Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 59-71
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

The Early Warning Signs of a Stroke: An Approach Using Machine Learning Predictions

Esraa H. Augi, Almabruk Sultan

Department of Computer Science, University of Benghazi, Benghazi, Libya
Email: esraa.augi@uob.edu.ly, almabruk.sultan@uob.edu.ly

How to cite this paper: Augi, E.H. and Sultan, A. (2024) The Early Warning Signs of a Stroke: An Approach Using Machine Learning Predictions. *Journal of Computer and Communications*, **12**, 59-71.
https://doi.org/10.4236/jcc.2024.126005

Received: May 7, 2024
Accepted: June 22, 2024
Published: June 25, 2024

Abstract

Early stroke prediction is vital to prevent damage. A stroke happens when the blood flow to the brain is disrupted by a clot or bleeding, resulting in brain death or injury. However, early diagnosis and treatment reduce long-term needs and lower health costs. We aim for this research to be a machine-learning method for forecasting early warning signs of stroke. The methodology we employed feature selection techniques and multiple algorithms. Utilizing the XGboost Algorithm, the research findings indicate that their proposed model achieved an accuracy rate of 96.45%. This research shows that machine learning can effectively predict early warning signs of stroke, which can help reduce long-term treatment and rehabilitation needs and lower health costs.

Keywords

Machine Learning, Stroke, k-Nearest Neighbors, Decision Tree, Random Forest, GXboost

1. Introduction

Stroke is a life-threatening condition that is more common in adults aged 45 and above [1] [2]. The brain is affected by a stroke similar to how a heart attack affects the heart. The cause of a stroke can be either the rupturing or bleeding of blood vessels in the brain or a restriction of blood supply to the brain. In both cases, the brain's tissues are deprived of blood and oxygen when a blockage or rupture occurs.

Brain strokes are a significant cause of death worldwide, ranking as the third leading cause of death [3] [4]. They can result from either a blockage of blood

supply to the brain or the rupture and hemorrhage of a blood vessel in the brain, leading to damage to the brain tissue due to the lack of oxygen and blood supply [5].

Early detection of brain stroke is crucial in reducing long-term death rates and costly medical treatment. When the blood flow to the brain is hindered, it causes damage to the brain tissues, affecting the functioning of physiological organs controlled by the affected nervous system section, resulting in various symptoms.

The chances of a stroke victim making a full recovery increase significantly if they receive medical attention as soon as possible. Some common signs of a stroke include hindrance in blood flow to the brain, which can result in damage to brain tissues and affect the functioning of physiological organs. The following are some stroke signs:

1) Numbness in the face, leg, and arm.
2) Communication difficulties.
3) Issues with response time.
4) Changes in behavior.
5) Vision problems.
6) Difficulties with mobility.
7) Feeling dizzy.
8) Headaches.
9) Feeling vomiting or nauseous.

A stroke requires immediate medical attention to prevent death, permanent disability, and brain damage [6]. The outcome for an individual after a stroke is dependent on the type of stroke, which is categorized into three categories:

1) Ischemic Stroke.
2) Hemorrhagic Stroke.
3) Transient Ischemic Attack (TIA).

A transient ischemic attack (TIA) occurs when the blood flow to the brain is temporarily disrupted. TIA patients often recover from this type of stroke within a few minutes. Blood clots are the most common cause of TIA and can serve as a warning sign for the individual experiencing it. According to data from the Centers for Disease Control and Prevention (CDC) [7], within a year of experiencing a transient ischemic attack, one-third of individuals are at risk of having a stroke.

The arteries that supply the brain with oxygen and nutrients narrow or close off during an ischemic stroke. Blood arteries become blocked due to blood clots and fragments of fractured plaque. According to the CDC, 87% of stroke victims experience the ischemic form of the disease [7]. An artery in the brain bursts, resulting in a hemorrhagic stroke.

Damage to brain cells and tissues can occur when the blood pressure in the artery exceeds that of the skull. According to the American Heart Association, hemorrhagic strokes account for 13% of all strokes [8].

There are various reasons why patients get stroke. According to the National Heart, Lung, and Blood Institute [9], stroke is primarily caused by diet, inactivity, alcohol, cigarettes, personal history, medical history, and complications. People consume foods that are extremely salty, saturated in fat, trans fat, and cholesterol in today's unbalanced diet. According to CDC, it is recommended to have at least 2.5 hours of physical activity per week. However, the risk of stroke can be significantly high due to insufficient exercise.

Alcohol and tobacco consumption, along with personal history, gender, family history, age, and geographical location, are important factors that contribute to stroke. Additionally, certain medical conditions like a history of TIA, high blood pressure, high cholesterol, obesity, heart valve defect, diabetes, and other diseases can also increase the risk of stroke [10].

Machine learning (ML) plays a crucial role in the digital age by predicting potential problems in advance. Early detection of diseases such as stroke can significantly increase the chances of successful treatment. ML is crucial in the healthcare industry for disease prediction and diagnosis. Robust data analysis techniques are needed to handle the increasing amount of medical data stored in patients' medical records in hospitals [11] [12].

Our research aims to create a machine-learning model for predicting strokes. We will utilize feature selection techniques to identify early warning signs of strokes from a medical record dataset.

This research makes a valuable contribution by utilizing multiple machine learning models on a publicly accessible dataset to predict the occurrence of stroke disease. The research involves analyzing the medical records of patients previously diagnosed with a stroke. The study will focus solely on patients who have had an ischemic stroke, excluding hemorrhagic stroke. This research doesn't aim to use all ML algorithms but focuses on those with promising results in prior studies. Four distinct models were used, and their results were compared with previous research.

The experimental methodology and procedures are detailed in Section 3, followed by a thorough analysis of the results in Section 4. The conclusions of the study are discussed in Section 5.

2. Related Work

In recent years, various works based on Machine Learning algorithms have been published. Some of these works are discussed below:

Govindarajan *et al.* [13]. utilized various machine learning algorithms including Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Tree, Logistic Regression, and ensemble methods (Bagging and Boosting) to classify stroke disease [13]. The data was collected from Sugam Multispecialty Hospital in India and consisted of information on 507 stroke patients aged between 35 to 90 years old. The novelty of their work lies in the data processing phase where they used an algorithm called novel stemmer to preprocess the da-

taset. In the collected dataset, 91.52% of patients were affected by ischemic stroke and only 8.48% of patients were affected by hemorrhagic stroke. Among the mentioned algorithms, Artificial Neural Networks with stochastic gradient descent learning algorithms achieved the highest accuracy of 95.3% for classifying stroke.

A model for stroke prediction based on Support Vector Machines was proposed by Jeena and Kumar [14]. They sourced their data from the International Stroke Trial Database [15], which contained 12 attributes. Their research utilized 350 samples, with 300 samples used for training and 50 for testing. Various kernel functions, including polynomial, quadratic, radial basis function, and linear functions, were applied. The most accurate results, with a balance measure F1-score F-measure of 91.7, were obtained using the linear kernel, which achieved an accuracy of 91%.

Singh and Choudhary [16] utilized Artificial Neural Network (ANN) to develop a model for predicting stroke. They gathered datasets from the Cardiovascular Health Study (CHS) database and constructed three datasets. These datasets contain 212 occurrences of stroke (all three) and 52, 69, and 79 non-stroke cases respectively. The final dataset comprises 357 attributes and 1824 entities. During feature selection, they used the C4.5 decision tree algorithm, and Principal Component Analysis (PCA) for dimension reduction. They implemented the Back Propagation learning method in ANN and achieved an accuracy of 95%, 95.2%, and 97.7% for the three datasets respectively.

Adam and colleagues [17] have created a classification model for ischemic stroke utilizing a decision tree algorithm and K nearest neighbor (k-NN). The dataset was collected from various hospitals and medical centers in Sudan, making it the first dataset for ischemic disease in the country. The dataset includes information about 400 patients and 15 features. The experiment results indicate that the decision tree classification outperforms the k-NN algorithm.

A study conducted by Sudha *et al.* [18] utilized the Decision Tree, Bayesian Classifier, and Neural Network to classify strokes. Their dataset consisted of 1000 records, and the PCA algorithm was utilized for dimensionality reduction. In ten rounds of each algorithm, the highest accuracy was achieved at 92%, 91%, and 94% in the Neural Network, Naive Bayes classifier, and Decision tree algorithm, respectively.

Al-Zubaidi *et al.* [19] used machine learning classification methods to predict the probability of stroke. They applied SMOTE to balance the dataset, which consisted of 5,000 samples and 16 features. The model was trained using five different classification algorithms, with the RF classifier achieving the best results at 94.6% accuracy. The authors suggest that machine learning could be useful in predicting stroke, but more research is needed to improve accuracy and test the model in a clinical setting.

Sailasya and Kumari [20] used a stroke dataset from Kaggle, consisting of 5110 rows and 12 columns. They balanced the data using under-sampling techniques and split it into 80% training data and 20% test data. Six machine-learning

classification algorithms were applied, and Naive Bayes had the highest precision accuracy of 82%. Finally, an HTML page was created for user-friendliness.

3. Methodology

In Figure 1, the proposed system is represented by this block diagram. Detailed information about each component of the block diagram is in this section.

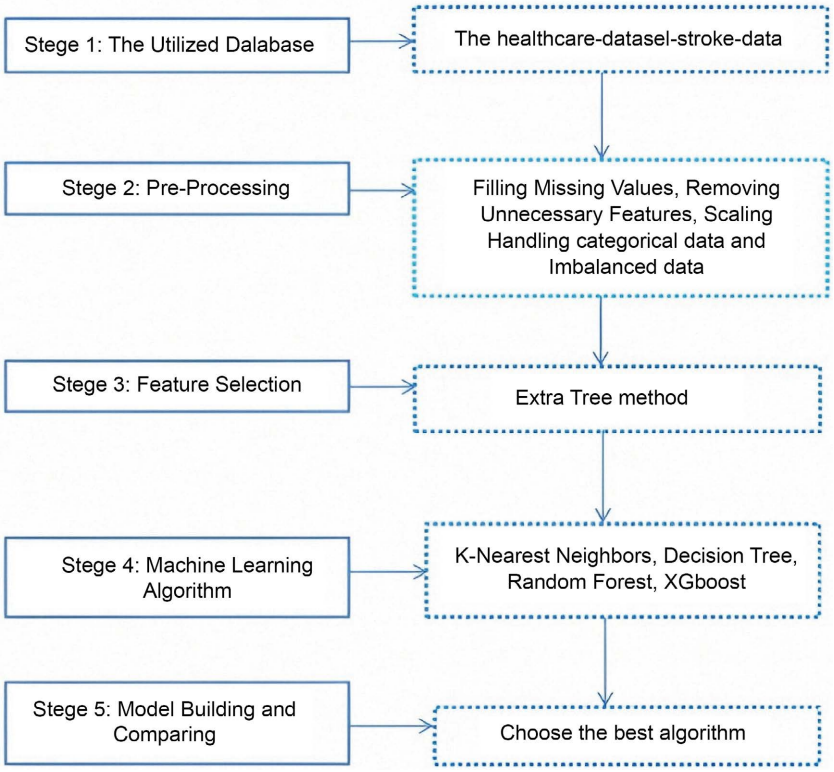

Figure 1. Proposed system's block diagram.

3.1. The Utilized Database

We utilized a publicly available dataset on stroke. The healthcare-dataset-stroke data used in this analysis was obtained from Kaggle [21] consisting of 5110 patients and 12 attributes, where the target variable is a binary classification of stroke. Out of the total patients, 249 had a stroke while 4869 did not. The dataset contained 201 null values in the "BMI" attribute, which were imputed using the median as it is less sensitive to outliers.

3.2. Preparing the Data

Data preprocessing is a necessary step in transforming raw data into formats that can be used and understood. Raw datasets often present challenges such as flaws, inconsistent behavior, lack of trends, and missing values. Consequently, it is essential to conduct preprocessing to address these issues [22]. The following points outline the various issues identified in the dataset and the steps taken to address them:

- Removing Unnecessary Features: The "ID" feature, which had no logical meaning, was removed.
- Filling Missing Values: Out of a total of 5110 records, the "BMI" feature had 201 missing values. To address this, the mean value of the "BMI" column was calculated and used to replace the null values.
- Handling categorical data (features): Categorical variables such as hypertension, heart disease, and stroke were converted to object data type. To facilitate the handling of categorical data by machine learning algorithms, the "Work Type" and "Smoking Status" features were converted into binary features using the One Hot Encoding method. For features with two classes such as "Ever Married", "Gender", and "Residence Type", label encoding was used to directly switch them to one and zero. A new dummy variable was created for each category, such as gender_male and gender_female.
- Removal of Unnecessary Features: It has been noticed that the "Gender" column has an "Other" category. To keep things simple, they are dropped to avoid model scattering during classification.
- Handling unbalanced data (label): The "Stroke" feature is the desired output of this research with an unbalanced dataset. The majority class has 4860 out of 5109 samples, while the minority class has only 249 samples. Addressing this imbalance is crucial for accurate results. This research uses SMOTE to handle the imbalance. SMOTE identifies new points in the minority class using the K-nearest-neighbor algorithm and creates line segments connecting them [19]. The resulting new samples are positioned close to the minority class's samples. **Figure 2** shows the boundary ratio before and after applying SMOTE.

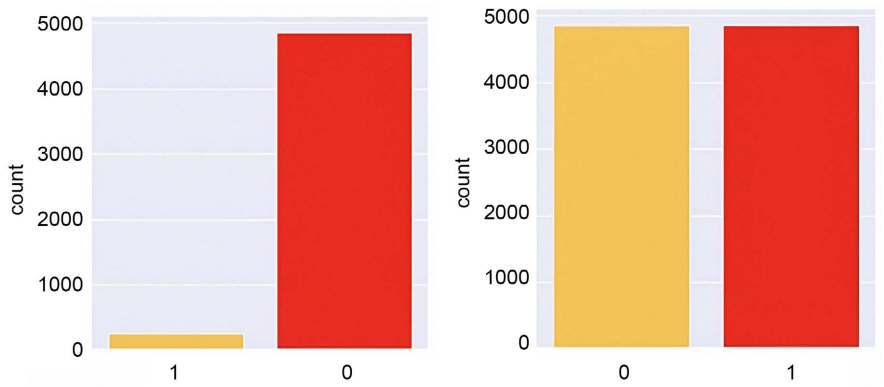

Figure 2. The stroke proportion before and after SMOTE.

- Scale the Features: Scale the Features: The standard scalar function scales the dataset's features for similar scales. Machine learning algorithms must consider varying feature scales to optimize performance. In this case, "BMI", "average glucose level", and "age" have significant scale differences. For example, "BMI" ranges from 15 to 50, "average glucose level" ranges from 70 to

300, and "age" ranges from 20 to 80. These differences impact algorithm accuracy, so addressing them appropriately is crucial. Differences in scale can hinder learning relationships between features for some algorithms due to data complexity.

After completing these pre-processing steps, our data set is now clean and ready for further analysis or model building.

3.3. Feature Selection

Extra Trees, a flow-learning algorithm, will be used for feature selection in this research. It works by generating multiple decision trees with random subsets of features. This machine learning algorithm identifies important features for predicting stroke risk by building decision trees on bootstrapped samples of training data and measuring feature importance [23]. The most predictive features are those with the highest importance scores, which are averaged across all trees.

As shown in **Figure 3** by the Extra Tree, the features are ranked in order of importance, according to the findings, BMI, average glucose level, and age, and were identified as the top three factors. Our observations from the correlation heat map support these findings, but the Extra Tree also takes into account nonlinear relationships and feature interactions, providing a more comprehensive view of feature significance.

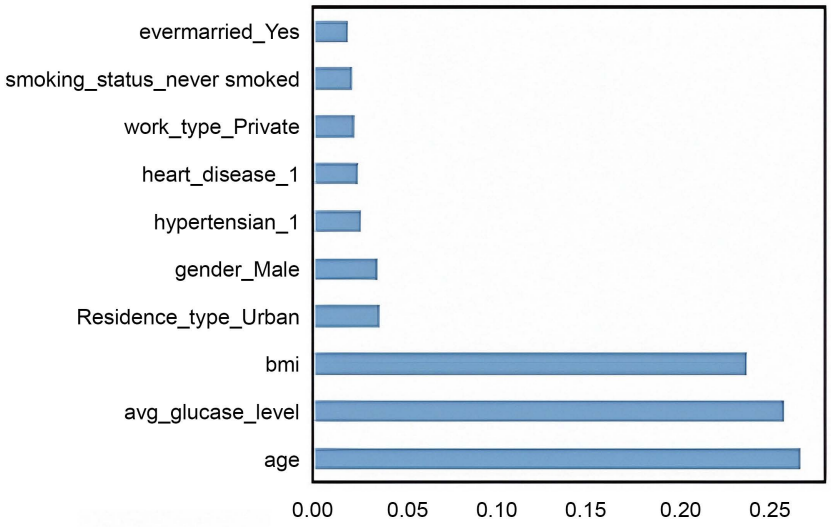

Figure 3. The feature importance of the extra tree.

Once the data preprocessing, handling of the imbalanced dataset, and feature selection are completed, the subsequent phase, to enhance accuracy and efficiency, we divide the under-sampled data into 80% training and 20% testing data Subsequently, various classification algorithms are utilized to train the models. with each category divided into training and test subsets.

The training group consisted of 3892 subjects with stroke and 3884 subjects without stroke, while the other test group consisted of 976 subjects without stroke

and 968 subjects with stroke.

3.4. Machine Learning Algorithm

In this section, four models were developed using four distinct machine learning algorithms, including Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost).The subsequent section provides a concise overview of each of these algorithms.

A decision tree is a type of supervised learning algorithm that builds a tree-like model to make predictions or classifications. Internal nodes represent attribute tests, branches represent test outcomes, and leaf nodes hold class labels. The training data is recursively divided into subsets based on attribute values until a stopping criterion is met [24]. The Decision Tree algorithm selects the optimal attribute during training by evaluating a metric like entropy or Gini impurity to maximize information gain or impurity reduction after the split.

In supervised learning, the k-nearest Neighbors (KNN) algorithm is a widely used non-parametric approach for both classification and regression tasks. It works by identifying the k most similar data points to a new data point and using this information to predict the class or value of the new data point based on its neighbors [25]. Despite its simplicity, KNN is often very effective.

Random Forest is a type of supervised machine learning algorithm that uses an ensemble of decision trees to create an accurate model. The algorithm trains multiple decision trees on different subsets of the data and then averages their predictions to obtain a final output. This approach helps to reduce overfitting and improves the generalization of the model, making it more accurate and reliable [26]. It is known for its high accuracy, robustness to overfitting, ability to handle missing data, ease of use, and outperformance of other machine learning algorithms.

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that uses an ensemble of decision trees for classification and regression tasks. It improves on the gradient boosting algorithm, which is an iterative algorithm that builds a model by adding new trees to correct the mistakes of the previous trees [27].

4. Discussion and Analysis of Experimental Results

In the proposed research methodology, the first step is to pre-process the data, due to the presence of missing and extreme data, as well as data imbalance, in addition to the presence of unnecessary columns that require deletion. In data processing, unnecessary features such as the “identifier” feature which has no logical meaning have been removed. The missing values were filled in the BMI feature, which had 201 blank values out of 5110. Numerical data is vital for machine learning algorithms to perform effectively. However, categorical features are not easily understood by these algorithms. Therefore, converting categorical features to numeric data is crucial to ensure the best possible outcome for your

machine-learning model.

The confusion matrix in Figure 4 is a helpful tool for evaluating the performance of machine learning classifiers. It helps identify correct and incorrect forecasts, with true positives and true negatives being accurately anticipated values and false positives and false negatives being poorly predicted values. To evaluate the model's accuracy, the matrix's predicted values were grouped.

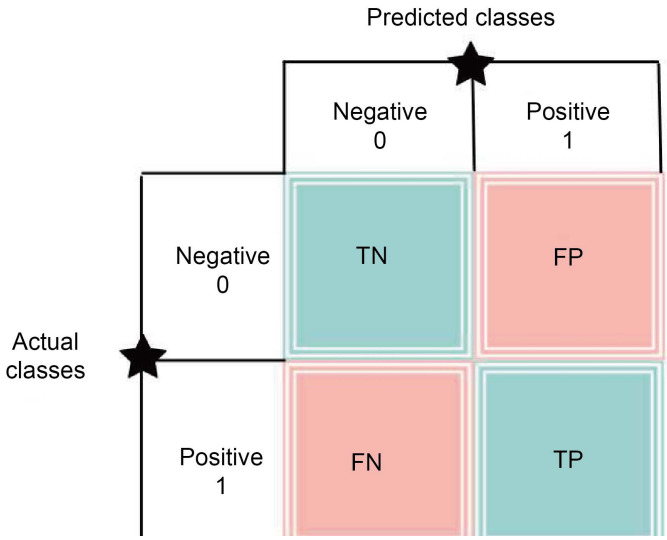

Figure 4. The confusion matrix block diagram.

Through the analysis, it is noticed that there is a category called "Other" in the "Gender" column, which will be deleted for simplicity. The last step in preprocessing involved unbalanced processing of the data (tagging). The stroke label was analyzed in the research, and it was found to have an unbalanced distribution. The majority of samples were from patients who had not experienced a stroke, with only 249 out of 5109 samples from the stroke group. To address this issue, the data underwent pre-processing and SMOTE was used to balance the data.

The standard scalar function is used to scale the features in a dataset so that they have a similar scale. Improving the performance of machine learning algorithms is essential. The data was split into a training set and a test set after performing feature selection with Extra Trees from which the three most important features of the prediction process were identified, with which the models would be trained. Four algorithms classification—RF, KNN, DT, and XGboost—were used to develop four models. The DT algorithm achieved an accuracy rate of 89.4% in the initial model training.

A KNN classification algorithm was used to build a second model with a value of K = 2, resulting in a 92.02% accuracy. The RF algorithm was used to build the third model, which achieved 94.54% accuracy. The XGBoost classification algorithm was utilized to build the fourth model resulting in a 96.45% accuracy rate. Figure 5 shows the true positives, false negatives, true negatives, and false positives

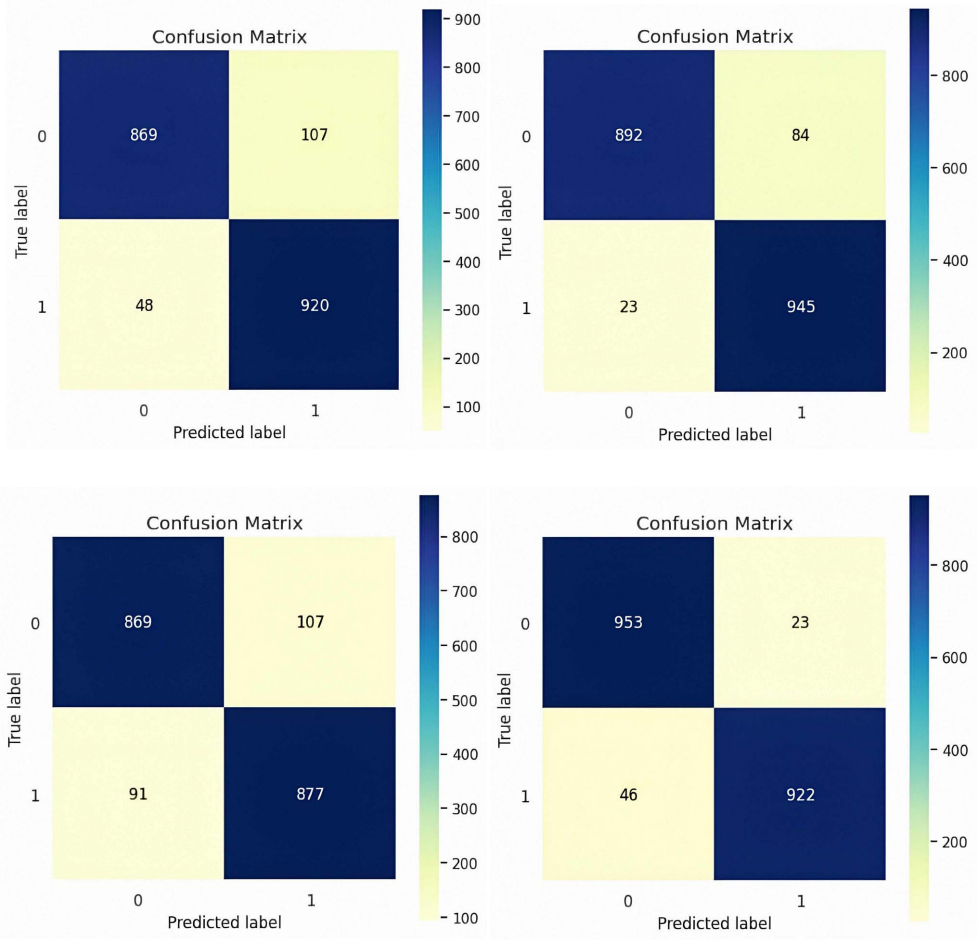

Figure 5. The confusion matrix for algorithm models.

that were obtained during the test. The performance evaluation measure used was accuracy, which is defined as the ratio of correct predictions to the total number of entries recorded. To calculate the accuracy rate, we used the following formula:

$$Accurcy = \frac{Correct\ Prediction}{Correct\ Prediction + Incorrectprediction} \times 100$$

Although our research used an equal number of datasets, our findings indicate According to the research methodology and data analysis, the XGboost model achieved the highest accuracy of 96.45% among all the algorithms utilized. **Figure 6** shows the accuracy of DT, KNN, RF, and XGBoost algorithms.

Table 1 illustrates the accuracy of implementing DT, KNN, RF, and XGboost algorithms. The results indicated that the XGBoost algorithm had the best accuracy and the highest accuracy outcome of 96.45%.

Meanwhile, **Table 2** provides a comparison of the proposed stroke data strategy with previous methods, using the same healthcare dataset, which shows higher discrimination accuracy. The accuracy is dependent on pre-processing, feature selection, and classification processes, which were approached differently in previous studies.

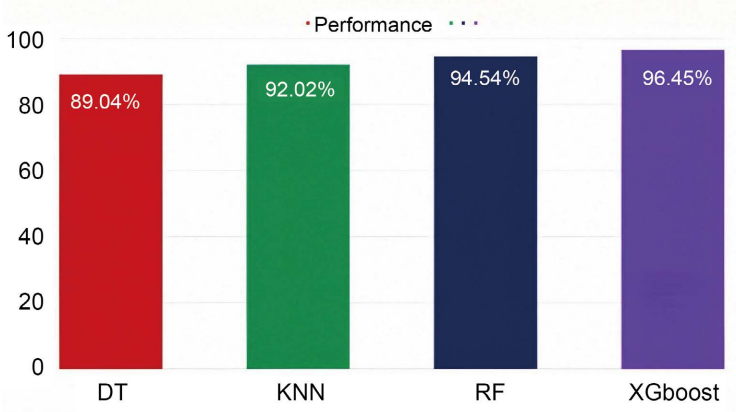

Figure 6. Illustrates the accuracy of implementing for all models.

Table 1. A Comparison of performance between all models.

Model	TP	FP	TN	FN	Accuracy rate %
DT	869	91	877	107	88%
KNN	869	48	920	107	92.02%
RF	892	23	945	84	94.54%
GXboost	953	46	922	23	96.45%

Table 2. Comparison of performance between previous studies and proposed study.

RelatedWorks	Year	Algorithm used	Accuracy
Our model	2023	XGboost	**96.45%**
Zubaidai *et al.* [13]	2022	RF	**94.6%**
Akter *et al.* [14]	2022	RF	**95.3%**
Sailasya *et al.* [15]	2021	NB	**82%%**
Hager *et al.* [16]	**2019**	**RF**	**90%**

5. Conclusion

Early detection of a brain stroke is crucial at an early stage to decrease long-term mortality rates, regardless of social or cultural background. Researchers have already utilized machine learning to forecast the likelihood of brain stroke with reasonable precision. In this paper, we have adopted a comparable technique but suggested a novel and enhanced strategy to further improve the accuracy of such predictions. According to research using the Healthcare Stroke Database, GXboost can accurately predict stroke. To ensure accuracy, outliers, and missing values were pre-processed and removed, and features important for the ordered classification process were selected using random forests. Compared to previous studies, the results of this study show that GXboost is a valuable tool for early stroke prediction. The XGBoost algorithm has demonstrated high accuracy in predicting stroke, achieving up to 96.45% accuracy in stroke risk prediction. The research indicates that GXboost could be a valuable tool for early stroke predic-

tion, which could ultimately improve the chances of survival and recovery for stroke patients. Data balancing and the selection of the important features enhanced the model's accuracy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Feigin, V. and Krishnamurthi, R. (2014) Epidemiology of Stroke. In: Norrving, B. Ed., *Oxford Textbook of Stroke and Cerebrovascular Disease*, Oxford University Press, 1-8. https://doi.org/10.1093/med/9780199641208.003.0001

[2] Lloyd-Jones, D., *et al.* (2009) Heart Disease and Stroke Statistics—2009 Update: A Report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*, **119**, e21-e181.

[3] World Health Organization (2021) The Top 10 Causes of Death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[4] Albers, G.W., Caplan, L.R., Easton, J.D., Fayad, P.B., Mohr, J.P., Saver, J.L., *et al.* (2002) Transient Ischemic Attack—Proposal for a New Definition. *New England Journal of Medicine*, **347**, 1713-1716. https://doi.org/10.1056/nejmsb020987

[5] Bamford, J., Sandercock, P., Dennis, M., Warlow, C. and Burn, J. (1991) Classification and Natural History of Clinically Identifiable Subtypes of Cerebral Infarction. *The Lancet*, **337**, 1521-1526. https://doi.org/10.1016/0140-6736(91)93206-o

[6] World Health Organization (2018) Global Health Estimates 2016: Disease Burden by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva.

[7] Zou, X., Li, Y., Xu, Y. and Yin, X. (2020) Clinical Risk Prediction of Intracranial Hemorrhage after Ischemic Stroke: The CHANCE Score. *Neurological Research*, **42**, 985-991.

[8] Larsson, S.C., Åkesson, A. and Wolk, A. (2015) Primary Prevention of Stroke by a Healthy Lifestyle in a High-Risk Group. *Neurology*, **84**, 2224-2228. https://doi.org/10.1212/wnl.0000000000001637

[9] Stroebele, N., Müller-Riemenschneider, F., Nolte, C.H., Müller-Nordhorn, J., Bockelbrink, A. and Willich, S.N. (2011) Knowledge of Risk Factors, and Warning Signs of Stroke: A Systematic Review from a Gender Perspective. *International Journal of Stroke*, **6**, 60-66. https://doi.org/10.1111/j.1747-4949.2010.00540.x

[10] Bhardwaj, R., Nambiar, A.R. and Dutta, D. (2017) A Study of Machine Learning in Healthcare. 2017 *IEEE* 41*st Annual Computer Software and Applications Conference* (*COMPSAC*), Turin, 4-8 July 2017, 236-241. https://doi.org/10.1109/compsac.2017.164

[11] Srinivas, S. and Ravindran, A.R. (2018) Optimizing Outpatient Appointment System Using Machine Learning Algorithms and Scheduling Rules: A Prescriptive Analytics Framework. *Expert Systems with Applications*, **102**, 245-261. https://doi.org/10.1016/j.eswa.2018.02.022

[12] Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R. (2020) RETRACTED ARTICLE: Classification of Stroke Disease Using Machine Learning Algorithms. *Neural Computing and Applications*, **32**, 817-828. https://doi.org/10.1007/s00521-019-04041-y

[13] Jeena, R.S. and Kumar, S. (2016) Stroke Prediction Using SVM. 2016 *International Conference on Control, Instrumentation, Communication and Computational Technologies* (*ICCICCT*), Kumaracoil, 16-17 December 2016, 600-602. https://doi.org/10.1109/iccicct.2016.7988020

[14] Sandercock, P.A.G., Niewada, M. and Członkowska, A. (2012) Erratum to: The International Stroke Trial Database. *Trials*, **13**, Article No. 24. https://doi.org/10.1186/1745-6215-13-24

[15] Singh, M.S. and Choudhary, P. (2017) Stroke Prediction Using Artificial Intelligence. 2017 8*th Annual Industrial Automation and Electromechanical Engineering Conference* (*IEMECON*), Bangkok, 16-18 August 2017, 158-161. https://doi.org/10.1109/iemecon.2017.8079581

[16] Yahiya, S., Yousif, A. and Bakri, M. (2016) Classification of Ischemic Stroke Using Machine Learning Algorithms. *International Journal of Computer Applications*, **149**, 26-31. https://doi.org/10.5120/ijca2016911607

[17] Sudha, A., Gayathri, P. and Jaisankar, N. (2012) Effective Analysis and Predictive Model of Stroke Disease Using Classification Methods. *International Journal of Computer Applications*, **43**, 26-31. https://doi.org/10.5120/6172-8599

[18] Al-Zubaidi, H., Dweik, M. and Al-Mousa, A. (2022) Stroke Prediction Using Machine Learning Classification Methods. 2022 *International Arab Conference on Information Technology* (*ACIT*), Abu Dhabi, 22-24 November 2022, 1-8. https://doi.org/10.1109/acit57182.2022.10022050

[19] Sailasya, G. and Kumari, G.L.A. (2021) Analyzing the Performance of Stroke Prediction Using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, **12**, 539-545. https://doi.org/10.14569/ijacsa.2021.0120662

[20] Fedesoriano (2020) Stroke Prediction Dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[21] Tanasa, D. and Trousse, B. (2004) Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems*, **19**, 59-65. https://doi.org/10.1109/mis.2004.1274912

[22] Rich Data (2021) SMOTE Explained for Noobs—Synthetic Minority Over-Sampling Technique Line by Line. https://rikunert.com/smote_explained

[23] Sanmorino, A., Marnisah, L. and Sunardi, H. (2023) Feature Selection Using Extra Trees Classifier for Research Productivity Framework in Indonesia. *Proceeding of the* 3*rd International Conference on Electronics, Biomedical Engineering, and Health Informatics*, Surabaya, 5-6 October 2022, 13-21. https://doi.org/10.1007/978-981-99-0248-4_2

[24] Witten, I.H. Frank, E. and Hall, M.A. (2016) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[25] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[26] Zhang, Z. (2016) Introduction to Machine Learning: K-Nearest Neighbors. *Annals of Translational Medicine*, **4**, 218-218. https://doi.org/10.21037/atm.2016.03.37

[27] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Zhou, T., *et al.* (2015) Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2, 1-4.