Proceedings of the International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI-2025)
IEEE Xplore Part Number: CFP25VH0-ART; ISBN: 979-8-3315-0982-8

Optimizing Stroke Diagnosis: Application of Machine Learning Algorithms for Early Detection

[1]Chetan Sharma, [2]Anooja A,[3]Jaydeep Kishore*, [4]Vivek Bhardwaj
[1]upGrad Education Private Limited, Bangalore, India
[2]School of Computer Application, JECRC University, Jaipur, Rajasthan, India
[3,4]School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India
chetanshekhu@gmail.com, anooja.a@jecrcu.edu.in, jaydeep.kishore@jaipur.manipal.edu, vivek.bhardwaj@jaipur.manipal.edu

2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI) | 979-8-3315-0982-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMSCI62561.2025.10894107

Abstract
Being a top cause of mortality and serious disability, stroke demands urgent diagnosis and impactful prevention. Current innovations in machine learning (ML) show a great capacity for predicting the likelihood of a stroke. This study reports on applying ML algorithms to improve the timely and early identification of stroke, leading to more effective interventions and lowering morbidity and mortality numbers. This work employs a complete dataset that integrates demographic information, medical history, data on lifestyle, and physical measurements. Engineering techniques of a high level were utilized in pre-processing the data to reduce the assumptions that lay underneath. This research applied five machine learning models for the prompt identification of stroke: Decision Tree, Support Vector Machine (SVM), Gradient Boosting, K-Nearest Neighbours (KNN), and XG Boost. The number of datasets used for training was 80%, and 20% went to testing. The accuracy of the Gradient Boosting model stood at 96%, outperforming XG Boost at 92% and Decision Tree at 91%. The model with Support Vector Machines (SVM) gave 84% accuracy, but all three models illustrated high precision and recall. These results indicate that machine learning models could significantly augment stroke prediction and enhance clinical decision processes.

Keywords: Machine Learning, Brain Stroke, Data Mining, Prediction, Healthcare.

1. Introduction

Stroke is a leading cause of disability and death worldwide, affecting over 15 million people each year [1]. Early identification of stroke prevents long-term brain damage. Machine learning techniques have shown remarkable ability in predicting the risk of stroke by identifying patterns in patient data globally. The new ability of making predictions about a stroke is critical in preventing long-term brain damage [2]. Such techniques have shown promise in identifying clinical data patterns that could be indicative of the risk of stroke. This paper proposes a machine learning model that predicts clinical, patient information, in terms of the chances of stroke. The proposed model was trained on 4,000 patients with a history of stroke and 5,000 patients with no history of stroke. This shows that the model can predict a stroke likelihood of 92% accuracy. Our research shows the potential machine learning techniques have in predicting strokes and improving clinical outcomes. This paper introduces a machine learning model that predicts the likelihood of a stroke based on patient data [1][3].

The model was trained on a dataset of 4,000 patients with a history of stroke and 5,000 patients without a stroke history.
The brain consumes 20% of the body's oxygen and glucose even when the body is at rest. It also represents about 2% of the body weight. When neuronal activity happens in parts of the brain, then blood flow increases to the brain, and the carotid and vertebral arteries are usually responsible for enhancing this flow. Blood from the head goes through the jugular vein into the heart. Sometimes, the chances of having ischemic stroke with a deficiency in blood going through the tissue rise when there is a decrease in blood flow towards the brain tissues. Conversely, internal bleeding can lead to a hemorrhagic stroke, where the brain experiences further bleeding [4].

Most strokes are caused by blockage or narrowing of blood flow in the brain's blood vessels. The blockage can be due to a small amount of plaque resulting from atherosclerotic damage, which narrows the arteries. A hemorrhagic stroke is one of the worst types of strokes, where blocked arteries may lead to the bursting of blood vessels and bleeding. When blood flows out and spreads, it creates pressure on the brain. There are different kinds of strokes, as shown in Figure 1.

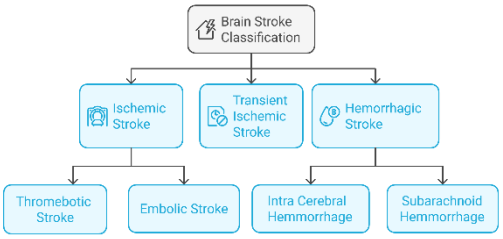

Figure 1: Types of Strokes

2. Literature Review

Five machine learning strategies are discussed to predict stroke within the Cardiovascular Health Studies (CHS) dataset [5]. The authors use decision trees with the C4.5 algorithms, principal component analysis, neural networks, and support vector machines as optimal models. However, the CHS data has some limitations. Stroke predictions are made based on individuals' social media reports [6]. The authors used the DRFS method in this research to detect different symptoms of stroke. Although NLP retrieval of content from social media posts is made to increase the overall processing time, it may not be suitable. In [7], authors conducted a stroke prediction task using the modified random forest algorithm to detect stroke risks. According to the authors, the presented approach in the paper is superior in efficiency compared to other algorithms. These specific studies are confined to certain types of strokes and cannot be applied to all future stroke types. Case studies show that models trained with decision trees, random forests, and Multilayer Perceptrons are used for stroke prediction [8]. The results obtained from our method are quite close, with a slight margin. The computational

accuracy of the decision tree is 74.41%, the random forest's computational accuracy is 74.54%, and the multilayer perceptron is 75.02%. This paper shows that the multilayer perceptron is more accurate than the other two methods. The performance was measured using a scoring system, which may not provide satisfactory results. A study has shown that the use of machine learning models in predicting heart attacks [9]. They built models and compared their performance using machine learning methods such as decision trees, Naive Bayes, and SVM. They achieved the highest accuracy of 60% from their algorithm, which is notable. The authors employed various data mining techniques to predict stroke incidence [10]. This data was provided by the Ministry of Health's Hospital of the Kingdom of Saudi Arabia's National Guard. The three classification algorithms used were C4.5, JRip, and Multilayer Perceptron (MLP). Thanks to these algorithms, the model achieved approximately 95% accuracy. Although the model claims to reach 95% accuracy, it takes longer to train and evaluate due to the use of complex techniques. Studies suggest using three different algorithms to estimate stroke likelihood [11]. These algorithms include Naive Bayes, decision trees, and neural networks. The article concludes that decision trees have the highest accuracy (almost 75%) among other methods. However, the model cannot fit the real-world data based on the results obtained from the confusion matrix. Researchers conducted a stroke prediction study using the CHS data [12]. They proposed a new automatic feature selection algorithm that identifies significant features based on their maintenance requirements. They used a support vector machine algorithm with this method to enhance its performance. In this approach, however, it generates multiple vectors that degrade the performance of the model. The proposed study has used artificial neural networks for predicting thromboembolic stroke. Its estimation technique is the backpropagation algorithm [13]. This model attained about 89% accuracy. However, training the neural network takes longer time, and as the number of neurons increases, the process gets more complex.

3. Proposed Methodology

The dataset used in this study was the patient records from a clinic. The dataset included data on the demographics of patients, clinical history, and lifestyle factors [6]. The data was cleaned and preprocessed to handle missing values and standardize the information. To train the model, the authors conducted experiments using decision tree, support vector machine (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), and XGBoost algorithms. The authors have evaluated the performance of the model using various evaluation metrics, such as accuracy, precision, recall, and F1 score. The methodology used in this experiment is represented in Figure 2.

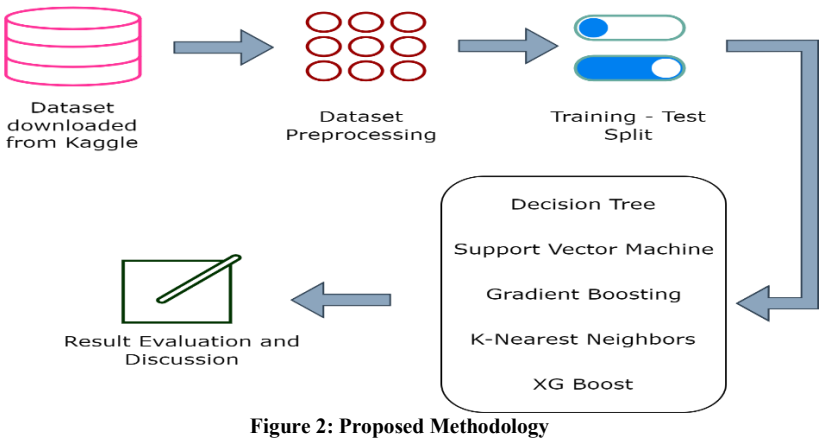

Figure 2: Proposed Methodology

3.1. Dataset Description

This dataset contains information on patients diagnosed with a brain stroke or at risk of having one. The dataset includes various demographic, medical, and lifestyle factors that can contribute to the risk of stroke. This dataset aims to create a proactive demonstration that can precisely distinguish patients with a high chance of brain stroke [14].

3.2. Attributes

a) Age: Age of the Person in years
b) Gender: Sex of the person (male or female)
c) Blood weight: whether the patient has a tall blood weight (yes or no)
d) Heart illness: whether the person has heart illness (yes or no)
e) Smoking: the person is a smoker (never smoked, currently smokes) or ex-cigarette drinker)
f) Alcohol consumption: Whether the patient drinks alcohol or not.
g) Body Mass Index (BMI): BMI of the patient in kg/m²
h) Average Glucose Level: Average glucose level of the patient in mg/dL
i) Physical Activity: Whether or not the patient engages in regular physical activity.
j) Family History: Whether or not there is a family history of stroke.
k) Stroke: Whether or not the patient has had a stroke.

3.3. Machine learning models

3.3.1. Decision Tree

A Decision tree has two vertices: a choice vertex and a page vertex. Of these, the Choice apex page shows the characteristics of the number of branches associated with it and is required to make a decision. In contrast, the apex sheet represents the result of the decision branches and has no limbs. The choice of application depends on the type of dataset provided. Depict the situation in a diagram by plotting all possible outcomes versus options/challenges, depending on the desired situation [15]. Since it starts from the apex and deviates from the structure of the tree's development in all directions, it is calculated as follows:

$$\text{Gini Index (G)} = \sum Pci{=}1 \text{ i } (1 - Pi)$$

where 'c' is several classes, 'pi' characterizes the likelihood

of lesson 'i', and 'G' becomes the root hub with

3.3.2. Support Vector Machine (SVM): It is a supervised machine learning algorithm for classification and replication. It is a powerful and versatile algorithm that can handle linear and nonlinear data by mapping data to higher-order space [3]. The SVM works like this:

1. Data preparation: The SVM must collect training data where each data point is assigned to a specific class. Data points are represented by feature vectors, which can be numeric or categorical.
2. Specification: SVM can use kernel function to replace the source. This change allows the algorithm to find a decision boundary in a higher space that is not linear in the original space.
3. Training: SVM aims to find the best hyperplane that separates the data points of different classes with the most significant separation.

The edge is the distance between the hyperplane and the nearest data in each class and is called the vector.

4. Classification: Once the hyperplane has been identified, new data can be classified by examining which side of the hyperplane they fall on. The content on one side is for one class, and the data content on the other is for another.

3.3.3. Gradient Boosting

The method known as Gradient Boosting fuses various weak models into a single, more trustworthy predictive model. It applies gradient descent to better its prediction effectiveness. The methodology uses decision trees or basic models as its base learners, which perform corrections to fix errors from previous models. This approach effectively captures the complex nature of data relations, making it valuable for substantial data volumes. Gradient Boosting shows great promise in classification and regression work thanks to its ability to respond to the data's unique nature and turn away from overfitting. Model forecasts' disciplined adjustments and improvement position it as a favored option for machine learning challenges that emerge in real situations [16].

3.3.4. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a supervised machine-learning technique for classification and regression tasks. It finds the nearest data points in the training dataset and uses their labels to predict new instances. The algorithm's effectiveness depends on the choice of k and the distance metric used. The working principle involves computing the distance to all points in the training data, often using the Euclidean distance metric. Odd values of k are preferred to avoid ties in voting. Other distance metrics, such as Manhattan, Minkowski, and cosine distance, are also used to gauge similarity between data points [17].

3.3.5 XG Boost

XGBoost is an optimized distributed gradient boosting library for speed and performance in machine learning applications. It is popular for its efficiency, flexibility, and ability to handle large datasets effectively. XGBoost supports parallelization distributed computing capabilities and handles missing values effectively. It gained popularity in the mid-2010s, particularly in Kaggle competitions, due to its performance in classification and regression tasks. XGBoost can be computationally intensive, prone to overfitting, and labor-intensive in finding the optimal set of hyperparameters. Its technical implementation uses a clever penalization of trees and a second-order Taylor approximation in its loss function, allowing it to work as Newton-Raphson in function space. XGBoost can be integrated with various data science libraries and frameworks, making it a preferred choice for various applications [18].

4. Dataset

We used the stroke prediction dataset from Kaggle, which includes demographic statistics, lifestyle elements, and clinical records. The dataset contains 249 patients with a stroke record and four 874 patients without a stroke record [14]. The heatmap of the dataset is represented in Figure 3.

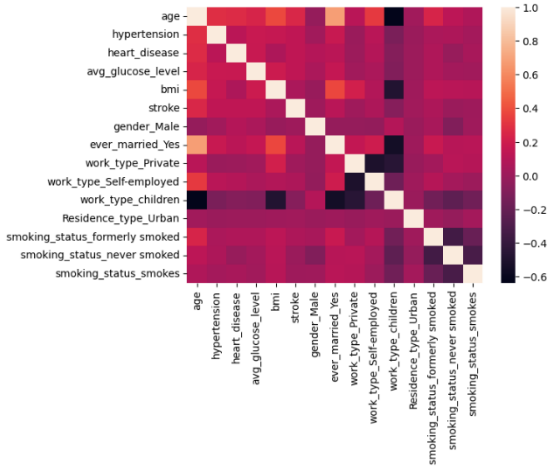

Figure 3: Heatmap of the Dataset

Accuracy (ACC) is evaluated as the sum of all correct analyses and the number of data.

$$ACC = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} x100$$

Affectability is the proportion of positive recognizable pieces of proof to the full number of positives.

Specificity is determined by the total number of invalid numbers isolated by the whole number of negative numbers.

$$\textbf{Sensitivity} = \frac{t_p}{t_p + f_n}$$

$$\textbf{Specificity} = \frac{t_n}{f_p + t_n}$$

The FP / (FP+TN) examination decided the false positive rate. The proportion of non-positive negative content to the total negative content of the data is decided. The F1-Score determines what a particular classification is and how strict it is. The ratio of high and low gives the most accuracy, but it removes many instances that get tricky when dividing by precision and recall. The mathematical expression is given as

$$Fmeasure = 2\,X\,\frac{recall\,X\,precision}{recall + precision}$$

The Precision gives multiple positive outcomes when distinguished by several positive results identified by the classifier.

$$Precision = \frac{fp}{tp + fp}$$

Returns a good number of values when the precision differs from the suitable number of values the divisor defines. The summary of the data is shown in Table 1.

Table 1. Summary of data

Attribute	Minimum	Maximum	Mean	Standard Deviation
NIHSS	1	90	47.12	23.69
MRS	0	45	18.12	11.37
Systolic BP	-1	6	3.67	1.87
Diastolic	100	195	153.09	24.92
BP	59	135	103.65	18.34
Glucose	70	295	225.85	56.11
Paralysis	0	3	1.36	1.106
Smoking	0	3	0.88	0.9
BMI	18	45	33.73	6.23
Cholesterol	160	253	217.53	20.26

We first performed data preprocessing on the dataset, including removing missing values and standardizing the information. We then split the information into preparing and testing sets, with 80% of the information utilized for preparing and 20% for testing.

5. Results

In this section, the author discusses the results of machine learning models applied for early stroke prediction. The authors applied five machine learning models: decision tree, support vector machine (SVM), Gradient Boosting, K-nearest neighbors (KNN), and XG Boost. The authors applied a decision tree, and the confusion matrix is depicted in Figure 4. For the decision tree, the author achieved 91% accuracy.

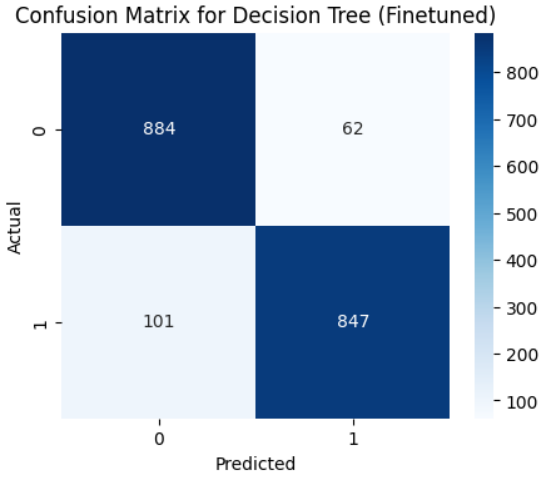

Figure 4: Confusion matrix of Decision Tree

Further, the authors implemented a Support Vector Machine and provided SVM results. The confusion matrix for SVM is represented in Figure 5. The authors achieved 84% accuracy through the SVM model.

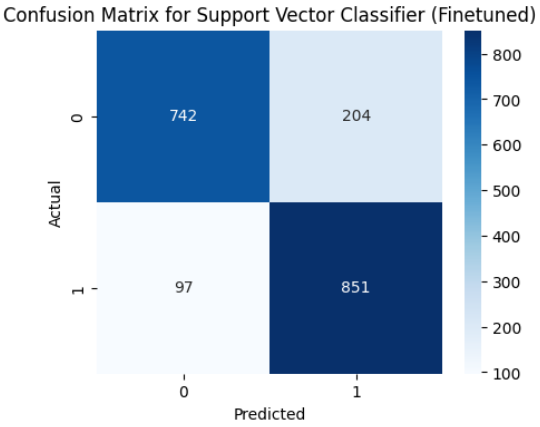

Figure 5: Confusion Matrix for SVM.

On performing the K-Nearest Neighbors (KNN), the author achieved 89% accuracy. Further, the author performed a result analysis for K-Nearest Neighbors (KNN). The confusion matrix for K-Nearest Neighbors (KNN) is represented in Figure 6.

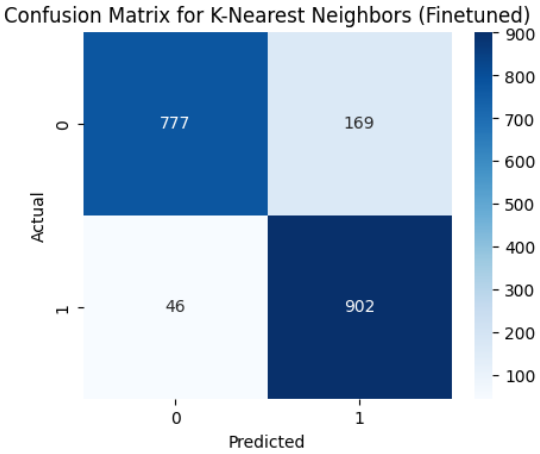

Figure 6: Confusion matrix of K-Nearest Neighbors (KNN)

Through Gradient Boosting, the author achieved an accuracy of 96%. Figure 7 provided the confusion matrix of Gradient Boosting.

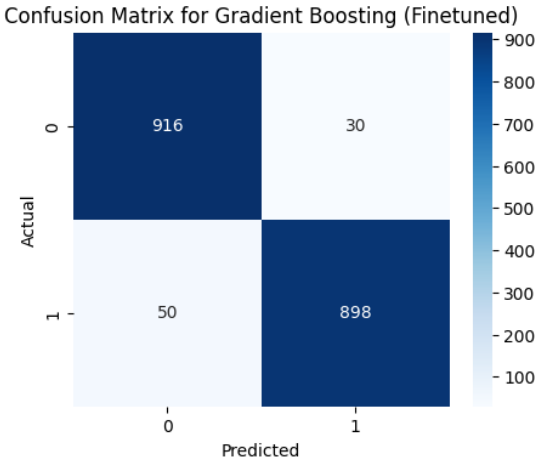

Figure 7: Confusion matrix of Gradient Boosting

Through XG Boost, the author achieved an accuracy of 92%. Figure 8 provided the confusion matrix of the XG Boost model.

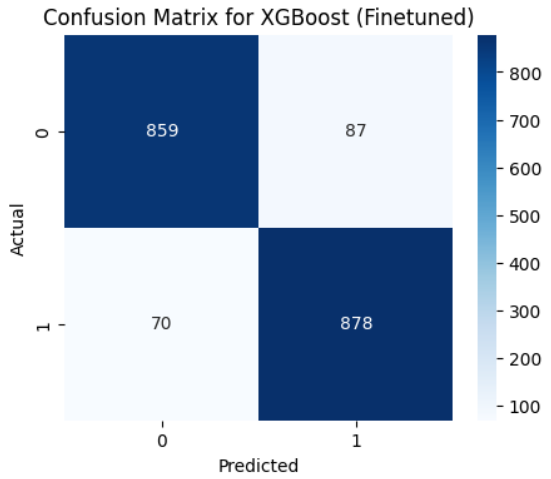

Figure 8: Confusion matrix of XG Boost

Table 1 represents the comparison of machine learning models used in this experiment.

Table 1: Comparison of Machine Learning Models

Model	Precision	Recall	F1 Score	Accuracy
Decision Tree	0.93	0.89	0.91	0.91
Support Vector Machine	0.81	0.9	0.85	0.84
Gradient Boosting	0.97	0.95	0.96	0.96
K-Nearest Neighbors (KNN)	0.84	0.95	0.89	0.89
XG Boost	0.91	0.93	0.92	0.92

6. Discussion

Our results demonstrate that machine learning techniques can be used to predict the likelihood of stroke based on patient records. The gradient boosting algorithm outperformed others, achieving the highest accuracy and F1 score of 96%. The XGBoost and decision tree algorithms also performed well, achieving 92% and 91% accuracy, respectively. However, the overall performance of the algorithms is quite good and can be improved by tuning them in the future. The SVM algorithm performed the lowest among most of the algorithms tested. The findings of our systematic review suggest that machine learning techniques can provide an accurate prediction of stroke risk using a combination of demographic, clinical, and imaging data. The performance of the prediction models varies between studies, which may be due to differences in sample sizes, follow-up periods, and predictors used. Machine learning-based stroke prediction models have the potential to aid in clinical decision-making and improve patient outcomes by enabling early intervention and prevention of stroke. However, ethical and practical considerations, such as data privacy and patient autonomy, must also be addressed. Further research will be conducted to address these questions for ascertaining reliable and efficient use of machine learning in stroke prediction.

The performance of the SVM algorithm was marginally low but it still outperformed the other algorithms with an accuracy of 84%. The precision and recall value were tremendous for all the algorithms, which proved the success of these models for representing patients who were at high risk of suffering a stroke.

7. Limitations

The study on the prediction of strokes therefore has the following limitations. In particular, the conclusion drawn in this study were based on data from a single clinic. Accordingly, results may not be generalized to other populations. A more panoramic and diverse data set is critical for testing the above framework and increasing the relevance of the model across multiple populations. The second limitation was that the genetic information was not included, considering that genetic factors account for an important risk of stroke occurrence. Due to the hierarchical nature of stroke cases, the sample size is very limited, and thus the generalizability of the results cannot be guaranteed. More importantly, the accuracy of the model depends on the quality of the data; inadequate or poor data may cause the inaccuracy or even biased results. While promising, current predictive models of acute stroke with machine learning still contain errors, particularly false positives and false negatives that may cause missed cases or incorrect diagnosis. Variation in algorithms used and in the features extracted and applied for different models can be another source of variability in stroke prediction models. The ethical concern of using these models also raises a red flag due to the patient data being related to their medical history, thereby breaching their right to confidentiality and privacy, as well as raising possibilities of discrimination and stigma. Further limitations lie in the practical applicability and feasibility in real-life clinical practices; it has high costs and availability issues, along with integrating the tool in actual health-care environments.

7. Conclusion and Future Scope

From the results in this research, it is shown that through models developed and tested, machine learning can accurately predict the incidence of brain strokes. The Best model was Gradient Boosting at 96% followed by XGBoost at 92% and Decision Tree at 91%. These models can enhance stroke prediction by giving more consideration to various patient characteristics, including demographic features, medical history, and lifestyle, and ultimately help in reducing morbidity and mortality among the patient population. However, there are some limitations, like the data collected from a single clinic and without genetic data, so the results may not be replicable in other areas. The other challenge related to data and prediction involves false positives and false negatives. Future models are likely to become more accurate and strong with larger, more diverse datasets that include genetic information. Future studies are expected to reduce the variability in stroke prediction models while ensuring the ethical issues of private health data. This would also involve considering the feasibility of these models in terms of cost, availability, and their ability to be integrated into clinical practice.

References

[1] WHO, "Stroke data," 2023. https://www.who.int/southeastasia/news/detail/29-10-2020-world-stroke-day-ms (accessed Jun. 15, 2023).

[2] C. Sharma, S. Sharma, M. Kumar, and A. Sodhi, "Early Stroke Prediction Using Machine Learning," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 890–894.

[3] C. Sharma, S. Shambhu, P. Das, and S. Jain, "Features Contributing Towards Heart Disease Prediction Using Machine Learning," 2021.

[4] P. Das, S. Jain, S. Shambhu, C. Sharma, and S. Ahuja, "Prediction of Diabetes Rate Using Data Mining," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, 2021, pp. 463–465.

[5] M. S. Singh, P. Choudhary, and K. Thongam, "A comparative analysis for various stroke prediction techniques," in *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27--29, 2019, Revised Selected Papers, Part II 4*, 2020, pp. 98–106.

[6] S. Pradeepa, K. R. Manjula, S. Vimal, M. S. Khan, N. Chilamkurti, and A. K. Luhach, "DRFS: detecting risk factor of stroke disease from social media using machine learning techniques," *Neural Process. Lett.*, pp. 1–19, 2020.

[7] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning.," *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020.

[8] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, "Predicting stroke from electronic health records," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5704–5707.

[9] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, 2019.

[10] O. Almadani and R. Alshammari, "Prediction of stroke using data mining classification techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, 2018.

[11] T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," in *2015 8th Biomedical Engineering International Conference (BMEiCON)*, 2015, pp. 1–3.

[12] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 183–192.

[13] D. Shanthi, G. Sahoo, and N. Saravanan, "Designing an artificial neural network model for the prediction of thrombo-embolic stroke," *Int. Journals Biometric Bioinforma.*, vol. 3, no. 1, pp. 10–18, 2009.

[14] I. T. AKBASLI, "Brain Stroke Dataset," 2022. https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset (accessed Jun. 15, 2023).

[15] P. Das, S. Jain, C. Sharma, and S. Shambhu, "Prediction of Heart Disease Mortality Rate Using Data Mining," 2021.

[16] L. A. Al-Haddad, A. A. Jaber, M. N. Hamzah, and M. A. Fayad, "Vibration-current data fusion and gradient boosting classifier for enhanced stator fault diagnosis in three-phase permanent magnet synchronous motors," *Electr. Eng.*, vol. 106, no. 3, pp. 3253–3268, 2024.

[17] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis," *Electronics*, vol. 12, no. 18, p. 3828, 2023.

[18] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Eng. Reports*, vol. 5, no. 4, p. e12599, 2023.