Impact Factor (SJIF): 5.158 E-ISSN: 2583-7206, P-ISSN: 3049-2432

International Journal of All Subject Research

Received: 28/August/2025 IJASR: 2025; 4(5):38-42 Accepted: 08/October/2025

A Dual-stage Machine Learning Framework for Heart and Stroke Prediction with Progression Path Modeling

[*1]Dr. S Krishnaveni and [2]Muthunayaki M

[*1]Assistant Professor, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

[2]PG Scholar, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

Abstract

This research introduces a machine learning framework designed to predict stroke risk while modelling its potential progression from heart disease. Using a structured healthcare dataset, the system applies supervised learning algorithms—Random Forest, Decision Tree, and Naive Bayes to analyse clinical, demographic, and lifestyle features. A key contribution is the development of a "Progression Path" logic that categorizes individuals into four stages: No Risk, Heart Only, Stroke Only, and Heart → Stroke. The models are evaluated using accuracy, confusion matrices, and feature importance analysis. Additionally, visual tools such as Sankey diagrams and patient-level probability simulations improve interpretability and clinical relevance. The proposed framework supports early diagnosis, personalized risk forecasting, and proactive intervention, demonstrating the practical value of machine learning in preventive healthcare.

Keywords: Stroke prediction, Sankey diagram, Random Forest, Naive Bayes, Machine learning, Disease progression.

1. Introduction

Stroke is a serious neurological condition that often occurs without warning and can lead to permanent disability or death. It is increasingly recognized as a disease that does not occur in isolation but is influenced by a combination of cardiovascular, metabolic, and lifestyle-related risk factors. Among these, heart disease stands out as a major precursor, sharing many overlapping causes such as hypertension, elevated glucose levels, obesity, and smoking habits. Despite advancements in medical diagnostics, early identification of stroke-prone individuals remains a significant challenge, especially in resource-limited healthcare environments [1].

Recent advancements in machine learning have opened up new possibilities in the field of preventive medicine. Complex patterns in healthcare data that may not be immediately obvious to physicians can be analyzed using machine learning algorithms. These models can assist in the early detection of risk by learning from historical patient data and making real-time predictions for new individuals.

This research aims to develop a two-stage machine learning framework that can predict both stroke and heart disease using patient features, including age, BMI, average glucose level, marital status, and lifestyle habits. Beyond binary classification, this study introduces a novel "Progression Path" that maps potential transitions from heart disease to stroke. This pathway categorizes patients into one of four stages: No Risk, Heart Only, Stroke Only, and Heart → Stroke, [3] thereby offering a broader understanding of disease progression.

To enhance the interpretability and practical value of the predictions, the study also incorporates visual tools such as Sankey diagrams and patient-level simulations [4]. These enable healthcare professionals to understand patient-specific outcomes better and take timely preventive action. This work contributes not only to stroke prediction but also to the broader goal of building predictive systems that support personalized healthcare and early intervention.

2. Review of Literature

Numerous studies have explored the application of machine learning to predict stroke using clinical and demographic datasets. Models such as Logistic Regression and Support Vector Machines have been commonly employed due to their interpretability and ease of use. Additionally, more advanced methods like Random Forest and XGBoost have shown improved performance by handling feature interactions and nonlinear patterns in the data [1, 2].

Despite these advancements, a majority of existing research focuses solely on stroke prediction without considering its possible links to preceding cardiovascular conditions. The isolation of stroke risk from heart disease in predictive models presents a critical gap in understanding disease progression [3]. Moreover, while some studies attempt to address class imbalance through resampling techniques such as oversampling and SMOTE, many still fall short in providing a

***Corresponding Author:** Dr. S Krishnaveni

comprehensive risk modelling pipeline that supports personalized healthcare [4].

In contrast to the above approaches, our study addresses stroke and heart disease independently and integrates them into a unified predictive system. The inclusion of a novel "Progression Path" logic makes our framework distinct, as it models transitions from heart disease to stroke and categorizes patients accordingly. This holistic view is further strengthened through the use of patient-level simulations and visualizations such as Sankey diagrams, which are largely absent in prior literature [5].

Therefore, our research advances the field by combining disease progression modeling with machine learning classification and enhancing transparency through feature importance analysis. This makes the proposed system more robust, interpretable, and suitable for integration into real-world clinical workflows.

3. Objective

The primary aim of this study is to develop a comprehensive dual-stage machine learning framework capable of predicting both stroke and heart disease risks. First, we seek to construct and compare supervised classifiers—including Random Forest, Decision Tree, and Naive Bayes—to determine which algorithm most effectively identifies individuals at risk based on clinical and lifestyle variables such as age, BMI, glucose level, hypertension, and smoking habits. Next, we introduce a novel "Progression Path" mapping logic that categorizes patients into four risk stages (No Risk, Heart Only, Stroke Only, and Heart → Stroke), illustrating the transition from cardiac conditions to cerebrovascular events. To ensure the model's applicability in real-world settings, we incorporate strategies like stratified train-test splits and balanced performance metrics (precision, recall, and F1-score) to address class imbalance. Moreover, we enhance model interpretability through feature importance extraction and deploy visualization techniques such as Sankey diagrams and patient-level probability simulations. Finally, by delivering personalized risk thresholds and actionable visual outputs, our work aims to support early intervention, improve clinical decision-making, and contribute a transparent, interpretable tool within preventive healthcare systems.

4. Methodology

Model Selection to achieve reliable and interpretable results, three well-established supervised learning algorithms were selected: Random Forest, Decision Tree, and Gaussian Naive Bayes. These models are widely used in healthcare prediction tasks due to their simplicity, speed, and robustness in handling both linear and nonlinear data.

Separate Modelling for Heart and Stroke, separate classification models were trained for predicting heart disease and stroke. The heart model incorporates variables including age, BMI, glucose level, hypertension, marital status, and smoking habits. The stroke model included all available predictors except the stroke label. Both models were trained using the same 80:20 train-test split.

Progression Path Logic: After generating predictions for heart disease and stroke, a novel

categorical variable called "Progression Path" was created. This variable defines four possible states:

i). No Risk
ii). Heart Only
iii). Stroke Only
iv). Heart → Stroke.

These labels provide insight into the probable sequence of health deterioration, enabling clinicians to prioritize patients based on combined risk. Visualization Tools to improve interpretability, results were visualized using (for feature importance). These tools offer comprehensive insights at both the global and individual patient levels. This methodology enables dual-disease prediction with progression awareness, improving the reliability and usability of the system for clinical settings. Sankey diagrams (to show transitions between risk stages), confusion matrices (for model performance), and bar plots.

5. Data Preprocessing Pipeline

i). **Dataset Description:** The dataset used in this study consists of 5,110 records, containing patient-level information across multiple categories, including demographic details (age, gender), clinical indicators (BMI, hypertension, average glucose level, heart disease), and behavioral features (smoking status, work type, residence type, and marital status). The outcome variable is dichotomous, signifying whether the individual has undergone a stroke event.

ii). **5.2 Handling Missing Data:** One of the major preprocessing steps involved addressing missing values. Specifically, the BMI column had null entries. Instead of discarding these rows—which could reduce the sample size—we replaced missing BMI values with the mean BMI of the dataset. This technique preserved data integrity while minimizing bias.

iii). **Encoding Categorical Variables:** Numerical input is essential for machine learning models. Thus, categorical features such as gender, ever married, work type, Residence type, and smoking status were converted into numerical form using Label Encoding. Each unique category was assigned a distinct integer value, preserving the integrity of the data while making it suitable for model training.

iv). **Class Separation:** The dataset was split into features (X) and target (y), where the stroke column served as the output variable. Additional target variables, such as heart disease, were used separately to train a heart prediction model, enabling us to model the risk progression from heart disease to stroke.

v). **Data Splitting:** An 80:20 ratio was used to divide the dataset into training and testing subsets to impartially assess the models' performance. This allowed us to train the models on one portion of the data and assess their accuracy on unseen data, simulating real-world predictions.

6. Implementation

The implementation was carried out in Python using the Jupyter Notebook environment. Key libraries included Pandas and NumPy for data handling, Scikit-learn for machine learning models and metrics, and Matplotlib, Seaborn, and Plotly for visualization.

i). Model Development

The implementation was carried out in Python using the Jupyter Notebook environment. Key libraries included Pandas and NumPy for data handling, Scikit-learn for building the machine learning models, and Matplotlib, Seaborn, and Plotly for data visualization. The models developed— Random Forest, Decision Tree, and Gaussian Naive Bayes—were trained using the processed dataset. Predictions were

generated separately for stroke and heart disease, and the classification performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices. Feature importance was also extracted from the Random Forest classifier to identify the most influential variables, including age, average glucose level, and BMI.

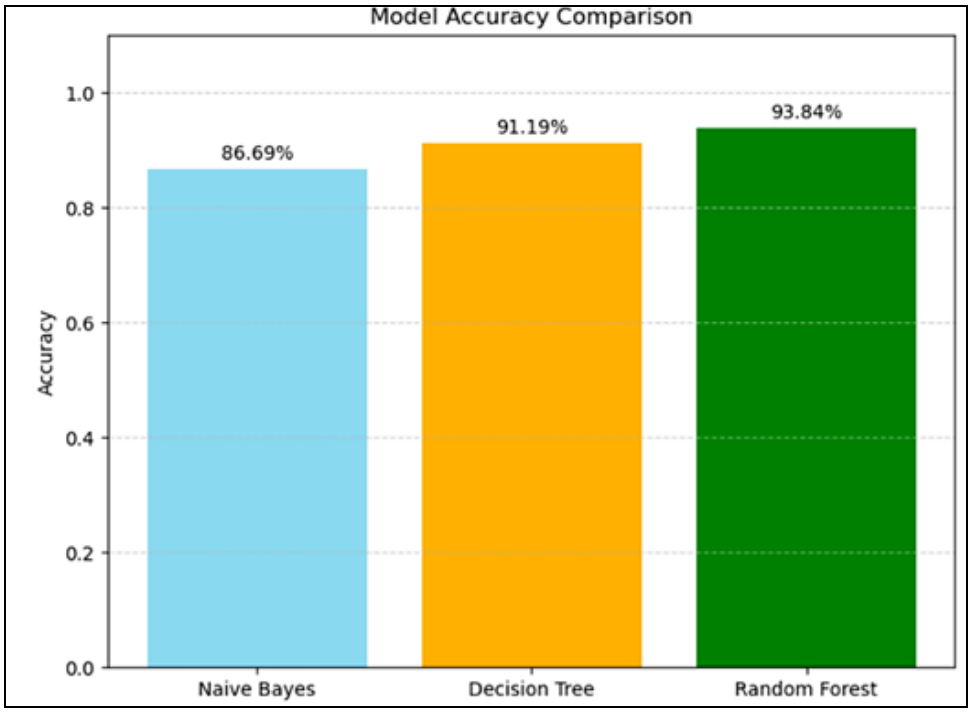


Fig 1: Model Accuracy Comparison

ii). Visualization and Patient Stimulation

To enhance interpretability, visual tools were used to represent model insights. A Sankey diagram was constructed using Plotly to illustrate transitions across the four risk stages: No Risk, Heart Only, Stroke Only, and Heart → Stroke. Confusion matrices were plotted using Seaborn to evaluate classification performance visually. Feature importance rankings were shown through bar plots. Patient-level simulation was conducted by selecting random samples from the dataset and predicting individual stroke probabilities. These simulations helped highlight high-risk cases and provided personalized output for analysis. This implementation bridges theory and application by integrating model performance, predictive simulation, and visualization within a unified framework. A Sankey diagram was constructed using Plotly to represent the transitions across the four stages of progression. Confusion matrices were plotted using Seaborn, and feature importance was visualized using bar plots to highlight influential predictors like age, glucose level, and BMI. The system also simulated predictions for randomly selected patient records and displayed individual stroke probabilities. These outputs provided meaningful insight into how underlying conditions may evolve and which patients are at highest risk. This implementation bridges theory and application by integrating model predictions, simulations, and interactive visual analysis in a single environment. This implementation bridges theory and application by integrating model predictions, simulations, and interactive visual analysis in a single environment.

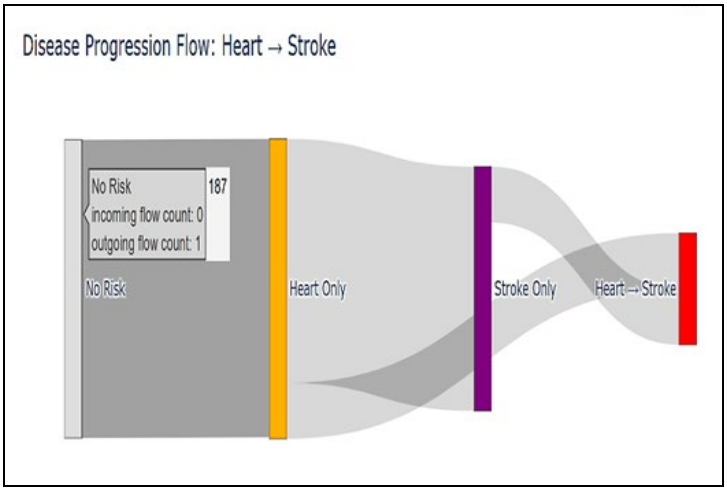


Fig 2: No risk

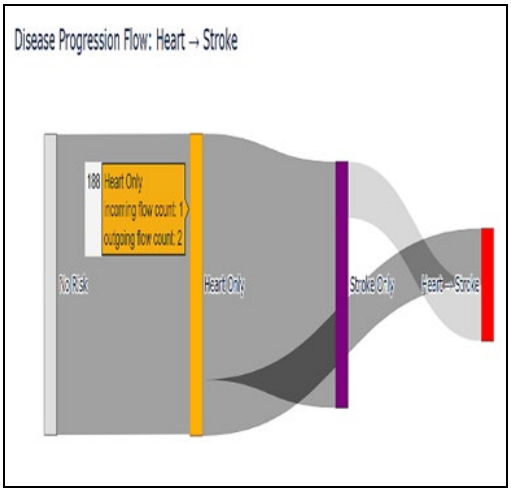


Fig 3: Heart disease

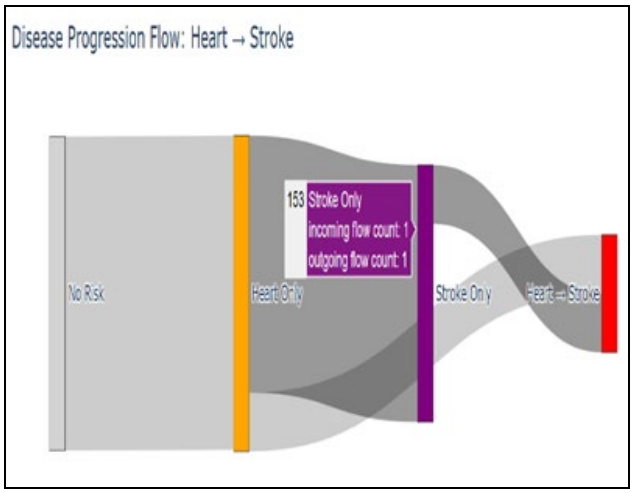
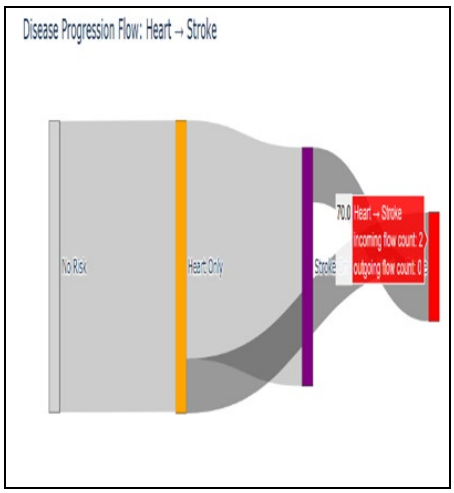


Fig 4: Stroke



Disease **Fig 5:** Heart and stroke disease

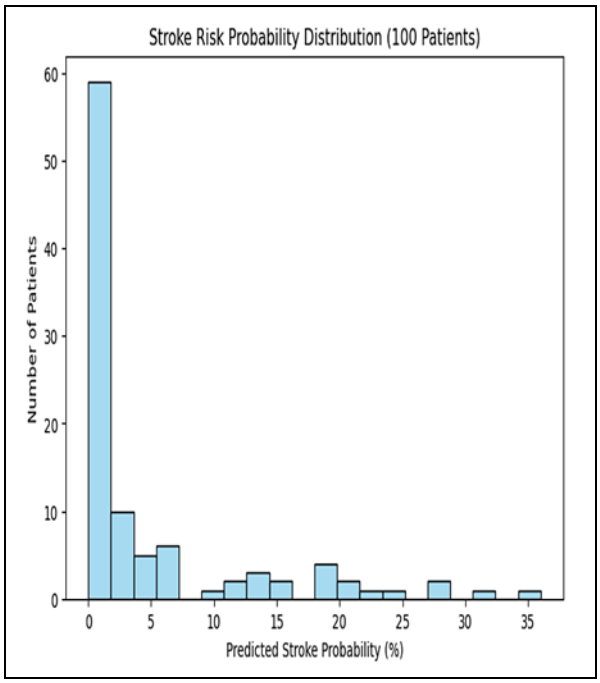


Fig 6: Stroke Risk Probability

7. Result

The evaluation of the proposed machine learning framework focused on accuracy, confusion matrix interpretation, and clinical relevance. Among the models tested, the Random Forest classifier achieved the highest accuracy in stroke prediction, followed by Decision Tree and Naive Bayes. The Random Forest model not only performed well in terms of overall accuracy but also exhibited better precision and recall for the minority stroke class, which is critical in imbalanced datasets.

The confusion matrix provided insights into false positives and false negatives, helping identify patterns of misclassification. The Decision Tree model, while interpretable, showed slightly lower performance, likely due to its sensitivity to overfitting. The Naive Bayes model had the lowest accuracy but offered speed and simplicity, making it suitable for preliminary assessments.

Age, BMI, and average glucose level were the main predictor s of stroke, according to feature importance analysis. These findings are consistent with established clinical knowledge and support the model's credibility.

The progression path simulation showed a clear transition from heart disease to stroke in certain high-risk individuals. An understandable summary of the course of the disease was provided by the Sankey diagram visualization, which showed how patient categories changed between danger states.

Simulated predictions for selected patients with ≥80% stroke probability helped flag critical cases and provided clinicians with focused intervention opportunities.

Overall, the results demonstrate that the integrated system not only performs well statistically but also enhances clinical decision-making through interpretability and individualized prediction. A Sankey diagram was constructed using Plotly to represent the transitions across the four stages of progression. Confusion matrices were plotted using Seaborn, and feature importance was visualized using bar plots to highlight influential predictors like age, glucose level, and BMI. The system also simulated predictions for randomly selected patient records and displayed individual stroke probabilities. These outputs provided meaningful insight into how underlying conditions may evolve and which patients are at highest risk. This implementation bridges theory and

application by integrating model predictions, simulations, and interactive visual analysis in a single environment.

This implementation bridges theory and application by integrating model predictions, simulations, and interactive visual analysis in a single environment.

8. Conclusion

This research presents a machine learning-based framework that predicts both stroke and heart disease while also analysing their potential progression relationship. The use of models like Random Forest, Decision Tree, and Naive Bayes allowed us to classify health risks based on important patient features such as age, BMI, glucose levels, and lifestyle habits.

A key feature of the framework is the "Progression Path" logic, which offers a clear understanding of how heart disease might lead to stroke in high-risk individuals. The visual tools and simulation results made the predictions easier to interpret and more relevant for clinical application.

In conclusion, this study provides an effective and interpretable system that supports early detection, personalized healthcare, and informed decision-making. The framework sets a strong foundation for further advancements in data-driven medical prediction systems. A Sankey diagram was constructed using Plotly to represent the transitions across the four stages of progression. Confusion matrices were plotted using Seaborn, and feature importance was visualized using bar plots to highlight influential predictors like age, glucose level, and BMI. The system also simulated predictions for randomly selected patient records and displayed individual stroke probabilities. These outputs provided meaningful insight into how underlying conditions may evolve and which patients are at highest risk. This implementation bridges theory and application by integrating model predictions, simulations, and interactive visual analysis in a single environment**.**

9. Future Work

In the future, the system can be enhanced by integrating real-time health data from wearable devices to enable continuous risk monitoring. Deep learning models may also be explored to improve prediction accuracy on larger and more complex datasets. Additionally, addressing class imbalance through advanced resampling techniques can further optimize stroke detection. The framework could be extended into a mobile or web-based application for practical use by doctors and patients. Incorporating explainable AI would also improve clinical trust in the system's predictions.

References

1. Agus Byna, M. M. Lak ulu, and I. Y. Panessai, “Current Critical Review on Prediction Stroke Using Machine Learning,” *Bulletin of Electrical Engineering and Informatics*. 2024 13(5):3470–3480. doi:10.11591/eei.v13i5.7435.
2. Q. Qin, X. Zhou, and Y. Jiang, “Prognosis Prediction of Stroke based on Machine Learning and Explanation Model,” *Int. J. Comput. Commun. Control*, 2021, 16(2). doi:10.15837/ijccc.2021.2.4108.
3. A. Tashkova, S. Eftimov, B. Ristov, and S. Kalajdziski, “Comparative Analysis of Stroke Prediction Models Using Machine Learning,” *arXiv preprint arXiv:2505.09812*, May 2025.
4. P. Chakraborty *et al*., “Predicting Stroke Occurrences: A Stacked Machine Learning Approach with Feature Selection and Data Preprocessing,” *BMC Bioinformatics*, 2024, 25(329). doi:10.1186/s12859-024-05866-8.
5. M.S. Sirsat, E. Fermé, and J. Câmara, “Machine Learning for Brain Stroke: A Review,” *Journal of Stroke and Cerebrovascular Diseases*. 2020; 29(10):105162-105170. doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.
6. L. Schwartz, R. Anteby, E. Klang, and S. Soffer, “Stroke Mortality Prediction Using Machine Learning: Systematic Review,” *Journal of the Neurological Sciences*. 2023; 444:120529-120540. doi: 10.1016/j.jns.2022.120529.
7. A. Panesar, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*, 2019. doi:10.1007/978-1-4842-3799-1.
8. F. Wang *et al*., “Personalized Risk Prediction of Symptomatic Intracerebral Hemorrhage After Stroke Thrombolysis Using a Machine-Learning Model,” *Ther Adv Neurol Disord*, vol. 13, 2020. doi:10.1177/1756286420902358.
9. H. J. Lee *et al*., “Risk of Ischemic Stroke in Metabolically Healthy Obesity: A Nationwide Population-Based Study,” *PLoS One*, 2018, 13(3). doi: 10.1371/journal.pone.0195210.
10. D. Richards *et al*., “Time to Diagnosis and Factors Affecting Diagnostic Delay in Amyotrophic Lateral Sclerosis,” *Journal of the Neurological Sciences*, 2020, 417. doi: 10.1016/j.jns.2020.117054.
11. J. Bajwa *et al*., “Artificial Intelligence in Healthcare: Transforming the Practice of Medicine,” *Future Healthcare Journal*, 2021, 8(2). doi:10.7861/fhj.2021-0095.
12. D. Kuriakose and Z. Xiao, “Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives,” *International Journal of Molecular Sciences*, 2020, 21(20). doi:10.3390/ijms21207609.
13. E. S. Donkor, “Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life,” *Stroke Res Treat*, vol. 2018, 2018. doi:10.1155/2018/3238165.
14. J. Benito-Lozano *et al*., “Diagnostic Process in Rare Diseases: Determinants Associated with Diagnostic Delay,” *Int J Environ Res Public Health*, vol. 19, no. 11, 2022. doi:10.3390/ijerph19116456.
15. F. Jiang *et al*., “Artificial Intelligence in Healthcare: Past, Present and Future,” *Stroke and Vascular Neurology*, BMJ Publishing Group. 2017; 2(4):230–243. doi:10.1136/svn-2017-000101.