Chapter 1

Book Chapter

Predictive Model for Brain Stroke Detection

Shaheena K V *[1], Prapitha Gopi K †[2], and Pallavi T H ‡[3]

[1]Assistant Professor, Dept. Of MCA, Acharya Institute of Technology, Bangalore, India
[2]Assistant Professor, PG Dept. of CS, Ansar Women's College, Perumpilavu, Thrissur, India
[3]Dept. Of MCA, Acharya Institute of Technology, Bangalore, India

Abstract

Strokes significantly impact the central nervous system and rank among the leading causes of death globally. The most damaging types are ischemic and hemorrhagic strokes, with the World Health Organization (WHO) reporting that 3% of people suffer from subarachnoid hemorrhage, 10% from intracerebral hemorrhage, and 87% from ischemic stroke. Strokes result from disrupted blood flow to the brain, often due to arterial blockages or damage. This project aims to develop a Python-based machine learning model for accurate stroke prediction, using classification algorithms such as Random Forest and Bagging Classifiers. These models offer promising tools for assisting medical professionals in diagnosing strokes, enabling earlier intervention and personalized care, potentially reducing the long-term effects and improving patient outcomes.

Keywords: Brain Stroke Prediction. Machine Learning. Random Forest. Medical Data Analysis. Healthcare Analytics.

*Email: shaheena2935@acharya.ac.in Corresponding Author
†Email: prapithakkm@gmail.com
‡Email: pallavih.22.mcav@acharya.ac.in

1 Introduction

Stroke is a serious worldwide health concern. It is a debilitating medical disorder marked by inadequate blood supply to the brain that results in cell death. There are two main kinds of stroke: hemorrhagic, which is defined by brain hemorrhage, and ischemic, which is brought on by insufficient blood supply. Impaired brain function can result from any type; symptoms may include speech difficulty, vertigo, unilateral body weakness, and visual loss. As the consequences of a stroke can be irreversible and lead to complications including pneumonia and loss of bladder control, early detection and intervention are essential.(Sangha & Ishida, 2021). The main cause of stroke risk is high blood pressure; other factors that could be involved include high cholesterol, smoking, obesity, diabetes, diabetes mellifluous, end-stage kidney disease, and atrial fibrillation (Boehme AK, Esenwa C, & Elkind MSV, 2017). Hemorrhagic strokes are brought on by bleeding inside or around the brain and are frequently associated with burst brain aneurysms, whereas ischemic strokes are typically caused by blood vessel blockage, but there are less common reasons as well. Physical tests and medical imaging, including CT and MRI scans, are essential for diagnosis. CT scans may not show symptoms of early ischemic strokes. To assist in determining risk variables and excluding alternative reasons, further procedures such as blood tests and electrocardiograms are performed. In some cases of hemorrhagic stroke, surgery may be beneficial. Ideally, stroke rehabilitation takes place in specialized stroke units, albeit these facilities are not always accessible. Rehab after a stroke is crucial to the recovery process.(Capriotti & Murphy, 2016).

The objective of this study is to leverage machine learning techniques, more specifically the Random Forest Classifier, to construct an advanced stroke diagnosis prediction system. Ultimately, this approach hopes to improve patient outcomes and lessen the burden of stroke-related disability by assisting medical professionals in early stroke identification and intervention using data analysis and predictive modeling. It is evident that the features of the dataset can all point to certain risk factors. Based on the information given, we can assess a patient's risk of stroke. To achieve the required accuracy for the project, we will employ decision trees and random forests.

A stroke is a disorder in which there is a rupture of blood arteries in the brain, leading to brain damage. It could also happen if there is a disruption in the blood and other nutritional supplies to the brain globally. The majority of research has been done on heart stroke prediction. The World Health Organization (WHO) states that stroke is the leading cause of disability and death one on brain stroke risk. In light of this, several machine learning models are developed to forecast the likelihood of a brain stroke.(Daidone et al., 2024; Rahman, Hasan, & Sarkar, 2023). This work has trained five distinct models for use machine learning algorithms to provide precise predictions such as Naive Bayes clas-

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

sification, K-Nearest-Neighbors, Support-Vector Machine, Random– Forest classification, Logistic Regression classification, and Decision-Tree classification. Variable physiological elements form the basis of the models. Naive Bayes, which yielded an accuracy of almost 82%, was the algorithm completed this task the best. Utilizing a data-driven technique to diagnose brain strokes has financial advantages. For the implementation of Clinical Decision Support System (CDSS), a straightforward method utilizing Machine Learning (ML) classification algorithms may yield sufficient accuracy. According to the Devaki and Rao's (2022), improving prediction performance can be achieved by creating an ensemble of numerous brain stroke prediction models. This is research's hypothesis, which served as inspiration for the work's execution and presentation. Another significant discovery from the literature is that the majority of ensemble techniques for brain stroke prediction are not data- driven strategies. By concentrating on an ensemble of data-driven prediction models, our work closes this research gap. Utilizing supervised machine learning techniques, we proposed a collaborative framework to enhance brain stroke prediction accuracy.

People's biological characteristics have changed as a result of swift changes in human lifestyles, increasing their susceptibility to specific illnesses like stroke. An irreversible illness that causes permanent disability is stroke. It is currently one of world's primary reasons why death. In addition, it ranks second in Jordan behind ischemic heart disease as a cause of mortality. Early identification of a stroke improves the prognosis, enhances patient care, and removes potential problems. We use Naive Bayes along with various machine learning methods. Decision Trees, and KNN to predict stroke in this work using patient data that we thought would be connected to the cause of stroke. Health care providers can forecast stroke disease and provide a better treatment plan by using Orange software, which automatically processes data and generates data mining models. According the results, the decision-tree-classifier outperformed other methods, achieving an accuracy of 94.2%.(Ghannam & Alwidian, 2022) Encased in the skull, the brain is most complex and fascinating organ in the human body, comprising the cerebrum, cerebellum, and brainstem. To prevent brain damage, it is crucial to treat strokes promptly, as they are the second leading cause of death worldwide.(National Institute of Neurological Disorders and Stroke, 2023). Reducing the severity or preventing brain strokes can decrease the associated fatality rates. Machine learning algorithms are an effective way to identify risk factors. The model presented in this research offers an accurate brain stroke prediction and includes a detailed methodology. For the proposed model to be successful, effective techniques for data collection, preprocessing, and transformation have been employed to ensure accurate information.

A stroke is a medical condition that occurs when a ruptured blood vessel damages the brain. Symptoms can arise if the brain's supply of blood and other nutrients is disrupted.

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

The World Health Organization (WHO) states that stroke is the leading cause of death and disability worldwide. Stroke severity can be decreased by recognizing the many warning symptoms of a stroke early on. Various machine learning (ML) models have been created to forecast the chance of a brain stroke. This study trains four distinct models for dependable. At almost 96% accuracy, Random Forest proved to be the most effective algorithm for this particular assignment. The open-access Stroke Prediction datasets was the source of data utilized to build the technique.(Tazin et al., 2021). This investigation's models have a far greater accuracy rate than those utilized in earlier investigations, suggesting that they are more trustworthy. The analysis of the study suggests the method, and multiple model comparisons have shown how robust it is.

2 Methodologies Used

I Methods

These days, a rising number of strokes are resulting in many untimely deaths. Many modern approaches have been developed in the modern era to use healthcare data analytics to predict stroke symptoms. These systems evaluate an individual's medical data to predict strokes, potentially saving lives, by utilizing machine learning algorithms.These tools preprocess the dataset to improve categorical data or remove missing values before supplying it to machine learning algorithms. A stroke predictor dataset frequently contains variables such as blood pressure, age, and gender, heart health, BMI, glucose tolerance, and smoking status. The dataset is partitioned into fractions for the purpose to train and test models. Models and predictions are created via machine learning techniques including Bagging Classifier, Random Forest, Decision Tree and Logistic Regression.

- Random Forest Classifier: The Random Forest classifier is a machine learning technique that constructs a huge number of ensemble decision trees from a random collection of data. In the forest of randomness, every single tree produces a class prediction using both row and column sampling. Random Forest's fundamental idea is straightforward but extremely effective.
  The random forest algorithm is classifier constructed by a group of tree- structured classifiers h(x, k), k = 1,2,..., where the k are each independently identical. Spread random vectors, and at input x, every tree votes with one unit for the most popular class. Each is to extract n samples from an assortment of N input samples, where N is typically the amount of training set samples. From these M features, m features are harvested using column sampling. K training sets will be provided as S1, S2,..., Sk after K times the was chosen at random. After that, training sets will be used

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

to create matching decision trees T1 (T2,..., Tk). There isn't pruning needed as each and every tree in the woodland is fully developed.

Many Decision trees are assembled to create a random forest for classification purposes. The overall amount of choices made trees, or N tree, is another important variable. A new after building the random -forest model, the illustration is incorporated into the model. Next, each decision tree will make a determination as to which group this sample should fall into. An individual's ultimate classification can be ascertained by counting the votes cast across all of the decision trees inside the forest.

Bagging, also known as bootstrap aggregating, is a general technique that is applied to tree learners in the random forest training process. Assuming X = x1,..., xn is a training set responses Y = y1,..., yn, bagging repeatedly (B times) chooses a random sample in place of the drill set and fits trees to these samples: If b = 1,..., B:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(\hat{x}) \tag{1}$$

- Sample n training examples from X, Y, with replacement; refer to these as Xb, Yb.
- Use Xb and Yb to train a classification tree fb. Predictions for sample that have yet to be observed can be made using a mean of the results made by every element of the tree of regression on x' after training:

This bootstrapping procedure enhances the model's performance since it lowers the model's variance without raising bias. Accordingly, even if a single tree's forecasts are quite susceptible to noise in it's given a training set and no correlation between the trees, a mean of several trees is not. For the purpose of de- correlate the trees, bootstrap sampling involves exposing them to several training sets. Provided multiple trees were taught on a single training set, therefore, highly correlated trees—or even the same tree repeatedly, provided the method of training is deterministic. The reason random forests work so well is multiple relatively uncorrelated trees working together as a model outperform any single tree. The key factor is the weak association among the models. Uncorrelated models can generate group forecasts that are more precise than individual predictions, similar to how a diversified portfolio of low-correlated assets, like stocks and bonds, is greater than the addition of its parts.

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

- Bagging Classifier: We implemented the Bagging Classifier Machine Learning Algorithms after achieving a 98% accuracy on the set of tests. Bagging, or Bootstrap Aggregating is a method of group learning where a multitude each base model is trained separately and in parallel on different sub sets of the training data (see figure 1). These subsets are created through bootstrap sampling, which involves identifying data points at random with replacement. When using the Bagging Classifier, the final forecast is produced by adding up all the forecasts. Of all base models using majority voting. For regression tasks, by averaging the forecasts from each base model, the ultimate result is produced, a method known as bagging regression.

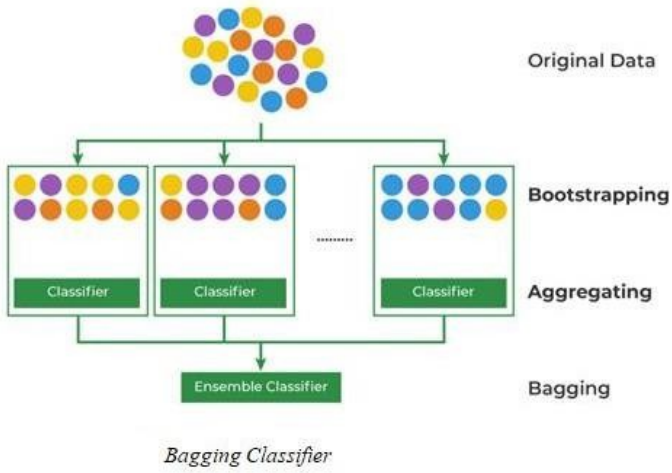

Figure 1. Bagging Classifier

- Decision Tree: A supervised learning machine technique known as decision trees might to be utilized using regression and classification issues alike. The Decision-Tree is a reversed tree. A Decision Tree makes judgments according to the circumstances revealed by the knowledge. It consists of decision leaves and decision nodes. The leaves represent the choices or outcomes, the nodes represent the features within the dataset and the branches represent the reasons for making those decisions. The training information set is used as the divided dataset's root to give a tree of decisions. Columns values should ideally be categorical before they are cleared up and preprocessed to discrete values, even if continuous values are more prevalent. Recursive record distribution is predicated on attribute values.

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

Entropy is a key component in decision tree construction. Entropy controls the data separation in terms of manner in which the decision tree builds its boundaries. Entropy values vary from 0 to 1, with less entropy denoting more reliability. The goal is to establish a training model utilize a decision tree by applying basic choice principle learned from historical data, can be applied to forecast the target variable's class or value.

3 Dataset Analysis

I Data Description

The dataset consists of 4982 individual data. There are 11 columns in the dataset, which are described below.

- gender: "Male", "Female" or "Other"
- age: patient's age
- Hypertension: Assigning 0 to patients without hypertension and 1 to those with hypertension.
- Heart disease: Assigning a value of 0 indicates absence of heart disease in the patient, while a number of 1 signifies presence of heart disease.
- Ever-married: "No" or "Yes"
- work type: "children", "Govtjob", "Never worked", "Private" or "Self-employed"
- Residence type: "Rural" or "Urban"
- avg glucose level: Mean blood glucose level
- BMI: Body mass index
- smoking status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- stroke: 1 if the patient had a stroke or 0 if not

II Data Cleaning

Certain values within the dataset can be character- based values, corrupt or erroneous entries, or missing altogether. For instance, the dataset had BMI NULL values, which required attention before being used. Since the model cannot read commas, brackets, or other special characters, they will all be eliminated from the data as soon as it loaded as a.csv file. With the values we currently have, we can additionally populate those null values. To fill in the blank BMI values in this stroke prediction dataset, we took the median among all BMI values. The ID column was likewise eliminated.

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

III Data Analysis

Data analysis is mostly necessary to understand the dataset. To improve our data visualization, we created several graphs (see figure 2). The two most significant numerical variables are Stroke and Normal.Figure 3 shows the data distribution for these variables.

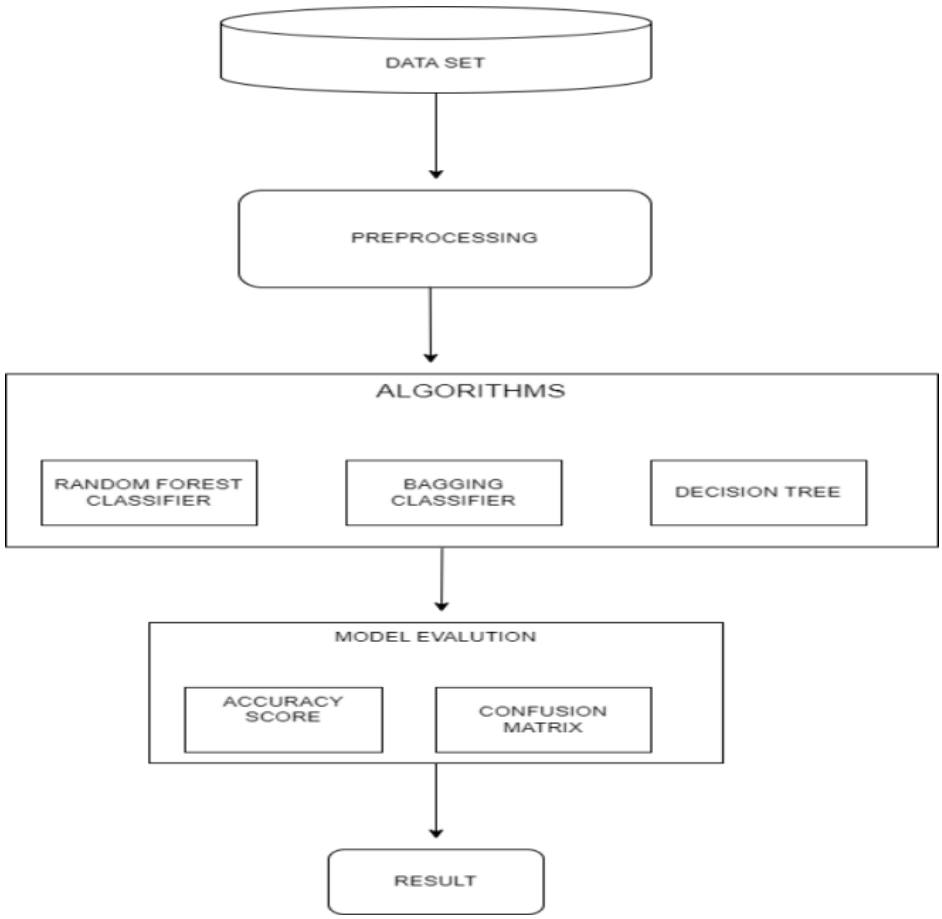

Figure 2. Stroke Prediction Model

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

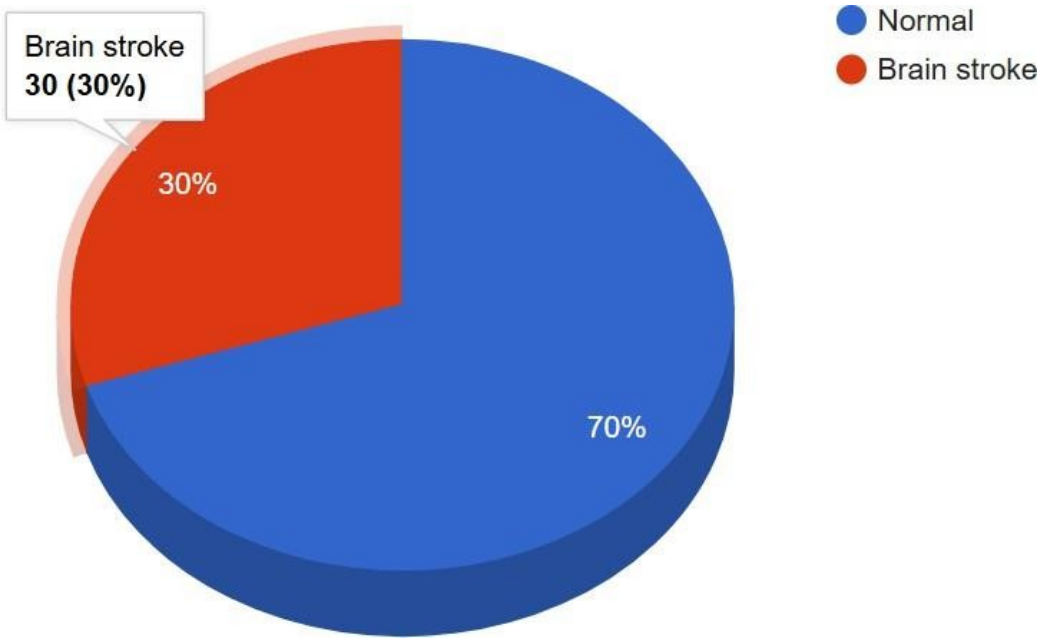

Figure 3. Pie chart of numerical variables are Stroke and Normal

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

IV Data Preprocessing The level of significance varies among the columns in the dataset. There are some that are completely unimportant. Our model does not take into account the ID column in the stroke prediction dataset. We shall so choose every feature barring ID.

4 Results

We used the stroke prediction dataset, which includes 4982 observations and 12 features, to predict strokes. This is how machine learning algorithms came out.

Table 1. Accuracy and Precision

Method	Accuracy	Precision
Random Forest	0.96	0.99
Bagging Classifier	0.94	0.96
Decision Tree	0.93	0.94

Table 1 shoes that Random-Forest achieved the high accuracy score of 0.96. However, before evaluating the result of the models, we must also review the other results. Our primary goal in this scenario is to identify individuals holding the greatest likelihood of being diagnosed with a stroke. To attain this, we compare the precision and accuracy of models that excel in predicting genuine advantages among patients. The Random Forest model makes accurate predictions 96.0% of the cases in the test set, demonstrating its superior ability to generalize from the usage data compared to the other models (see figure 4). Random Forest model's precision of 99.0% indicates that, in 99.0% of cases, its predictions about strokes are accurate. Reducing false positives is critical in medical diagnostics, and this incredibly high precision makes all the difference.

The Bagging Classifier achieves an precision of 94.0%, which exceeds which of the Decision-Tree but falls short of the Random Forest (see figure 5). Based on comparison, Ensemble Decision Trees demonstrates superior generalization performance compared to the Bagging Classifier. Precision (96.0%): The Bagging Classifier shows superior accuracy over the Decision Tree in predicting real-world stroke cases, but it lags behind Ensemble Decision Trees in this aspect (see figure 6). The Decision Tree model achieves an precision of 93.0%, correctly predicting the outcome in the majority of cases. While this accuracy is Bagging Classifiers and Random Forest, it remains notably high. Precision (94.0%): The Decision Tree's accuracy in predicting

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

positive strokes is comparatively less than the Random Forest and Bagging Classifier, with a precision of 94.0%. This suggests a higher incidence of false positives. If we take a look at the confusion matrix, we may get a clear idea

Here, True Negative = 0/0 False Negative: 0/1 False Positive = 1/0 1/1 = Veritable Positive

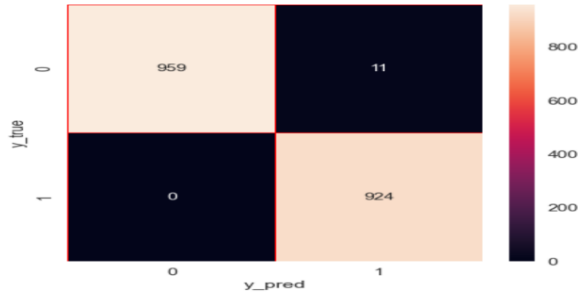

Figure 4. Random Forest Classifier for Confusion Matrix

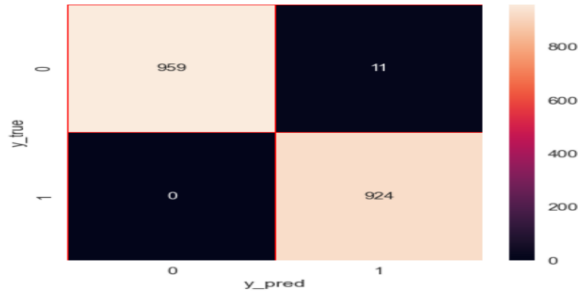

Figure 5. Bagging Classifier for Confusion Matrix

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

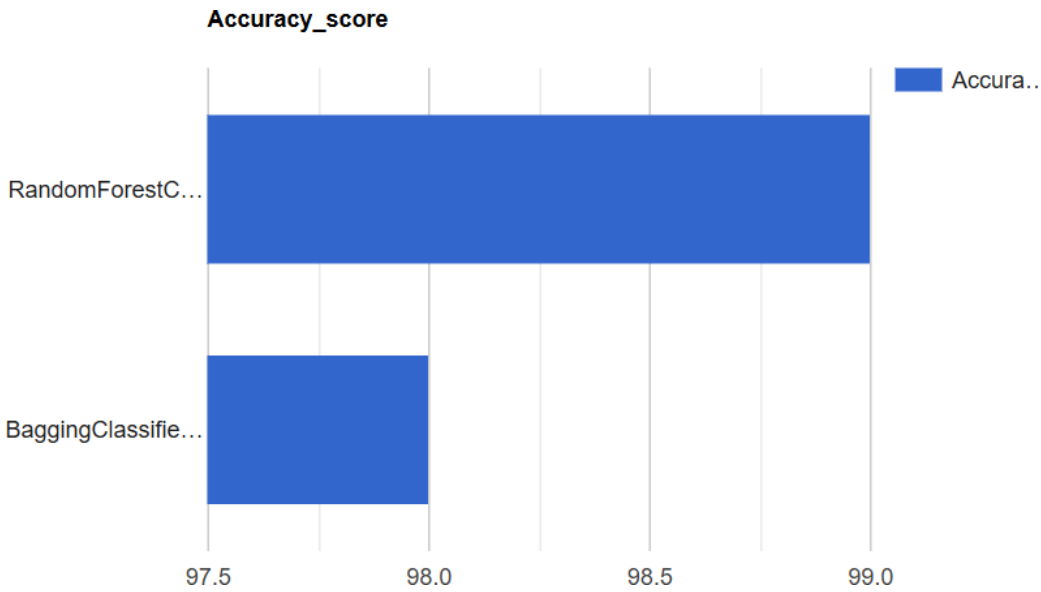

Figure 6. Accuracy of Bagging Classifier and Random Forest

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

5 Conclusion

The progression of "A Machine Learning Model for Brain Stroke Prediction" marks a notable stride in medical diagnostics and healthcare. This research sought to establish a robust and accurate tool for early stroke detection through the utilization of the Random Forest Classifier. We diligently gathered, processed, and curated features for an all-inclusive dataset used in model training. Leveraging the ensemble learning method of the Ensemble of Decision Trees facilitated precise and reliable predictions of stroke diagnoses. Its robustness to data imbalances and interpretability made it a valuable asset in clinical decision-making. Moreover, ethical considerations and fairness were at the forefront of our system's design. Bias detection and mitigation techniques were employed to ensure favorable healthcare scores for all patient groups, while strict privacy compliance measures were implemented to safeguard patient confidentiality. Collaboration with healthcare experts played a crucial part in refining the system's features and evaluation criteria, aligning it closely with the requirements for real-world medical practice. The system's scalability and user-friendly interfaces make it practical for deployment in clinical settings, where it can assist healthcare professionals in making timely and informed decisions. In conclusion, the "Machine Learning Model to Predict Brain Strokes" offers a practical method for early stroke detection. Its advantages in precision, robustness, interpretability, ethical considerations, and cooperation with healthcare professionals make it's an essential instrument for enhancing patient outcomes and lessening the effects of stroke-related disabilities. This initiative underscores the potential ML applications in health care emphasizing the significance of ethical considerations in developing medical diagnostic tools.

References

Boehme AK, Esenwa C, & Elkind MSV. (2017). Stroke Risk Factors, Genetics, and Prevention. Circulation Research, 120(3), 472–495.

Capriotti, T., & Murphy, T. (2016). Ischemic STROKE. Home Healthcare Now, 34(5), 259–266. https://doi.org/10.1097/NHH.0000000000000387

Daidone, M., Ferrantelli, S., Tuttolomondo, A., Daidone, M., & Daidone, M. (2024). Machine learning applications in stroke medicine: Advancements, challenges, and future prospective. Neural Regeneration Research, 19(4), 769–773. https://doi.org/10.4103/1673-5374.382228

Devaki, A., & Rao, C. V. (2022). An Ensemble Framework for Improving Brain Stroke Prediction Performance. 2022 1st International Conference on Elec-

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5

trical, Electronics, Information and Communication Technologies, ICEEICT 2022. https://doi.org/10.1109/ICEEICT53079.2022.9768579

Ghannam, A., & Alwidian, J. (2022). A Predictive Model of Stroke Diseases using Machine Learning Techniques. International Journal of Recent Technology and Engineering (IJRTE), 11(1), 53–59. https://doi.org/10.35940/ijrte.a6900.0511122

National Institute of Neurological Disorders and Stroke. (2023). Brain Basics: Know Your Brain. National Institute of Neurological Disorders and Stroke, 1–6. https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-know-your-brain

Rahman, S., Hasan, M., & Sarkar, A. K. (2023). Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. European Journal of Electrical Engineering and Computer Science, 7(1), 23–30. https://doi.org/10.24018/ejece.2023.7.1.483

Sangha, N., & Ishida, K. (2021). Acute Stroke. Seminars in Neurology, 41(1), 3. https://doi.org/10.1055/s-0041-1722922

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering, 2021. https://doi.org/10.1155/2021/7633381

DOI:10.48001/978-81-966500-2-5-1 | ISBN: 978-81-966500-2-5