UMYU Scientifica, Vol. 2 NO. 1, March 2023, Pp 76– 87
https://doi.org/10.56919/usci.2123.011
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Comparative analysis of resampling algorithms in the prediction of stroke diseases

Abdullahi Sani Dauda[1]*, Muhammad Sirajo Aliyu[2], Musa Usman Abdullahi [3]
[1] Department of Computer Science, Federal University of Kashere, Gombe, Nigeria
[2] Department of Cyber Security, Federal university Dutse, Dutse, Nigeria
[3] Department of Computer Science, Federal university Dutse, Dutse, Nigeria

ABSTRACT
Stroke disease is a serious cause of death globally. Early predictions of the disease will save a lot of lives, but most of the clinical datasets are imbalanced in nature, including the stroke dataset, making the predictive algorithms biased towards the majority class. The research aims to compare different data resampling algorithms on the stroke dataset to improve the prediction performances of the machine learning models. This paper considered five (5) resampling algorithms, namely, Random over Sampling (ROS), Synthetic Minority oversampling Technique (SMOTE), Adaptive Synthetic (ADASYN), hybrid techniques like SMOTE with Edited Nearest Neighbor (SMOTE-ENN) and SMOTE with Tomek Links (SMOTE-TOMEK). The datasets are trained on six (6) machine learning classifiers, namely, Logistic Regression (LR), Decision Tree (DT), K-nearest Neighbor (KNN), Support Vector Machines (SVM), Random Forest (RF), and XGBoost (XGB). The hybrid technique SMOTE-ENN influences the best machine learning classifiers, followed by the SMOTE technique, while the combination of SMOTE and XGB performs better with an accuracy of 97.99% and G-mean score of 0.99, and auc_roc score of 0.99. Resampling algorithms balance the dataset and enhance machine learning algorithms' predictive power. Therefore, we recommend resampling the stroke dataset in predicting stroke disease than modeling on the imbalanced dataset.

ARTICLE HISTORY
Received February 5, 2023
Accepted March 15, 2023
Published March 30, 2023

KEYWORDS
Stroke, imbalanced data, resampling algorithms, machine learning classifiers, SMOTE.

INTRODUCTION
Stroke disease is caused by the breakage of the blood vessel or the presence of a blood clot in the blood vessel, resulting in an insufficient supply of nutrients to the brain (Donnan *et al.,* 2008). Studies show that several risk factors contribute to stroke disease and are preventable (Feigin *et al.*, 2016); usually, they are grouped into modifiable risk factors and unmodifiable risk factors (Markus & Brainin, 2020). The clinical dataset is imbalanced in nature, and so also the stroke datasets, where the distribution of the minority class is lower than the majority class; this has profound effects on the algorithms to be used, as the algorithms will pay more attention to the majority class during training (Wu & Fang, 2020). Researchers resolve this problem either at the data preprocessing level or classifier level. At the data preprocessing level, we can under-sample the majority class to have the same distribution ratio as the minority class or oversample the minority class to have the same distribution ratio as the majority class, thereby creating more synthetic data (Sridharan *et al.*, 2021). However, during the data sampling phase, there is no right sampling technique to use, but having a technique that samples the dataset efficiently is the optimal goal.

Many researchers have contributed to applying various sampling algorithms and machine learning models to predict stroke. The work of Ahmed *et al.*, (2019) proposed distributed machine learning algorithms using apache spark to predict stroke, and a random oversampling technique was used to sample the dataset. Cross-validation and hyperparameter tuning were applied to the logistic regression, decision tree, random forest, and support vector machine. Among all, random forest is the best model with an accuracy of 90%. Applying the Synthetic Minority oversampling Technique (SMOTE) on the dataset to sample the data by creating synthetic data points from the minority class to level up with the majority class and reducing the feature subspace will help improve the model performance and also take less time to train the model (Ray *et al.*, 2020). Statistical methods like chi-squared were used and reduced the feature space to six (6): age, heart_disease, average_glucose_level, hypertension, work_type, and ever_married with a performance accuracy of 97.6% using two-class Boosted Decision Tree as shown by Ray *et al.*, (2020). Sailasya & Kumari, (2021) have employed an under-sampling technique that works by reducing the ratio of the majority class to label with the minority class, making the dataset balanced but facing the

Correspondence: Abdullahi S. D. Department of Computer Science, Federal University of Kashere, Gombe, Nigeria.
✉ daudasaniaa008@gmail.com ; +2347033398707
How to cite: Abdullahi Sani Dauda, Muhammad Sirajo Aliyu, Musa Usman. (2023). Comparative analysis of resampling algorithms in the prediction of stroke diseases. UMYU Scientifica, 2(1), 76 – 87. https://doi.org/10.56919/usci.2123.011

issue of insufficient data to train the models. Decision trees, logistic regression, random forest, k-nearest neighbor, support vector machine, and Naïve Bayes classification algorithms are designed and compared. Out of all the algorithms used, Naïve Bayes performs best with an accuracy of 82% which is less compared to other literature. In recent studies, AdaBoost and gradient boosting algorithms along with eight traditional methods like decision tree, logistic regression, multi-layer perceptron, k-nearest neighbor, random forest, and naïve Bayes are used in predicting stroke, a random oversampling technique is used, and a web page and mobile application are developed to calculate the result of the prediction based on real-time inputs. Random forest, and support vector machine performed the best with an accuracy of 99.87% and 99.99% respectively (Biswas *et al.*, 2022). Using a hybrid sampling technique like in the work of Abdullahi & Muhammad, (2022) has shown significant performance in predicting stroke using xgboost, lightgbm, catboost, and adaboost classifiers. The highest performance reach is by catboost with 99.7% accuracy and 0.99 auc_roc score. None of the pieces of literature compare the influences of these sampling techniques on the model performance. In this work, we examine the different data resampling techniques with conjunction of engineering features like age, bmi, and average_glucose_level to improve the predictive performance of stroke prediction.

Therefore, this study aims to compare various resampling techniques in balancing stroke datasets and examine their effects in predicting stroke disease using different machine learning algorithms.

MATERIALS AND METHODS

In this study, we compare various data resampling algorithms in conjunction with machine learning algorithms to predict stroke disease. The following are the steps involved in achieving the objective of the study.

Data collection

The dataset is obtained from kaggle open-source data repository (Fedesoriano, 2021). It comprises of 10 stroke risk factors and target output, which signifies stroke or no stroke (Alberto & Rodríguez, 2021; Emon *et al.,* 2020; Sailasya & Kumari, 2021). Detail description of the stroke risk factors is given in Table 1.

Table 1: Description of the stroke risk factors

S/N	Variable Name	Data Type & Value	Descriptive Statistics
1	**gender**	Object, ['male', 'female', 'other']	Counts: 5110 Male: 2994, female: 2115, other: 1
2	**age**	Float, in years/months	Counts: 5110, min: 0.08, max: 82.0, mean: 43.2, std: 22.6
3	**hypertension**	Integer, [0, 1]	Counts: 5110, 'No hypertension (0)': 4612, 'hypertension (1)': 498
4	**heart_disease**	Integer, [0,1]	Counts: 5110, 'No heart Disease (0)': 4834, 'Heart Disease (1)': 276
5	**ever_married**	Object, ['Yes', 'No']	Counts: 5110, 'Yes': 3353, 'No': 1757
6	**work_type**	Object, ['private', 'Self-employed', 'children', 'Govt_job', 'Never_worked']	Counts: 5110, private: 2925, Self-employed: 819, children: 687, Govt_job: 657, Never_worked: 22
7	**Residence_type**	Object, ['Rural', 'Urban']	Counts: 5110, Rural: 2514, Urban: 2596
8	**Avg_glucose_level**	Float	Counts: 5110, min: 55.12, max: 97.6, mean: 106.14, std: 45.28
9	**bmi**	Float	Counts: 4909, min: 10.30, max: 97.60, mean: 28.89, std: 7.85
10	**smoking_status**	Object, ['Smoked', 'Never_smoked', 'formerly smoked', 'Unknown']	Counts: 5110, Never_smoked: 1892, Unknown: 1544, formerly smoked: 885, smokes: 789
11	**stroke**	Integer, [0,1]	Counts: 5110, 'No Stroke (0)':4861, 'Yes Stroke (1)': 249

Data preparation

Each feature is transformed to algorithm-based form in terms of numeric attributes, the followings involve steps taken in preparing the dataset.

i. Handling Missing Value: The stroke dataset has only one variable with a missing value which is the bmi variable. Firstly, we employ the KNNImputer algorithm to fit the data and replace the missing values. Secondly, the smoking_status variable has a value of 'Unknown' which can signify a null value; we applied the statistical mode of the 'smoking_status' variable to replace the 'Unknown'.

ii. Discretization: Discretization of features involved creating categorical features from numerical features. We feature engineered age, bmi, and avg_glucose_level risk factors to create other categorical variables, namely; age_cat with values [children, teen, adults, mid-adults, elderly], bmi_cat with values [underweight, ideal, overweight, obesity], and avg_glucose_level_cat with values [Very Low, Low, Normal, High, Very High].
iii. Handling outliers: outliers affect the performance of predictive models hence the need for handling them can improve model performance. Here, we handle outliers in bmi and avg_glucose_level, using the interquartile range method, data points greater than the threshold (1.5*IQR + Q3) is an outliers, so also any data point less than (Q1 - 1.5*IQR) is an outlier. Where IQR stands for interquartile range and is defined as IQR = Q3 – Q1, while Q1 and Q3 stand for the first quartile and third quartile, respectively.
iv. Label Encoding: This involves converting discrete features with values other than numbers to contain a numerical value. We converted all categorical features into numerical values using LabelEncoder.

Data resampling

Data resampling involves the use of algorithms and techniques to handle class imbalances in datasets. In this study, we use the following resampling algorithms and compare their effect on machine learning algorithms.

i. Random Over Sampling (ROS): it is an over-sampling algorithm that randomly selects instances from the minority class with replacement, and adds to the training set until the class ratio is balanced. Duplication of instances can cause overfitting (He & Garcia, 2009).
ii. Synthetic Minority Oversampling Technique (SMOTE): it is an over-sampling algorithm that creates synthetic data from the minority class. It does that by interpolating between samples of the same target class and creating a new instance in between. The newly created instance is different from the existing instances of the minority class. Though it has shown good performance in balancing the data, it can introduce noise in the data (Chawla *et al.*, 2002)
iii. Adaptive Synthetic (ADASYN): it is an over-sampling technique proposed by (He *et al.*, 2008) It creates synthetic data by using different weighted distributions for the minority class instances based on the difficulty of learning, thereby, generating more synthetic data for instances that are difficult to learn.
iv. Synthetic Minority Oversampling Technique with Edited Nearest Neighbor (SMOTE-ENN): It comprises under-sampling and over-sampling Techniques. The SMOTE is an over-sampling Technique that creates synthetic data and balances the distribution while the ENN is an under-sampling technique that performs the task of removing instances from overlapping regions. Here, it removes misclassified instances from both minority and majority classes (Lamari *et al.*, 2021).
v. Synthetic Minority Oversampling Technique with (SMOTE-TOMEK) comprises under-sampling and over-sampling Techniques. The SMOTE is an over-sampling Technique that creates synthetic data and balances the distribution while the TOMEK removes data instances from the majority class that has minimum Euclidean distance from the minority class data instances (More, 2016).

Model training

The dataset is further divided into training and validation sets using the ratio 80:20. A 10-fold cross-validation was used to prevent overfitting. We used six (6) machine learning algorithms to train on the 80 ratios and validate it on the 20 ratios. The algorithms are; logistic regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Extreme gradient Boosting (XGBoost). The model's hyperparameter setting is given in Table 2.

Table 2: machine learning classifiers and their parameter settings

MODEL	PARAMETER SETTINGS
LR	C: 1.0, solver: 'lbfgs', penalty: l2
K-NN	K=6, p=2
DT	max_depth: 10, min_samples_leaf: 20, criterion: 'entropy'
SVM	C: 1, gamma: 'scale', kernel: 'rbf, degree: 3
RF	n_estimators: 35, max_depth: 5, max_features: 'auto', min_samples_split: 2, min_samples_leaf: 1
XGB	Default hyperparameters

Evaluation metrics

In this study, we used only three metrics of measure to compare the algorithms and they are:

i. Accuracy: Accuracy measures how often a classifier correctly predicts. It is the ratio of the number of correct predictions to the total number of predictions (Jason, 2020). Mathematically defined in Eq. (1);

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

ii. auc_roc score: The AUC-ROC Score stands for Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC), it is the area under the probability curve that plots True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values (Sun *et al.*, 2009). It is best suited for imbalanced data problems.
iii. Geometric mean score: This measure tries to maximize the accuracy on positive and negative instances while balancing these accuracies. For binary classification, G-mean is the squared root of the product of the sensitivity and specificity. It is an excellent metric to use when the data is been

resampled. And it is mathematically broken down in Eq. (2).

$$S = \sqrt{sensitivity * specificity} \qquad (2)$$

$$where\ sensitivity\ (true\ positive\ rate) = \frac{TP}{TP + FN}$$

$$specificity\ (true\ negative\ rate) = \frac{TN}{TN + FP}$$

$TP = True\ Positive,$
$TN = True\ Negative,$
$FP = False\ Negative,$
$and\ FN = False\ Negative$

RESULT

Resampling algorithms result

The collected stroke dataset is highly imbalanced with only 5% of the minority (1) class and 95% of the majority (0) class as are given in Figure 1. While other distributions of the resampling algorithms is given in Figure 2 for ROS where both classes are 50% each. Figure 3 for SMOTE where both classes are 50% each. Figure 4 for ADASYN where class (1) is 50.1% of the total dataset and class (0) is 49.9%. Figure 5 for SMOTE-ENN where class (1) is 54.0% of the total dataset and class (0) is 46.0%., and Figure 6 for SMOTE-TOMEK with 50% each, for both classes.

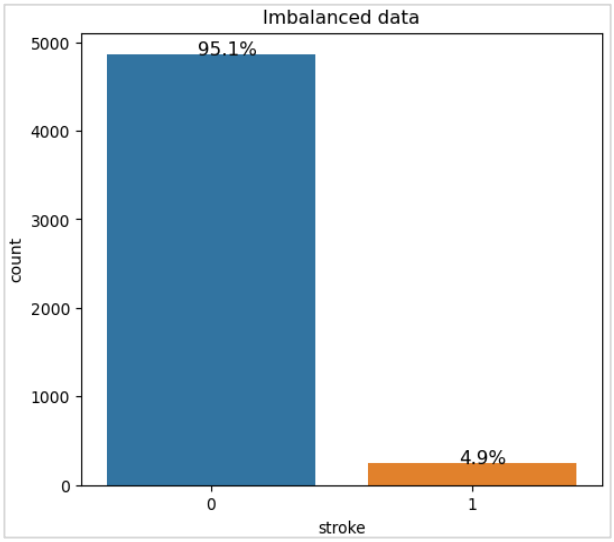

Figure 1: Imbalanced stroke dataset (original dataset)

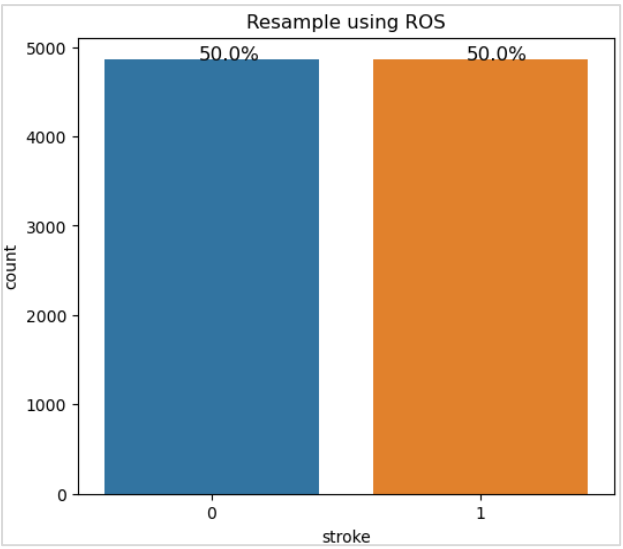

Figure 2: resampled dataset using ROS

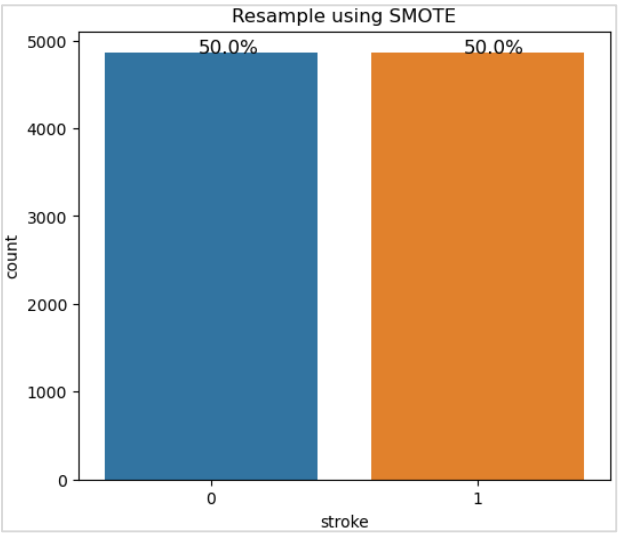

Figure 3: resampled dataset using SMOTE

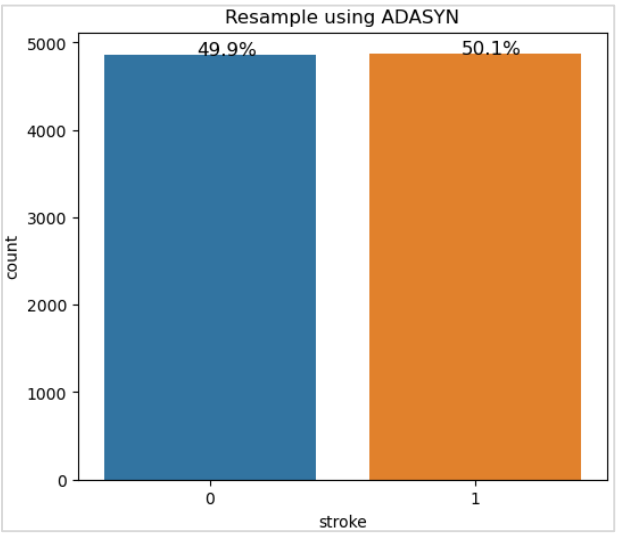

Figure 4: resampled dataset using ADASYN

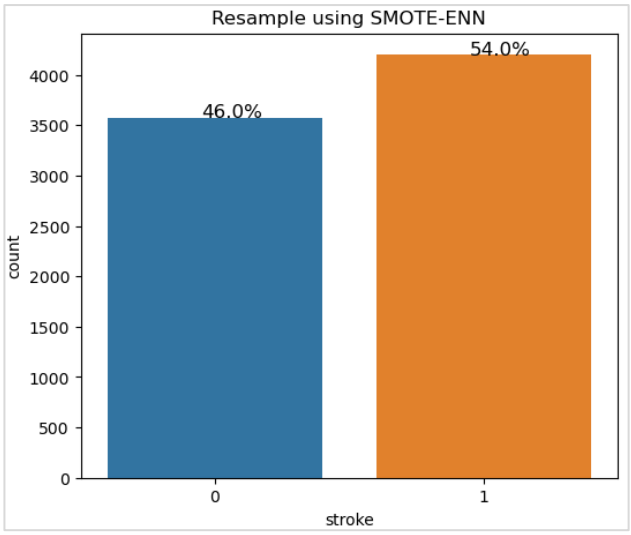

Figure 5: resampled dataset using SMOTE-ENN

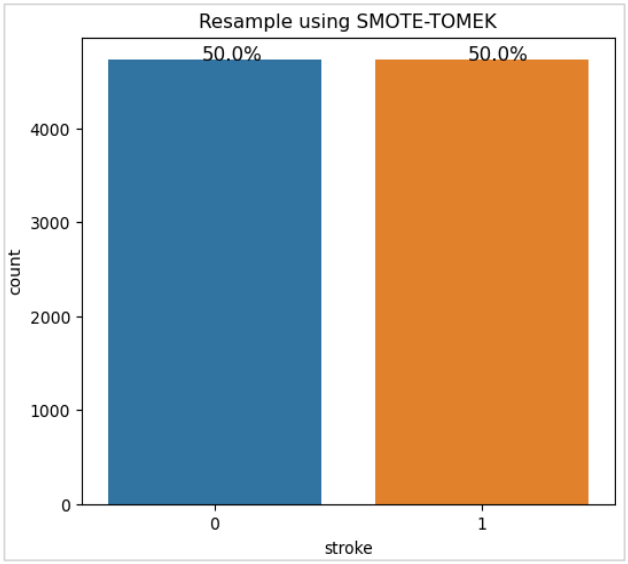

Figure 6: resampled dataset using SMOTE-TOMEK

Analysis of the resampling algorithms based on the time it takes to balance the dataset in seconds is given in Table 3.

Table 3: Execution time of the Resampling algorithms in seconds

Resampling algorithm	Execution time (sec)
ROS	0.0160
SMOTE	0.0320
ADASYN	0.1080
SMOTE-ENN	2.5622
SMOTE-TOMEK	5.5557

The result shows that hybrid algorithms like SMOTE-TOMEK and SMOTE-ENN took more time to balance the data than the individual oversampling algorithms.

Machine learning classifiers result

After applying the resampling algorithm, we trained six (6) machine learning classifiers on the resample data, and their result is given in table 4 and table 5. Three (3) evaluation metrics namely; accuracy, G-mean scores, and auc_roc score are employed to compare between this classifiers and know the model's performance.

Table 4: performance result of the machine learning classifiers on the imbalanced dataset

Model	Accuracy (%)	G-mean score	auc_roc score
LR	95.01	0.00	0.49
KNN	94.72	0.14	0.51
DT	90.80	0.31	0.52
SVM	95.12	0.00	0.50
RF	95.01	0.14	0.51
XGB	94.52	0.28	0.53

Table 5: performance result of the machine learning classifiers on the resample dataset

Model	Evaluation metrics	ROS	SMOTE	ADASYN	SMOTE-ENN	SMOTE-TOMEK
LR	Accuracy (%)	79.64	92.23	92.39	94.86	92.65
	G-Mean score	0.79	0.92	0.92	0.95	0.93
	Auc_roc score	0.80	0.92	0.92	0.95	0.93
KNN	Accuracy (%)	93.21	95.58	94.92	95.18	94.98
	G-Mean score	0.93	0.96	0.95	0.95	0.95
	Auc_roc score	0.93	0.96	0.95	0.95	0.95
DT	Accuracy (%)	97.79	94.36	94.92	94.92	95.14
	G-Mean score	0.98	0.94	0.95	0.95	0.95
	Auc_roc score	0.98	0.94	0.95	0.95	0.95
SVM	Accuracy (%)	86.63	94.19	94.56	95.12	93.76
	G-Mean score	0.86	0.94	0.95	0.95	0.94
	Auc_roc score	0.86	0.94	0.95	0.95	0.94
RF	Accuracy (%)	99.28	96.86	96.92	96.98	96.88
	G-Mean score	0.99	0.97	0.97	0.97	0.97
	Auc_roc score	0.97	0.97	0.97	0.97	0.97
XGB	Accuracy (%)	96.99	97.99	97.28	97.55	97.20
	G-Mean score	0.97	0.98	0.97	0.98	0.97
	Auc_roc score	0.97	0.98	0.97	0.98	0.97

LR – Logistic regression, KNN – K-Nearest Neighbor, DT – Decision Tree, SVM – Support Vector Machine, RF – Random Forest, XGB – XGBoost

After we have trained the classifiers on imbalanced data we can see that all of them perform well based on the accuracy. But one of the issues with the accuracy metric is that if the dataset is not balanced, it pays more attention to the majority class thereby making it seems well but is not. In this case, we look at the auc_roc score, which clearly shows that our classifiers barely performed above average. This is the worst performance, especially logistic regression. So, resampling the data could solve the problem.

From Table 5, the result of the classifiers trained on the resample data shows good results as both the accuracy and auc_roc score are encouraging.

DISCUSSION

The study aims to compare the influence of various resampling algorithms in predicting stroke diseases using different machine learning classifiers. Accuracy, G-mean score, and auc_roc score metrics are utilized in the comparison.
In Table IV, all the machine learning classifiers perform worst as they are not better than a random guess classifier, which signifies the need to handle the dataset's imbalanced nature.
In Table V, all the resampling algorithms perform well in balancing the dataset; the G-mean score has proven that. The hybrid resampling algorithms, especially SMOTE-ENN influence the most in balancing the data because it makes the machine learning classifiers predict well with good accuracy score, G-mean score, and auc_roc score.
The combination of SMOTE with XGBoost gives us the best result followed by SMOTE-ENN with XGBoost in terms of accuracy score, G-mean Score, and auc_roc score.

CONCLUSION

In conclusion, the measure of accuracy score on imbalanced data is misleading and machine learning classifiers performed above average as measured using the auc_roc score. The result shows that the machine learning classifiers performed well on all the resampling algorithms with a hybrid technique like SMOTE-ENN performing the best in terms of the entire machine learning classifiers' result. The combination of SMOTE and XGBoost produced the best result in predicting stroke disease.

REFERENCES

Abdullahi, S. D., & Muhammad, S. A. (2022). Early Prediction of Cerebrovascular Disease using Boosting Machine Learning Algorithms to Assist Clinicians. *Journal of Applied Sciences and Environmental Management*, *26*(6), 1031–1037. **[Crossref]**

Ahmed, H., Abd-El Ghany, S. F., Youn, E. M. G., Omran, N. F., & Ali, A. A. (2019). Stroke prediction using distributed machine learning based on apache spark. *International Journal of Advanced Science and Technology*, *28*(15), 89–97. **[Crossref]**

Alberto, J., & Rodríguez, T. (2021). *Stroke prediction through Data Science and Machine Learning Algorithms*. *Ml*.

Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. *Healthcare Analytics*, *2*(October), 100116. **[Crossref]**

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357. **[Crossref]**

Donnan, G. A., Fisher, M., Macleod, M., & Davis, S. M. (2008). Stroke. *The Lancet*, *371*(9624), 1612–1623. **[Crossref]**

Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Mamun, M. S. Al, & Kaiser, M. S. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 1464–1469. **[Crossref]**

Fedesoriano. (2021). *Stroke prediction dataset*. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., Mensah, G. A., Norrving, B., Shiue, I., Ng, M., Estep, K., Cercy, K., Murray, C. J. L., & Forouzanfar, M. H. (2016). Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet Neurology*, *15*(9), 913–924. **[Crossref]**

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1322–1328. **[Crossref]**

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. **[Crossref]**

Jason, B. (2020). *Tour of Evaluation Metrics for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Lamari, M., Azizi, N., Hammami, N. E., Boukhamla, A., Cheriguene, S., Dendani, N., & Benzebouchi, N. E. (2021). SMOTE--ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for

Imbalanced Medical Data Classification. In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Eds.), *Advances on Smart and Soft Computing* (pp. 37–49). Springer Singapore.

Markus, H. S., & Brainin, M. (2020). COVID-19 and stroke—A global World Stroke Organization perspective. *International Journal of Stroke*, *15*(4), 361–364. **[Crossref]**

More, A. (2016). *Survey of resampling techniques for improving classification performance in unbalanced datasets*. *10000*, 1–7. **[Crossref]**

Ray, S., Alshouiliy, K., Roy, A., Alghamdi, A., & Agrawal, D. P. (2020). Chi-Squared Based Feature Selection for Stroke Prediction using AzureML. *2020 Intermountain Engineering, Technology and Computing, IETC 2020*. **[Crossref]**

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 539–545. **[Crossref]**

Sridharan, M., Mantyla, M., Rantala, L., & Claes, M. (2021). Data balancing improves self-admitted technical debt detection. *Proceedings - 2021 IEEE/ACM 18th International Conference on Mining Software Repositories, MSR 2021*, 358–368. **[Crossref]**

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719. **[Crossref]**

Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older chinese. *International Journal of Environmental Research and Public Health*, *17*(6), 1–11. **[Crossref]**