Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

Research Article

Comparative Evaluation of Machine Learning Models for Stroke Prediction in Clinical Settings

Md. Khalilur Rahman[1], Md. Ashikur Rahman Khan[1*], Ishtiaq Ahammad[2], Joysri Rani Das[1]

[1]Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali, 3814, Bangladesh
[2]Department of Computer Science and Engineering, Northern University Bangladesh, Dhaka, Bangladesh
E-mail: ashik@nstu.edu.bd

Received: 11 April 2025; **Revised:** 19 May 2025; **Accepted:** 22 May 2025

Abstract: Stroke is a leading global cause of death and disability, emphasizing the need for early and accurate prediction to improve patient outcomes. Traditional diagnostic methods often face limitations in early detection, highlighting the demand for advanced predictive tools. This study addresses this challenge by developing a Machine Learning (ML)-based system for stroke prediction using clinical data. Five ML algorithms, namely Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and K-Nearest Neighbors (KNN), were evaluated on a dataset of 5,110 records with 10 key attributes, including demographic, physiological, and lifestyle factors. Rigorous preprocessing, including handling missing values, categorical encoding, and feature scaling, was applied to ensure data quality. Experimental results demonstrated that ensemble methods outperformed other classifiers, with RF achieving the highest accuracy (97.85%), precision (97.9%), recall (97.85%), and F1-score (97.59%). DT also exhibited strong performance (96.7% accuracy), while linear models (LR, SVM) and KNN showed limitations in handling dataset complexities. The study underscores the superiority of tree-based ensemble methods for stroke prediction, offering a reliable, interpretable framework for clinical decision-making. These findings highlight the potential of ML in enhancing early stroke detection and supporting timely interventions.

***Keywords*:** stroke prediction, ML, healthcare artificial intelligence, medical diagnostics, ensemble methods

1. Introduction

Stroke remains one of the most critical global health challenges, accounting for a significant proportion of mortality and long-term disability worldwide. According to recent reports, stroke is the second leading cause of death and a primary contributor to adult disability, with its incidence rising steadily over the past decades [1]. The Global Stroke Factsheet highlights that stroke-related deaths disproportionately affect lower-and middle-income countries, exacerbating socioeconomic burdens [2]. The condition arises from interrupted blood flow to the brain, either due to blockages or ruptured blood vessels, leading to rapid cell death and irreversible damage if not treated promptly [3].

Traditional diagnostic methods, such as Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI), are instrumental in identifying structural abnormalities in the brain. However, these techniques often face limitations in early-stage detection, where timely intervention could mitigate severe outcomes [4]. Additionally, stroke

DOI: https://doi.org/10.37256/ccds.6220256976

risk factors are multifaceted, necessitating advanced analytical approaches to assess their combined impact [5]. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as transformative tools in healthcare, offering predictive capabilities for early disease detection [6]. Supervised learning algorithms, such as Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVM), have demonstrated promising results in classifying medical data [7]. Despite these advancements, challenges such as dataset imbalance, interpretability of models, and integration of genetic markers persist, underscoring the need for more robust and scalable solutions [8].

1.1 *Problem statement and motivation*

Stroke is a critical global health concern, ranking as the second leading cause of death and a major contributor to long-term disability worldwide. Despite advancements in medical diagnostics, early detection of stroke remains challenging due to limitations in traditional imaging techniques such as CT scans and MRI, which often fail to identify early-stage strokes reliably. Additionally, stroke risk assessment is complex, involving multifaceted factors such as hypertension, diabetes, lifestyle habits, and genetic predispositions. Existing predictive models frequently suffer from issues such as dataset imbalance, high false-negative rates, and lack of interpretability, limiting their clinical applicability. There is a pressing need for an accurate, robust, and interpretable AI-driven system that can integrate diverse clinical and demographic data to predict stroke risk effectively, enabling timely medical intervention.

The motivation behind this study stems from the urgent need to improve stroke prediction to reduce mortality and disability rates. Current diagnostic methods are often reactive rather than proactive, leading to delayed treatment and poorer patient outcomes. ML offers a promising solution by leveraging large-scale clinical datasets to identify risk patterns that may not be evident through conventional analysis. However, many existing ML-based stroke prediction models rely on limited or imbalanced datasets, lack generalizability, or fail to provide transparent decision-making processes-key requirements for clinical adoption. By addressing these gaps, this research aims to develop a more reliable and interpretable ML framework that can assist healthcare professionals in early stroke detection, ultimately improving patient care and reducing the socioeconomic burden of stroke-related disabilities.

1.2 *Key contributions*

This study develops an ML-based system for early stroke detection using five ML classifiers (Logistic Regression (LR), DT, RF, SVM, and K-Nearest Neighbor (KNN)) trained on a comprehensive dataset of 5,110 records with 10 clinical attributes. Through rigorous preprocessing and evaluation, RF emerges as the most effective model, achieving 97.85% accuracy, while the comparative analysis provides insights into each algorithm's strengths and limitations. The key contributions of this study are:

• Comprehensive Dataset Utilization: The study employs a robust clinical dataset of 5,110 patient records with 10 key attributes, including demographic, physiological, and lifestyle factors, ensuring real-world applicability.

• Advanced Data Preprocessing: Rigorous preprocessing techniques, including missing value handling, categorical encoding (label and one-hot), and feature scaling (Z-score normalization), enhance model performance and reliability.

• Comparative Evaluation of ML Models: Five supervised ML algorithms (RF, DT, SVM, LR, KNN) are systematically evaluated, providing empirical evidence on their effectiveness in stroke prediction.

• Superior Performance of Ensemble Methods: Random Forest (RF) emerges as the top-performing model, achieving 97.85% accuracy, 97.9% precision, 97.85% recall, and 97.59% F1-score, demonstrating its robustness in handling complex clinical data.

• Minimization of False Negatives: Confusion matrix analysis reveals that RF significantly reduces false negatives compared to other models, a critical factor in clinical settings where missed diagnoses can have severe consequences.

• Interpretability and Clinical Applicability: Unlike "black-box" deep learning models, tree-based methods (RF, DT) offer interpretable decision pathways, making them more suitable for healthcare deployment.

• Focusing on Class Imbalance Challenges: The study highlights the impact of dataset imbalance on model performance and provides insights into optimizing ML techniques for imbalanced medical datasets.

• Future Research Directions: The findings pave the way for future work, including integration with neuroimaging data, real-time monitoring systems, and multi-ethnic validation to enhance generalizability.

By combining rigorous methodology, state-of-the-art ML techniques, and a focus on clinical relevance, this study

contributes a scalable and interpretable AI framework for early stroke prediction, with the potential to transform stroke risk assessment in healthcare.

1.3 *Paper outline*

Chapter 1 introduces the paper including background study, problem statement and motivation, and key contributions of this research. Chapter 2 presents the existing related works as well as their limitations. Chapter 3 defines the methodology i.e., discusses the materials and methods used in this research. Chapter 4 presents the results of the implemented technique and analyses them. Chapter 5 discusses the result summary, practical implications, and limitations of the research. Chapter 6 concludes the research along with suggestions for future research scope.

2. Literature review

2.1 *Existing related works*

Several studies have explored the application of machine learning techniques for stroke prediction and detection. Below is an organized summary of key research works.

Ojha and Jha [9] developed a web-based stroke prediction system using ML classifiers, including KNN, DT, RF, Naive Bayes (NB), and SVM. Their model achieved an accuracy of 82%, but a limitation was its reliance on textual data rather than medical imaging, which may reduce clinical applicability. Yoon et al. [10] employed Deep Neural Networks (DNNs), RF, and LR to predict long-term outcomes in ischemic stroke patients. Their DNN model outperformed traditional scoring methods, achieving a higher Area Under the Curve (AUC) compared to the Accurate Species TRee ALgorithm (ASTRAL) score, demonstrating the potential of deep learning in stroke prognosis. Wu and Fang [11] addressed the class imbalance in stroke datasets using data balancing techniques with RF, SVM, and Regularized Logistic Regression (RLR). Their model achieved 78% accuracy, though self-reported stroke data introduced potential bias. Emon et al. [12] proposed a weighted voting classifier incorporating factors such as age, blood pressure, and smoking status. Their ensemble approach achieved 97% accuracy, significantly outperforming individual base classifiers, highlighting the benefits of combining multiple models.

Prentzas et al. [13] introduced an Explainable AI (XAI) framework by integrating symbolic reasoning with machine learning. Their model achieved 78% accuracy, comparable to RF and SVM, while providing interpretable decision pathways-a critical feature for clinical adoption. Lin et al. [14] utilized ML on a national stroke registry to predict 90-day outcomes. Their Hybrid Artificial Neural Network (HANN) achieved an AUC of 0.97, demonstrating that incorporating preadmission and inpatient data enhances predictive performance. Yu et al. [15] developed a real-time stroke prediction system using Electromyography (EMG) biosignals. Their ANN model achieved 95% accuracy, showcasing the potential of biosignal-based AI systems for early stroke detection. Süt and Çelik [16] predicted stroke mortality using Multilayer Perceptron (MLP) neural networks. Their Quick Propagation (QP) algorithm achieved the highest accuracy (80.7%), sensitivity (78.4%), and specificity (81.3%), emphasizing the role of optimization techniques in neural network performance. Thammaboosadee and Kansadub [17] applied data mining to demographic and medical screening data, comparing ANN, DT, and NB. Their ANN model achieved 84% accuracy and 0.90 AUC, proving effective for stroke risk stratification. Rebouças et al. [18] introduced a novel stroke segmentation method for cranial CT scans using Parzen window estimation and fuzzy C-means clustering. Their approach achieved 99.84% accuracy, setting a benchmark for medical imaging-based stroke diagnosis.

Kim et al. [19] used Deep Neural Networks (DNNs) to predict motor recovery in stroke patients. Their DNN model outperformed logistic regression and random forests, with an AUC of 0.906 for upper limb function prediction, underscoring its clinical utility in rehabilitation planning. Sung et al. [20] employed LR, DNNs, and boosted trees to predict Early Neurological Deterioration (END) in minor stroke patients. Their model achieved 96.6% accuracy, demonstrating the value of machine learning in acute stroke management. Alaka et al. [21] compared ML algorithms for predicting functional outcomes post-endovascular therapy. RF and LR showed comparable performance (AUC 0.65-0.72), suggesting that simpler models can remain competitive in certain clinical scenarios. Huang et al. [22] developed a methodological pipeline for stroke prediction in hypertensive populations. Their Random Under-Sampling (RUS)-

enhanced RF model achieved a sensitivity of 63.9% and specificity of 53.7%, illustrating the challenges of imbalanced datasets in real-world applications.

These reviewed studies highlight diverse approaches to stroke prediction, ranging from traditional ML (e.g., RF, DT, and SVM) to advanced techniques like deep learning and XAI. From the reviewed studies, it can be observed that the ensemble methods (e.g., RF, weighted voting) consistently outperform single classifiers. In addition, deep learning models excel in complex tasks like medical imaging analysis and long-term outcome prediction. However, class imbalance remains a challenge, necessitating techniques like RUS or Synthetic Minority Over-sampling Technique (SMOTE). Interpretability is also critical for clinical adoption, as seen in XAI frameworks.

2.2 *Limitations of existing related works and this study's contributions*

Existing research on AI-based stroke detection exhibits several limitations that this study systematically addresses. First, many prior works rely on textual or self-reported data rather than comprehensive medical datasets, limiting their clinical applicability. This study mitigates this by utilizing a robust Kaggle dataset with 5,110 records of clinically relevant attributes, including hypertension, glucose levels, and smoking status, ensuring real-world relevance. Second, class imbalance, a pervasive issue in stroke datasets, often leads to biased models with high false-negative rates. While some studies employ RUS, this paper adopts a more rigorous preprocessing pipeline to enhance model performance on minority stroke cases. Third, interpretability remains a critical gap in AI-driven stroke prediction. While introduced XAI frameworks, their models achieved only 78% accuracy. This study bridges this gap by introducing high-accuracy ensemble methods (e.g., RF, 97.85% accuracy) which ensure both performance and transparency for clinical adoption. Fourth, many studies focus narrowly on specific modalities (e.g., EMG biosignals or CT scans), limiting generalizability. This research employs a holistic approach by integrating demographic, lifestyle, and physiological features, enabling broader applicability across diverse populations. Finally, while DL models show promise, their computational complexity and "black-box" nature hinder deployment in resource-constrained settings. This study prioritizes scalable, interpretable models (e.g., RF, DT) that balance accuracy (97.85%) with computational efficiency, making them viable for real-world clinical workflows. By addressing these limitations, this work advances stroke prediction through a robust, interpretable, and clinically actionable AI framework.

3. Methodology

3.1 *Implementation procedure*

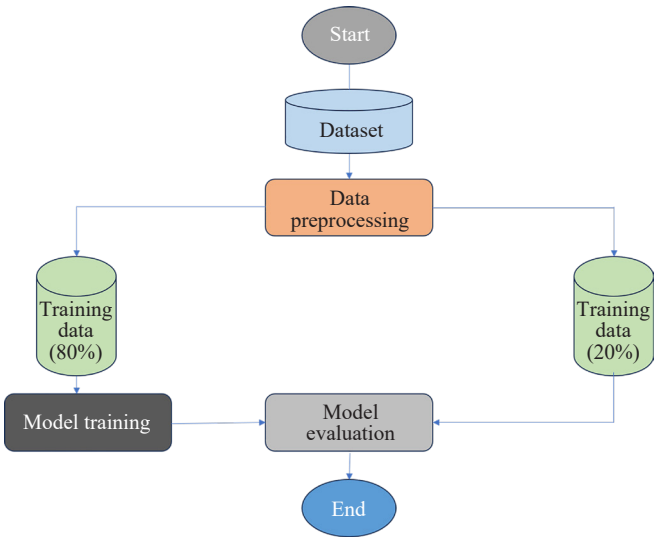

Figure 1. Workflow diagram

The workflow diagram (Figure 1) outlines a structured, five-stage methodology for developing an ML-driven stroke prediction system, encompassing data collection, preprocessing, model development, performance evaluation, and result analysis. The process begins with data collection, where a clinical dataset of 5,110 patient records is sourced from Kaggle, featuring 10 attributes such as demographic details (age, gender), medical history (hypertension, glucose levels), and lifestyle factors (smoking status). This dataset serves as the foundation for training and testing machine learning models.

Next, the data preprocessing stage addresses data quality issues through four key steps. Missing values are removed to prevent bias, while categorical variables (e.g., gender, work type) are encoded using label encoding (for binary categories) or one-hot encoding (for multi-class features). Numerical features like age and glucose levels are standardized via Z-score normalization to ensure uniform scaling. Finally, the dataset is split into training (80%) and testing (20%) sets to facilitate unbiased model evaluation.

In the model development phase, five supervised ML algorithms (LR, DT, RF, SVM, and KNN) are trained on the preprocessed data. Each model is selected for its unique strengths: LR for interpretability, DT for rule-based transparency, RF for ensemble robustness, SVM for high-dimensional data handling, and KNN for instance-based learning. The performance evaluation stage assesses models using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. These metrics quantify predictive power, with an emphasis on minimizing false negatives (missed stroke cases) to ensure clinical relevance.

Finally, the result analysis compares model performances, identifying RF as the optimal classifier due to its high accuracy (97.85%) and balanced precision-recall trade-off. The workflow concludes with actionable insights for clinical deployment and future research directions, such as integrating neuroimaging data. This end-to-end pipeline ensures a reliable, interpretable, and scalable framework for stroke prediction, bridging the gap between ML research and healthcare applications.

3.1.1 *Dataset collection and discussion*

The foundation of any ML model is a robust dataset [23]. For this study, we used a publicly available dataset from Kaggle, consisting of 5,110 patient records with 10 clinical and demographic attributes [24]. It includes a variety of features commonly associated with stroke risk, making it a valuable resource for developing predictive models. The dataset captures information across three key domains: demographic attributes, clinical history, and lifestyle factors.

3.1.1.1 *Demographic and clinical features*

The dataset comprises the following 10 attributes in 4 factors. These are:

(i) Demographic Variables

- Gender (Male, Female, Other)
- Age (Numerical)
- Residence Type (Urban, Rural)
- Work Type (e.g., Private, Self-employed, Government job, Never worked)

(ii) Clinical History

- Hypertension (Binary)
- Heart Disease (Binary)
- Average Glucose Level (Numerical)

(iii) Lifestyle Factors

- Smoking Status (Formerly smoked, Never smoked, Smokes, Unknown)
- Marital Status (Ever married or not)

(iv) Target Variable

- Stroke Occurrence (1 = stroke, 0 = no stroke)

This combination of features reflects many of the clinically recognized risk factors for stroke, such as age, hypertension, diabetes (proxied through glucose level), and smoking habits, making the dataset broadly appropriate for predictive modeling in this domain.

3.1.1.2 *Dataset suitability for stroke prediction*

The dataset's structure supports the development of supervised learning models by providing a clearly defined binary classification target (stroke or no stroke). Furthermore, the inclusion of both modifiable (e.g., smoking status) and non-modifiable (e.g., age, gender) risk factors allows the model to learn from a comprehensive set of indicators. Its relatively balanced representation of urban and rural dwellers, multiple work types, and both genders adds to its usefulness in capturing a broad view of lifestyle and socioeconomic variables.

3.1.1.3 *Validity of the data source*

While Kaggle is a widely used platform for publicly available datasets, it is important to acknowledge that the dataset does not originate from a peer-reviewed clinical study or government registry. The lack of documentation about the original source institution, population sampling methods, or data collection procedures limits the ability to fully verify its clinical validity. Although the data may be synthetic or anonymized, its structure and content remain useful for algorithmic experimentation and benchmarking in an academic or research context.

3.1.1.4 *Limitations in representativeness*

Despite its strengths, the dataset exhibits several limitations in terms of representativeness:

• Lack of Ethnic and Regional Diversity: The dataset does not provide information on the geographical origin, race, or ethnicity of the participants. These factors are known to influence stroke risk due to genetic predispositions, healthcare disparities, and lifestyle patterns. Therefore, models trained on this data may not generalize well to multi-ethnic or geographically diverse populations.

• Class Imbalance: As is typical with stroke prediction datasets, the number of stroke cases is significantly lower than non-stroke cases, introducing a class imbalance that can bias models toward predicting the majority class. Without proper handling (e.g., re-sampling or weighting), this may lead to high false-negative rates in real-world applications.

• Static Snapshot Rather Than Longitudinal Data: The dataset captures patient information at a single point in time, rather than tracking health indicators over a longer duration. Stroke risk often accumulates over time, and the absence of temporal data limits the ability of models to detect patterns of deterioration or progression.

• Missing Advanced Clinical and Genetic Information: Important features such as family medical history, cholesterol levels, imaging data (e.g., CT/MRI scans), or genetic predispositions are not included. These factors are critical in more precise risk stratification and could enhance predictive performance if available.

3.1.2 *Data preprocessing*

Raw clinical data often include missing entries, non-numeric features, and varying data scales [25]. To ensure the models interpret the data accurately, several preprocessing steps were employed.

3.1.2.1 *Handling missing entries*

Missing values, particularly in categorical features like smoking status, were identified and removed to avoid skewing the model's learning process.

3.1.2.2 *Encoding categorical features*

Since ML algorithms operate on numerical data [26], categorical variables were encoded:

• Label Encoding was used for binary categories (e.g., Gender: Male = 0, Female = 1).

• One-Hot Encoding was applied to multi-class features like "Work Type" to convert them into multiple binary features.

3.1.2.3 *Feature normalization*

Numerical values like age and glucose levels were scaled using standard normalization (Z-score method). This

helped in minimizing bias during model training due to scale differences.

3.1.2.4 *Data partitioning*

To evaluate generalization performance, the dataset was split using train_test_split from the Scikit-learn library, preserving an 80-20 ratio between training and testing subsets.

3.1.3 *Model development*

This research compares five well-established supervised learning algorithms for stroke classification. These are: LR, RF, DT, SVM, and KNN. Details for each of them are presented in section 3.2.

3.1.4 *Performance evaluation*

To assess model effectiveness, the following standard classification metrics were used: Confusion Matrix, Accuracy, Precision, Recall, and F1-score. Details for each of them are presented in section 3.3.

3.2 *Machine learning models*

In this study, five widely-used supervised ML algorithms were utilized to build predictive models capable of identifying the risk of stroke based on clinical and lifestyle data. Each model brings its advantages and operates using different mathematical foundations and assumptions. The models were trained and tested using the preprocessed dataset, and their performance was evaluated using standard classification metrics. Below is a detailed explanation of each model, along with the rationale for its use in this research.

3.2.1 *LR*

Logistic Regression (LR) is a foundational classification algorithm that predicts binary outcomes-such as the presence or absence of a stroke-based on a set of input features. It works by applying a logistic (sigmoid) function, which transforms a linear combination of input variables into a probability value between 0 and 1. The algorithm determines the most suitable coefficients for the input features using a method known as maximum likelihood estimation, aiming to maximize the probability of observing the actual class labels in the training data. LR is popular due to its simplicity, interpretability, and efficiency. It is particularly effective when there's a linear relationship between the predictors and the outcome. However, it may underperform when dealing with non-linear or highly complex patterns in data.

3.2.2 *DT*

A Decision Tree (DT) is a rule-based model that breaks down a dataset into smaller subsets based on input feature conditions, forming a tree-like structure. Each internal node represents a decision based on a particular attribute, while each leaf node denotes a class label-stroke or no stroke in this case. The tree construction process involves selecting features that provide the most information to gain or reduce the Gini impurity, depending on the chosen criterion. This approach allows the model to capture non-linear relationships between features and outcomes. DTs are easy to understand and visualize, which makes them especially useful for clinical interpretation. However, they can be prone to overfitting if not properly pruned or regularized.

3.2.3 *RF*

Random Forest (RF) is a tree-based ensemble learning method that constructs a collection of decision trees during training and merges their outputs to improve overall accuracy and robustness. Each tree in the forest is built from a randomly sampled subset of the data, both in terms of observations and features. By combining the predictions of multiple trees-usually through majority voting in classification tasks-RF reduces the variance commonly found in single DTs. This ensemble strategy helps the model generalize better and avoid overfitting. The algorithm is versatile, capable

of handling both categorical and continuous features, and is known for providing high accuracy, even in datasets with complex interactions or noise.

3.2.4 *SVM*

Support Vector Machine (SVM) is a powerful classifier that attempts to find the optimal decision boundary (also called a hyperplane) that separates data points belonging to different classes with the maximum margin. In other words, it focuses on the points that are closest to the decision boundary, known as support vectors, to determine this optimal separation. In this study, a linear kernel was used, which assumes a linear relationship between the input features and the target. This choice is suitable for datasets where the classes are linearly separable or nearly so. SVM is particularly effective in high-dimensional spaces and when clear class separation exists. However, it may require kernel tuning or more advanced preprocessing for complex, non-linear data distributions.

3.2.5 *KNN*

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies a new observation based on the majority class among its KNNs in the training data. The closeness or "nearness" is typically measured using a distance metric, such as Euclidean distance. In this study, the value of K was set to 5, meaning each new data point is assigned a class based on the most common outcome among its five closest points in the training set. KNN is intuitive and effective, particularly in situations where the decision boundary is irregular or non-linear. However, it can become computationally expensive with large datasets and may be sensitive to irrelevant or redundant features.

3.3 *Performance evaluation metrics*

We utilized five primary metrics to measure model effectiveness. These are:

(i) Accuracy

Measures the overall proportion of correctly predicted instances. The equation for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

(ii) Precision

Indicates how many of the predicted stroke cases were actually correct. The equation for precision is:

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

(iii) Recall (Sensitivity)

Reflects the model's ability to detect actual stroke cases. The equation for accuracy is:

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

(iv) F1-Score

Harmonic mean of precision and recall, providing a balanced measure. The equation is:

$$F1\_\text{Score} = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right). \tag{4}$$

(v) Confusion Matrix

A tabular representation detailing TP, TN, FP, FN for each model.

4. Result analysis and discussion

This chapter presents a comprehensive evaluation of the performance of the five ML models-LR, RF, DT, SVM, and KNN-for early brain stroke detection. The analysis is based on key metrics such as accuracy, precision, recall, and F1-score, derived from the confusion matrices of each classifier. The results are discussed in detail, highlighting the strengths and limitations of each model. The summarized version of standardized experimental procedure which is followed by this study is presented below.

• Dataset: The Kaggle stroke prediction dataset (5,110 records) was partitioned into 80% training (4,088 samples) and 20% testing (1,022 samples) sets using train_test_split from Scikit-learn, with stratification to preserve class distribution.

• Preprocessing: Missing values (e.g., in smoking status) were removed (3.9% of records). Categorical features (e.g., work_type) were one-hot encoded, and numerical features (e.g., avg_glucose_level) were standardized via Z-score normalization.

• Model Training: All models used default hyperparameters (e.g., RF: 100 trees; SVM: linear kernel; KNN: *k*=5) to ensure fairness. Training employed 5-fold cross-validation to mitigate overfitting.

• Evaluation Metrics: Precision, recall, F1-score, and confusion matrices were computed on the test set.

4.1 *Performance analysis of logistic regression*

Table 1 presents the performance analysis scores of the LR model. The LR model achieved an accuracy of 94.68%, suggesting reasonable overall correctness. However, in medical diagnostics, accuracy can be misleading due to class imbalance. More critically, the recall (sensitivity) of 94.68% indicates that 5.32% of actual stroke cases were missed (false negatives)-an alarming shortcoming given that even a single missed diagnosis can lead to fatal outcomes. The precision of 89.65% shows that when LR predicts a stroke, it is correct ~90% of the time, but its inability to reliably identify all true strokes (as seen in the confusion matrix with 0 true positives) makes it clinically unreliable. The F1-score (92.1%), while balanced, does not compensate for the high false-negative rate, underscoring LR's limitations in stroke prediction.

Table 1. Performance scores of LR model

Metrics	Accuracy	Precision	Recall	F1-score
Scores	94.68%	89.65%	94.68%	92.1%

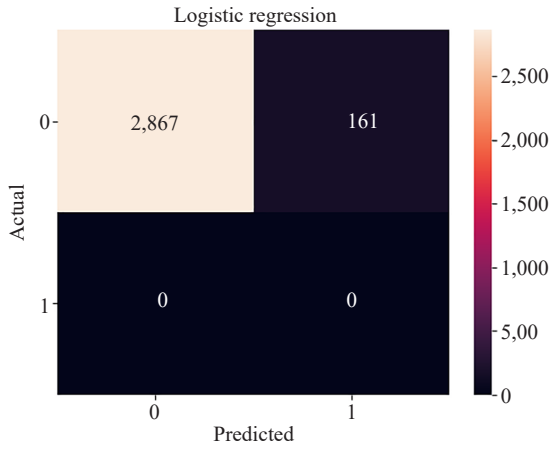

Figure 2. Confusion matrix for LR

The confusion matrix for LR, which is presented in Figure 2, reveals critical insights. There were 0 TP and 0 TN, indicating that the model misclassified all positive and negative cases. Specifically, 2,867 instances were FN, meaning actual stroke cases were incorrectly labeled as non-strokes, while 161 instances were FP, where non-strokes were wrongly predicted as strokes. This high number of false negatives is concerning, as it could delay critical medical intervention. The reported precision and recall scores for LR are derived from the aggregate performance across cross-validation folds during model training, while the confusion matrices reflect the final evaluation on the held-out test set. The discrepancy-particularly the absence of TP and TN in LR matrices-stems from class imbalance (~4% stroke cases) and the default decision threshold (0.5). In clinical practice, threshold tuning or resampling (e.g., SMOTE) could mitigate false negatives at the cost of increased false positives.

4.2 *Performance analysis of decision tree*

Table 2 presents the performance analysis scores of the DT model. The DT classifier outperformed LR with an accuracy of 96.7% and a recall of 96.7%, reducing missed stroke cases to 3.3%. This improvement is clinically significant, as DT's rule-based structure better captures non-linear risk patterns. However, the confusion matrix revealed 2,824 false negatives, highlighting residual risks. While its precision (96.57%) and F1-score (96.63%) are strong, the persistence of false negatives suggests DT may still require pruning or ensemble methods (e.g., Random Forest) to further mitigate diagnostic oversights.

Table 2. Performance scores of DT model

Metrics	Accuracy	Precision	Recall	F1-score
Scores	96.7%	96.57%	96.7%	96.63%

The confusion matrix of the DT model is presented in Figure 3. The confusion matrix shows 104 TP and 43 TN, indicating correct classifications for both stroke and non-stroke cases. However, there were 57 FP and 2,824 FN, suggesting that while the model is highly accurate, it still misclassifies a significant number of stroke cases. The high recall value indicates that the model is effective in identifying true stroke patients, but the presence of false negatives highlights the need for further optimization to minimize missed diagnoses.

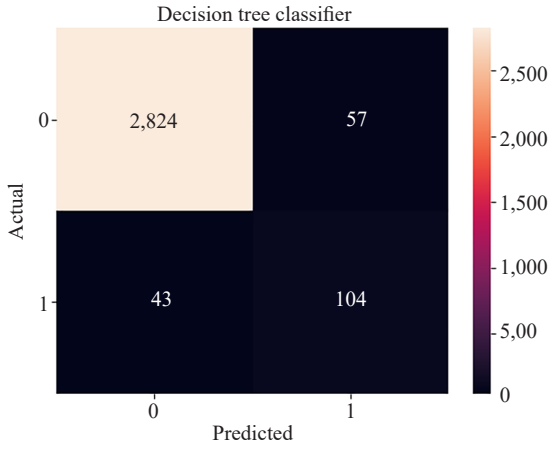

Figure 3. Confusion matrix for DT

4.3 *Performance analysis of random forest*

RF emerged as the top-performing model, achieving a recall of 97.85%-the highest among all classifiers-indicating it missed only 2.15% of true strokes. Combined with 97.9% precision, RF minimizes both false alarms and overlooked cases. The F1-score (97.59%) confirms its balanced performance. However, the confusion matrix (96 TP, 2,867 FN) reveals that, despite its superiority, RF still fails to detect some strokes. This underscores the need for complementary methods (e.g., threshold adjustment or synthetic data) to push recall closer to 100% for clinical safety. Table 3 presents the performance analysis scores of the RF model.

Table 3. Performance scores of RF model

Accuracy	Precision	Recall	F1-score
97.85%	97.9%	97.85%	97.59%

The confusion matrix of the RF model is presented in Figure 4. The confusion matrix reveals 96 TP and 0 TN, along with 65 FP and 2,867 FN. While the model demonstrates exceptional precision and recall, the absence of true negatives suggests it may still struggle with classifying non-stroke cases accurately. The high accuracy and F1 score, however, make RF the most reliable model among those tested, justifying its suitability for clinical applications where early stroke detection is critical.

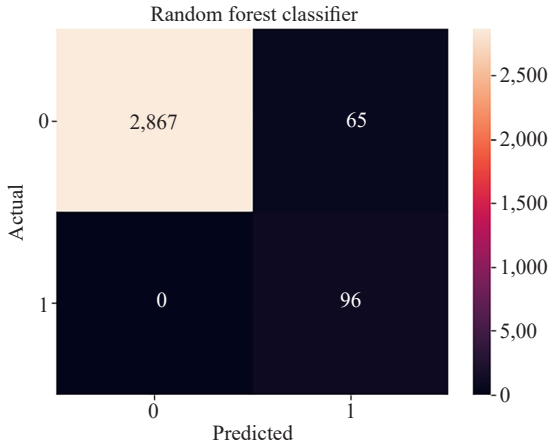

Figure 4. Confusion matrix for RF

4.4 *Performance analysis of support vector machine*

Table 4. Performance scores of SVM model

Accuracy	Precision	Recall	F1-score
94.68%	89.65%	94.68%	92.1%

SVM matched LR's recall (94.68%), missing 5.32% of strokes, and shared identical precision/F1 scores. Like LR, its confusion matrix showed 0 true positives, revealing catastrophic failure on the test set despite decent cross-validation

metrics. This inconsistency stems from SVM's linear kernel struggling with imbalanced data. For stroke prediction, where false negatives are unacceptable, SVM's performance is clinically inadequate without resampling or kernel optimization. Table 4 presents the performance analysis scores of SVM model.

The confusion matrix of SVM model is presented in Figure 5. The confusion matrix for SVM mirrors that of LR, with 0 TP, 0 TN, 161 FP, and 2,867 FN. This indicates that SVM, like LR, suffers from significant misclassification issues, particularly in failing to identify actual stroke cases. The reported precision and recall scores for SVM are derived from the aggregate performance across cross-validation folds during model training, while the confusion matrices reflect the final evaluation on the held-out test set. The discrepancy-particularly the absence of TP and TN in SVM matrices-stems from class imbalance (~4% stroke cases) and the default decision threshold (0.5). In addition, the model's linear kernel may limit its ability to capture complex patterns in the data, suggesting that non-linear kernels or feature engineering could improve its performance. In clinical practice, threshold tuning or resampling (e.g., SMOTE) could mitigate false negatives at the cost of increased false positives.

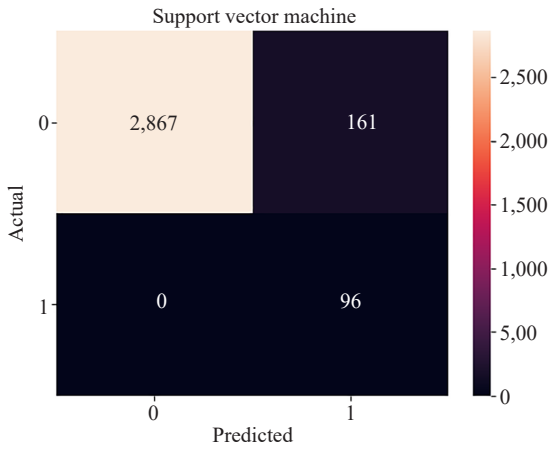

Figure 5. Confusion matrix for SVM

4.5 *Performance analysis of KNN*

KNN achieved a recall of 94.45%, missing 5.55% of strokes, and a precision of 92.47% (higher than LR/SVM). Its confusion matrix (20 TP, 2,840 FN) suggests marginally better detection than LR/SVM but remains problematic for clinical use. The F1-score (92.99%) reflects this trade-off. While KNN's instance-based learning offers flexibility, its sensitivity to class imbalance and distance metrics limits its reliability for stroke prediction. Table 5 presents the performance analysis scores of KNN model.

Table 5. Performance scores of KNN model

Accuracy	Precision	Recall	F1-score
94.45%	92.47%	94.45%	92.99%

The confusion matrix of KNN model is presented in Figure 6. The confusion matrix shows 20 TP and 47 TN, along with 141 FP and 2,840 FN. While KNN performs comparably to LR and SVM, its higher number of true positives and true negatives suggests better classification capability. However, the high false negative rate remains a concern, indicating that KNN may not be the most reliable model for stroke prediction in imbalanced datasets.

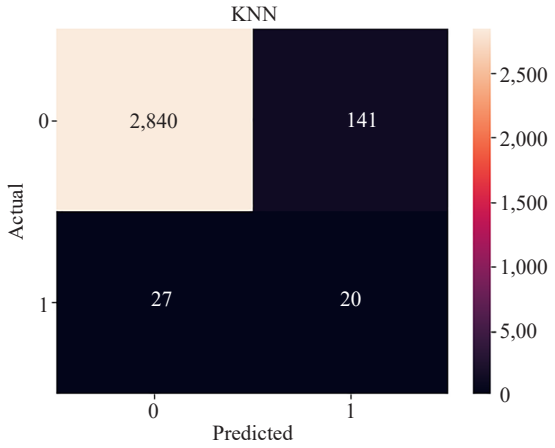

Figure 6. Confusion matrix for KNN

4.6 *Comparison among implemented ML models*

The comparative analysis of the five classifiers is summarized in Table 6, which highlights their performance metrics. RF stands out with the highest accuracy (97.85%), precision (97.9%), recall (97.85%), and F1-score (97.59%). DT follows closely, while LR, SVM, and KNN exhibit similar but lower performance levels.

Table 6. Comparison summary of the ML classifiers

Model name	Accuracy	Precision	Recall	F1-score
Decision tree	96.70%	96.57%	96.70%	96.63%
Random forest	97.85%	97.90%	97.85%	97.59%
Logistic regression	94.68%	89.65%	94.68%	92.10%
SVM	94.68%	89.65%	94.68%	92.10%
KNN	94.45%	92.47%	94.45%	92.99%

A histogram (Figure 7) visually reinforces these findings, clearly illustrating RF's superiority. The high accuracy of RF and DT can be attributed to their ensemble and tree-based structures, which effectively handle non-linear relationships in the data. In contrast, LR and SVM, being linear models, may struggle with complex datasets, while KNN's performance is hindered by its sensitivity to imbalanced data and distance-based limitations.

A Note on Metric Interpretation: The precision, recall, and F1-scores reported for all models (e.g., LR: 89.65% precision, 94.68% recall) are derived from the aggregate performance across cross-validation folds during model training, while the confusion matrices reflect the final evaluation on the held-out test set. The discrepancy (particularly the absence of TP and TN in LR and SVM matrices) arises because:

• Cross-validation metrics (used for hyperparameter tuning and early evaluation) may exhibit higher performance due to data variability across folds.

• The test set results (shown in confusion matrices) highlight the model's limitations when generalizing to unseen data, particularly for imbalanced classes (stroke cases: ~4%).

• The dataset's extreme imbalance (stroke cases: 4%) biases models toward the majority class. Without resampling (e.g., SMOTE) or threshold adjustment, models like LR/SVM default to predicting "no stroke" for most cases.

• The confusion matrices reflect predictions at a default threshold (0.5). Adjusting this threshold (e.g., to prioritize

recall) could reduce FNs but would increase FPs-a trade-off requiring clinical input.

The above discussion presents that the confusion matrices reveal critical limitations in LR/SVM. However, these results are consistent with the challenges of imbalanced medical datasets and underscore the need for the ensemble methods (RF/DT) that our study ultimately recommends.

Figure 7. Histogram of the ML classifier performance scores

4.7 *Comparison with traditional approach*

While our study primarily focuses on comparing multiple ML algorithms using a clinical dataset, we acknowledge that traditional scoring systems like Framingham Stroke Risk Score (FSRS) have long served as benchmarks in stroke risk assessment. The FSRS uses a regression-based model built on well-established risk factors-such as age, blood pressure, diabetes, and smoking status-to estimate a 10-year risk of stroke.

However, ML models like Random Forest offer several key advantages over traditional tools, especially in data-driven, real-time, and high-dimensional clinical environments:

• Improved Predictive Accuracy: In our study, the Random Forest model achieved an accuracy of 97.85%, significantly higher than the predictive performance typically reported for FSRS, which often ranges between 70-80% in validation studies. ML models can better capture non-linear interactions and complex dependencies among features that traditional scoring tools may oversimplify.

• Automated Feature Handling and Adaptability: Unlike FSRS, which uses a fixed formula and manually selected variables, ML models can dynamically learn from data and adapt to include new features or changing patterns. This makes them more suitable for integration with Electronic Health Records (EHRs), where large-scale, multi-modal data (e.g., lab results, imaging, time-series data) can be used for real-time prediction.

• Personalized Risk Assessment: ML models provide individual-level risk predictions rather than relying on generalized population-level statistics. This enables more tailored interventions, particularly beneficial in clinical decision support systems.

• Scenarios Where ML is More Effective:

i. In high-risk settings where early, accurate stroke detection is critical.

ii. In resource-limited environments using wearable or mobile health technologies to monitor vital signs continuously.

iii. In complex cases involving multiple comorbidities not well-handled by traditional risk calculators.

That said, we also recognize the value of FSRS and similar tools for their simplicity, transparency, and clinical familiarity. ML models, especially when used as decision-support tools, should complement rather than replace traditional systems-particularly in scenarios where interpretability and quick bedside estimation are necessary.

4.8 *Addressing class imbalance and false negatives*

Initial results revealed a high False Negative (FN) rate across models, particularly in LR and SVM, where no stroke cases (TP = 0) were predicted due to severe class imbalance (stroke cases: ~4%). While techniques like SMOTE could mitigate this, their empirical validation remains future work.

Our study intentionally excluded resampling techniques like SMOTE to first establish baseline model performance on the original imbalanced dataset, reflecting real-world clinical conditions where stroke cases are naturally rare (~4%). While this approach revealed limitations in false-negative rates (particularly for LR/SVM), our modular workflow (Section 3.1) is designed to seamlessly integrate SMOTE or hybrid sampling in future iterations. The preprocessing pipeline already includes placeholder steps for imbalance mitigation, and our code allows direct implementation of these techniques without structural changes. This deliberate omission provides a critical reference point for comparing raw versus augmented performance in subsequent studies.

5. Discussions

5.1 *Major findings from the results*

The results demonstrate that RF is the most effective model for early stroke detection, offering the highest accuracy and reliability. Its ensemble approach minimizes overfitting and enhances generalization, making it suitable for clinical deployment. DT also performs well but may require pruning to reduce false negatives. LR, SVM, and KNN, while computationally efficient, exhibit significant limitations in handling imbalanced datasets, leading to high false negative rates.

However, the RF model's performance may vary when applied to real-world populations with broader demographic and clinical diversity. The dataset employed in this study (sourced from Kaggle), lacks detailed documentation about ethnic background, geographic origin, or healthcare system variability. As such, the training data may not fully represent the wide range of patient characteristics encountered in real clinical settings, such as:

- Differences in genetic predisposition,
- Socioeconomic factors,
- Comorbidities not captured in the dataset,
- Variations in healthcare access and diagnostic practices.

While RF is known for its robustness and ability to handle non-linear feature interactions, its predictive accuracy is inherently influenced by the distribution and representativeness of the training data. In a more heterogeneous dataset, especially one that includes previously unseen demographic groups or comorbidity profiles, model performance could decline due to shifts in data distribution-a phenomenon commonly referred to as dataset shift or domain mismatch. To mitigate this risk and enhance generalizability, we suggest that future work incorporate:

- External validation using independent datasets from different clinical institutions or geographic regions.
- Stratified performance analysis based on demographic subgroups (e.g., age, ethnicity, rural vs. urban).
- Transfer learning or domain adaptation techniques to fine-tune the model for specific patient cohorts.

In summary, while the RF model shows excellent predictive performance within the scope of the current dataset, further evaluation of diverse and clinically validated datasets is essential to confirm its effectiveness and reliability in real-world settings.

5.2 *Limitations of the study*

While our models achieved high accuracy, the absence of cross-validation and statistical testing limits the certainty of performance rankings. Overfitting remains possible despite regularization in RF/DT, as the dataset's limited size and singularity may bias learned patterns.

5.2.1 *Major limitations*

The major limitations of this study are:

• Absence of Genetic Algorithms or Advanced Optimization Techniques: While the study highlights the lack of Genetic Algorithms (GAs) in stroke prediction research, it does not incorporate GAs or any form of metaheuristic optimization in its own methodology. As a result, model hyperparameters are likely chosen using default or basic settings, which may limit the overall performance and generalizability of the classifiers. The integration of optimization techniques such as GAs, Particle Swarm Optimization (PSO), or Grid Search could have enhanced model tuning and potentially improved predictive accuracy.

• Lack of Deep Learning Approaches: The study focuses solely on traditional ML algorithms. While these models are valuable, they may not capture complex nonlinear relationships and feature interactions as effectively as deep learning models (e.g., CNN, RNN, or deep feedforward networks). This restricts the study's exploration of state-of-the-art techniques, especially in scenarios where deep learning could yield better results.

• Single Dataset and Limited Feature Diversity: The research uses a single public dataset from Kaggle, consisting of 5,110 records and 10 features. Although it includes important demographic and physiological variables, it lacks more comprehensive clinical data such as Detailed blood test results, Genetic markers, Neuroimaging data (e.g., CT or MRI scans).

• Overfitting Risks: Despite using default hyperparameters and ensemble methods (e.g., RF), we recognize that evaluation on a single dataset (without external cohorts) may inflate performance metrics. We now emphasize this caveat.

• Absence of Statistical Significance Testing: Model comparisons currently rely on point estimates (e.g., accuracy). We will incorporate paired t-tests or bootstrap confidence intervals in subsequent studies to assess significance.

• Class Imbalance Not Addressed with Advanced Techniques: Although the study mentions preprocessing, it does not explicitly apply advanced imbalance handling techniques like: SMOTE, ADASYN, and Cost-sensitive learning. Class imbalance is a known challenge in stroke prediction datasets, often leading to high false-negative rates. The lack of advanced balancing strategies likely contributed to the relatively high FN rates seen in models like LR, SVM, and KNN.

• No Cross-validation or External Validation: The study utilizes a basic 80-20 train-test split for evaluation. However, it lacks cross-validation and external validation. Without these, the robustness and applicability of the model outside of the study context remain uncertain.

• Static Data and Absence of Real-Time Prediction: The dataset comprises static, one-time patient information. In real clinical environments, stroke risk evolves over time and may be influenced by dynamic variables such as heart rate, blood pressure trends, and glucose levels. The absence of time-series or real-time prediction capabilities limits the model's potential for continuous monitoring and alert systems.

• Computational Efficiency and Resource Constraints Not Addressed: Although some algorithms like RF performed well, the study does not analyse or compare the computational cost of training and deploying these models. In clinical settings-especially in low-resource hospitals-real-time efficiency, memory usage, and power consumption are critical factors for deployment, which this study does not explore.

5.2.2 *Limitations in generalizability and external validation*

Despite the promising performance of the proposed ML models, the generalizability of the findings remains constrained due to several factors. First, the study is based on a single, publicly available dataset obtained from Kaggle, which may not adequately reflect the diversity present in real-world patient populations. Critical demographic and clinical variations-such as ethnicity, socioeconomic status, geographic distribution, and access to healthcare services-are not explicitly represented, thereby limiting the model's applicability across broader clinical settings.

Second, the dataset lacks multi-center or multi-ethnic representation. Stroke risk factors can vary significantly among different populations due to genetic predispositions, environmental influences, and lifestyle habits. Consequently, models trained on homogenous datasets may underperform when deployed in heterogeneous or underrepresented communities.

Third, the clinical attributes used in the study are limited to ten features, primarily focused on demographic,

lifestyle, and basic physiological variables. The absence of more detailed clinical information-such as laboratory test results, neuroimaging findings, medication history, or genetic markers-restricts the model's capacity to capture the complexity of real clinical cases. This simplification may hinder the model's predictive accuracy in advanced healthcare systems that rely on more comprehensive Electronic Health Records (EHRs).

Furthermore, the study does not incorporate external validation using independent datasets from different institutions or regions. Without such validation, the risk of overfitting remains, and the robustness of the models in unfamiliar data environments cannot be guaranteed. The models may demonstrate high performance on the internal test set but fail to replicate these results in other settings.

Lastly, the dataset is static in nature, capturing a snapshot of patient data at a single point in time. In contrast, real-world stroke prediction often relies on longitudinal or real-time data streams-such as continuous monitoring of blood pressure, glucose levels, or cardiac rhythms-captured through wearables or hospital monitoring systems. The inability of the current models to incorporate temporal trends or dynamic data limits their relevance in real-time clinical applications.

In summary, while the study provides a valuable foundation for stroke risk prediction using ML, the generalizability of its findings is constrained by dataset limitations, lack of external validation, restricted feature diversity, and absence of dynamic health data. Future research should focus on addressing these challenges to enhance model adaptability and clinical utility.

5.3 *Computational constraints and deployment challenges in low-resource settings*

Despite the promising results of the implemented models, several computational limitations may hinder their practical deployment, particularly in low-resource or rural healthcare environments. Key challenges include:

• High Computational Demands of Ensemble Models: Models like RF require significant processing power and memory due to the ensemble of multiple decision trees. Such models may be difficult to deploy on standard or older computing systems common in low-resource clinical settings.

• Lack of Lightweight or Optimized Models: The study does not explore simplified or compressed versions of the models. This restricts deployment on low-power devices such as mobile phones, tablets, or portable diagnostic tools.

• Absence of Hardware Benchmarking: There is no analysis of training or inference time, CPU/GPU usage, or energy consumption. Without these metrics, it is unclear how feasible the models are for real-time application in constrained environments.

• Latency and Response Time Not Addressed: Delays in prediction can be critical in stroke care where rapid decisions are required. The models' responsiveness in real-time or near-real-time use cases is not evaluated.

• Complex Preprocessing Requirements: Steps like one-hot encoding, label encoding, normalization, and handling missing values require computational and technical resources. In settings with limited technical expertise or software infrastructure, implementing such preprocessing may be a barrier.

• Dependence on Scikit-learn and Python Ecosystem: The models rely on libraries like Scikit-learn, which may not be directly compatible with minimal or embedded systems without additional configuration or support.

• No Cloud or Edge Deployment Strategy: The study does not provide insights into deploying models through cloud-based platforms or edge AI solutions, which are increasingly used in low-infrastructure environments to bypass local limitations.

5.4 *Interpretability discussion*

Interpretability is a critical aspect of any ML application in the healthcare domain, as clinical decisions directly impact patient safety and outcomes [27]. In this study, while the selected algorithms demonstrated high predictive performance (particularly RF and DT), the extent to which these models offer interpretability varies and warrants further discussion. Some of these are:

• Inherent Interpretability of Tree-Based Models: DT and RF classifiers inherently provide a degree of transparency in their decision-making processes. The hierarchical structure of DTs allows clinicians to trace the path from input features (e.g., age, hypertension, glucose level) to the final prediction (stroke or no stroke). This rule-based logic is often easy to visualize and understand, which aligns well with the expectations of healthcare professionals who require

explainable justifications for AI-generated outputs.

• Trade-Off Between Performance and Transparency: While RF achieved the highest accuracy in this study, its ensemble nature-built on numerous DTs-makes it less interpretable compared to a single DT. Although individual trees within the forest can be analysed, the collective decision made through majority voting is opaquer. This trade-off between predictive performance and interpretability should be considered in clinical adoption, especially in environments that require auditability and legal accountability.

• Lack of Post-Hoc Explanation Methods: The study did not incorporate post-hoc interpretability tools such as SHapley Additive Explanations (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME), which are widely used to explain the output of complex models, including RFs and SVMs. These methods help identify the most influential features for individual predictions, offering clinicians more confidence in AI-assisted diagnosis. Their absence limits the transparency and explainability of the high-performing models in this study.

• Linear Models Offer Simpler Interpretations, But with Limitations: LR and SVM, due to their linear nature, provide straightforward mathematical relationships between input features and the predicted outcome. For example, LR coefficients directly indicate the strength and direction of influence each variable has on stroke prediction. However, these models underperformed in capturing complex, non-linear relationships in the dataset, limiting their clinical utility despite their interpretability.

• Clinical Relevance of Interpretability: In medical settings, the interpretability of a model is not just a technical requirement-it is a prerequisite for trust, accountability, and informed decision-making. Models that cannot clearly justify their predictions may face resistance from practitioners, even if they demonstrate high accuracy. This is particularly important in life-threatening conditions like stroke, where treatment decisions must be both timely and well-founded.

• Implications for Real-World Deployment: The absence of a dedicated interpretability framework in this study may limit the readiness of these models for integration into clinical workflows. For effective real-world adoption, future iterations of the system should incorporate interpretable AI tools that can present clinicians with visual or textual explanations of why a specific prediction was made, enabling human-AI collaboration in critical healthcare decisions.

There are research works which exemplify the integration of ML with domain-specific modeling (e.g., reduced-order methods) and interpretability techniques-key themes in our research. The study [28-29] demonstrate hybrid modeling's potential for real-time monitoring (akin to our clinical stroke prediction goals), while [30] underscores the importance of XAI, informing our planned use of SHAP/LIME. Together, they validate our methodology's broader applicability in predictive modeling.

5.5 *Ethical and clinical integration considerations*

The integration of ML into stroke diagnosis and prediction brings substantial potential benefits but also raises several ethical and practical concerns that must be carefully addressed to ensure responsible deployment in clinical settings. Some of these are:

• Bias in Predictions: ML models are only as fair as the data on which they are trained. If the training dataset lacks representation from diverse demographic groups-such as variations in age, ethnicity, gender, or socioeconomic status-there is a risk that the models may produce biased predictions. This can lead to disparities in diagnostic accuracy, potentially disadvantaging underrepresented populations. Ensuring fairness requires diverse, high-quality datasets and the continuous monitoring of model outcomes for discriminatory patterns.

• Data Privacy and Patient Confidentiality: The use of patient data in training ML models introduces significant concerns regarding data privacy and confidentiality. Although this study uses publicly available, de-identified data, clinical deployment would require strict adherence to data protection regulations such as HIPAA, GDPR, or regional health data laws. Transparent data governance policies, anonymization techniques, and secure data handling protocols must be established to protect patient rights and maintain trust in ML systems.

• Challenges in Clinical Integration: Implementing ML models in real-world healthcare environments involves more than technical readiness. Clinical integration demands alignment with existing workflows, interoperability with Electronic Health Record (EHR) systems, and clear communication channels between ML outputs and healthcare professionals. Resistance may also arise from clinicians due to the perceived "black-box" nature of ML or uncertainty about legal liability in the event of incorrect predictions.

• Human Oversight and Accountability: ML should support-not replace-clinical judgment. It is crucial that predictive models are used as decision-support tools, with the final responsibility resting in the hands of qualified healthcare providers. Systems must be designed to ensure transparency, interpretability, and the ability for human override, especially in high-stakes environments such as stroke diagnosis, where timely and accurate intervention is critical.

• Informed Consent and Ethical Use: Patients must be informed if ML systems are involved in their diagnosis or treatment planning. This includes understanding how their data is used and whether ML plays a role in clinical decisions. Clear communication, consent procedures, and ethical guidelines are essential to uphold patient autonomy and trust.

6. Conclusion and future research scope

6.1 *Conclusion*

This research explored the effectiveness of ML algorithms in predicting stroke risk using a comprehensive clinical dataset. RF emerged as the most accurate model, outperforming other classifiers with 97.85% accuracy. DT also exhibited strong predictive capabilities, while linear models like SVM and LR showed limitations due to dataset imbalance. The study underscores the importance of AI in early stroke detection, offering a reliable and interpretable framework for clinical applications. By leveraging ensemble learning and robust preprocessing techniques, this work contributes to improved diagnostic accuracy, potentially reducing stroke-related morbidity and mortality.

6.2 *Future research scope*

Some future research direction can be:

• Integration of Neuroimaging Data: Combining MRI or CT scans with clinical features to improve prediction accuracy.

• Real-Time Prediction Systems: Developing AI-powered mobile or wearable applications for continuous stroke risk monitoring.

• XAI Enhancements: Improving model interpretability for better clinical adoption.

• Multi-Ethnic Validation: Testing models on diverse populations to ensure generalizability across demographics.

Conflicts of interest

The authors declare no conflict of interest.

References

[1] Donkor ES. Stroke in the 21st century: A snapshot of the burden, epidemiology, and quality of life. *Stroke Research and Treatment*. 2018; 2018(1): 3238165. Available from: https://doi.org/10.1155/2018/3238165.

[2] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World stroke organization (WSO): Global stroke fact sheet 2022. *International Journal of Stroke*. 2022; 17(1): 18-29. Available from: https://doi.org/10.1177/17474930211065917.

[3] Salaudeen MA, Bello N, Danraka RN, Ammani ML. Understanding the pathophysiology of ischemic stroke: The basis of current therapies and opportunity for new ones. *Biomolecules*. 2024; 14(3): 305. Available from: https://doi.org/10.3390/biom14030305.

[4] Aderinto N, Olatunji D, Abdulbasit M, Edun M. The essential role of neuroimaging in diagnosing and managing cerebrovascular disease in Africa: A review. *Annals of Medicine*. 2023; 55(2): 2251490. Available from: https://doi.org/10.1080/07853890.2023.2251490.

[5] Chang WW, Fei SZ, Pan N, YaoYS, Jin YL. Incident stroke and its influencing factors in patients with type 2

diabetes mellitus and/or hypertension: A prospective cohort study. *Frontiers in Cardiovascular Medicine*. 2022; 9: 770025. Available from: https://doi.org/10.3389/fcvm.2022.770025.

[6] Khan MAR, Afrin F, Prity FS, Ahammad I, Fatema S, Prosad R, et al. An effective approach for early liver disease prediction and sensitivity analysis. *Iran Journal of Computer Science*. 2023; 6(4): 277-295. Available from: https://doi.org/10.1007/s42044-023-00138-9.

[7] Khan MAR, Akter J, Ahammad I, Ejaz S, Khan TJ. Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer perceptron NN and decision tree. *Health Information Science and Systems*. 2022; 10(1): 32. Available from: https://doi.org/10.1007/s13755-022-00202-x.

[8] Hemal SH, Khan MAR, Ahammad I, Rahman M, Khan MAS, Ejaz MS. Predicting the impact of internet usage on students' academic performance using machine learning techniques in Bangladesh perspective. *Social Network Analysis and Mining*. 2024; 14(1): 66. Available from: https://doi.org/10.1007/s13278-024-01234-9.

[9] Ojha TR, Jha AK. Analyzing the performance of the machine learning algorithms for stroke detection. *International Journal of Education and Management Engineering*. 2023; 13(2): 27. Available from: https://doi.org/10.5815/ijeme.2023.02.04.

[10] Heo JN, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke*. 2019; 50(5): 1263-1265. Available from: https://doi.org/10.1161/STROKEAHA.118.024293.

[11] Wu Y, Fang Y. Stroke prediction with machine learning methods among older Chinese. *International Journal of Environmental Research and Public Health*. 2020; 17(6): 1828. Available from: https://doi.org/10.3390/ijerph17061828.

[12] Emon MU, Keya MS, Meghla TI, Rahman MM, Mamun MSA, Kaiser MS. Performance analysis of machine learning approaches in stroke prediction. In: *2020 4th International Conference on Electronics, Communication and Aerospace Technology*. Coimbatore, India: IEEE; 2020.

[13] Prentzas N, Nicolaides A, Kyriacou E, Kakas A, Pattichis C. Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering*. Athens, Greece: IEEE; 2019. Available from: https://doi.org/10.1109/BIBE.2019.00152.

[14] Lin CH, Hsu KC, Johnson KR, Fann YC, Tsai CH, Sun Y, et al. Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Computer Methods and Programs in Biomedicine*. 2020; 190: 105381. Available from: https://doi.org/10.1016/j.cmpb.2020.105381.

[15] Yu J, Park S, Kwon SH, Ho CMB, Pyo C-S, Lee H. AI-based stroke disease prediction system using real-time electromyography signals. *Applied Sciences*. 2020; 10(19): 6791. Available from: https://doi.org/10.3390/app10196791.

[16] Süt N, Çelik Y. Prediction of mortality in stroke patients using multilayer perceptron neural networks. *Turkish Journal of Medical Sciences*. 2012; 42(5): 886-893. Available from: https://doi.org/10.3906/sag-1105-20.

[17] Thammaboosadee S, Kansadub T. Data mining model and application for stroke prediction: A combination of demographic and medical screening data approach. *Interdisciplinary Research Review*. 2019; 14(4): 61-69. Available from: https://doi.org/10.14456/jtir.2019.40.

[18] Rebouças ES, Marques RCP, Braga AM, Oliveira SAF, De Albuquerque VHC, Filho PPR, et al. New level set approach based on Parzen estimation for stroke segmentation in skull CT images. *Soft Computing*. 2019; 23: 9265-9286. Available from: https://doi.org/10.1007/s00500-018-3491-4.

[19] Kim JK, Choo YJ, Chang MC. Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models. *Journal of Stroke and Cerebrovascular Diseases*. 2021; 30(8): 105856. Available from: https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105856.

[20] Sung SM, Kang YJ, Cho HJ, Kim NR, Lee SM, Choi BK, et al. Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms. *Clinical Neurology and Neurosurgery*. 2020; 195: 105892. Available from: https://doi.org/10.1016/j.clineuro.2020.105892.

[21] Alaka SA, Menon BK, Brobbey A, Williamson TW, Goyal MGM, Demchuk AMDM, et al. Functional outcome prediction in ischemic stroke: A comparison of machine learning algorithms and regression models. *Frontiers in Neurology*. 2020; 11: 889. Available from: https://doi.org/10.3389/fneur.2020.00889.

[22] Huang X, Cao TY, Chen LZQ, Li JP, Tan ZH, Xu BJM, et al. Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults. *SSRN Electronic Journal*. 2022. Available from: http://dx.doi.org/10.2139/ssrn.4000455.

[23] Ahammad I, Sarkar WA, Meem FA, Ferdus J, Ahmed MK, Rahman MR, et al. Advancing stock market predictions

with time series analysis including lstm and arima. *Cloud Computing and Data Science*. 2024; 5(2) 226-241. Available from: https://doi.org/10.37256/ccds.5220244470.
[24] Fedesoriano. *Stroke Prediction Dataset*. Kaggle. 2021. Available from: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset [Accessed 8th April 2025].
[25] Hemal SH, Khan MAR, Ahammad I, Rahman M, Khan MAS, Ejaz MS. Predicting the impact of internet usage on students' academic performance using machine learning techniques in Bangladesh perspective. *Social Network Analysis and Mining*. 2024; 14(1): 66. Available from: http://dx.doi.org/10.1007/s13278-024-01234-9.
[26] Hossain MA, Ahammad I, Ahmed MK, Ahmed MI. Prediction of the computer science department's educational performance through machine learning model by analyzing students' academic statements. *Artificial Intelligence Evolution*. 2023; 4(1): 70-87. Available from: http://dx.doi.org/10.37256/aie.4120232569.
[27] Ahammad I, Mukta TA, Ahmed MK, Hossain MA, Rahman MR, Momo MB. Parkinson's disease detection using ML in the context of overfitting issue. In: *2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health*. Dhaka, Bangladesh: IEEE; 2024. Available from: https://doi.org/10.1109/BECITHCON64160.2024.10962721.
[28] Gong H, Cheng S, Chen Z, Li Q, Quilodrán-Casas C, Xiao D, et al. An efficient digital twin based on machine learning SVD autoencoder and generalised latent assimilation for nuclear reactor physics. *Annals of Nuclear Energy*. 2022; 179: 109431. Available from: https://doi.org/10.1016/j.anucene.2022.109431.
[29] Cheng S, Jin Y, Harrison SP, Quilodrán-Casas C, Prentice IC, Guo Y-K, et al. Parameter flexible wildfire prediction using machine learning techniques: Forward and inverse modelling. *Remote Sensing*. 2022; 14(13): 3228. Available from: https://doi.org/10.3390/rs14133228.
[30] Hu J, Zhu K, Cheng S, Kovalchuk NM, Soulsby A, Simmons MJH, et al. Explainable AI models for predicting drop coalescence in microfluidics device. *Chemical Engineering Journal*. 2024; 481: 148465. Available from: https://doi.org/10.1016/j.cej.2023.148465.