The 2nd International Conference on Multidisciplinary Engineering and Applied Sciences (ICMEAS-2023)

Review of Machine Learning Techniques For Class Imbalance Medical Dataset

Hauwa Ahmad Amshi
D*eptment of Computer Science*
Federal University
Gashua, Nigeria
hauwa.amshi@gmail.com

Asmau Usman
Department of Compter Science
Abdu Gusau Polytechnic
Talata Mafara, Nigeria
asmee08@gmail.com

Rajesh Prasad
Department of Computer Science
African University of Science and Technology
Abuja,Nigeria
rprasad@aust.edu.ng

Ibrahim Anka Salihu
Department of Software Engineering
Nile University of Nigeria
Abuja, Nigeria
ibrahim.salihu@nileuniversity.edu.ng

Abubakar Muhammad Dadile
Department of Computer Science
Umar Sulaiman College of Education
Gashua, Nigeria
abubakarmdadile@gmail.com

***Abstract*— Data imbalance threatens a medical dataset where the dominant class is typically viewed as unfavorable. In contrast, the minority class is supposed to be the positive one, affecting the machine learning prediction performance. This aims to examine how resampling strategies in Machine Learning (ML) have recently been used in medical data sets. Many researchers used the preprocessing stage's data-level approach to resample the imbalanced medical data. Thirty-two sources were reviewed in which data level techniques of balancing the imbalanced data were applied to medical datasets spanning 2018 to 2023, with oversampling methods outperforming the under-sampling methods.**

Keywords— Data-level approach, Medical data, Oversampling, SMOTE, Under-sampling.

I. INTRODUCTION

The class imbalance (CI) distribution is an issue of real-world medical data [1]. The unequal quality of the minority and majority classes is a severe problem when employing data analysis for diagnosis and therapy [2]. When comparing data sets that are imbalanced, representative samples spread unevenly across classes, and some classes have fewer samples than others.

A class-imbalanced dataset has one class that has a disproportionately smaller samples than the other [3], [4]. Rare minority samples could be considered noise, which might be a mistaken label for the sample. Such an imbalance issue is relatively common in the medical industry. Class imbalance approaches have previously been implemented in various applications within the realm of various instances, such as gene expression and medical diagnostics. Real-life data are naturally imbalanced, which is the primary reason of decreased generalization in machine learning (ML) algorithms.

In imbalanced classification problems, the dominating class is often measured as the negative class, while the minority class (MNC) is typically viewed as the positive class because it is frequently of more interest in medical datasets, and classification errors for the MNC can have more severe consequences than errors in the majority class (MJC). The imbalance between the classes can pose a challenge for ML algorithms, as they are often geared towards the MJC examples, which results in poor performance in predicting the MNC. To address the issue, various techniques can be used, including data-level, algorithm-level, and hybrid approaches [5].

In the data-level approaches, objects are redistributed throughout the data space to generate the class balance before any input data is processed. Either increasing the minority class size or decreasing the size of the majority class is used to establish balance [6].

To address the problem of imbalanced data sets, the learning algorithm is either changed or developed from scratch in algorithm-level approaches. Several methods have relied on biased existing classifiers at the algorithmic level rather than rebalancing the class distribution at the data level [3].

Both the data-level and algorithm-level methods are used to handle class imbalance in medical datasets in the hybrid approach. The approaches aim to leverage the strengths of both approaches to achieve better results by handling data externally and modifying the distribution of classes in the sample using data-level approaches [7].

II. RELATED WORK

Class imbalance is a prevalent problem in ML, particularly in medical diagnosis jobs when the majority of patients are healthy and disease detection is more important. Imbalanced training data can have a significant detrimental impact on performance, necessitating the use of non-standard ML approaches to attain desirable outcomes. Various techniques have been developed to address this issue, including data-level approaches, algorithm-level approaches, and hybrid approaches.

A survey conducted by [8] identified several machine-learning techniques for handling class imbalance, including resampling, cost-sensitive learning, and evaluation metrics. A systematic mapping study [9] identified various preprocessing methods to balance imbalanced datasets for ML training.

Some authors presented a review of approaches used to classify imbalanced datasets and their various application areas [19]. Classification of imbalanced data review [10] revealed that leading ML strategies were discovered to handle imbalanced datasets by focusing on avoiding the minority class and reducing inaccuracy for the majority class. High-

979-8-3503-5883-4/23/$31.00 ©2023 IEEE

quality review papers on imbalanced data and related topics have been published over the past years. However, to our knowledge, a review has yet to be made on imbalanced problems/solutions to medical data. The inspiration for this survey is to introduce the bigger picture of the issues and developments of the imbalanced behavior of medical data and to investigate the different resampling strategies used.

III. Methods of Balancing Imbalanced Data

The strategies for solving the imbalance problems are: data level, algorithm level, and hybrid approach.

A. Data-level approach

Data-level approaches are used to externally manipulate data and govern the distribution of groups in the sample. To address the problem of class imbalance, sampling procedures have been employed to either eliminate some data from the majority class (under-sampling) or add some intentionally generated or duplicated data to the minority class (oversampling). *[15]*.

1) Under-sampling

In order to reach a predetermined balanced-ratio by under-sampling, specific samples that are repeated from the initial data set are removed. The proper categorization of the minority sample results from the second method's effective reduction of the space allotted to the majority class which might lead to data loss [3, 4]. The frequently used techniques include Random Under-Sampling (RUS), Condensed Nearest Neighbor (CNN), Tomek Links (TL) and the Edited Nearest Neighbors (ENN).

2) Oversampling

This technique addresses the initial uneven dataset by generating fresh artificial samples and then add them to the minority group. The major drawback of this approach is that over-fitting may occur. In order to solve this shortcoming, approaches for producing additional synthetic samples are being explored in potential regions [4]. The most general approaches for oversampling are Random Over-Sampling (ROS), Adaptive Synthetic sampling technique (ADASYN) and the Synthetic Minority Over-Sampling Technique (SMOTE).

3) Hybrid method

Hybrid resampling, which uses both oversampling and under sampling, is proposed to get reliable results from data processing. To correct the sample imbalance, the main idea is to increase the number of samples from minorities while decreasing the number of samples from the majority [16].

B. Algorithm Level Approach

Instead of modifying data, these approaches create new algorithms or enhance existing ones, working from the inside out to achieve optimal performance [17]. The algorithmic level approaches include threshold, learning techniques (one-class and cost-sensitive learning), and ensemble-based strategies [18-21].

C. Hybrid Approach

Hybrid methodologies encompass a fusion of both data-level and algorithmic techniques for addressing class imbalance in medical datasets [22]. These approaches leverage data-level methods for external data processing, thereby adapting the distribution of categories within the sample. Subsequently, algorithmic methods come into play to internally modify the learning process [23]. This combination of data and algorithmic strategies within hybrid approaches proves instrumental in enhancing the efficacy of machine learning models when confronted with imbalanced datasets. Such strategies may encompass a diverse set of techniques including oversampling, undersampling, cost-sensitive learning, and adaptations to learning algorithms [24].

IV. Methodology

Before Articles published recently identified different machine learning strategies utilized in balancing out. It ensures that the search process entails finding good digital libraries, specifying search terms, and choosing the time gap between published articles.

We looked at the essential online digital resources for computer science and a Medline digital library that produced peer-reviewed publications. ACM, ScienceDirect, IEEE, Springer, PubMed, PLOS and HINDAWI are the digital libraries examined. Several relevant journals and conference proceedings in AI were also reviewed. The search was restricted to items published between 2018 and 2023. In this search, the following search string was used: *"disease imbalanced data" OR "imbalanced medical data" OR "imbalanced clinical data" AND " artificial intelligence" OR "machine learning" OR "data mining"*.

V. State-Of-The-Art for Balancing Imbalanced Medical Datasets

Medical data is available from various sources in this era of big data, including disease prediction/forecasting, public health records, biometric data, and medical imaging, which is rapidly expanding [25]. The data level approach for balancing is more effective and is applied before the learning process at the data preparation step. Hence, many researchers utilized this approach in the preprocessing stage [17]. Table I, Table II and Table III summarizes the papers that utilized data-level approaches for resampling the imbalanced data, Algorithm-level techniques and Hybrid Techniques.

Researchers employed many forms of medical databases to study various diseases/medical cases with data imbalanced problems. The main criteria used for evaluating the classification of imbalanced medical data based on the reviewed document are sensitivity, specificity, and accuracy [39-41]. Some researchers [3,33,35], use Area Under Curve (AUC) and Mathew correlation coefficient (MCC) as part of the evaluation metrics. The equations below describe the performance metrics used.

$$Accuracy = \frac{number\ of\ accurate\ predictions}{total\ number\ of\ predictions} \quad [4]$$

$$Recall/sensitivity = \frac{true\ positive}{True\ Negative + False\ Positive} \quad [5]$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad [6]$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [7]$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad [8]$$

$$Gmean = \sqrt{sensitivity \times specificity} \quad [9]$$

$$AUC = \frac{1 + True\ positive_{rate} - False\ positive_{rate}}{2} \quad [10]$$

Mathew correlation coefficient (MCC) =

$$\frac{((TP.TN) - (FP.FN))}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}} \quad [11]$$

TABLE I. SUMMARY OF REVIEWED PAPERS THAT USES DATA-LEVEL TECHNIQUES

Authors	Method	Datasets used	Findings	Evaluation Metrics
[26]	SMOTE	Heart	Extra Trees (ET) and SMOTE exhibited excellent performance compared to other models in the diagnosis of CVD.	ACC, PRC R, F, MCC
[27]	Borderline-SMOTE with ANN-MLP	Parkinson	Borderline-SMOTE outperformed other techniques.	ACC, Gmean, IBA, AUC
[28]	D-SMOTE and BP-SMOTE	COVID-19, Breast and Framingham	BP-SMOTE and D-SMOTE exhibit improved accuracy over standard SMOTE.	ACC, PRC, R, ROC
[29]	SMOTEEN	Heart, breast cancer, Pima, Indian liver patient, diabetes, and coronary kidney	SMOTEEN outperformed all other data-balancing techniques.	ACC, R, F
[30]	(STL)	Epilepsy	Achieved robust high sensitivity and AUC, outperforming other methods currently in use.	F, SEN, ACC, AUC
[31]	SMOTE	Heart Disease	SMOTE-based ANN outperform other models.	PRC, R, F
[32]	SMOTEEN	lung cancer	SMOTEEN standard deviation is substantially lower than that of under-sampling.	AUC
[33]	RCSMOTE	15 medical data from UCI repository	Addresses the issue of over-generalization problem caused by over-sampling.	AUC
[34]	SMOTE	Heart	Improved the predictive capabilities of tree-based classifiers in determining the survival of individuals with cardiac conditions.	ACC, PRC, R, F
[35]	ADASYN	Epilepsy	Adaptive Synthetic Sampling (ADASYN) demonstrated an outstanding result outperforming other method.	AUC
[36]	Outlier SMOTE	COVID 19	Outlier-SMOTE performs significantly better than the classic SMOTE algorithm.	R, PRC, F
[37]	SMOTE	Heart	High accuracy obtained showed that the model beat existing models and earlier study findings.	ACC, PRC, F, MCC, FPR, FNR, TNR, R
[38]	SMOTE	Athletes	Enhanced cardiovascular risk assessment by increasing the area under the curve.	TPR, TNR, J, AUC
[39]	SMOTE	Parkinson's Disease	Improved in terms of accuracy and AUC.	ACC, Kappa, PRC, R, F, AUC
[22]	ADASYN and Borderline-SMOTE.	Cancer	Oversampling balancing techniques result in better outcomes than under sampling ones.	AUC
[40]	ADASYN	Cholera	It improved the overall performance of the prediction result	SEN, SPE, ACC
[41]	SMOTE	Cholera	The overall performance of the model is improved	SEN, F1, ACC
[42]	SMOTE and SPIDER	Epilepsy	Improved sensitivity in detecting epileptic seizure detection.	ACC, PRC, R, F
[43]	SMOTE	Lasa	Improved the accuracy of the medical data due to look-alike sound-alike (LASA) mix-ups.	R, PRC, F, ACC
[44]	SMOTE	CKD	Improve classification algorithm performance, and learning rate on multilayer perceptron performance.	PRC, R, F, RMSE
[3]	SMOTE	Osteoporosis	Highest efficiency was obtained when the SMOTE and Random Forest classifier were used.	SPC, SEN, AUC, GMean, BAcc, MCC
[45]	ROS	Cholera	ROSE resampling with random forests method results in a high (AUC), balanced sensitivity and specificity.	AUC, SPC, SEN
[46]	SMOTE	Heart	SMOTE technique improved the performance of the KNN-algorithm	Accuracy
[47]	Borderline-smote	Appendicitis, Breast	Improved the accuracy of AIRS performance due to the influence of SMOTE techniques.	ACC, Gmean, SEN, SPE

ACC: Accuracy, SEN: Sensitivity, SPC: Specificity, R: Recall, F: F1 score, PRC: precision, RMSE: Root Mean Squared Error, AUC: Area Under Curve, MCC: Matthew Correlation coefficient.

Table 1 provides a comprehensive overview of the datasets and techniques employed at the data-level in the analyzed articles. The results reveal that the heart disease dataset was utilized in 9 articles, breast cancer dataset in 6 articles, and cholera, epilepsy, and Parkinson's datasets in 3 articles each. Hepatitis, Pima, COVID-19, diabetes, Haberman, and liver datasets were featured in 2 articles each, while various other datasets were the focus of 1 article.

In terms of dataset balancing techniques, the majority of articles (17 in total) employed SMOTE, while 3 articles utilized SMOTE-Tomek. Additionally, 2 articles each utilized ADASYN, SMOTEEN, and Borderline SMOTE, and 1 article each used RUS, ROS, and SPIDER.

Regarding performance metrics, the analysis revealed that Recall was the most commonly employed metric, appearing in 16 articles. Accuracy closely followed, featured in 14 articles, while the F1 score was used in 11 articles. MCC (Matthews Correlation Coefficient) and G-mean were each utilized in 8 articles, while SPC (Specificity) was the chosen metric in 4 articles. G-mean appeared in 3 articles, and other accuracy measures were employed in 5 articles. All this are shown in Fig 1, Fig 2 and Fig 3 respectively.

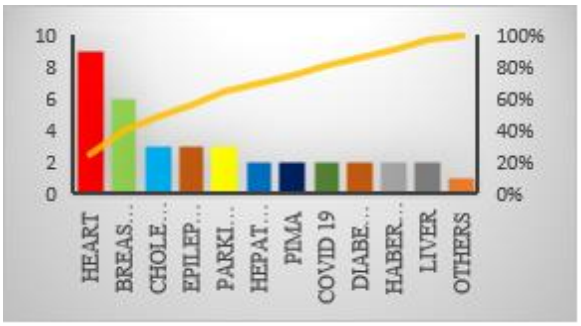

Fig.1: Datasets used

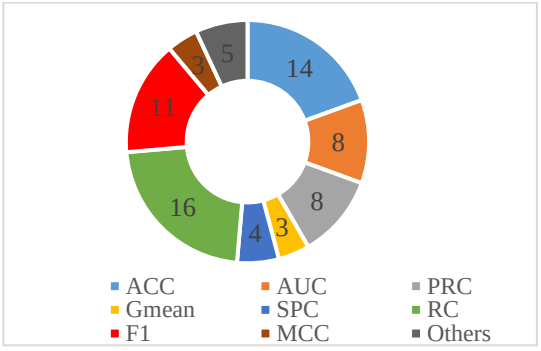

Fig. 3: Performance Metrics Frequently Used

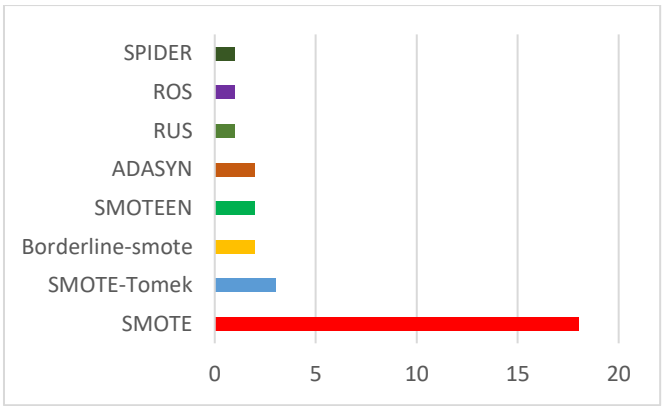

Fig 2: Balancing Technique used

Table II below provides an overview of the datasets and techniques employed at the algorithm in the analyzed articles. The results reveal that the breast cancer Disease was utilized in 2 articles while various other datasets were the focus of 1 article.

In terms of dataset balancing techniques, the majority of articles (2 in total) employed Modified SMOTE coupled with stacked deep learning algorithm. Regarding performance metrics, the analysis revealed that all the 3 articles analyzed used Accuracy, sensitivity, specificity and F1-score.

TABLE II. SUMMARY OF REVIEWED PAPERS THAT USES ALGORITHM LEVEL APPROACH

Authors	Method	Datasets used	Findings	Evaluation Metrics
[29]	Modified SMOTE coupled with stacked deep learning algorithms	Breast Cancer Disease, Coronary Kidney Disease, Indian Liver Patient, Coronary Heart Disease, and Pima Indians Diabetes	The approach surpassed other methods in terms of accuracy, sensitivity, specificity, and F1-score.	ACC, SEN, SPC, F1
[48]	Modified SMOTE coupled with stacked deep learning algorithms	Framingham, breast cancer and COVID-19	The proposed method outperformed other methods in terms of sensitivity, specificity, and F1-score.	ACC, SEN, SPC, F1
[49]	Hybrid approach combining data-level and algorithm-level techniques	Healthcare	When compared with other techniques. It outperformed them all in terms of sensitivity, specificity, and F1-score.	ACC, SEN, SPC, F1

TABLE III. SUMMARY OF REVIEWED PAPERS THAT USES HYBRID APPROACH

Authors	Method	Datasets used	Findings	Evaluation Metrics
[50]	SMOTE-ENN with Linear Regression	Kidney	Resampling was able to solve the problem of an imbalanced data structure.	ACC, RC PRC, F1
[51]	SMOTE-ENN with Random Forest	Heart failure	Presents how the SMOTE-ENN algorithm aids ML algorithms perform better on datasets that are imbalanced in terms of	ACC, PRC, F1, RC, ROC_AUC
[52]	TRIM-Smoothed Bootstrap Resampling (TRIM-SBR)	COVID-19	The result shows that the approach surpasses all alternative oversampling techniques in terms of AUC, sensitivity, specificity, and F1-score.	ACC, SEN, SPC, Gmean, F1, AUC
[53]	NCL+SMOTE	Breast Cancer, ILPD, Pima Indians, Fertility, Haberman	NCL+SMOTE technique revealed that all adjusted dataset causes the recall measure to increase.	RC
[16]	M-SMOTE and ENN	Haberman	The experimental findings indicate that RFMSE yields notably better results on 10 UCI datasets compared to alternative sampling algorithms.	F-1, MCC, SEN, SPC

[19]	SMOTE, BLSMOTE, kmUnder, OBU, AdaOBU, BoostOBU	Epilepsy and Parkinson's	OBU show significant improvement in sensitivity and outperformed other methods in terms of F1-score and G-mean.	SEN, SPC, Gmean, F1
[54]	HUSDOS-Boost	Stomach cancer	The findings indicate that HUSDOS-Boost exhibited better performance compared to existing methods for handling imbalanced data, specifically in terms of G-mean and AUC.	SEN, SPC, Gmean, AUC, AUPRC
[55]	EUSBoost	Breast cancer	EUSBoost other methods for solving imbalanced medical problem in terms of sensitivity and AUC.	SEN, Gmean, AUC
[56]	SMOTE, ROS, ADASYN along with two RUS and Near Miss paired with 3 data reduction and cleaning methods namely Tomek Links, One Sided Selection and ENN	PIMA Indians Diabetes Dataset based.	The pairing of ADASYN with One-Sided Selection demonstrates enhanced performance for large-sized datasets, while the combination of ROS with Tomek Link exhibits superior performance for smaller-sized datasets.	

ACC: Accuracy, SEN: Sensitivity, SPC: Specificity, R: Recall, F: F1 score, PRC: precision, MCC: Matthew Correlation coefficient, AUC: Area Under Curve, RMSE: Root Mean Squared Error.

Table III provides a comprehensive overview of the datasets and techniques employed that uses the Hybrid approach in the analyzed articles. The results reveal that the breast cancer dataset was used in 2 articles while various other datasets were the focus of 1 article.

In terms of dataset balancing techniques, the majority of articles (6 in total) employed SMOTE as the oversampling technique together with other undersampling and the algorithm technique. Additionally, 1 article SMOTE, ROS, ADASYN along with two RUS and Near Miss paired with cleaning methods and 3 data reduction namely Tomek Links, One Sided Selection and ENN. One study, each utilized HUSDOS – Boost, EUSBoost, TRIM-Smoothed Bootstrap Resampling (TRIM-SBR), SMOTE, BLSMOTE, kmUnder, OBU, AdaOBU, and BoostOBU.

Regarding performance metrics, the analysis revealed that Recall, F1-score, Gmean and Accuracy were utilized in 3 articles each, while sensitivity and specificity are featured in 7 articles. AUC appeared in 2 articles and AUCPR in 1 article.

VI. Conclusion

This paper aims to examine how resampling strategies in ML have recently been used in medical data sets. Approaches of balancing the imbalanced data were applied to medical datasets spanning through the years 2018 to 2023, based on the selection criteria used for these papers. In addition, the publications that met the survey requirements are shown in Tables I, II, and III, together with the general subject of the application strategy employed in the study.

The contributions of this paper can be summarized as follows:

- The type of databases and the resampling strategies are identified, and it is observed that the resampling strategy generally improves the performance of the prediction.
- The oversampling technique generally performs better in terms of improving performance than the under-sampling techniques.
- Hybrid approaches performed better and are more appropriate to be applied for the resampling of the imbalanced data.

Based on the results of this review, some recommendations are provided for future work as follows:

- There is a need to explore further the capability of algorithm-level techniques as these might help overcome the problem of the data-level technique.
- A limited number of investigations have been conducted in the area of the under-sampling technique. Existing studies have shown that under-sampling technique benefits the resampling of imbalanced medical data.
- There is a need to look into multi-class imbalanced data. When multiple classes are considered, the data imbalance problem becomes considerably more problematic.

References

[1] A. Majid, S. Ali, M. Iqbal, and N. Kausar, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 792–808, 2014.

[2] M. H. Tahan and S. Asadi, "EMDID: Evolutionary multi-objective discretization for imbalanced datasets," *Inf. Sci. (Ny).*, vol. 432, pp. 442–461, 2018.

[3] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," *Inf. Sci. (Ny).*, vol. 384, pp. 174–190, 2017.

[4] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, no. December, pp. 1–11, 2018.

[5] H. Ali, M. Najib, M. Salleh, K. Hussain, and A. Ahmad, "A review on data preprocessing methods for class imbalance problem," vol. 8, no. 3, pp. 390–397, 2019.

[6] E. Hemlata and G. Kaur, "A REVIEW ON DIFFERENT ON DIFFERENT TECHNIQUES FOR BALANCING OF THE IMBALANCED DATA .," vol. 5, no. 5, pp. 521–524, 2018.

[7] S. M. Abd Elrahman and A. Abraham, "Class imbalance problem using a hybrid ensemble approach," *Int. J. Hybrid Intell. Syst.*, vol. 12, pp. 219–227, 2015.

[8] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019.

[9] R. Walsh and M. Tardy, "A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer," *Diagnostics*, vol. 13, no. 1. 2023.

[10] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Softw.*, vol. 13, no. 6, pp. 479–496, 2019.

[11] D. Ramyachitra and P. Manikandan, "IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS : A REVIEW," 2014.

[12] A. Wong and M. S. Kamel, "Classification of imbalanced data : a

review CLASSIFICATION OF IMBALANCED DATA : A REVIEW," no. March 2015, 2011.

[13] H. Patel, D. S. Rajput, G. T. Reddy, C. Iwendi, A. K. Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," vol. 16, no. 4, 2020.

[14] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012077, 2021.

[15] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, no. November 2018, p. 103089, 2019.

[16] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, no. May, p. 103465, 2020.

[17] N. Zhang, "Imbalanced Data Classification Based on Hybrid Methods," pp. 16–20, 2018.

[18] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, p. 58, 2023.

[19] P. Vuttipittayamongkol and E. Elyan, "Improved Overlap-based Undersampling for Imbalanced Dataset Classification with Application to Epilepsy and Parkinson ' s Disease," vol. 30, no. 8, 2020.

[20] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high - class imbalance in big data," *J. Big Data*, 2018.

[21] S. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Handling imbalanced datasets : A review Handling imbalanced datasets : A review," no. May 2014, 2005.

[22] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, no. October 2017, p. 103089, 2019, doi: 10.1016/j.jbi.2018.12.003.

[23] M. Khushi, K. Shaukat, T. M. Alam, X. Yang, and M. C. Reyes, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," vol. 9, 2021.

[24] P. Kazienko and E. Lughofer, "Hybrid and Ensemble Methods in Machine Learning J . UCS Special Issue," vol. 19, no. 4, pp. 457–461, 2013.

[25] A. Garg and V. Mago, "Role of machine learning in medical research : A survey," *Comput. Sci. Rev.*, vol. 40, p. 100370, 2021.

[26] A. Abdellatif and H. Abdellatef, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods," *IEEE Access*, vol. 10, no. July, pp. 79974–79985, 2022.

[27] G. Almahadin, A. Lotfi, M. Mc, and C. Philip, "Enhanced Parkinson ' s Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques," *SN Comput. Sci.*, vol. 3, no. 1, pp. 1–21, 2022.

[28] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Appl. Nanosci.*, no. 0123456789, 2022.

[29] V. Kumar, G. S. Lalotra, P. Sasikala, and D. S. Rajput, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," pp. 1–28, 2022.

[30] Z. Jiang and W. Zhao, "Seizure : European Journal of Epilepsy Fusion Algorithm for Imbalanced EEG Data Processing in Seizure Detection," *Seizure Eur. J. Epilepsy*, vol. 91, no. June, pp. 207–211, 2021.

[31] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," vol. 2021, 2021.

[32] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, no. August, pp. 109960–109975, 2021.

[33] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE : Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Inf. Sci. (Ny).*, vol. 542, pp. 92–111, 2021.

[34] A. Ishaq, S. Sadiq, M. Umer, and S. Ullah, "Improving the Prediction of Heart Failure Patients ' Survival Using SMOTE and Effective Data Mining Techniques," pp. 39707–39716, 2021.

[35] G. Varotto, G. Susi, L. Tassi, F. Gozzo, S. Franceschetti, and F. Panzica, "Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning : Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy," vol. 15, no. November, pp. 1–21, 2021, doi: 10.3389/fninf.2021.715421.

[36] V. Pavan, K. Turlapati, and M. Ranjan, "Intelligence-Based Medicine Outlier-SMOTE : A re fi ned oversampling technique for improved detection of COVID-19," *Intell. Med.*, vol. 3–4, no. November, p. 100023, 2020, doi: 10.1016/j.ibmed.2020.100023.

[37] N. Latif, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM : An Effective Heart Disease Prediction Model for a Clinical Decision Support System," vol. 8, 2020.

[38] D. Barbieri, N. Chawla, L. Zaccagni, and M. Coklo, "Predicting Cardiovascular Risk in Athletes : Resampling Improves Classification Performance," 2020.

[39] K. Polat, "A Hybrid Approach to Parkinson Disease Classification using speech signal : The combination of SMOTE and Random Forests," *2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci.*, pp. 1–3, 2019.

[40] J. Leo, E. Luhanga, and K. Michael, "Machine Learning Model for Imbalanced Cholera Dataset in Tanzania," *Sci. World J.*, vol. 2019, 2019.

[41] A. M. Campbell, M. F. Racault, S. Goult, and A. Laurenson, "Cholera risk: A machine learning approach applied to essential climate variables," *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, pp. 1–24, Dec. 2020.

[42] S. Haldar, S. Banerjee, R. Mukherjee, S. Chaudhury, P. Chakraborty, and S. Chatterjee, "Improved Epilepsy Detection method by addressing Class Imbalance Problem," pp. 934–939, 2018.

[43] Y. Zhao, Z. S. Wong, and K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events ' Classification : A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," vol. 2018, no. 2010, 2018.

[44] P. Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron," 2017.

[45] N. H. Chau, "Enhancing Cholera Outbreaks Prediction Performance in Hanoi, Vietnam Using Solar Terms and Resampling Data," 2017, vol. 10448 LNAI, pp. 266–276.

[46] N. Khateeb and M. Usman, "Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique," pp. 21–26, 2017.

[47] K. Wang, A. Melani, K. Chen, and K. Wang, "A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer : A case study in Taiwan," *Comput. Methods Programs Biomed.*, vol. 119, no. 2, pp. 63–76, 2015.

[48] M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Appl. Nanosci.*, vol. 13, Feb. 2022, doi: 10.1007/s13204-021-02063-4.

[49] A. S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arab. J. Sci. Eng.*, vol. 46, pp. 3853–3864, 2021.

[50] X. Shi, T. Qu, G. Van Pottelbergh, M. Van Den Akker, and B. De Moor, "A Resampling Method to Improve the Prognostic Model of End-Stage Kidney Disease : A Better Strategy for Imbalanced Data," vol. 9, no. March, pp. 1–9, 2022.

[51] M. M. Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," vol. 2022, no. Cvd, 2022.

[52] P. Wibowo and C. Fatichah, "Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of Covid-19," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021.

[53] N. Junsomboon and T. Phienthrakul, "Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset," no. 1, pp. 243–247.

[54] K. Fujiwara *et al.*, "Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis," *Front. Public Heal.*, vol. 8, no. May, pp. 1–15, 2020.

[55] B. M. G. Ł. J. F. Krawczyk, "Evolutionary Undersampling Boosting for Imbalanced Classification of Breast Cancer Malignancy," 2015.

[56] D. A. Ofori *et al.*, "Forecast and prediction of COVID-19 using machine learning," *Molecules*, vol. 2, no. 1, pp. 1–12, 2020.