Early Stroke Prediction by Machine Learning

S. Janakiraman[1] & N. Sathish[2]

[1]Assistant Professor, Department of Master of Computer Applications
Er. Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India
[2]II MCA, Department of Master of Computer Applications
Er. Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India
DOI: *doi.org/10.34293/iejcsa.v4i2.101*

Abstract*- Stroke is a major global health concern and one of the leading causes of mortality and long-term disability. Early prediction of stroke risk is essential for timely medical intervention and prevention of severe complications. This paper presents a machine learning-based predictive framework designed to assess the likelihood of stroke occurrence using patient health records. The proposed system utilizes key clinical and lifestyle parameters such as age, hypertension, heart disease, glucose level, body mass index (BMI), and smoking status. Several classification algorithms, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), are implemented and evaluated for performance comparison. The proposed model is evaluated using a healthcare stroke dataset containing demographic and clinical attributes of patients. Experimental analysis shows that the Random Forest classifier achieved the highest prediction accuracy of 96.2% compared to other machine learning algorithms. The system incorporates data preprocessing techniques, feature engineering, and model evaluation using standard performance metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that ensemble methods, particularly Random Forest, outperform other models in terms of predictive accuracy and robustness. The proposed framework aims to support healthcare professionals in early diagnosis, risk assessment, and preventive care strategies.*

Keywords: *Machine Learning, Stroke Prediction, Healthcare, Classification Algorithms, Data Analysis, Early Diagnosis*

INTRODUCTION

Stroke is a life-threatening medical condition that occurs when the blood supply to the brain is interrupted, either due to a blockage (ischemic stroke) or rupture (hemorrhagic stroke) of blood vessels. This disruption leads to insufficient oxygen supply to brain tissues, resulting in irreversible damage or death if not treated promptly. According to global health statistics, millions of people suffer from stroke annually, with a significant percentage experiencing permanent disability. Early identification of individuals at high risk is crucial for reducing mortality rates and improving quality of life.

Traditional stroke diagnosis relies on clinical examination and imaging techniques such as CT scans and MRIs. However, these methods are often reactive rather than predictive. With the advancement of Machine Learning (ML), predictive models can now analyze large volumes of patient data and detect hidden patterns that may indicate stroke risk before symptoms appear.

This research focuses on developing a machine learning-based system that predicts stroke risk efficiently and accurately, thereby assisting healthcare providers in making informed decisions.

Despite advancements in healthcare analytics, many existing stroke prediction systems suffer from low prediction accuracy, poor scalability, and lack of real-time implementation. Therefore, there is a strong need for an intelligent predictive framework capable of providing reliable early stroke risk assessment using machine learning techniques.

Importance of Stroke Prediction

Early stroke prediction offers several benefits:

- **Preventive Healthcare:** Enables lifestyle and medical interventions before stroke occurrence
- **Reduced Mortality:** Early detection significantly lowers fatality rates
- **Clinical Decision Support:** Assists doctors with data-driven insights
- **Cost Efficiency:** Reduces long-term healthcare costs associated with stroke treatment
- **Improved Patient Outcomes:** Enhances recovery chances and reduces complications

Problem Statement

Existing healthcare systems primarily depend on manual diagnosis and clinical expertise, which may not always provide early warnings of stroke risk. These methods are often time-consuming and prone to subjective errors. Additionally, handling large-scale patient data manually is inefficient. Therefore, there is a need for an automated, intelligent system that can:

- Analyze multiple risk factors simultaneously
- Provide accurate and real-time predictions
- Assist healthcare professionals in early diagnosis and prevention

EXISTING SYSTEM

In traditional healthcare settings, stroke risk is assessed through:

- Patient medical history
- Physical examinations
- Laboratory test results
- Imaging techniques

While these approaches are effective for diagnosis, they lack predictive capabilities and are often reactive.

Limitations of Existing System

- Lack of predictive analytics
- Time-consuming manual processes
- High dependency on expert knowledge
- Limited scalability for large datasets
- Increased chances of human error
- Absence of real-time monitoring systems

PROPOSED SYSTEM

The proposed system introduces a machine learning-based approach to predict stroke risk using structured patient data. The system automates data analysis and generates predictions with high accuracy.

Key Features

- Automated data processing and prediction
- Integration of multiple ML algorithms
- Real-time prediction capability
- Scalable and adaptable system

Advantages

- Improved prediction accuracy
- Faster processing time
- Reduced human intervention
- Early detection of high-risk individuals
- Enhanced decision-making support for clinicians

RELATED WORK

Several studies have explored the application of machine learning in healthcare, particularly for disease prediction. Logistic Regression has been widely used for binary classification problems due to its simplicity and interpretability.

Decision Trees provide intuitive decision-making structures, while Random Forest improves prediction accuracy by combining multiple trees and reducing overfitting. Support Vector Machines are effective for high-dimensional data and complex classification boundaries.

Recent advancements include deep learning techniques and hybrid models that combine statistical and machine learning approaches. However, many existing systems lack real-time applicability and comprehensive feature integration.

METHODOLOGY

Dataset Description

The dataset used for this research is collected from publicly available healthcare repositories such as Kaggle/UCI Machine Learning Repository. The dataset contains patient health records with attributes including age, gender, hypertension, heart disease, BMI, glucose level, smoking status, and stroke outcome.

Data Collection

The dataset used in this study includes both demographic and clinical features:

- Age
- Gender
- Hypertension status
- Heart disease history

- Average glucose level
- Body Mass Index (BMI)
- Smoking status

These features are selected based on medical research indicating their strong correlation with stroke risk.

Data Preprocessing

Data preprocessing ensures data quality and improves model performance:

- **Handling Missing Values:** Imputation using mean/median for numerical data
- **Normalization:** Scaling features to a standard range
- **Encoding:** Converting categorical variables using label encoding or one-hot encoding
- **Outlier Removal:** Eliminating extreme values that may affect model accuracy

Feature Engineering

Feature engineering enhances model performance by selecting and transforming relevant attributes. Correlation analysis is performed to identify the most influential features contributing to stroke risk.

Model Building

The following machine learning models are implemented:

Logistic Regression

A probabilistic model that estimates the likelihood of stroke occurrence

$P(Y=1)=1/(1+e^{-(\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn)})$

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Decision Tree

A rule-based model that splits data into branches based on feature values

Random Forest

An ensemble learning method combining multiple decision trees to improve accuracy and reduce overfitting

Support Vector Machine (SVM):

A model that finds the optimal hyperplane for classification

Training and Testing

- Dataset is split into training (80%) and testing (20%) sets
- Models are trained using the training dataset
- Predictions are evaluated on the testing dataset
- Models are trained using the training dataset

Evaluation Metrics

The performance of models is evaluated using:

- **Accuracy:** Proportion of correct predictions

- **Precision:** Accuracy of positive predictions
- **Recall:** Ability to detect true positives
- **F1-Score:** Balance between precision and recall

SYSTEM ARCHITECTURE

The system architecture consists of the following components:

1. **Data Input Layer** – Collects patient data
2. **Preprocessing Layer** – Cleans and transforms data
3. **Model Layer** – Applies machine learning algorithms
4. **Prediction Layer** – Generates stroke risk prediction
5. **Visualization Layer** – Displays results and insights

SYSTEM IMPLEMENTATION

The system is implemented using Python and its data science libraries:

- **Pandas:** Data manipulation and analysis
- **NumPy:** Numerical computations
- **Scikit-learn:** Machine learning models and evaluation
- **Matplotlib/Seaborn:** Data visualization

Development is carried out using Jupyter Notebook and Visual Studio Code, providing an interactive environment for experimentation and analysis.

RESULTS AND ANALYSIS

Among all classification models, Random Forest achieved superior performance due to its ensemble learning capability and reduced overfitting characteristics. The model demonstrated better generalization and classification accuracy compared to individual classifiers. The experimental results show:

- Random Forest achieved the highest accuracy due to ensemble learning
- Logistic Regression performed well with interpretable results
- SVM provided strong classification but required parameter tuning
- Decision Tree was easy to interpret but prone to overfitting

Comparative analysis indicates that Random Forest offers the best balance between accuracy, robustness, and generalization.

Algorithm| Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 89.4%| 88%| 87%| 87.5%|
| Decision Tree| 91.2%| 90%| 89%| 89.4%|
| SVM | 93.5%| 92%| 91%| 91.5%|
| Random Forest| 96.2%| 95%| 94%| 94.5%|

CONCLUSION

This study demonstrates the effectiveness of machine learning techniques in predicting stroke risk. The proposed system successfully identifies high-risk individuals based on health data, enabling early intervention and preventive measures.

The integration of machine learning into healthcare systems can significantly improve diagnostic accuracy, reduce mortality rates, and enhance patient care.

Future Enhancement

Future work can focus on:

- Implementing deep learning models such as Artificial Neural Networks
- Integrating IoT-based wearable devices for real-time monitoring
- Developing mobile and web-based healthcare applications
- Incorporating real-time data streaming and cloud computing
- Expanding datasets for improved model generalization

REFERENCES

1. Yoo, H. Y. *et al.* 2026. 'Machine learning for predicting stroke risk stratification using multiomics data: Systematic review', *Journal of Medical Internet Research*, vol. 28.
2. Khanum, U. *et al.* 2026. 'Predictive performance of machine learning models in acute ischemic stroke: A systematic review and meta-analysis', *Frontiers in Neurology*.
3. Soladoye, A. A. *et al.* 2025. 'Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps', *International Journal of Medical Informatics*, vol. 203.
4. Tang, X. *et al.* 2025. 'Explainable machine learning for stroke risk prediction: A comparative study with SHAP-based interpretation', *Frontiers in Neurology*, vol. 16.
5. Desor, K. *et al.* 2026. 'Machine learning in the diagnosis and prognosis of transient ischaemic attack: A systematic review', *BMC Neurology*, vol. 26, no. 299.
6. Byna, A. 2024. 'Machine learning-based stroke prediction: A critical analysis', *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 5, pp. 1609-1618.
7. Dev, S. *et al.* 2022. 'A predictive analytics approach for stroke prediction using machine learning and neural networks', *arXiv*.
8. Lu, J. *et al.* 2022. 'Performance of multilabel machine learning models and risk stratification schemas for predicting stroke and bleeding risk in patients with non-valvular atrial fibrillation', *arXiv*.
9. Kapoor, A. K. *et al.* 2026. 'Large language models predict functional outcomes after acute ischemic stroke', *arXiv*.
10. Tashkova. *et al.* 2025. 'Comparative analysis of stroke prediction models using machine learning', *arXiv*.
11. World Health Organization. 2024. 'Stroke, cerebrovascular accident', *World Health Organization*.
12. Pedregosa, F. *et al.* 2023. 'Scikit-learn: Machine learning in python', *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830.
13. Chollet, F. 2023. *Deep learning with python*. Manning Publications.
14. Géron, A. 2023. *Hands-on machine learning with scikit-learn, keras, and tensorflow*. O'Reilly Media.
15. Bishop, C. M. 2022. *Pattern recognition and machine learning*. Springer.