptive K-Means Clustering,” *Am. Assoc. Artif. Intell.*, no. June, 2014, doi: 10.13140/2.1.4197.9845.

[16] J. A.-F. et Al, “KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *J. Mult. Log. Soft Comput*, vol. 17, no. 2–3, pp. 255–287, 2011.

[17] I. Triguero *et al.*, “KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining,” *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, p. 1238, 2017, doi: 10.2991/ijcis.10.1.82.

Author Biography

Mrs. Rose Mary Mathew is a Research Scholar in the Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore. Currently she is working as Assistant Professor (Special Grade) in the Department of Computer Applications, Federal Institute of Science and Technology, Angamaly. She has more than ten years of teaching experience. She obtained her Master of Computer Applications degree from Mahatma Gandhi University, Kottayam in 2009 and MBA from Bharathiyar University, Coimbatore in 2018. Her area of specialization is Machine Learning and Artificial Intelligence. She had attended several National and International seminars and conferences and published articles in several international journals.

Dr.R. Gunasundari is presently working as Professor in the Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore. She has more than fifteen years of teaching experience. She has participated and presented several papers in National and International conferences. Her research interests include Data Mining, Machine Learning, Cryptography and Network Security.