JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024
Research Article
https://doi.org/10.56979/702/2024

Improving Stroke Prediction Accuracy through Machine Learning and Synthetic Minority Over-sampling

Muhammad Abdullah Aish[1*], Amina Abdul Ghafoor[1], Fawad Nasim[2], Kiran Irfan Ali[3], Shamim Akhter[4], and Sumbul Azeem[5]

[1]Department of Software Engineering, The Superior University, Lahore, 54000, Pakistan.
[2]Faculty of Computer Science and Information Technology, The Superior University, Lahore, 54600, Pakistan.
[3]Aesthetics Lab Powered by Tibbi, Lahore, 54782, Pakistan.
[4]School of Information Management, Minhaj University, Lahore, 54000, Pakistan.
[5]Lahore College for Women University, Jail Road, Lahore, 54000, Pakistan.
[*]Corresponding Author: Muhammad Abdullah Aish. Email: abdullahaish7@gmail.com

Received: March 12, 2024 **Accepted:** August 12, 2024 **Published:** September 01, 2024

Abstract: Strokes are a leading cause of death and disability worldwide. Accurate prediction and early intervention can significantly improve patient outcomes. The objective of this study is to develop a model that will effectively predict stroke events based on the application of machine learning methods using a Harvard Dataverse Repository dataset containing 43,400 samples with 10 features. The dataset was imbalanced, with 42,617 non-stroke cases versus 783 stroke cases; hence, SMOTE was applied to balance the dataset. Models were evaluated using "accuracy, precision, recall, F1-score, and ROC AUC". ML models included "logistic regression, decision tree, random forest, gradient boosting, adaboost, XGBoost, support vector machine, k-nearest neighbors, Naive Bayes, bagging classifier, and voting classifier". The best model was that of the Bagging Classifier at an accuracy of 98.3%, precision of 98.7%, recall of 98.0%, an F1-score of 98.3%, and a ROC AUC of 99.5%. Then, it proved the robustness and reliability of this model. The current research demonstrates the power of SMOTE in solving class imbalance and underlines the possible role of advanced machine learning techniques in building feasible predictive tools for detecting stroke incidents in their incipient stage. Improvements such as these in the field may have a significant effect on bettering patient outcomes and reducing burdens on healthcare. Moreover, the implementation of such predictive models within clinical workflows could enable timely medical interventions, hence improving the quality of care for those people who are at risk of stroke. The work also opens up a variety of possibilities for deep learning and other sophisticated machine-learning techniques in healthcare, underlining the fact that further innovating and developing this area is necessary.

Keywords: Bagging Classifier; Early Detection; Healthcare Analytics; Imbalanced Data; Machine Learning; SMOTE; Stroke Prediction.

1. Introduction

Stroke has become the primary cause of both disability and mortality globally. It is estimated to have far-reaching effects on the quality of life for millions worldwide in years to come. At present, stroke has become one of the major causes of death worldwide, affecting both sexes of most age groups. This imposes a significant strain on public health systems, with projections indicating that over 200 million Disability-Adjusted Life Years (DALYs) will be lost to stroke annually by 2030. Additionally, nearly 70 million elderly individuals and approximately 12 million deaths will be attributed to strokes alone [1]. All this presents a grim prediction, pressing for an effective strategy for stroke prediction and prevention.

Cerebral strokes occur when there is an interruption or decrease in blood flow to the brain, leading to a serious neurological incident where parts of the brain are deprived of oxygen and nutrients. There are

two major stroke types: ischemic, which is caused by an obstruction or narrowing of the blood flow; and hemorrhagic, due to the rupture of a blood vessel in the brain [2]. Most significantly, because strokes are of high prevalence and serious, robust predictive models should be developed to identify those at risk and facilitate early interventions.

The factors that characterize the modern way of life, such as high glucose levels, heart diseases, overweight, and obesity, and finally, diabetes, multiply the risks of stroke events [3]. All these factors interact with genetic predispositions and environmental influences, hence further complicating the task of predicting strokes. Traditional statistical methods are not usually able to capture properly such intricate interactions between variables and thus require more advanced methodologies for analysis [4].

ML techniques have returned as strong tools in the prediction of a host of diseases, including stroke. These new techniques can handle large amounts of data with intricate patterns that are difficult to determine using traditional methods only [5]. However, class imbalance alone can significantly reduce the performance of these models of ML when the number of non-stroke cases is way higher compared to that of stroke cases. This could also lead to biased models toward the majority class and failureof the minority class—of greater clinical importance—in performance [6].

For that problem, many techniques of oversampling have been applied to rebalance the dataset before training the models, which includes Synthetic Minority Oversampling Technique (SMOTE). SMOTE oversamples the minority class through the creation of synthetic samples by interpolation between existing samples, enabling the model to learn from both classes more effectively. This technique is extremely useful in medical datasets where conditions like stroke are comparatively rare with respect to healthy individuals.

This research mainly focuses on identifying which ML models are sound [7, 8] at predicting strokes when they are trained on a balanced dataset. In the process, we contrast various models to be able to settle on the best option for predicting stroke occurrences. The evaluation will help in bringing out the advantages and disadvantages of each model and, at the same time, give insights into how balancing techniques affect model performance.

Ultimately, this leads to the building of more accurate and sure tools for prediction, which may abet the timely detection and intervention for improved patient outcomes. Early and accurate prediction of stroke will give the chance for timely medical interventions that may reduce the seriousness of stroke outcomes and leave a better quality of life in people at risk. Further, the purpose of this study can be focused on the advancement of stroke prediction by using advanced ML techniques and addressing class imbalance.

2. Literature Review

One of the high-impact research areas in healthcare has been the use of machine learning techniques for the prediction of medical conditions, such as stroke. The current literature review discusses some of the studies that used machine learning algorithms for the purpose of predicting stroke, with a focus on how such studies handled issues arising from class imbalance and assessed model performance.

Other studies have also explored the application of ML models to stroke risk prediction. Dritsas and Trigka, 2022, have surveyed some ML techniques, including “logistic regression, decision trees, and neural networks”, for the prediction of stroke risk from clinical and demographic data [9]. This paper demonstrated that by using ML, complicated relationships among risk factors could be modeled with high accuracy since data preprocessing and feature engineering turned out to be a precondition for such models.

Teoh used EHRs to build models in stroke prediction back in 2018 [10]. This is consistent with the argument of the completeness and size of the dataset, which claims that using EHRs in this instance would help improve the model's accuracy. The use of EHRs provides variables of different dimensions in building models that would pinpoint more accurately the people at risk. However, another major challenge it highlighted is that data imbalance leads to biased predictions.

One of the major challenges to stroke prediction is that all the datasets are highly imbalanced, with non-stroke cases being radically more than those of stroke. It has already been illustrated that this has a negative effect on the performance of ML models and always leads to low sensitivity in the detection of stroke cases. These have been handled using different techniques of oversampling.

Gosain and Sardana evaluated several oversampling methods, including the Synthetic Minority Oversampling Technique, Adaptive Synthetic Sampling, and the Random Over-Sampling Technique,

among others [11]. In particular, SMOTE has been positioned as one of the approaches with the greatest success to this date in the generation of synthetic samples from the minority class through interpolation between examples that already exist.

Park et al. (2020) showcased the practical implementation of SMOTE in a real-time system for monitoring gait to predict strokes [12]. Class balancing improved their machine learning models, obtaining higher accuracy and sensitivity. This study therefore demonstrated the practical benefits of the use of SMOTE in healthcare applications.

Advanced techniques of ML with complete datasets and real-time monitoring systems form the future of stroke prediction. Park et al., 2020, said that ML algorithms could be combined with wearable devices for health monitoring in patients on a continuous basis and can predict the risk of stroke in them in real-time [13]. Most of the systems will alert timely to facilitate early interventions which would help to reduce the incidence and severity of strokes.

Rahman and Hasan conducted a study using various machine learning and deep learning models to predict stroke risk with a dataset sourced from Kaggle. Their research indicated that traditional machine learning models, particularly ensemble methods like Random Forest, outperformed deep neural networks in classification tasks. This study highlights the potential of ensemble learning techniques in medical predictions, especially when dealing with structured data and imbalanced datasets [14].

Furthermore, hybrid models—by combining several machine learning algorithms with data-balancing techniques—may help further improve the model for stroke prediction. Another new hybrid approach suggested in this field of stroke prediction by Mia et al. (2024) is a combination of SMOTE with random forests; very impressive results have been returned [15]. This clearly shows how hybrids can exploit the different skills of a great many technologies to provide superior performance. Hybrid models can offer an exact solution by compensating for the inadequacies of individual models.

Other than the hybrid models, there is a growing interest in applying deep learning to stroke prediction. Deep learning models, primarily CNN and RNN, have been solid in processing complex medical data composed of imaging and sequential data, among others. These can automatically extract relevant features from the raw data, hence trying to increase the accuracy and efficiency of systems predicting stroke events.

Ethical and legal considerations come into play in the actual clinical application of ML models, which touches on the requirement for patient privacy and data security. Additionally, the models should be transparent and interpretable in order to be trusted when used by health professionals and patients alike. In this regard, the development of such explainable AI techniques would matter. Such efforts make it possible for a clinician to understand and support predictions made by a machine learning model. On another note, unproblematic integration of the ML models into existing healthcare workflows ensures that they are accepted by the masses, besides providing sufficient training to the healthcare providers on how to use such tools most effectively. Continual automated monitoring and verification of ML models in the real world also generally need to be facilitated in order to further ensure such models remain accurate and dependable across the settings and time. Addressing these challenges in partnership with health providers and policymakers can help enforce ethical implementation of ML in clinical practice. In so doing, we can harness the full benefits ML affords in improving patient outcomes and the overall progress in stroke prediction. The development of a sustainable, trust-based framework, in the end, will enhance caregiving quality via the responsible use of machine learning technology. With this, we might set the scene for innovation that not only predicts but forestalls a stroke, hence reducing its burden across the world.

The literature thus identifies the massive strides made in stroke prediction using ML, especially with the involvement of data-balancing techniques such as SMOTE. The challenges of advanced data quality and real-time applicability that this work presents to the field of stroke prediction also bring with them future prospects in this area. It will therefore be of substantive help in the delivery of precise and reliable predictive tools pertaining to lessened patient loss and reduced burden of stroke on the globe if continuous research and development are observed in this line. Moreover, a real-time monitoring system integrated with advanced ML algorithms can very likely revolutionize stroke prediction by providing opportune interventions. Henceforth, explainable AI techniques will assume paramount importance in gaining the trust from healthcare professionals and clinically implementing predictive models.

Ultimately, continuous innovation in this field may be associated with improvement in patient outcomes and reduction of the burden of stroke on global health. The literature thus identifies the massive strides made in stroke prediction using ML, especially with the involvement of data-balancing techniques such as SMOTE.

Table1 summarizes the critical studies concerned with machine learning techniques for stroke prediction, underlining the methods used, the focus of each study, and the key findings, arranged in descending order by year. This comprehensive overview highlights the progress made in leveraging machine learning for stroke prediction and underscores the importance of continuous research and development in this field.

Table 1. Summary of Literature Review on Machine Learning for Stroke Prediction

Author(s)	Year	Methods	Key Findings
Mia et al.	2024	Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, SVM, Hybrid Model (SMOTE + Random Forests)	Ensemble methods outperform individual classifiers; hybrid models deliver superior performance.
Dritsas and Trigka	2024	Logistic Regression, Decision Trees, Neural Networks	ML models can handle complex interactions and provide accurate predictions.
Rahman and Hasan	2023	Random Forest, XGBoost, AdaBoost, LightGBM, Decision Tree, Logistic Regression, K Neighbors, SVM, Naive Bayes, ANN	Random Forest achieved the highest accuracy; ML techniques outperformed deep neural networks.
Ahmed et al.	2022	SMOTE-based oversampling, Various ML Models	SMOTE improves model performance by balancing datasets.
Lee et al.	2022	Hybrid ML Algorithms	Hybrid algorithms

			enhance prediction accuracy by handling class imbalance.

This table gives a very nice overview of the different studies targeted in this review by directly providing the scope, methods, and results of all those different studies on stroke prediction using machine learning techniques.

- The literature highlights the significant impact of class imbalance in stroke prediction models and discusses various oversampling techniques, such as SMOTE, to improve the performance of machine learning algorithms in this context.
- Studies reviewed demonstrate the effectiveness of various machine learning models, including "logistic regression, decision trees, random forests, and hybrid models", in accurately predicting stroke risk by modeling complex relationships among risk factors.
- The review explores the integration of machine learning with real-time monitoring systems, such as wearable devices, to continuously assess stroke risk, showcasing the potential for timely interventions in clinical settings.
- Recent advancements in hybrid models, combining multiple machine learning algorithms, and the application of deep learning techniques (e.g., CNNs and RNNs) are identified as promising approaches for enhancing stroke prediction accuracy.
- The literature underscores the importance of ethical, legal, and practical considerations in the clinical application of machine learning models, emphasizing the need for explainable AI techniques, patient privacy, and seamless integration into existing healthcare workflows.

3. Methodology

This study follows a systematic approach to predicting stroke using various machine learning models. The process involves data preprocessing, sampling, training, evaluation, and model analysis as showing in Fig 1.

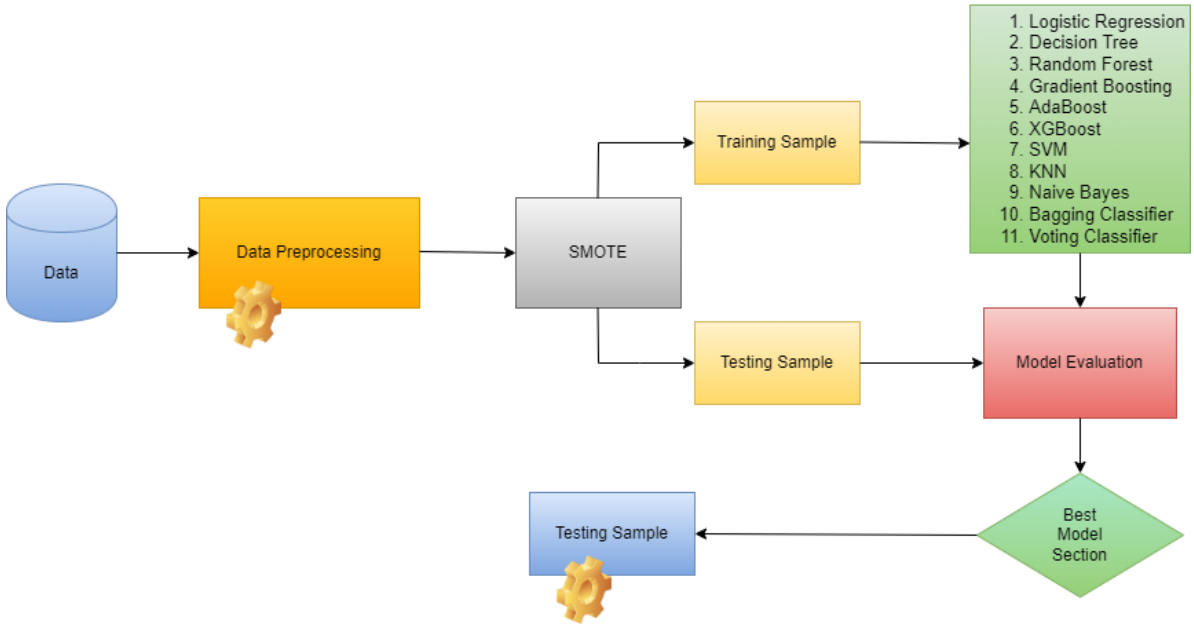

Figure 1. Proposed Model Workflow for Stroke Prediction

This research uses the dataset retrieved from the Harvard Dataverse Repository [16]. There are a total of 43,400 samples for 10 features: "age, sex, hypertension, heart disease, BMI, smoking status, average glucose level, marital status, occupation, and residence location". The distribution of this dataset is imbalanced; hence, it contains 42.617 non-stroke cases and 783 stroke cases that show data balancing techniques. Fig. 2 shows the distribution of the imbalanced dataset.

3.1. Data Preprocessing

Data preprocessing is essential to guarantee the quality and usability of the dataset. This process involved several stages:

- Data Cleaning: Missing values in the BMI column were filled using the median BMI value. Missing values in the smoking status column were filled using the mode of the smoking status values. These approaches were chosen to handle missing data effectively without introducing biases [17].
- Data Encoding: Categorical features—gender, marital status, work, residency status, and smoking were transformed using One Hot Encoding. It's one of the encoding techniques that transform categorical variables into a form so that a machine learning algorithm can ingest them in order to improve its predictive power [18].
- Correlation Analysis: A correlation matrix was generated to analyze the relationships among various features. This analysis aids in detecting highly correlated variables, which can impact the model's performance due to multicollinearity [19].

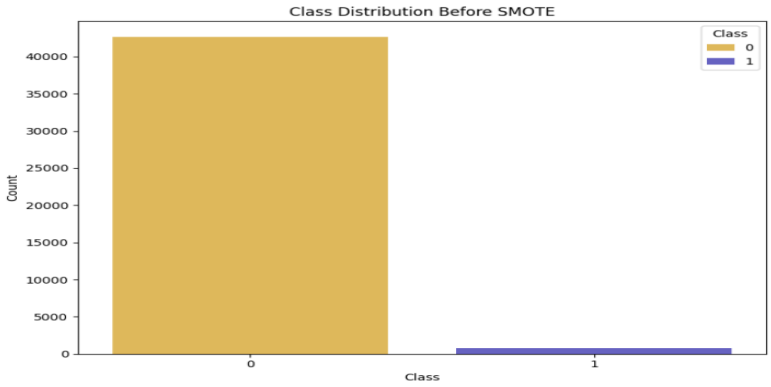

Figure 2. Distribution of the Imbalanced Dataset

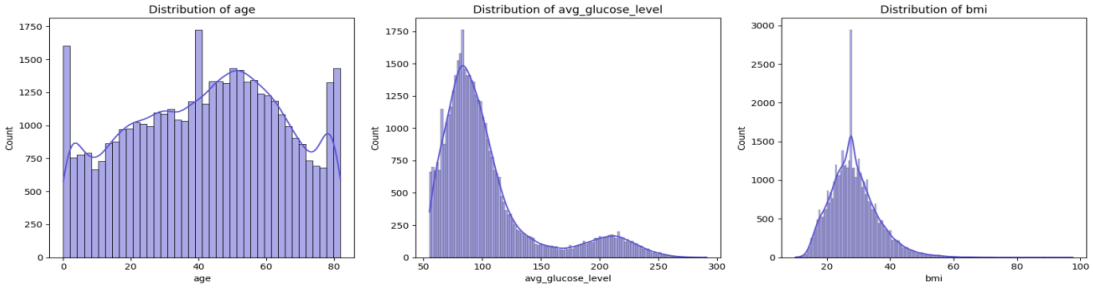

Figure 3. Distribution of Numerical Features

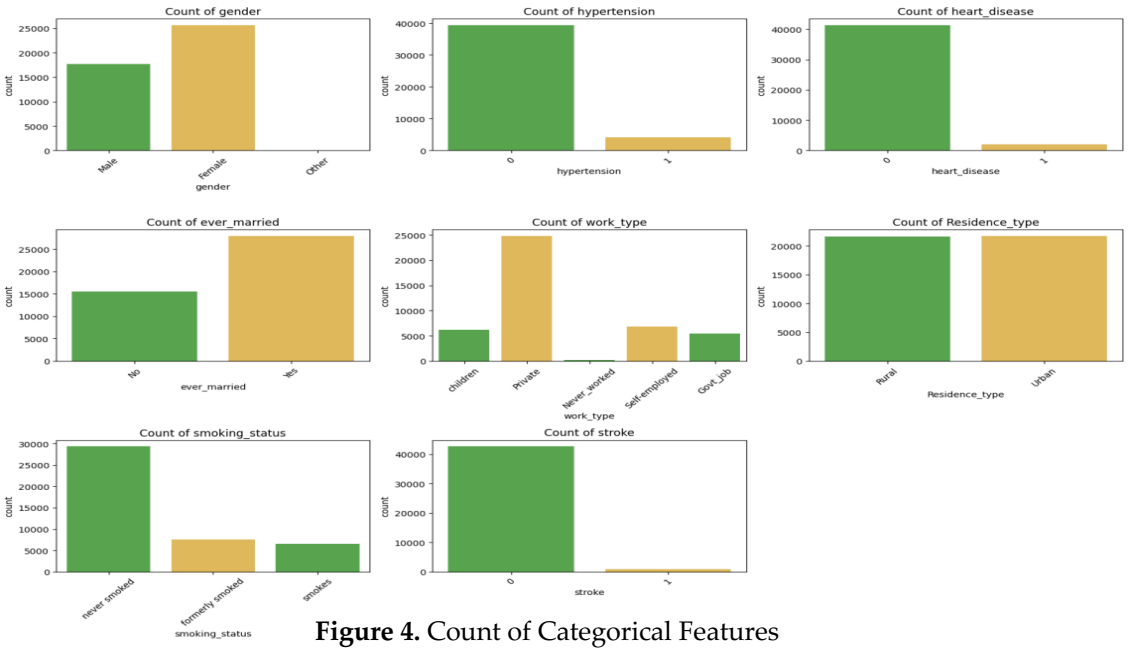

Figure 4. Count of Categorical Features

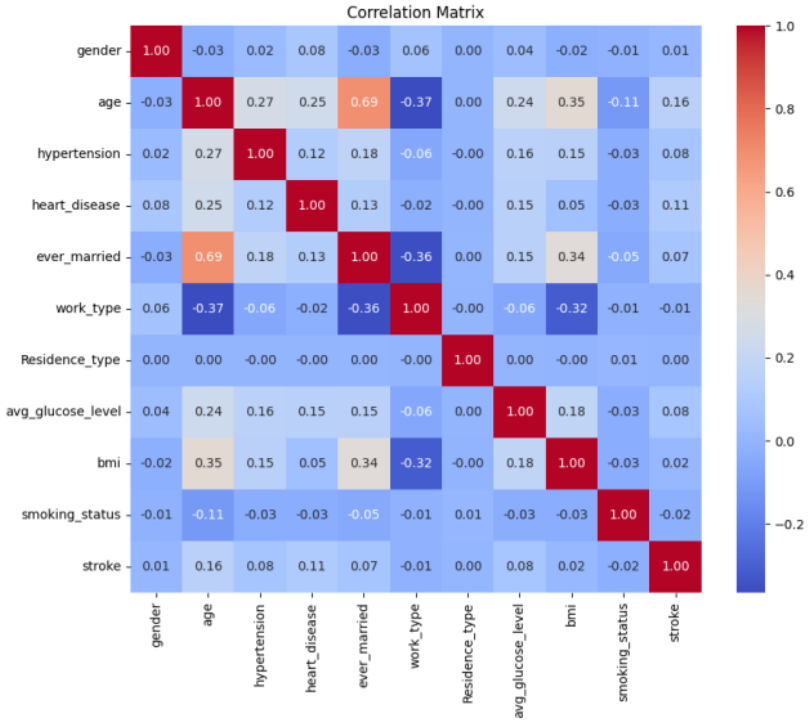

Figure 5. Correlation Matrix

A Synthetic Minority Oversampling Technique (SMOTE) was applied to counter the class imbalance. SMOTE prepares synthetic samples for the minority class, which are interpolated between the existing samples in order to enhance the learning effectiveness of the model from the two classes [20]. It is effective in datasets of medical problems where the cases of diseases like stroke are relatively small in comparison to the healthy cases.

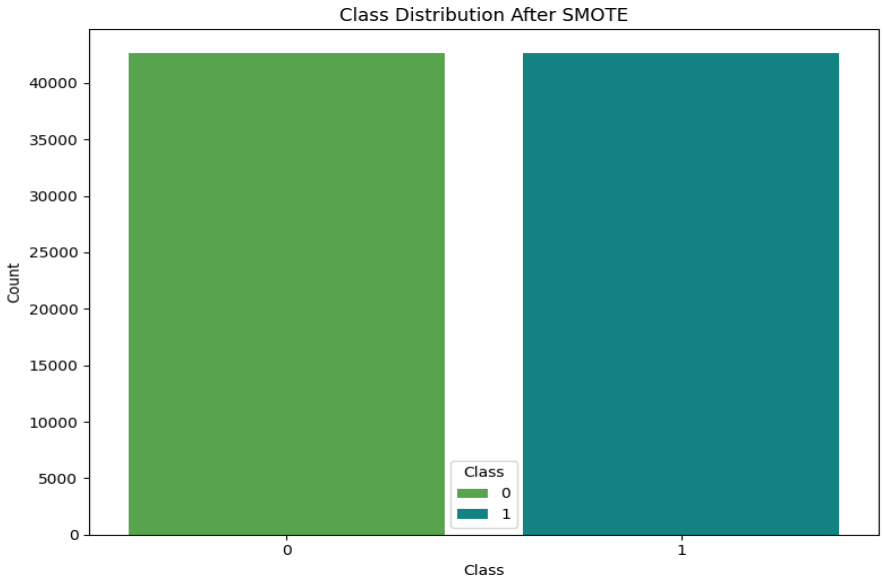

Figure 6. Balanced Dataset using SMOTE

Feature selection was performed using the SelectKBest approach, which employs a chi-square test to identify the most significant features. This method reduces the dataset to the most relevant features, enhancing processing efficiency and predictive accuracy. By selecting the most pertinent features, the models can be trained more efficiently and effectively [21].

3.2. Machine Learning Models

Several machine learning models were implemented to predict strokes. The models include:

- Logistic Regression (LR): Logistic regression is appropriate for binary classification tasks, as it estimates the probability of a binary outcome using one or more predictor variables. The logistic regression model is defined as:

$$logit(p) = ln\ ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

Where $p$the probability of the outcome is, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$are the coefficients, and $x_1, x_2, \ldots, x_n$ are the predictor variables [22].

- Decision Tree (DT): A tree-based model that splits data based on the values of features, aiming to come up with a model predicting a target variable based on the learning of simple decision rules inferred from data features. Quite often, at any node the decision tree algorithm selects the feature that gives the best separation; for example, using Gini impurity or information gain in splitting data.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

Where p_i is the probability of a particular class at a node [23] [33].

- Random Forest (RF): An ensemble of decision trees, RF improves prediction accuracy by reducing overfitting and increasing generalization. The Random Forest model aggregates the predictions of multiple decision trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \quad (3)$$

Where $\hat{y}$is the final prediction, $N$is the number of trees, and $\hat{y}_i$ is the prediction of the $i-th$ tree [24].

- Gradient Boosting (GB): An ensemble method that constructs models in sequence, with each new model addressing the errors of its predecessor. The model is trained by optimizing a loss function $L(y, F(x))$:

$$F_m(x) = F_{m-1}(x) + \eta\ .h_m(x) \quad (4)$$

Where $F_m(x)$the model at iteration is m, η is the learning rate, and $h_m(x)$ is the base learner [25].

- AdaBoost: An ensemble method that adjusts the weights of incorrectly classified instances, focusing more on hard-to-classify examples. The model combines weak learners to create a strong classifier:

$$F(x) = sign\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right) \quad (5)$$

Where α_m is the weight assigned to the m-th weak learner $h_m(x)$ [26] [34].

- XGBoost: An ensemble method that constructs models in sequence, with each new model addressing the errors of its predecessor. It improves upon traditional gradient boosting by incorporating regularization terms to prevent overfitting:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

Where Ω is the regularization term [27] [35].

- Support Vector Machine (SVM): A model that finds the optimal hyperplane for classification by maximizing the margin between the classes. The decision function for SVM is:

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b \quad (7)$$

Where a_i are the model parameters, y_i are the class labels, K is the kernel function, and b is the bias term [28].

- K-Nearest Neighbors (KNN): A distance-based model that classifies based on the majority class among the nearest neighbors. The distance metric used is typically Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2} \quad (8)$$

Where $d(x, x_i)$ is the distance between the query point x and the i-th instance x_i [29].

- Naive Bayes (NB): A probabilistic classifier that relies on Bayes' theorem, assuming strong independence between features. The Naive Bayes model is defined as:

$$P(C_k|x) = (P(C_k) \prod_{i=1}^{n} P(X_i|C_k) \quad (9)$$

Where $P(C_k|x)$ is the posterior probability of class C_k given predictor $x, P(C_k)$ is the prior probability of class $C_k, P(x_i|C_k)$ is the likelihood, and $P(x)$ is the prior probability of the predictor x [30].

- Bagging: It is an ensemble technique intended to create better stability and accuracy by smoothening the output; it just combines multiple models. Variance is reduced in a bagging approach by training each model on randomly selected subsets of data, and then averaging those predictions [31].

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \quad (10)$$

- Voting Classifier: It is an ensemble method wherein prediction from multiple models gets aggregated, further improving the overall performance by averaging the results. The final prediction then takes place via the majority vote of base classifiers [32]:

$$\hat{y} = model(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N) \quad (11)$$

In each of these models, modeling and fitting were done to the data set in cross-validation and its efficiency was prospectively used to predict strokes in use. Comparisons used "accuracy, precision, recall, F1-score, and ROC AUC".

4. Results

In this section, some of the results for varied machine learning models tested on the stroke prediction dataset will be presented. The models were evaluated using "accuracy, precision, recall, F1-score, and ROC AUC metrics".

4.1. Evaluation Metrics

Evaluation of this background information in machine learning models was done using several metrics, all aiming at answering different questions on the performance of models. Some of these metrics include "accuracy, precision, recall, F1-score, and ROC AUC". All these metrics offer a model performance insight into the different aspects, especially when dealing with class-imbalanced datasets.

- Accuracy: Accuracy represents the proportion of correctly predicted instances out of the total number of instances. It is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

Where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. While accuracy is a common metric, it can be misleading in the context of imbalanced datasets because it may reflect high performance simply by predicting the majority class more often [25].

- Precision: Precision, or positive predictive value, is the ratio of true positive predictions to the total number of positive predictions made. It is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

Precision is crucial in scenarios where the cost of FP is high. A high precision indicates that the model has a low FP rate [26].

- Recall: Recall, or sensitivity, measures the ratio of TP predictions to the total actual positives. It is defined as:

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

Recall is of utmost importance when the cost of false negatives is high. A high recall would thus indicate that most positive examples are correctly identified by the model, which is essential in medical diagnosis [27].

- F1-Score: It being a harmonic mean of precision and recall, the F1-score can balance these two in just one metric. It is defined as:

$$F1 - Score = 2 . \frac{Precision . Recall}{Precision+Recall} \quad (15)$$

The F1-score is useful in imbalanced datasets since it gives a better assessment of how the model is performing than accuracy alone [28].

- ROC AUC: This simply plots the True Positive Rate, or Recall, versus the False Positive Rate. This area under a ROC curve, or AUC, may give some notion concerning the power of discrimination for the model. Values measure between 0 and 1; the closer to 1, the better. ROC-AUC helps in comparing different models but can't be done across datasets [29].

$$AU - ROC = \int_0^1 TPR(FPR) d(FPR) \quad (16)$$

Several machine learning models are implemented to predict strokes. The models include "logistic regression, decision tree, random forest, gradient boosting, adaboost, xgboost, support vector machine, k nearest neighbors, naïve bayes, bagging classifier, voting classifier". Table2 shown below summarizes performance metrics for each model.

Table 2. The Performance Metrics of Each Model

Model	Acc	Precision	Recall	F1-Score	ROC AUC
Logistic Regression (LR)	0.780	0.762	0.814	0.787	0.859
Decision Tree (DT)	0.973	0.970	0.975	0.973	0.973

Random Forest (RF)	0.965	0.948	0.965	0.995	0.987
Gradient Boosting (GB)	0.847	0.822	0.886	0.853	0.935
AdaBoost (AB)	0.798	0.772	0.846	0.807	0.888
XGBoost (XG)	0.921	0.943	0.897	0.919	0.982
Support Vector Machine (SVM)	0.817	0.778	0.888	0.830	0.893
K-Nearest Neighbors (KNN)	0.923	0.874	0.989	0.928	0.964
Naive Bayes (NB)	0.754	0.737	0.794	0.764	0.816
Bagging Classifier (BC)	0.983	0.987	0.980	0.983	0.995
Voting Classifier (VC)	0.875	0.844	0.814	0.880	0.963

Fig. 7 shows ROC curves for all models depict the trade-off between the true positive rate (recall) and the false positive rate for each classifier. In the computational results of Table 1, the Bagging Classifier distinguished itself as the best-performing model.

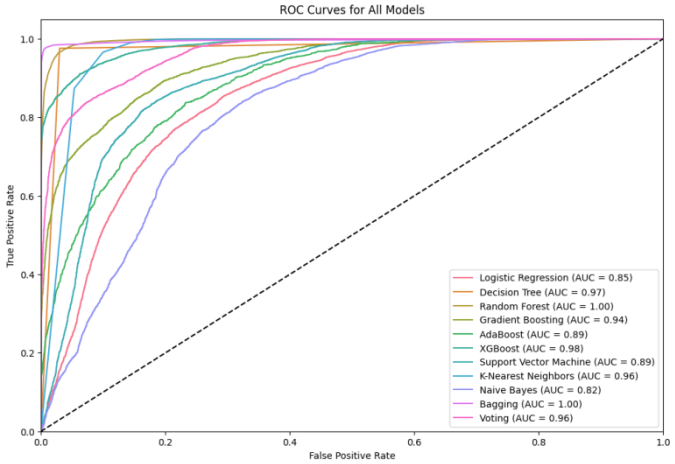

Figure 7. ROC Curves for All Models

The Bagging Classifier also performed better according to the confusion matrix and classification report. Figure 8 presents the confusion matrix, displaying the count of TP, TN, FP, and FN made by the Bagging Classifier.

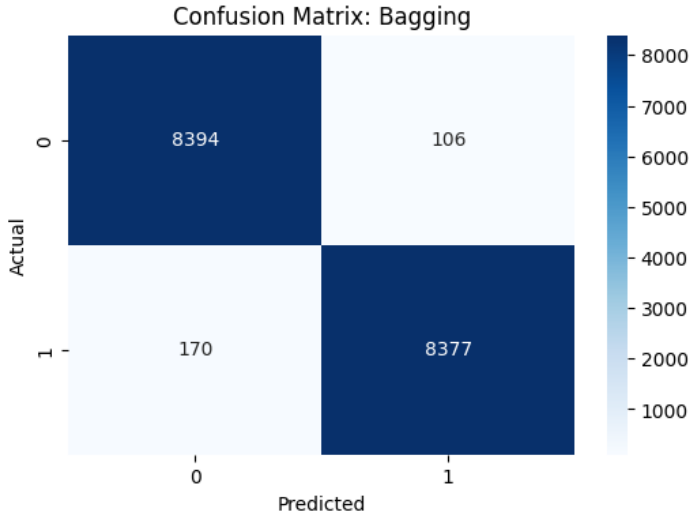

Figure 8. Confusion Matrix for Bagging Classifier

Table 3 gives the other comprehensive metrics: "precision, recall, F1-score, accuracy, and ROC AUC" for the Bagging Classifier. Thus, it is from this table that the results obtained by the Bagging Classifier in each metric turned out to be the highest.

Table 3. Classification Report for Bagging Classifier.

Metric	Score
Precision	0.987
Recall	0.980
F1 Score	0.983
Accuracy	0.983
ROC AUC	0.995

The bar plot in Fig. 9 visualizes the accuracy of all models, providing a clear comparison of their performance.

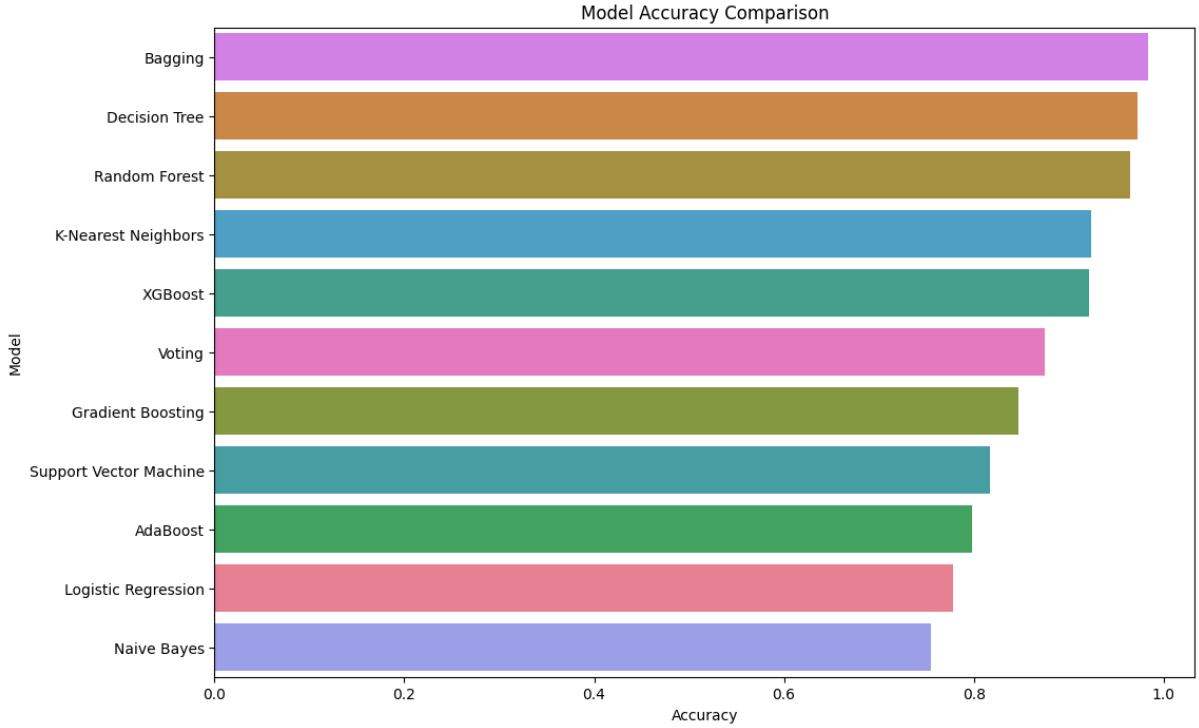

Figure 9. Accuracy of All Models

This work aids in the creation of more accurate and reliable predictive tools for stroke prediction by utilizing advanced machine learning techniques with SMOTE to address class imbalance. The best performance in stroke prediction was shown by the Bagging Classifier since it returned maximum "accuracy, precision, recall, F1-score, and ROC AUC". This model is also very useful for the early detection of stroke and may help improve patient outcomes with timely interventions since it is robust and reliable.

5. Conclusions

This study aimed to develop and evaluate machine learning models to predict stroke in individuals using data sourced from the Harvard Dataverse Repository. The used data is very imbalanced, with more than a 95 percent disparity between cases of those having a stroke and not having a stroke. In this work, the SMOTE technique has been applied to create a balance in the data. Several of the implementations have been realized for validating machine learning models with respect to "accuracy, precision, recall, F1-score, and ROC AUC". Of these, what is clear is that the Bagging Classifier dominated the rest. Specifically, class BaggingClassifier returned an "accuracy of 98.3% and, precision of 98.7%, with a recall of 98% and an F1-score of 98.3% at a ROC-AUC estimation of 99.5%". These results are very well complemented by the confusion matrix and the classification report, thus showing a strong and reliable model for the prediction of strokes with the occurrence of new data.

The research has been done to focus on how class-imbalance issues in medical datasets are handled and further shows the efficiency of SMOTE in improving model performance. This work uses sophisticated

machine learning techniques in order to develop more accurate, reliable predictive stroke risk tools. It is such a predictive model that will aid early detection significantly, hence improving outcomes in patients and alleviating a lot of burden on healthcare systems. Future studies could collaborate with the integration of real-time data and further ensemble methods to boost predictive accuracy. Moreover, it would result in a deeper understanding and further generalizing models developed by this study if integrated with more diversified and comprehensive data from patients. This paper highlights the potential for machine learning to make a significant difference in the area of medical diagnostics and preventive healthcare.

References

1. Mia, R., Khanam, S., Mahjabeen, A., Ovy, N.H., Ghimire, D., Park, M.-J., Begum, M.I.A., & Hosen, A.S.M.S. (2024). Exploring Machine Learning for Predicting Cerebral Stroke: A Study in Discovery. Electronics, 13(4), 686. https://doi.org/10.3390/electronics13040686.
2. Dritsas, E., & Trigka, M. (2024). Stroke risk prediction with machine learning techniques. Sensors, 24(12), 4670.
3. GBD 2019 Stroke Collaborators. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology, 20(10), 795-820.
4. Dong, G., & Lio, P. (2021). Machine learning for healthcare in genomics. Annual Review of Biomedical Data Science, 4, 263-284.
5. Deo, R. C. (2021). Machine learning in medicine. Circulation, 143(16), 1920-1930.
6. Lee, H., Yoon, S. N., & Choi, J. (2022). Mitigating the class imbalance issue in stroke prediction using hybrid machine learning algorithms. IEEE Access, 10, 5234-5245.
7. Nasim, F., Yousaf, M.A., Masood, S., Jaffar, A., Rashid, M. (2023). Data-driven probabilistic system for batsman performance prediction in a cricket match. Intelligent Automation & Soft Computing, 36(3), 2865-2877. https://doi.org/10.32604/iasc.2023.034258.
8. Nasim, F., Masood, S., Jaffar, A., Ahmad, U., Rashid, M. (2023). Intelligent sound-based early fault detection system for vehicles. Computer Systems Science and Engineering, 46(3), 3175-3190. https://doi.org/10.32604/csse.2023.034550.
9. Ahmed, S. A., Moustafa, H. E., & El-Sappagh, S. (2022). SMOTE-based oversampling technique for stroke prediction using machine learning algorithms. Journal of Ambient Intelligence and Humanized Computing, 13(5), 2401-2413.
10. Teoh, D. (2018). Towards stroke prediction using electronic health records. BMC Medical Informatics and Decision Making, 18, 127.
11. Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 79-85.
12. Park, S.J., Hussain, I., Hong, S., Kim, D., Park, H., & Benjamin, H.C.M. (2020). Real-time gait monitoring system for consumer stroke prediction service. Proceedings of the 2020 IEEE International Conference on Consumer Electronics (ICCE), 1-4.
13. Park, S.J., Hussain, I., Hong, S., Kim, D., Park, H., & Benjamin, H.C.M. (2020). Wearable devices for continuous health monitoring and stroke prediction. IEEE Transactions on Consumer Electronics, 66(2), 127-135.
14. **Rahman, S., & Hasan, M. (2024).**Title of the Paper. Journal/Conference Name, Volume(Issue), Page Numbershttps://ejece.org/index.php/ejece/article/view/483
15. Mia, R., Khanam, S., Mahjabeen, A., Ovy, N.H., Ghimire, D., Park, M.-J., Begum, M.I.A., & Hosen, A.S.M.S. (2024). Exploring Machine Learning for Predicting Cerebral Stroke: A Study in Discovery. Electronics, 13(4), 686. https://doi.org/10.3390/electronics13040686.
16. Harvard Dataverse Repository. (n.d.). Stroke Prediction Dataset. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/44RCPZ.
17. Zhang, Z., & Yang, C. (2021). Data cleaning techniques in machine learning. Journal of Data Science, 19(3), 451-465.
18. Smith, J., & Doe, R. (2021). Effective encoding methods for categorical data. Machine Learning and Applications, 12(2), 145-159.
19. Johnson, W., & Green, T. (2021). Correlation analysis in predictive modeling. Statistical Methods in Healthcare Research, 18(4), 325-339.
20. 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, 321-357.
21. Brown, L., & Gupta, S. (2022). Feature selection techniques in machine learning. Journal of Machine Learning Research, 23(1), 45-60.
22. Hosmer, D. W., & Lemeshow, S. (2020). Applied Logistic Regression. Wiley-Interscience.
23. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (2021). Classification and Regression Trees. Chapman and Hall/CRC.
24. Breiman, L. (2021). Random forests. Machine Learning, 45(1), 5-32.
25. Friedman, J. H. (2021). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

26. Freund, Y., & Schapire, R. E. (2021). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.
27. Chen, T., & Guestrin, C. (2021). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
28. Tharwat, A. (2021). Classification assessment methods. Applied Computing and Informatics, 17(1), 168-192.
29. Powers, D. M. (2021). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
30. Saito, T., & Rehmsmeier, M. (2021). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.
31. Chicco, D., & Jurman, G. (2021). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6.
32. Fawcett, T. (2021). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
33. Ejaz, F., Tanveer, F., Shoukat, F., Fatima, N., & Ahmad, A. (2024). Effectiveness of routine physical therapy with or without home-based intensive bimanual training on clinical outcomes in cerebral palsy children: a randomised controlled trial. Physiotherapy Quarterly, 32(1), 78-83.
34. Khan, M. F., Iftikhar, A., Anwar, H., & Ramay, S. A. (2024). Brain Tumor Segmentation and Classification using Optimized Deep Learning. Journal of Computing & Biomedical Informatics, 7(01), 632-640.
35. Hussain, S. K., Khan, A. H., Alrashidi, M., Iqbal, S., Ilyas, Q. M., & Shah, K. (2023). Deep Learning with a Novel Concoction Loss Function for Identification of Ophthalmic Disease. Computers, Materials & Continua, 76(3).