SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I1P102

Volume 13 Issue 1, 18-26, January 2026

Original Article

Hybrid Sampling Approach for Multiclass Imbalanced Data

Madhura Prabha[1], Sasikala[2]

[1,2]*Department of Computer Science, University of Madras, Tamil Nadu, India.*

[1]*Corresponding Author : madhura.prabha@gmail.com*

Received: 20 September 2025 Revised: 08 October 2025 Accepted: 06 November 2025 Published: 14 January 2026

Abstract *- Machine learning algorithms may create skewed results in classification. In some of the scenarios, like machine fault detection, fraud detection, and disease diagnosis, the intent class or the focused class instances may be very few when compared to the other instances or non-focused instances. In these cases, the classifier is trained based on the other non-focused classes and creates a skewed result. This will be high for false negatives. This paper proposes Stratified Near Miss Undersampling with Deep Neural Network (DeepSNMU) to investigate the imbalanced dataset using the Spark framework. The proposed DeepSNMU method uses IQR outlier detection for pre-processing, near-miss undersampling for balancing the dataset, and a Deep Neural Network for classification. Experiments were conducted using four highly imbalanced datasets from the KEEL repository. These datasets have multiple target classes, which are experimented with using DeepSNMU. In this research work, DeepSNMU produces high classification accuracy. The result of DeepSNMU is compared with existing balancing techniques and classifiers. The outcome of this research work has shown higher accuracy in predicting the minority intent class than the existing methods.*

Keywords *- Deep Neural Network, Hybridization, Imbalanced data, Multi-Class, Near Miss Undersampling.*

1. Introduction

Datasets that have a huge difference in the instance count for intent and non-intent classes are called an imbalanced dataset. When there are n-number of intent and non-intent classes, the dataset is called a multiclass imbalanced dataset. Normally, a classifier can be trained with the dataset irrespective of the instance count. This result may create a skewed one, which is based on the non-intent class or the majority class (J. L. Leevy et al., 2018).

In a multiclass dataset, each instance is classified as a class or a category, which is any one of the values from a defined set of classes. A multiclass imbalanced dataset is a category that consists of an uneven number of instances for each category. If there is a huge number of differences between all categories, it is called a highly imbalanced multiclass dataset (M. Sahare and H. Gupta, 2012). Generally, an imbalanced dataset classification may be biased towards a class that has high instances. This bias can be rectified by balancing the dataset before doing multi-classification (J. Tanha et al., 2020).

Generally, high-dimensional and imbalanced datasets can be classified using an ensemble classifier, which produces good accuracy when compared to the standard classifier (T.N. Rincy and R. Gupta, 2020). Ongoing research is being conducted for the reduction of bias in an imbalanced dataset. Among other ensemble classifiers, the Random Forest classifier method produces high accuracy (H. A. Salman et al., 2024)

Class imbalance is common in all real-time datasets. And the nature of imbalance is varied from slight to very high (M. Sahare and H. Gupta, 2012). Another problem of class imbalance is class difference and overlaps. These problems are degrading the performance of classifiers (M. S. Santos et al., 2022).

Class imbalance arises when there are an uneven number of majority classes and minority classes. For data scientists, minority class prediction is more important than classifier accuracy (J. Gao et al., 2021). Identifying minority classes is crucial in the field of cybersecurity, sentiment analysis, rare disease diagnosis, and healthcare (G. Haixiang et al., 2017). This makes the imbalance learning a crucial part in decision-making systems (W. Chen et al., 2024).

K-Nearest Neighbors (k-NN), Neural Networks, Support Vector Machines (SVMs), and gradient decision tree methods are widely used in research on classification tasks (M. Bansal et al., 2022). Even though we have several classifiers, no single method performs better than the other

methods for highly imbalanced data (C. M. Van der Walt and E. Barnard, 2021). Noisy data, outliers, skewed class distributions, and missing values are the factors for a highly imbalanced dataset (L. Wang et al., 2021).

1.1. Research Gap

Though many methods are available to balance and classify the multiclass imbalanced dataset, there is a research gap to achieve high accuracy in multi-classification of imbalanced data. This paper approaches with a hybrid approach to bridge the research gap.

2. Literature Review

T. K. Hasanin et al. (2019) experimented with three learners, namely Gradient-Boosted Trees, Logistic Regression, and Random Forest, and five sampling techniques, namely ADAptive SYNthetic (ADASYN), Synthetic Minority Over-Sampling Technique (SMOTE), Random Undersampling (RUS), Random Oversampling (ROS), and SMOTE-borderline for two case studies. The experimental results show that Random Undersampling produces good results in terms of computation and training time.

M. Rachmatullah (2022) proposed Iterative SMOTE for balancing multiclass imbalanced datasets and k-NN, Naive Bayes, Random Forest, and neural networks for classification. This experiment iteratively applied the SMOTE procedure on the six-class Glass dataset. In this experiment, the Random Forest classifier produced a higher level of performance with an accuracy of 86.27%, sensitivity 86.18%, and specificity 86.18%.

E.M. Hassib et al. (2019) proposed the LSH-SMOTE (Locality Sensitive Hashing Synthetic Minority Oversampling Technique) algorithm to address the imbalanced classes problem. This algorithm produced good results when compared with SMOTE and its nine variations. In the suggested framework, this method makes use of locality-sensitive hashing for two reasons: firstly, it helps to create subsets by categorizing instances into buckets through hashing. This approach simplifies the identification of the global optimum within each bucket. Secondly, the implementation of locality-sensitive hashing in conjunction with large datasets produces significant reductions in the operational time.

K. Jiang (2016) proposed a genetic algorithm-based SMOTE oversampling technique, called GASMOTE. This algorithm used a different combination of sampling rates for oversampling different minority classes. Finally, the optimal sampling rates for each minority class are found and applied. According to the performance evaluation, the GASMOTE method produced higher classification accuracy than the SMOTE and Borderline-SMOTE algorithms.

N. B. Abdel-Hamid (2018) proposed a Spark-Based Mining Framework (SBMF) to solve the imbalanced problem. For this use, there are two primary modules. The first module is the Border Handling Module (BHM). BHM oversamples minority class instances while undersampling majority border instances with no impact. The second module is the Selective Border Instances (SBI) Module. The second module improves the first module's output. The SBMF framework's performance is assessed and contrasted with that of other modern systems. A few tests with large and moderate datasets with various skewed ratios were carried out. In comparison to recent efforts, the findings derived from the SBMF framework demonstrate higher performance for various datasets and classifiers.

C. F. Tsai et al. (2024) proposed a method to combine feature selection and an oversampling technique for a multiclass imbalanced dataset. This experiment tests single and ensemble feature selection with the SMOTE technique for ten multiclass imbalanced datasets with low-to high feature dimensions. The result shows that XGBOOST with SMOTE gives a good performance result for classification.

A. Cano and B. Krawczyk (2022) proposed the Kappa Updated Ensemble (KUE) technique, which has a set of classifiers that is updated dynamically. The ROSE approach uses background classifiers, which are activated after drift is identified, and undersamples the majority classes. ROSE essentially uses self-adjusting bagging to balance training sets by keeping an eye on the accuracy and statistical product.

J. Zhang et al. (2023) proposed differential evolution for highly imbalanced datasets using a novel oversampling technique called SS_DEBOHID. This method is compared with a few existing methods on a significant number of highly imbalanced datasets. The experimental result shows that the SS_DEBOHID method produces high classification performance as 8.07% with 24.34% average AUC metric and 6.96% to 45.37% average G_mean metric.

M. Mohammed Al Sameer et al. (2021) suggested using cartographic data for predicting forest cover type quickly than the existing methods. In comparison to the current model, the model trained by analyzing exploratory data and applying feature engineering and ensemble learning techniques increased the model's overall accuracy by 93%. The suggested methodology has provided the best means of determining the type of forest cover.

B. Jabir et al. (2023) proposed the Ensemble Partition Sampling (EPS) for imbalanced multiclass classification. This method produced higher performance results than the existing state-of-the-art methods, which include OvA, SMOTE, k-means SMOTE, Bagging-RB, DES-MI, OvO-EASY, and OvO-SMB.

I. Naglik et al. (2024) proposed GMMSampling to learn multiclass imbalanced data. This method removes the class overlapping region of the majority classes and creates synthetic instances of minority classes. This method produces balanced accuracy, which is better than existing methods. According to the widely accepted literature, the neural network algorithm and the undersampling strategy seem to be effective and capable of striking a balance between different approaches in the big data analytics space.

3. Materials and Methods

3.1. Data Source

This section outlines the suggested method for multiclass classification that addresses the imbalance problem by utilizing DeepSNMU within the Spark framework. Experiments were conducted using four datasets from the KEEL repository. All four datasets have multiple target classes. These are tested with the proposed DeepSNMU method, and the results were compared with existing methods.

3.2. Dataset Description

The size and quantity of features in the datasets vary; some have as few as four features, while others have up to thirty-four. There are both nominal and numerical attributes in the datasets, and the imbalance ratios range from 5.55 to 71.5. Table 1 shows the experimental dataset description. The experiment is carried out in a Spark environment with 1 TB of storage on HDD and 16 GB of RAM.

Imbalance Ratio (IR) = majority class instance count/minority class instance count

3.3. Proposed Model

In the proposed model, the IQR method is used for preprocessing the dataset. After preprocessing, the Hybridization of Deep Neural Network with Stratified Near Miss Undersampling (DeepSNMU) is applied on an imbalanced dataset to balance and classify. The proposed model is shown in Figure 1.

3.4. Hybridization of Deep Neural Network with Stratified Near Miss Undersampling (DeepSNMU)

In the first phase of the proposed system, an imbalanced dataset is preprocessed using the Inter Quartile Range (IQR) method (A. Q. Md et al. 2023). This will remove the outliers that are farther from other instances of the same class. In the second phase, the stratified near-miss undersampling has been applied to balance the dataset. In the third phase, a balanced dataset will be classified by a deep neural network with ReLU activation function in hidden layers. 64 neurons are used in a single hidden layer. The output layer utilizes the softmax function.

3.4.1. Preprocessing

Preprocessing is done by the Inter-Quartile Range (IQR), which is the difference between Q1 (first quartile) and Q3 (third quartile) (A. Alabrah, 2023). When datasets contain outliers based on the density property, this data normalization technique is used to concentrate the denser points.

To remove the effect of outliers on model training, the approach makes use of median and interquartile ranges rather than mean and standard deviation (H. P. Vinutha et al. 2018). Additionally, this approach works well with datasets that contain skewed and abnormal results.

3.4.2. Stratified Random Sampling

According to sampling theory, effective stratification should arrange the samples in a way that minimizes differences within each stratum while maximizing differences between the strata. The methodology of stratified random sampling represents a practical approach for the selection of a sample, whereby each individual within the population possesses an opportunity to be selected. The best allocation depends on the absence of labels to determine the budget assignment for each stratum, but this makes it impossible to observe the variance of accuracy in each stratum (Z. Wu et al. 2024).

3.4.3. Near Miss Undersampling

One of the most popular strategies to deal with class imbalance is the undersampling approach, which chooses the majority instances using the distance method. Nevertheless, information loss is a natural consequence of conventional undersampling-based techniques. Using distance-based localization techniques, they frequently undersample the majority class while utterly disregarding the density of the minority class data.

Table 1. Dataset description

Dataset name	Instance count (Samples)	Features count	Target classes	Imbalance Ratio (IR)
BalanceScale	625	4	3	5.88
Car	1728	6	2	18.62
Dermatology	366	34	6	5.55
Ecoli	336	7	8	71.5

'Clinical trial number: not applicable.'

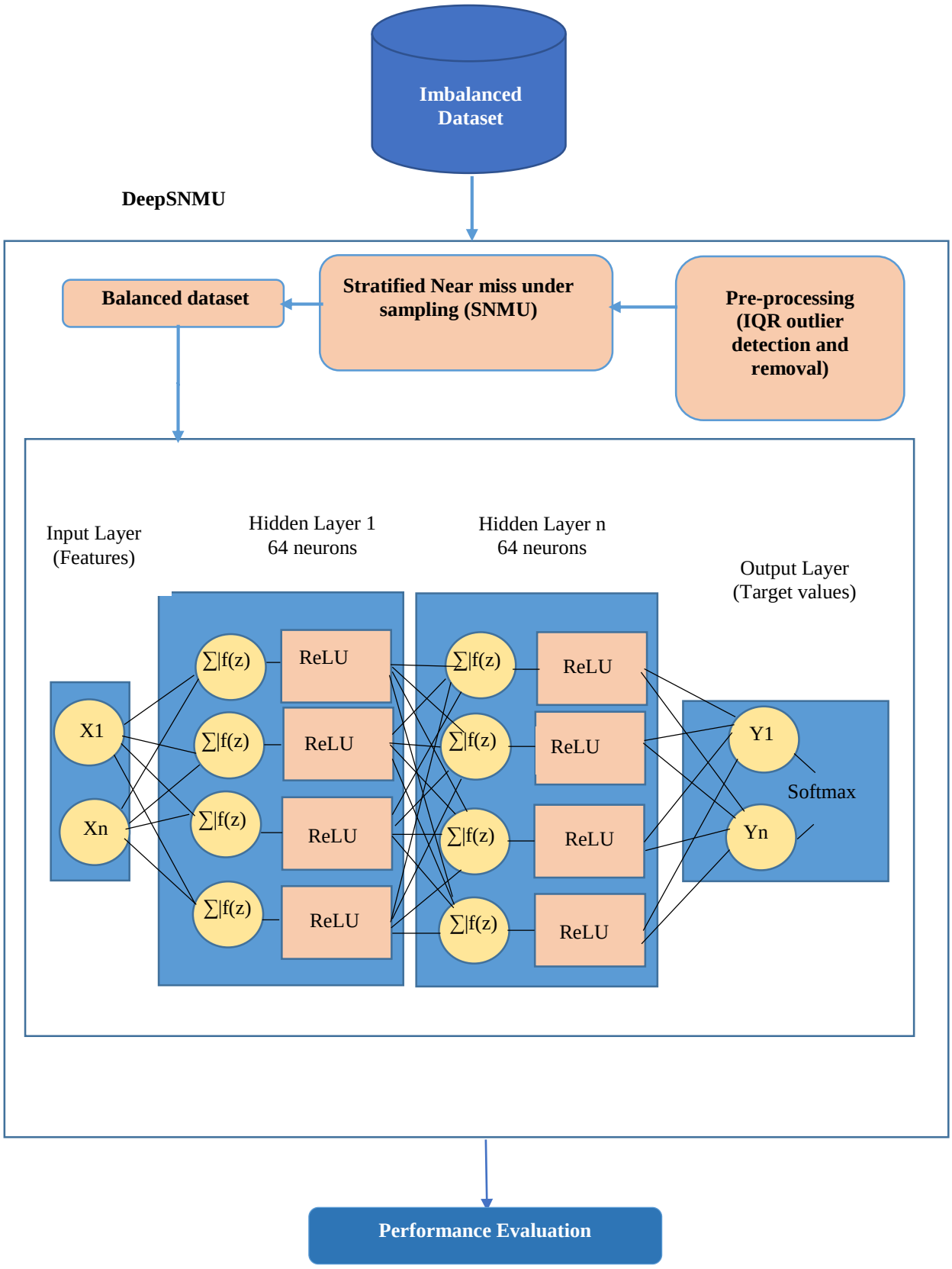

Fig. 1 Proposed model – DeepSNMU

In order to choose information-rich samples, sampling fitness is suggested as a way to assess each majority class sample's desirable value. An undersampling technique is suggested, and the algorithm's complexity is examined, taking into account the minority class density.

This method is used with more extensive data. Based on the minority class's distribution density, a novel approach to determining fitness is put forth. Wherein the roulette operator specifically preserves the majority class's distribution properties when choosing the instances (Z. Sun et al. 2024).

Models trained on unbalanced data can perform noticeably better thanks to the potent NearMiss method. It is simple to use and can be combined with other methods, such as oversampling, to enhance model performance even further. One such undersampling method is NearMiss, which reduces the overlap between the two classes. There are three variations of the Near Miss technique. Sampling can also alter the presence of noise, particularly in the presence of marginal outliers. Noise will have less of an impact on NearMiss-3 due to the first-step sample selection, and also suits a multiclass imbalanced dataset.

Apart from the three variations, NearMiss features several adjustable characteristics that can be used to enhance its functionality. These include the sampling approach used to decide how many samples to retain, the number of neighbors to take into account when choosing samples, and the distance metric used to gauge how close samples are to one another.

All things considered, NearMiss is a helpful technique for dealing with unbalanced data, and it may be used in conjunction with other strategies like ensemble methods or oversampling to further enhance classifier performance.

3.4.4. Deep Neural Network

Training and inference are the two main stages of DNNs, which are modeled after the human brain. The percentage of DNN outputs that match the intended output is known as accuracy. Neurons are building blocks of DNNs. After receiving certain activation signals, each neuron multiplies them by the appropriate weights. The weighted activations are then added up and sent to the output. A layer is formed by a collection of neurons that may perform additional tasks, such as batch normalization (max or average), activation function (ReLu, sigmoid, etc.), pooling, and so on.

3.4.5. Algorithm 1: DeepSNMU Pseudocode

Step 1: Preprocessing – IQR Outlier detection and removal
  For each instance in the dataset
    Check is the instance is an outlier or not
    If it is an outlier, remove the instance; else, keep it
  End For
Step 2: Remove the majority class instances using the nearmiss ver-3 undersampling technique
  For each majority class
    Iterate until the majority class instance count is equal to the minority class instance count.
  End For
  Return a balanced dataset
Step 3: A Balanced dataset is trained using a DNN model
Step 4: Testing can be done using a DNN model
Step 5: Classification output is displayed
Step 6: Stop

4. Results and Discussion

The experiment is conducted in two phases: data balancing and classification. Figure 2 depicts the class distribution of all datasets before and after balancing. All datasets' class distribution is equal after balancing.

The performance results are analysed based on precision, recall, F-measure, and accuracy. All four datasets are experimented with the proposed methodology, DeepSNMU.

Tables 2, 3, 4, and 5 consolidate the performance of the datasets BalanceScale, Car, Dermatology, and Ecoli. The proposed DeepSNMU results are compared with the EPS (B. Jabir et al. 2023) method.

Table 2. Comparative analysis for the balancescale dataset

Metric	EPS	DeepSNMU
Accuracy	81.87 [26]	96
Precision	84.01 [26]	96.87
Recall	81.13 [26]	96
F1-score	73.69 [26]	96.21

Table 3. Comparative analysis for the car dataset

Metric	EPS	DeepSNMU
Accuracy	98.3 [26]	100
Precision	96.52 [26]	100
Recall	99.87 [26]	100
F1-score	98.67 [26]	100

Table 4. Comparative analysis for dermatology dataset

Metric	EPS	DeepSNMU
Accuracy	94.18 [26]	97.2973
Precision	93.59 [26]	97.7477
Recall	94.97 [26]	97.2973
F1-score	93.92 [26]	97.2973

Table 5. Comparative analysis for ecoli dataset

Metric	EPS	DeepSNMU
Accuracy	71.89 [26]	89.0625
Precision	69.31 [26]	91.5649
Recall	71.17 [26]	89.0625
F1-score	67.82 [26]	89.4812

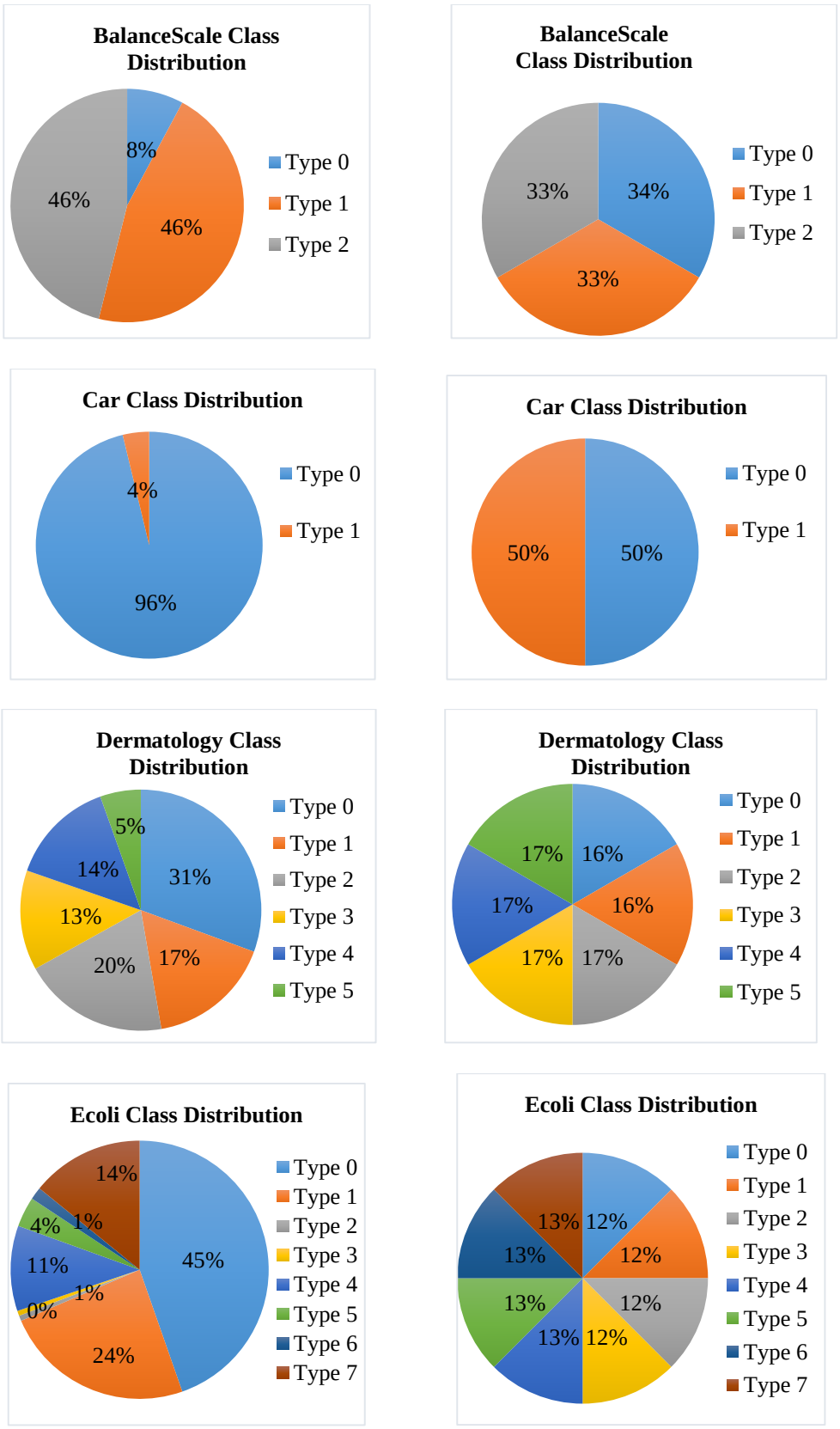

Fig. 2 Datasets before and after balancing

The DeepSNMU method is applied to the balanceScale dataset. The performance results are presented in Table 2. In this, the DeepSNMU accuracy is 96%, precision is 96.8%, recall score is 96% and f1 score is 96.2%.

In Figure 3, DeepSNMU and EPS (B. Jabir et al. 2023) performance results are compared for the balanceScale dataset. In this, DeepSNMU performs well in accuracy, precision, recall, and f1-score.

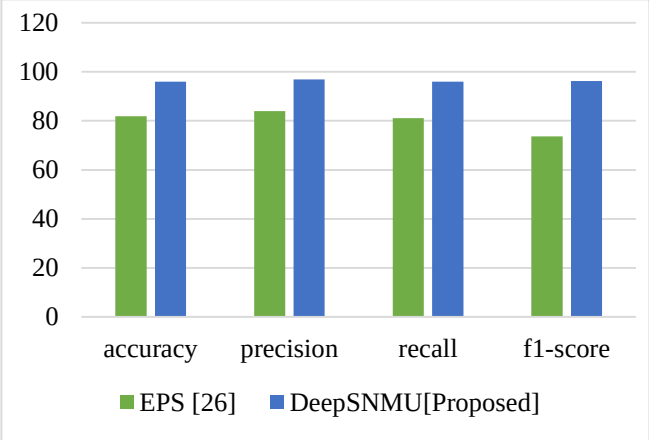

Fig. 3 Performance comparison for balancescale dataset

The DeepSNMU method is applied to the car dataset. Performance results are shown in Table 3. In this, the DeepSNMU accuracy is 100%, precision is 100%, recall score is 100% and f1 score is 100%.

In Figure 4, DeepSNMU and EPS (B. Jabir et al. 2023) performance results are compared for the car dataset. In this, DeepSNMU performs well in accuracy, precision, recall, and f1-score.

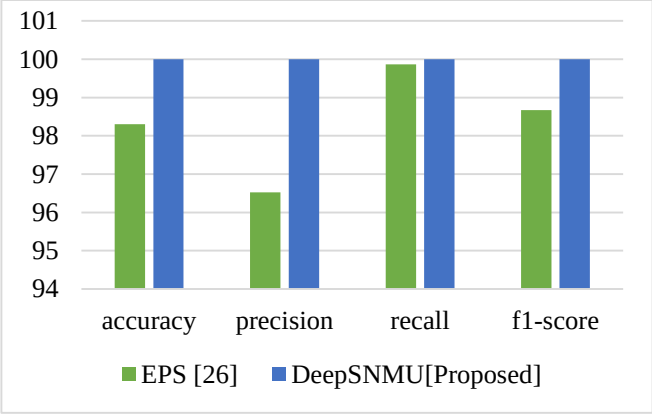

Fig. 4 Performance comparison for the car dataset

The DeepSNMU method is applied to the dermatology dataset. Performance results are shown in Table 4. In this, the DeepSNMU accuracy is 97.2%, precision is 97.7%, recall score is 97.2% and f1 score is 97.2%.

In Figure 5, DeepSNMU and EPS (B. Jabir et al. 2023) performance results are compared for the dermatology dataset. In this, DeepSNMU performs well in accuracy, precision, recall, and f1-score.

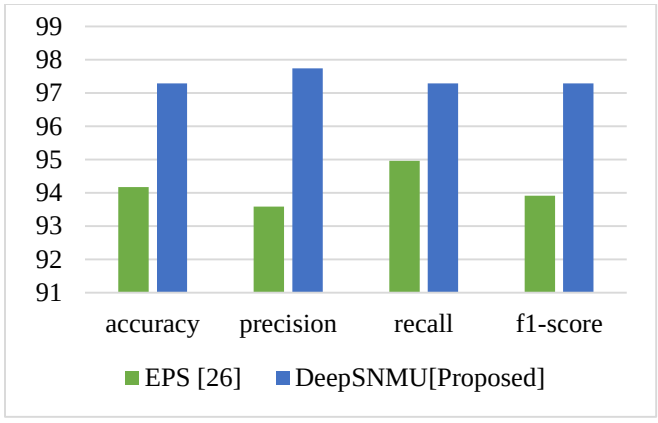

Fig. 5 Performance comparison for dermatology dataset

The DeepSNMU method is applied to the Ecoli dataset. Performance results are shown in Table 5. In this, the DeepSNMU accuracy is 89%, precision is 91.5%, recall score is 89% and f1 score is 89.5%.

In Figure 6, DeepSNMU and EPS (B. Jabir et al. 2023) performance results are compared for the Ecoli dataset. In this, DeepSNMU performs well in accuracy, precision, recall, and f1-score.

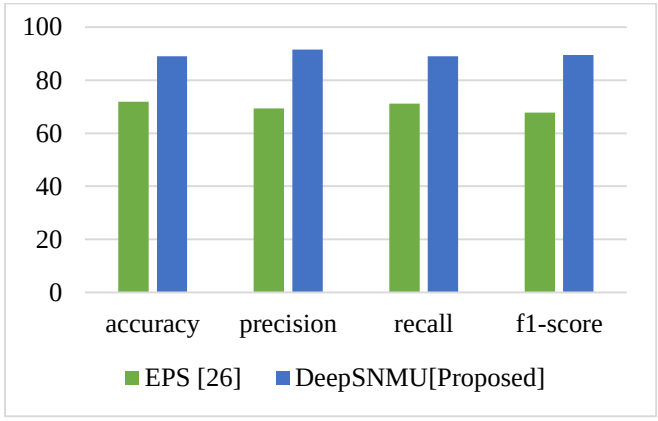

Fig. 6 Performance comparison for the ecoli dataset

Among the existing methods, namely EPS, OvA, k-means SMOTE, SMOTE, and Bagging-RB (B. Jabir et al. 2023), DeepSNMU produces an enhanced performance. Accuracy comparison results are shown in Table 6.

In this, for all the datasets, DeepSNMU attained higher accuracy than EPS, OvA, k-means SMOTE, SMOTE, and Bagging-RB (B. Jabir et al. 2023). Figure 7 is a graphical representation of the accuracy comparison of all methods. In all datasets, DeepSNMU has higher accuracy, and the proposed model's performance is higher than existing methods.

Table 6. Accuracy comparison

Datasets	EPS	OvA	k-means SMOTE	SMOTE	Bagging-RB	DeepSNMU (Proposed)
balanceScale	81.87 [26]	70.33 [26]	81.21 [26]	79.58 [26]	76.58 [26]	95.83
car	94.52 [26]	91.83 [26]	92.66 [26]	92.34[26]	92.28 [26]	100
dermatology	94.18 [26]	91.82 [26]	91.52 [26]	89.3 [26]	94.62 [26]	97.29
ecoli	71.89 [26]	70.56 [26]	69.25 [26]	66.31[26]	67.86 [26]	89.06

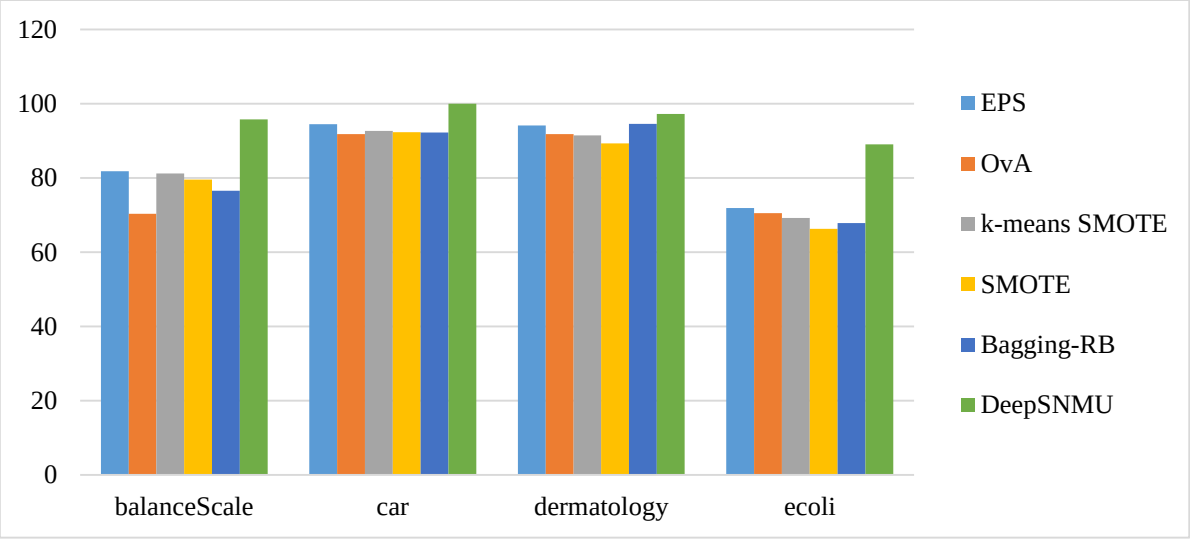

Fig. 7 Accuracy comparison

4.1. Novelty and Contributions

Existing state-of-the-art classification algorithms apply to the standard dataset. But those algorithms may produce biased classification when the dataset is imbalanced. The proposed method will balance and classify the dataset using the DeepSNMU method. This DeepSNMU produces higher accuracy than the existing methods, such as EPS, OvA, SMOTE, k-means SMOTE, and Bagging-RB. The existing EPS (B. Jabir et al. 2023) method produces the accuracy for the datasets balanceScale, car, dermatology, and ecoli are 81.87, 94.52, 94.18, and 71.89. But the proposed DeepSNMU method produces the accuracy for the datasets balanceScale, car, dermatology, and ecoli are 95.83, 100, 97.29, and 89.06. In this way, the proposed method outperforms existing methods.

5. Conclusion

Most of the domains have an imbalanced dataset. Classifying an imbalanced dataset without bias is challenging one. The challenge is greater when the dataset has multiple target classes. This problem can be addressed by balancing techniques like oversampling, undersampling, etc. In this research, the multiclass imbalanced problem is addressed by the DeepSNMU technique, which is a hybridization of Stratified Nearmiss undersampling with a Deep Neural Network. This proposed technique is applied to four imbalanced datasets. For all the datasets, DeepSNMU attained higher accuracy than the existing methods, which are EPS, OvA, k-means SMOTE, SMOTE, and Bagging-RB.

Acknowledgments

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a broader audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

References

[1] Hasan Ahmed Salman, Ali Kalakech, and Amani Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69-79, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Thomas N. Rincy, and Roopam Gupta, "Ensemble Learning Techniques and Its Efficiency in Machine Learning: A Survey," *Proceedings of the IEEE 2nd International Conference on Data, Engineering and Applications*, Bhopal, India, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mahendra Sahare, and Hitesh Gupta, "A Review of Multi-Class Classification for Imbalanced Data," *International Journal of Advanced Computer Research*, vol. 2, no. 5, pp. 163-168, 2012. [Google Scholar] [Publisher Link]

[4] Miriam Seoane Santos et al., "On the Joint-Effect of Class Imbalance and Overlap: A Critical Review," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6207-6275, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Lu Gao, Pan Lu, and Yihao Ren, "A Deep Learning Approach for Imbalanced Crash Data in Predicting Highway-Rail Grade Crossings Accidents," *Reliability Engineering & System Safety*, vol. 216, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Guo Haixiang et al., "Learning from Class-Imbalanced Data: Review of Methods and Applications," *Expert Systems with Applications*, vol. 73, pp. 220-239, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Wuxing Chen et al., "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," *Artificial Intelligence Review*, vol. 57, no. 6, pp. 1-51, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Malti Bansal, Apoorva Goyal, and Apoorva Choudhary, "A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning," *Decision Analytics Journal*, vol. 3, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] C. M. Van der Walt, and E. Barnard, "Data Characteristics that Determine Classifier Performance," *SAIEE Africa Research Journal*, vol. 98, no. 3, pp. 87-93, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Le Wang et al., "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606-64628, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Tawfiq Hasanin et al., "Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches," *Journal of Big Data*, vol. 6, no. 1, pp. 1-25, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Muhammad Ibnu Choldun Rachmatullah, "The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data," *MATRIK: Journal of Management, Information Technology and Computer Engineering*, vol. 22, no. 1, pp. 13-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Eslam Mohsen Hassib et al., "An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance," *IEEE Access*, vol. 7, pp. 170774-170795, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kun Jiang, Jing Lu, and Kuiliang Xia, "A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE," *Arabian Journal for Science and Engineering*, vol. 41, pp. 3255-3266, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Nahla B. Abdel-Hamid, "A Dynamic Spark-Based Classification Framework for Imbalanced Big Data," *Journal of Grid Computing*, vol. 16, pp. 607-626, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] M. Mohammed Al Sameer, T. Prasanth, and R. Anuradha, "Rapid Forest Cover Detection using Ensemble Learning," *Proceedings of the International Virtual Conference on Industry 4.0: (IVCI4.0)*, pp. 181-190, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Chih-Fong Tsai, Kuan-Chen Chen, and Wei-Chao Lin, "Feature Selection and Its Combination with Data Over-Sampling for Multi-Class Imbalanced Datasets," *Applied Soft Computing*, vol. 153, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Alberto Cano, and Bartosz Krawczyk, "ROSE: Robust Online Self-Adjusting Ensemble for Continual Learning on Imbalanced Drifting Data Streams," *Machine Learning*, vol. 111, no. 5, pp. 1-32, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jiaoni Zhang et al., "A New Oversampling Approach Based Differential Evolution on the Safe Set for Highly Imbalanced Datasets," *Expert Systems with Applications*, vol. 234, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Abdul Quadir Md et al., "Enhanced Preprocessing Approach using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] H.P. Vinutha, B. Poornima, and B.M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," *Advances in Intelligent Systems and Computing*, vol. 701, pp. 511-518, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Amerah Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Zhuo Wu et al., "Stratified Random Sampling for Neural Network Test Input Selection," *Information and Software Technology*, vol. 165, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Zhongqiang Sun et al., "Undersampling Method Based on Minority Class Density for Imbalanced Data," *Expert Systems with Applications*, vol. 249, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Joffrey L. Leevy et al., "A Survey on Addressing High-Class Imbalance in Big Data," *Journal of Big Data*, vol. 5, no. 1, pp. 1-30, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Brahim Jabir et al., "Ensemble Partition Sampling (EPS) for Improved Multi-Class Classification," *IEEE Access*, vol. 11, pp. 48221-48235, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jafar Tanha et al., "Boosting Methods for Multi-Class Imbalanced Data Classification: An Experimental Review," *Journal of Big Data*, vol. 7, no. 1, pp. 1-47, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Iwo Naglik, and Mateusz Lango, "GMMSampling: A New Model-Based, Data Difficulty-Driven Resampling Method for Multi-Class Imbalanced Data," *Machine Learning*, vol. 113, no. 8, pp. 5183-5202, 2024. [CrossRef] [Google Scholar] [Publisher Link]