Cureus Journal of Computer Science

Open Access Original Article

Detecting Brain Stroke Using Enhanced Ensemble Models in Data-Driven Healthcare Systems for the Future Factories

Bikram Shit [1] , Tamosa Chakraborty [1, 2] , Nashreen Nesa [1]

1. Computer Science and Engineering, Aliah University, Kolkata, IND 2. Computer Science and Engineering, Haldia Institute of Technology, Kolkata, IND

Corresponding author: Tamosa Chakraborty, tamosa.hit.it@gmail.com

Received 05/01/2025
Review began 05/26/2025
Review ended 06/23/2025
Published 06/23/2025

DOI:
https://doi.org/10.7759/s44389-025-06211-7

Abstract

Brain stroke, a severe cerebrovascular phenomenon, is often caused by a lack of blood supply to the brain, leading to the death of brain cells and possibly causing permanent disabilities. According to the World Health Organization, stroke is one of the major causes of global mortality. As a result, many predictive models using machine learning (ML) have become essential for early detection due to interlinked data-driven environments. In this paper, we explore the possibility of detecting brain stroke through ML algorithms such as Decision Tree, Random Forest, K-Nearest Neighbors, Logistic Regression, Support Vector Machines, and Naive Bayes. We have demonstrated their application in future factories, as they require advanced healthcare systems that provide real-time health monitoring. We have developed accurate stroke prediction models using physiological data that align with the idea behind future factories, where proactive healthcare is enhanced through the collection of real-time data. Moreover, we have used ensemble techniques such as bagging to enhance the accuracy of the ML algorithms. After ensembling, the Decision Tree algorithm was found to produce the best performance, demonstrated by its 97% accuracy rate. The high accuracy presented by these models indicates their application value for future factory systems, providing smarter health monitoring systems and improved safety protocols at industrial sites.

Categories: AI/ML-based decision support systems, Machine Learning (ML), Health Informatics
Keywords: brain stroke, machine learning, health-care, future factory, data analysis

Introduction

A person's health is among the vital components of life, necessitating effective systems to track and manage health data. One of the critical health issues is stroke, a severe cerebrovascular accident resulting from a lack of blood supply to the brain, causing loss of functions in areas controlled by that region. Globally, strokes are one of the leading causes of death, with two major types: ischemic strokes due to blockage of blood vessels and hemorrhagic strokes that arise from bleeding inside or around the brain [1]. Annually, about 15 million people suffer strokes globally, according to the World Health Organization (WHO) [2], with 5 million dying and an additional 5 million suffering permanent disability, as shown in Figure *1*.

How to cite this article
Shit B, Chakraborty T, Nesa N (June 23, 2025) Detecting Brain Stroke Using Enhanced Ensemble Models in Data-Driven Healthcare Systems for the Future Factories. Cureus J Comput Sci 2 : es44389-025-06211-7. DOI https://doi.org/10.7759/s44389-025-06211-7

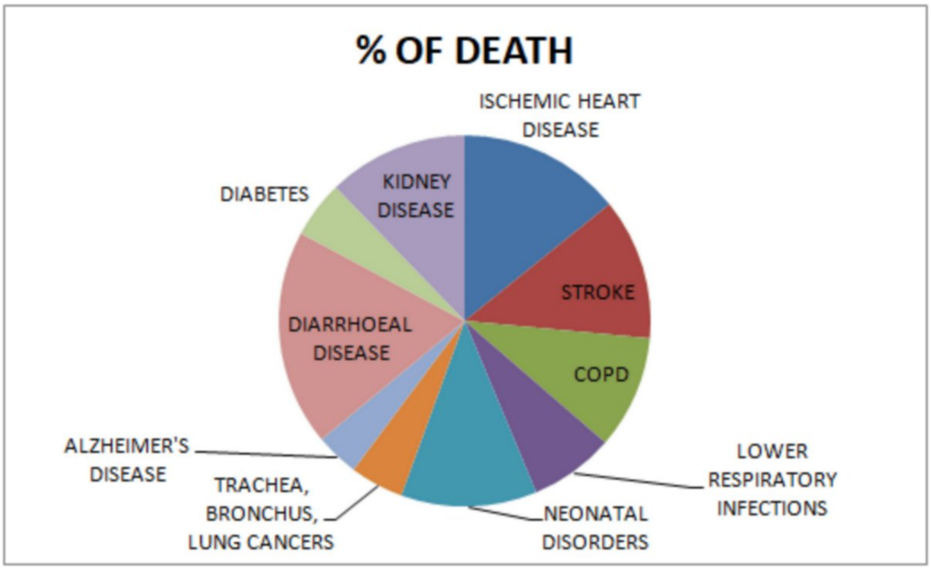

FIGURE 1: Major Global Causes of Mortality as of 2021

Source: World Health Organization [2]

COPD, Chronic Obstructive Pulmonary Disease

We can find 86% of stroke-related deaths and 89% of disability-adjusted life years (DALYs) in poorer regions. In the last 10 years, there has been a significant increase in the number of cases; this includes a 70% rise in incidence rates, a 43% increase in mortality rates, and a 143% increment in DALYs [3]. Among Indians, strokes are one of the main causes of death, ranking third, as demonstrated in Figure *2*.

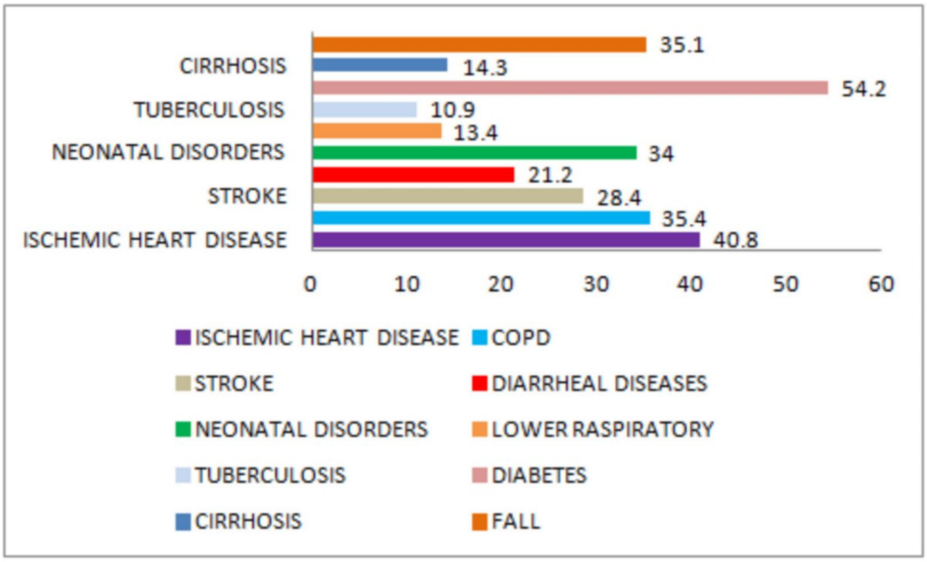

FIGURE 2: Major Causes of Mortality in India as of 2021

Source: World Health Organization [3]

COPD, Chronic Obstructive Pulmonary Disease

Stroke is the leading cause of disability and death worldwide, making it a significant health issue. The ability to detect and predict stroke early is crucial for improving patient outcomes and ensuring timely medical interventions. Modern healthcare has identified machine learning (ML) as a fundamental instrument that presents an opportunity for early detection and treatment of illnesses. Numerous studies have explored various risk factors and employed predictive techniques to forecast stroke occurrences.

However, many of the existing prediction models face challenges in achieving high levels of accuracy and reliability. There is a clear need for more robust models that can leverage a broader range of physiological parameters and incorporate advanced ML techniques. Addressing these limitations can provide healthcare professionals with a more reliable tool for early detection, risk assessment, and the development of preventive measures, ultimately helping to reduce the overall impact of stroke. In order to fill this gap, this study examines six different classification algorithms in predicting brain stroke using physiological parameters such as hypertension, heart disease, average glucose level, body mass index, smoking status, etc., using ML techniques [4]. The results of this research will help advance predictive healthcare systems in future factories.

The main motivation for this work stems from the urgent need for more accurate and timely detection of brain strokes, a leading cause of death and disability worldwide. Traditional detection methods, such as physical exams and medical imaging, have limitations in terms of accuracy, availability, and cost-effectiveness, highlighting the necessity for alternative approaches. ML has shown great potential in medical diagnostics, capable of analyzing large datasets, identifying patterns, and making accurate predictions, which can significantly improve early stroke detection. This study aims to develop an ML model to enhance stroke diagnosis, contributing to better patient outcomes through timely interventions and preventive measures. Additionally, it holds promise for optimizing healthcare resources by focusing on individuals at higher risk, thereby upgrading the overall competency of healthcare systems.

Literature review

Numerous studies have examined the application of ML algorithms for brain stroke detection, employing various datasets, techniques, and performance metrics to develop reliable and accurate predictive models. Heo et al. [5] examined stroke prediction using deep neural networks, random forest (RF), and logistic regression (LR), with deep neural networks achieving an 88% accuracy, particularly for ischemic or acute stroke patients. Yu et al. [6] utilized the C4.5 decision tree (DT) algorithm to classify stroke severity based on real-time National Institutes of Health Stroke Scale scores, reporting an 88.9% accuracy with the RF model, thus enhancing stroke severity prediction and aiding in timely medical intervention. Furthermore, Venkatesh et al. [7] employed support vector machines (SVM), achieving a 91% accuracy rate for stroke analysis, showcasing the effectiveness of pre-processing techniques and SVM in improving diagnosis. Kansadub et al. [8] analyzed stroke prediction using DT, Naive Bayes (NB), and neural network models, with DT achieving the highest accuracy and neural network being preferred for its safety in minimizing false negatives. Meanwhile, Letham et al. [9] introduced Bayesian Rule Lists, a model that enhances accuracy and interpretability in stroke prediction, demonstrating its utility in clinical decision-making.

Other notable studies include those focusing on the analysis of electronic health records and risk factors, where models like neural networks, DT, and RF were used, achieving classification accuracies of around 75%. Additionally, researchers investigating heart stroke prediction using DT, NB, and SVM found room for improvement, as the highest accuracy achieved was only 60%. Collectively, these studies highlight the potential of ML in improving stroke risk prediction and patient outcomes, while emphasizing the need for ongoing research to enhance accuracy and the clinical applicability of these models.

Materials And Methods

Dataset description

This research utilized a stroke prediction dataset [10] comprising 5,110 records with 12 attributes. These attributes include ID, Gender, Age, Hypertension, Heart Disease, Marital Status, Employment Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke. The "Stroke" column, serving as the target variable, is binary, where a value of 1 indicates the presence of stroke risk, and 0 represents no risk. The dataset is notably imbalanced, with 4,861 instances classified as 0 and only 249 instances labeled as 1. To address this imbalance and enhance the predictive capabilities of the model, various data preprocessing techniques were employed. Nine features were selected as inputs for machine learning algorithms, with the "Stroke" column designated as the output variable.

Data processing

Data Cleaning: Data cleaning can be performed to eliminate unwanted data, which includes missing values, duplicate entries, and outliers. The common approach that deals with missing values is either mean, median, or interpolation. In this case, missing values within the column 'bmi' have been replaced by the mean value of the column. After handling the missing data, there is a need to perform label encoding for categorical variables.

Label Encoding: Label encoding is such a process through which categorical string data is converted into numbers by the use of various encoding rules that make the data ML friendly. Most models work on numeric data, so it becomes inevitable to convert categorical strings into numbers. For example, in the column 'gender' in the dataset, the string values have been replaced with equivalent numeric values using label encoding. At this stage, the new dataset would carry both types of values, numerical as well as categorical values.

Handling Noisy/Imbalanced Data: The dataset for stroke prediction is heavily imbalanced, with 5,110 entries in total, of which only 249 signify stroke while 4,861 signify no stroke. Such an imbalance can generally affect the performance of the model and lead to incorrect predictions. One way to handle this is through oversampling, which balances the data by increasing the instances of the minority class. In this work, we used the random oversampling technique to address the class imbalance issue. The stroke occurrence class '1' is oversampled to match the number of cases in the non-stroke class '0', as shown in Figure *3*, so that during training, there are 4,861 instances in each class.

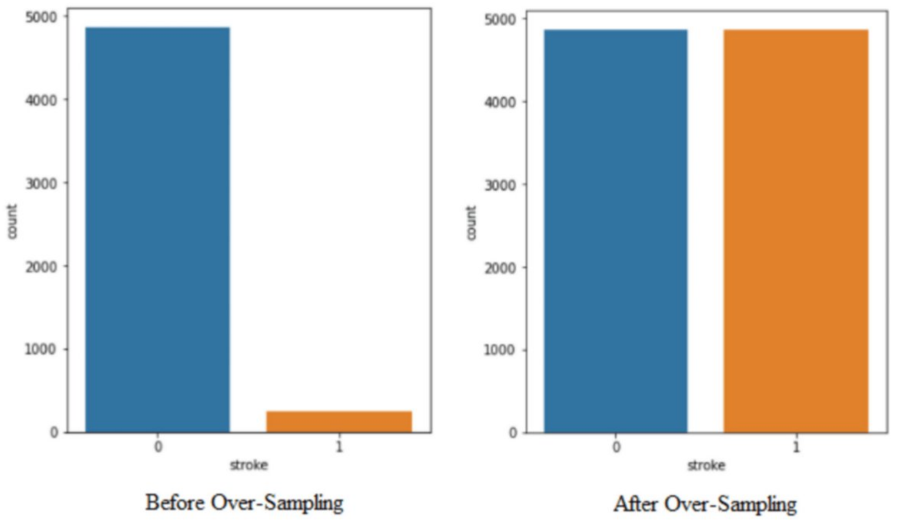

FIGURE 3: Data Visualization Before and After Sampling

Data Transformation: It converts data to conform to the assumptions of an analysis or modeling process. Of the many transformations in wide use, one of the most popular is standardization, wherein numerical features are scaled on some common scale. It standardizes features such that all of them have a mean value of 0 and a standard deviation of 1, in order to be able to compare them with models done with different kinds of ML algorithms.

Data Splitting: After completing data preprocessing and rectifying the class imbalance, it is finally time to prepare for model development. To optimize both accuracy and efficiency, the balanced dataset is split into training and testing sets in a 70:30 ratio. Multiple classification algorithms are then used to train the model. These include LR, DT, RF, K-Nearest Neighbors (K-NN), SVM, and NB.

Classification algorithms

In the vision of future factories, where real-time health monitoring is critical, several classification techniques are employed to predict health-related risks, particularly the risk of stroke. The following ML methods have emerged as effective tools for enhancing predictive analytics in industrial healthcare systems:

K-Nearest Neighbors: This is a straightforward supervised learning algorithm. It works by comparing new data points to existing ones, classifying the new point based on its similarity to known categories. The algorithm stores all available data and, when presented with new input, assigns it to the category most similar to those in the stored dataset. During training, K-NN memorizes the data, and upon receiving new input, it classifies it based on the closest match [11].

Logistic Regression: LR is a statistical technique that predicts a binary outcome through a relationship between an independent variable and one or more independent variables. LR does not predict continuous values, but categorical outcomes based on the possibility of whether something takes place or not [11].

Decision Tree: DT is a type of supervised learning algorithm, which can be used to carry out classification as well as regression tasks. It creates a tree-like model, where the internal node depicts the decision being taken based on attributes, and branches depict the outcome of these decisions, resulting in the final prediction or output at the leaf node [11].

Random Forest: RF is an extremely popular supervised ML algorithm for classification and regression tasks. It is basically an ensemble learning technique, whereby it aggregates the results from multiple DTs to generate predictions [11]. In this case, it builds up a classifier by training several DTs on different subsets of data and averages their results to increase the accuracy of the results. However, taking the majority vote from all trees contributes to the end result of RF, which comes out with much more reliable results compared to the use of single DT.

Support Vector Machine: SVM is a popular supervised learning algorithm utilized for both classification and regression tasks, with its primary application being in classification problems [11]. It operates by identifying an optimal hyperplane that divides data points into distinct classes within an n-dimensional space. The algorithm emphasizes selecting key data points, known as support vectors, which are crucial in defining and constructing this hyperplane.

Naïve Bayes: NB is a supervised learning method rooted in Bayes' theorem, commonly employed to solve classification tasks. This algorithm is known for its simplicity and efficiency, making it a popular choice for building rapid machine learning models that deliver swift predictions. As a probabilistic classifier, it makes predictions by estimating the likelihood of an object belonging to a particular class [11].

Results

Experimental setup

The experiment code was written in Python, using a computer with an Intel Core i5 processor.

Raw data may contain errors such as missing entries, inconsistencies, duplicates, or improper formatting, which can weaken analysis and reduce model performance. To enhance data quality and reliability, preprocessing steps such as missing value imputation and normalization were applied. The cleaned dataset was then used for further analysis.

Result

A comparison of six ML models was done, comparing their performances with respect to predicting brain strokes based on the given dataset. The models used include LR, DT, RF, K-NN, SVM, and NB. Their performance was assessed using accuracy, precision, recall, and F1 score, providing insights into each model's ability to correctly identify stroke occurrences (true positives) and avoid misclassifications or missed detections (false negatives). The performance analysis for each model is presented in Figure *4* and Table *1*.

Model	Accuracy (%)	Recall (%)	F1-Score (%)
Decision Tree (DT)	97.12	98	97
Random Forest (RF)	94	99	97
Support Vector Machine (SVM)	92	99	95
K-Nearest Neighbors (KNN)	87	100	93
Logistic Regression (LR)	76	84	80
Naive Bayes (NB)	76	78	77

TABLE 1: Model Performance Comparison Based on Accuracy, Recall, and F1 Score for Stroke Prediction

Overall, the DT model stood out as the most accurate and reliable for stroke prediction, making it the best candidate for deployment in real-time health monitoring systems within future factories. Its ability to handle imbalanced datasets and deliver high precision and recall ensures that workers' health risks are accurately predicted with minimal false alarms. The ensemble models like RF and SVM also showed strong performance in accuracy of 94% and 92%, respectively, offering valuable redundancy and stability in prediction systems. While KNN was highly sensitive (100%), its false positive rate suggests it could be useful as a backup model rather than a primary system. Finally, LR and NB provided lower accuracy (76%) and should be considered only for specific use cases requiring simplicity and speed rather than high accuracy. This evaluation of model performance demonstrates the effectiveness of ML in healthcare applications for future factory environments, where proactive health monitoring and real-time interventions are critical.

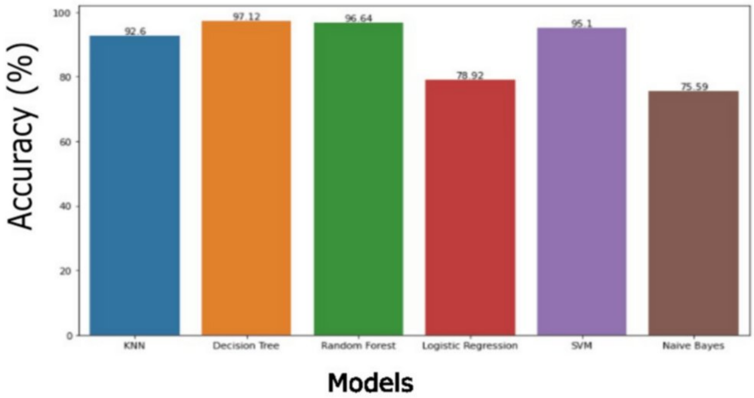

FIGURE 4: Comparison of Accuracies Amongst All the Models

Discussion

Experimental analysis

The evaluation of the ML models was performed using a dataset split into training and testing sets in a 70:30 ratio to optimize both accuracy and efficiency. Multiple classification algorithms were used to train the models, including LR, DT, RF, K-NN, SVM, and NB. Each model's accuracy was assessed on both the training and testing sets to examine the generalization ability and robustness of the algorithms.

Integrating these prediction models into future factories would enhance real-time health monitoring through wearable devices or sensors. The high accuracy of the DT and RF models makes them ideal for providing early warnings and preventing critical health incidents, ensuring worker safety in smart factory environments. This study demonstrates that ML models, particularly DT, are effective tools for proactive healthcare in future factories. Early stroke detection can reduce downtime and enhance worker safety by enabling timely interventions, such as immediate medical alerts. This approach aligns with the "Factory of the Future" vision, extending predictive maintenance to human health.

Conclusions

This study highlights the potential of enhanced ML models in predicting brain strokes. In the context of future factories, such models can improve worker safety through real-time health monitoring and early intervention. Implementing these predictive systems will improve operational efficiency and health outcomes, making future factories smarter and safer. In this work, we used bagging technique to improve the accuracy of existing ML algorithms such as LR, DT, RF, K-NN, SVM, and NB. Each model's accuracy was evaluated on both the training and testing datasets using the standard 70:30 split to examine the generalization ability and robustness of the algorithms. Among the models tested, the DT outperformed all others, achieving an accuracy of 97%. Although the existing models show promise, further refinements are needed to enhance their adaptability in various scenarios. Future improvements may involve expanding datasets with blood pressure readings, real-time glucose levels, and other health metrics from wearable devices. Additionally, employing advanced approaches such as reinforcement learning or hybrid models could further boost predictive accuracy in dynamic factory settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Nashreen Nesa, Bikram Shit, Tamosa Chakraborty

Acquisition, analysis, or interpretation of data: Nashreen Nesa, Bikram Shit

Drafting of the manuscript: Nashreen Nesa, Bikram Shit

Critical review of the manuscript for important intellectual content: Nashreen Nesa, Bikram Shit, Tamosa Chakraborty

Supervision: Nashreen Nesa

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Medical News Today: What are the different types of stroke?. (2023). Accessed: May 27, 2023: https://www.medicalnewstoday.com/articles/types-of-strokes.
2. World Health Organization: The top 10 causes of death. (2024). Accessed: May 27, 2023: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
3. World Health Organization: Global health estimates: Leading causes of death. (2021). http://who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death.
4. Shoily TI, Islam T, Jannat S, Tanna SA, Alif TM, Ema RR: Detection of stroke using machine learning algorithms. 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019 , 1-6. 10.1109/ICCCNT45670.2019.8944689
5. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH: Machine learning-based model for prediction of outcomes in acute stroke. Stroke. 2019, 50:1263-1265. 10.1161/STROKEAHA.118.024293
6. Yu J, Kim D, Park H, et al.: Semantic analysis of NIH stroke scale using machine learning techniques. 2019 International Conference on Platform Technology and Service (PlatCon). 2019, 1-5. 10.1109/PlatCon.2019.8668961
7. Venkatesh R, Mithun V, HariVenkatesh M, Hari Prathiunun R, Kishore S: Accuracy assessment of machine learning algorithm for prediction of heart stroke. Journal of Basic Science and Engineering. 2024, 21:406-413.
8. Kansadub T, Thammaboosadee S, Kiattisin S, Jalayondeja C: Stroke risk prediction model based on demographic data. 2015 8th Biomedical Engineering International Conference (BMEiCON). 2015, 1-3. 10.1109/BMEiCON.2015.7399556
9. Letham B, Rudin C, McCormick TH: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics. 2015, 1350-1371. 10.1214/15-AOAS848
10. Stroke Prediction Dataset. Accessed: September 30, 2024: http://kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.
11. James G, Witten D, Hastie T, Tibshirani R, Taylor J: An Introduction to Statistical Learning: With Applications in Python. Springer, Cham; 2023. 10.1007/978-3-031-38747-0