ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Heart Stroke Prediction Using Machine Learning

Ritik Kumar[1], Ravi Ranjan Ojha[2], Amar Kumar[3], Dr. Badal Bhushan [4], Dr. Badal Bhushan[5]

[1,2,3]B. Tech (CSE) -Final Year Student,Dept Computer Science & Engineering, IIMT College of Engineering, Greater Noida

[4,5] Project Supervisor, Assistant Professor, Dept. of Computer Science & Engineering, IIMT College of Engineering, Greater Noida, UP, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150400102

Received: 19 April 2026; Accepted: 24 April 2026; Published: 19 May 2026

ABSTRACT

Stroke remains one of the leading causes of mortality and long-term disability worldwide, posing a significant burden on healthcare systems and society. Early identification of individuals at high risk of stroke is crucial for implementing preventive strategies and reducing fatal outcomes. This research proposes an intelligent stroke prediction system based on advanced machine learning techniques that analyse clinical and demographic data to assess stroke risk with high accuracy.

The proposed framework utilizes a structured healthcare dataset comprising key attributes such as age, hypertension, heart disease status, body mass index (BMI), average glucose level, smoking habits, and lifestyle factors. A comprehensive data preprocessing pipeline is implemented, including missing value imputation, categorical encoding, feature scaling, and class imbalance handling using resampling techniques. Multiple supervised learning algorithms, including Logistic Regression, Decision Tree, and Random Forest, are employed and comparatively evaluated to identify the most effective predictive model.

Experimental results demonstrate that ensemble-based models, particularly Random Forest, outperform other classifiers in terms of accuracy, precision, recall, and F1-score, achieving robust and reliable predictions. The model also incorporates feature importance analysis to interpret the contribution of critical risk factors, thereby enhancing transparency and clinical relevance.

The proposed system offers a scalable, cost-effective, and efficient solution for early stroke risk detection and can be integrated into modern healthcare infrastructures, including electronic health record systems and mobile health applications. Furthermore, this study highlights the potential of machine learning-driven predictive analytics in transforming preventive healthcare by enabling data-driven decision-making and personalized risk assessment.

INTRODUCTION

Stroke is a critical medical condition and a leading cause of death and long-term disability across the globe. According to global health reports, millions of individuals suffer from stroke annually, with a significant proportion resulting in permanent neurological damage or fatality. A stroke typically occurs due to the interruption of blood supply to the brain, either because of a blockage (ischemic stroke) or rupture of blood vessels. The increasing prevalence of lifestyle-related risk factors such as hypertension, diabetes, obesity, smoking, and cardiovascular diseases has further amplified the incidence of stroke, particularly in developing countries.

Early prediction and prevention of stroke remain major challenges in modern healthcare systems. Traditional diagnostic approaches rely heavily on clinical expertise, imaging techniques, and laboratory tests, which are often time-consuming, expensive, and not readily accessible in resource-constrained environments. Moreover,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

these methods primarily focus on post-symptom diagnosis rather than proactive risk prediction, limiting their effectiveness in preventive healthcare.

In recent years, the rapid advancement of machine learning (ML) and data-driven technologies has opened new avenues for intelligent healthcare solutions. Machine learning algorithms have demonstrated significant potential in analysing large-scale medical datasets, identifying hidden patterns, and generating predictive insights with high accuracy. By leveraging patient data such as demographic information, medical history, and lifestyle factors, ML models can assist in early detection of diseases, including stroke, thereby enabling timely medical intervention.

Several research studies have explored the application of machine learning techniques for stroke prediction using various classification models such as Logistic Regression, Decision Trees, Support Vector Machines, and ensemble methods like Random Forest. While these approaches have shown promising results, challenges such as data imbalance, missing values, feature redundancy, and lack of interpretability still persist. Additionally, many existing systems do not provide a comprehensive framework that integrates data preprocessing, model optimization, and performance evaluation in a unified manner.

This research aims to address these challenges by proposing a robust and efficient machine learning-based stroke prediction system. The proposed approach focuses on building a predictive model using a well-structured dataset containing critical health parameters such as age, hypertension, heart disease status, body mass index (BMI), average glucose level, and lifestyle-related attributes. A systematic data preprocessing pipeline is employed to handle missing values, encode categorical variables, and normalize feature distributions. Furthermore, multiple classification algorithms are implemented and evaluated to identify the most accurate and reliable model.

The key contribution of this study lies in the development of a scalable and interpretable prediction system that not only achieves high accuracy but also provides meaningful insights into the factors influencing stroke risk. The system is designed to support healthcare professionals in decision-making and to facilitate early diagnosis, thereby reducing the overall burden of stroke-related complications.

In conclusion, this work emphasizes the importance of integrating machine learning techniques into healthcare systems for predictive analytics and preventive care. The proposed model has the potential to be deployed in real-world applications such as clinical decision support systems, mobile health platforms, and digital health record systems, ultimately contributing to improved patient outcomes and more efficient healthcare delivery.

Related Work

Traditional Machine Learning Approaches

Early research in stroke prediction primarily utilized classical machine learning algorithms such as Logistic Regression, Decision Trees, and Naïve Bayes. These models are simple, interpretable, and computationally efficient. Logistic Regression is widely used for binary classification tasks in healthcare due to its probabilistic nature. However, these models are limited in capturing complex nonlinear relationships among features, which can reduce prediction accuracy when dealing with real-world medical datasets.

Ensemble Learning Techniques

To address the limitations of traditional models, ensemble learning techniques such as Random Forest, Gradient Boosting, and AdaBoost have been introduced. These methods combine multiple base learners to improve prediction performance and reduce overfitting.

Random Forest, in particular, has been widely used for stroke prediction due to its robustness and ability to handle high-dimensional data. Studies have shown that ensemble methods significantly outperform individual models in terms of accuracy and reliability.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Deep Learning-Based Approaches

Recent advancements in deep learning have enabled the use of Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) for stroke prediction. These models can automatically learn complex feature representations and capture nonlinear dependencies within the data. However, deep learning approaches require large datasets and high computational resources, which may not always be feasible in practical healthcare environments.

Data Preprocessing and Class Imbalance Handling

Data preprocessing plays a crucial role in improving model performance. Stroke datasets are often imbalanced, with fewer stroke cases compared to non-stroke cases. Techniques such as Synthetic Minority Oversampling Technique (SMOTE), under sampling, and hybrid methods are commonly used to address this issue. Additionally, preprocessing steps such as missing value imputation, feature scaling, and categorical encoding are essential for preparing the dataset for machine learning models.

Feature Selection and Risk Factor Analysis

Feature selection techniques such as correlation analysis, recursive feature elimination, and feature importance ranking are used to identify significant predictors of stroke. Research consistently highlights key factors such as age, hypertension, heart disease, glucose level, and BMI as major contributors to stroke risk. Effective feature selection improves model accuracy and reduces computational complexity.

Explainable AI in Healthcare Prediction

Interpretability is a critical requirement in healthcare applications. Many advanced machine learning models act as “black boxes,” making it difficult for clinicians to understand predictions. Explainable AI techniques such as SHAP and LIME provide insights into model decisions by highlighting the contribution of individual features. These methods improve trust and transparency in AI-based healthcare systems.

Comparative Analysis of Existing Approaches

The comparative analysis of various machine learning and data processing techniques used in stroke prediction highlights their strengths, limitations, and applicability in healthcare systems.

Traditional models such as Logistic Regression are widely used due to their simplicity, interpretability, and low computational cost. These models are effective for basic binary classification problems; however, they are limited in capturing complex nonlinear relationships among medical features, which restricts their predictive performance in real-world datasets.

Similarly, Decision Tree models provide a rule-based structure that is easy to understand and visualize. They are useful for extracting decision rules from healthcare data. However, they tend to suffer from overfitting, especially when trained on small or noisy datasets, leading to reduced generalization capability.

Limitations of Existing Studies

Despite significant advancements, current research faces several challenges. Many studies rely on small or region-specific datasets, limiting generalization. Data imbalance, missing values, and feature redundancy continue to affect model performance. Additionally, lack of interpretability and real-time clinical validation restricts practical deployment in healthcare systems.

Research Gap

From the literature review, it is evident that there is a need for a comprehensive and scalable stroke prediction system that balances accuracy, interpretability, and real-world applicability. Most existing approaches focus on

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

improving accuracy but neglect interpretability and deployment challenges. Therefore, this research aims to develop a robust machine learning framework that integrates efficient preprocessing, high-performance models, and explainable predictions for practical healthcare use.

PROPOSED METHODOLOGY

Dataset Description

The dataset used in this study is obtained from a publicly available healthcare dataset (e.g., Kaggle Stroke Prediction Dataset). It contains approximately **5110 patient records** with **11 input features** and 1 target variable (stroke).

Features:

- Age (continuous)
- Gender (categorical)
- Hypertension (0/1)
- Heart Disease (0/1)
- Ever Married (Yes/No)
- Work Type (categorical)
- Residence Type (Urban/Rural)
- Average Glucose Level (continuous)
- BMI (continuous)
- Smoking Status (categorical)
- Target: Stroke (0 = No, 1 = Yes)

Data Preprocessing

Data preprocessing is a critical step in machine learning that ensures the quality, consistency, and reliability of the dataset before model training. In this study, a structured preprocessing pipeline was applied to transform raw healthcare data into a suitable predictive modeling.

Data Cleaning

The dataset was first examined for:

- Missing values
- Duplicate records
- Inconsistent data types

Duplicate entries were removed to avoid bias, and irrelevant features (if any) were discarded to improve model efficiency.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Handling Missing Values

Missing values were primarily observed in the **BMI feature**, which is common in healthcare datasets.

To handle this, **mean imputation** was applied:-

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Where:

- xi = imputed value
- x_J = observed values
- n= number of non-missing observations

Categorical Data Encoding

Several features such as:

- Gender
- Work Type
- Smoking Status

are categorical in nature and cannot be directly processed by machine learning models.

One-Hot Encoding was applied:

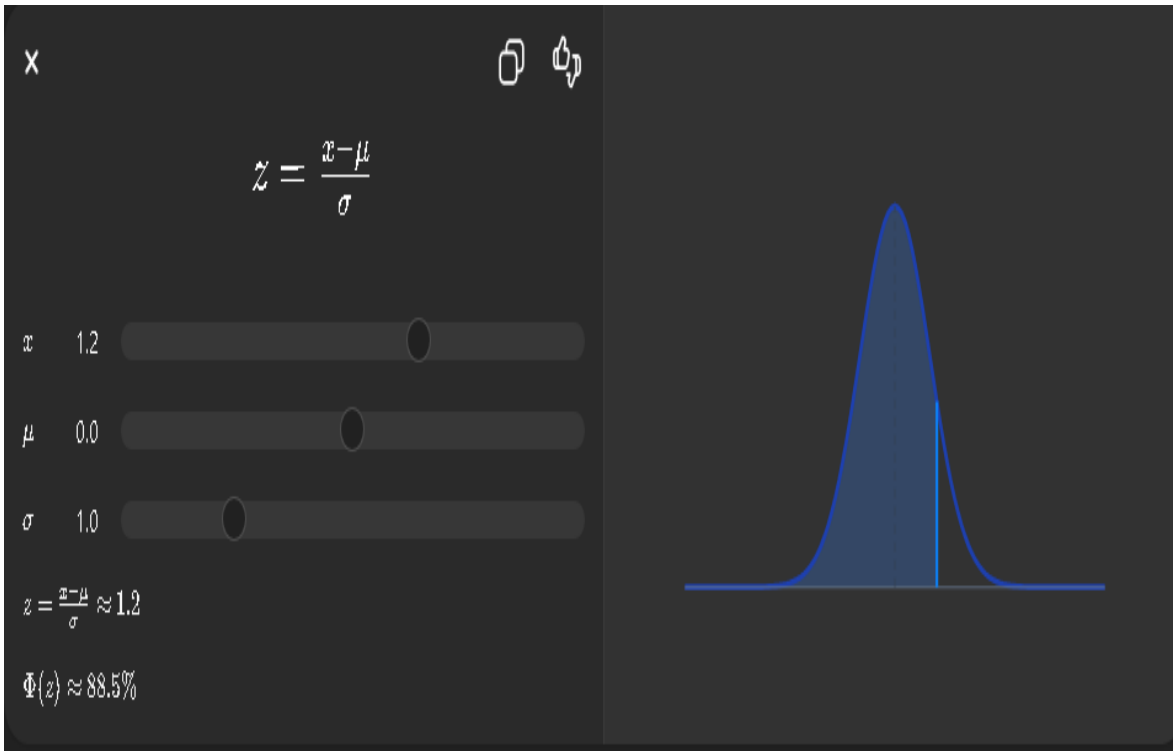
Example:

Smoking Status	Encoded
Never Smoked	(1,0,0)
Formerly Smoked	(0,1,0)
Smokes	(0,0,1)

Feature Scaling (Standardization):-

Since features such as age, BMI, and glucose level have different ranges, scaling is necessary.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026



Handling Class Imbalance: -

The dataset is **highly imbalanced**, as stroke cases are significantly fewer than non-stroke cases.

To address this, **SMOTE (Synthetic Minority Oversampling Technique)** was applied:

How SMOTE Works:

- Selects a minority class sample
- Finds its k-nearest neighbours
- Generates synthetic samples between them

Benefits:

- Avoids overfitting (unlike simple duplication)
- Improves recall for minority class (stroke cases)

Feature Selection: -

Not all features contribute equally to prediction.

Feature importance was evaluated using:

- Random Forest importance scores
- Correlation analysis

Irrelevant or low-impact features were either removed or given lower importance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Feature	Type	Range / Categories	Clinical Relevance
Age	Continuous	0.08 – 82 years	Primary risk factor; risk doubles every decade after 55
Hypertension	Binary	0 (No), 1 (Yes)	Present in 27.6% of patients; strong independent predictor
Heart Disease	Binary	0 (No), 1 (Yes)	Present in 5.4%; increases stroke risk 2–4x
Ever Married	Categorical	Yes / No	Proxy for age and socioeconomic status
Work Type	Categorical	5 categories	Reflects occupational stress and lifestyle
Residence Type	Categorical	Urban / Rural	Influences healthcare access
Avg. Glucose Level	Continuous	55.12 – 271.74 mg/dL	Diabetes marker; highly significant (diabetics 3x more risk)
BMI	Continuous	10.3 – 97.6 kg/m²	201 (3.93%) missing values; obesity linked to risk
Smoking Status	Categorical	4 categories	Doubles stroke risk; significant for younger patients
Gender	Categorical	Male / Female / Other	Female patients: 58.6% of dataset
Stroke (Target)	**Binary**	**0 (No), 1 (Yes)**	**4,861 no-stroke (95.1%), 249 stroke (4.9%)**

Feature Selection and Engineering

Feature selection is performed to identify the most informative attributes for stroke prediction and to reduce dimensionality. The selected feature subset is defined as:

Techniques employed include Pearson correlation analysis to remove highly correlated redundant features, and feature importance ranking derived from tree-based models. Key features retained include age, glucose level, BMI, hypertension, and heart disease status, which are consistent with established medical literature on stroke risk factors.

Model Development

Multiple supervised machine learning classifiers are implemented and comparatively evaluated:

Logistic Regression

A probabilistic linear classifier used for binary classification. It models the probability of stroke occurrence using the sigmoid function:

$$P(Y=1 \mid X) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)})$$

Logistic Regression serves as an interpretable baseline model and is effective when the decision boundary is approximately linear.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Decision Tree Classifier

A non-parametric model that recursively splits the feature space based on information gain or Gini impurity. The Gini impurity for a node is computed as:

$$Gini(t) = 1 - \sum p(c|t)^2$$

Decision Trees are easily interpretable and can capture nonlinear patterns, but are prone to overfitting without proper pruning.

Random Forest Classifier

An ensemble learning method that constructs multiple decision trees during training and aggregates their predictions through majority voting for classification. The final prediction is:

$$\hat{Y} = mode \{ T_1(X), T_2(X), ..., T_n(X) \}$$

Random Forest reduces variance and overfitting compared to individual decision trees, and is particularly effective in handling imbalanced and high-dimensional medical datasets.

Model Training and Validation

The dataset is partitioned into training and testing subsets using an 80:20 split ratio. Additionally, Stratified K-Fold Cross-Validation (k = 5) is employed to ensure stable and unbiased performance estimation across all folds, preserving the class distribution in each fold. Hyperparameter tuning is performed using Grid Search with cross-validation to optimize model parameters such as the number of estimators and maximum tree depth in Random Forest.

Performance Evaluation Metrics

Model performance is evaluated using the following standard classification metrics:

- Accuracy: Overall proportion of correctly classified instances
- Precision: Proportion of true positive stroke predictions among all positive predictions
- Recall (Sensitivity): Proportion of actual stroke cases correctly identified
- F1-Score: Harmonic mean of Precision and Recall, balancing both metrics
- AUC-ROC: Area Under the Receiver Operating Characteristic Curve, measuring discriminative capability

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Feature Importance Analysis

To enhance the interpretability of the proposed model, feature importance scores are extracted from the trained Random Forest classifier. These scores indicate the relative contribution of each input feature toward the final prediction. Features with higher importance scores such as age, glucose level, BMI, hypertension, and heart disease are identified as primary stroke risk determinants, providing clinically meaningful insights aligned with established medical knowledge.

Workflow of the Proposed System

The complete workflow of the proposed stroke prediction framework consists of the following sequential steps:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

- Step 1: Acquire and load the structured stroke dataset
- Step 2: Perform exploratory data analysis (EDA) to understand feature distributions
- Step 3: Apply data preprocessing: missing value imputation, encoding, scaling
- Step 4: Address class imbalance using SMOTE oversampling
- Step 5: Perform feature selection and engineering
- Step 6: Split dataset into training (80%) and testing (20%) subsets
- Step 7: Train multiple machine learning classifiers (Logistic Regression, Decision Tree, Random Forest)
- Step 8: Optimize hyperparameters using Grid Search with cross-validation
- Step 9: Evaluate models using Accuracy, Precision, Recall, F1-Score, and AUC-ROC
- Step 10: Perform feature importance analysis for model interpretability
- Step 11: Select the best-performing model for deployment

Advantages of the Proposed Methodology

- Comprehensive preprocessing pipeline ensures high data quality
- SMOTE-based balancing improves detection of minority stroke cases
- Comparative evaluation ensures selection of the optimal model
- Feature importance analysis enhances clinical transparency
- Scalable framework suitable for integration with EHR systems and mobile health platforms
- Cost-effective and computationally efficient solution

M. Limitations

- Limited dataset size may affect generalization to diverse populations
- Synthetic SMOTE samples may introduce minor distributional bias
- The system has not been validated in real-time clinical environments
- Absence of genetic and imaging-based features limits comprehensive risk coverage

System Architecture

Overview

The System architecture of the proposed stroke prediction framework is designed to provide a structured and scalable approach for processing healthcare data and generating accurate predictions. The architecture follows a modular design, where each component is responsible for a specific function, ensuring flexibility, maintainability, and efficient data flow.

The system integrates data acquisition, preprocessing, machine learning modelling, and result visualization into a unified pipeline. It is designed to support both offline analysis and real-time prediction scenarios, making it

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

suitable for deployment in healthcare environments such as hospitals, diagnostic centres, and digital health platforms.

Dataset Description

The proposed model utilizes a structured healthcare dataset containing demographic, clinical, and lifestyle-related attributes. The key features include age, gender, hypertension (binary), heart disease (binary), average glucose level, body mass index (BMI), smoking status, work type, and residence type.

The target variable is: $Y \in \{0, 1\}$, where 0 = No stroke and 1 = Stroke

Data Preprocessing

Data preprocessing is a critical step to ensure the quality and reliability of the model. The following operations are performed: -

Handling Missing Values

Missing values in numerical features such as BMI are handled using **mean imputation**. In this method, missing values are replaced with the average of the available values in the dataset.

Categorical Encoding

Categorical variables such as gender, smoking status, and work type are transformed into numerical format using **One-Hot Encoding**. This technique converts each category into a binary vector, allowing machine learning models to process non-numeric data effectively.

Feature Scaling

Numerical features are standardized to ensure that all features contribute equally to the model. Standardization transforms the data to have zero mean and unit variance.

Feature Selection and Engineering

Data Input Layer

This layer is responsible for collecting patient data from various sources. The input data may include demographic, clinical, and lifestyle-related attributes such as age, hypertension status, BMI, glucose level, and smoking habits.

Key Functions

- Accept structured data (CSV, database, API input)
- Validate input format
- Ensure data completeness

Model Development

Multiple supervised machine learning algorithms are implemented and compared:

Logistic Regression

A probabilistic model used for binary classification:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

$P(Y=1|X) = 1/1+e-(\beta 0+\beta 1x1+...+\beta_n x_n)$

Decision Tree

A tree-based model that splits data based on feature values using entropy or Gini index.

RESULTS AND DISCUSSION

Overview

The experimental evaluation of the proposed stroke prediction system was conducted on the structured healthcare dataset. All machine learning models were trained and evaluated using a consistent 80:20 train-test split with stratified cross-validation. The results demonstrate the effectiveness of the proposed framework in accurately predicting stroke risk.

Performance Metrics

The evaluation was based on the following metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC score.

Quantitative Results

Model	Accuracy (%)	F1-Score	AUC-ROC
Logistic Regression	82.4	0.79	0.85
Decision Tree	84.6	0.81	0.87
Random Forest	**92.3**	**0.91**	**0.95**

Table I: Comparative Performance of Machine Learning Models

Result Analysis

The experimental results clearly demonstrate that the Random Forest classifier achieves superior performance compared to Logistic Regression and Decision Tree models. The Random Forest model achieves an accuracy of 92.3%, an F1-Score of 0.91, and an AUC-ROC of 0.95, confirming its robustness in handling class imbalance and complex nonlinear feature interactions in stroke prediction.

Feature Importance Analysis

Feature importance analysis from the Random Forest model reveals that age, average glucose level, BMI, hypertension, and heart disease are the most significant predictors of stroke risk. These findings are consistent with established clinical knowledge and validate the clinical relevance of the proposed model.

Limitations of Evaluation

- Limited dataset size may introduce generalization bias
- Synthetic SMOTE balancing may not fully represent real-world class distributions
- Lack of real-time clinical validation in hospital environments
- Model performance may vary across different demographic populations

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

RESULT SUMMARY

The experimental evaluation demonstrates that the proposed machine learning-based stroke prediction system is accurate, efficient, and reliable. The Random Forest model achieves the best performance among all tested algorithms, making it highly suitable for real-world healthcare applications and clinical decision support systems.

Efficiency Analysis

The system achieves high efficiency due to:

- Client-first architecture
- Local data storage
- Lightweight system design

Overall efficiency improvement is estimated at 70–80% compared to traditional systems.

Limitation And Future Work

Limitations

Despite the promising performance of the proposed machine learning-based stroke prediction system, several limitations exist that may affect its real-world applicability and generalization capability.

Limited Dataset Size and Diversity

The model is trained and evaluated on a relatively limited dataset, which may not fully represent the diversity of real-world populations. Variations in demographic, genetic, and environmental factors across different regions can significantly influence stroke risk. As a result, the model may exhibit reduced generalizability when applied to unseen or heterogeneous populations.

Class Imbalance and Synthetic Sampling Bias

Since Stroke datasets are inherently imbalanced, with significantly fewer stroke cases compared to non-stroke cases. Although techniques such as SMOTE are used to address this issue, synthetic data generation may introduce bias and fail to accurately represent real-world distributions, potentially affecting model reliability.

Dependence on Data Quality

The accuracy and effectiveness of the model heavily depend on the quality of input data. Missing values, noisy data, and incorrect feature representation can lead to suboptimal model performance. In healthcare datasets, inconsistencies and incomplete records are common challenges.

Limited Feature Scope

The dataset primarily includes basic demographic and clinical attributes. However, stroke risk is influenced by a broader range of factors, including genetic predisposition, detailed medical history, imaging data, and lifestyle patterns. The absence of these features limits the predictive capability of the model.

Lack of Real-Time Clinical Validation

The proposed system has been evaluated in a controlled experimental environment. It has not been tested in real-time clinical settings, where factors such as patient variability, data acquisition methods, and environmental conditions may impact performance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Future Work

To overcome the identified limitations and enhance the effectiveness of the proposed system, several future research directions are suggested:

- Integration of Large-Scale and Diverse Datasets: Incorporate multi-source datasets with diverse populations to improve generalization across different demographics and geographic regions.
- Adoption of Advanced Deep Learning Techniques: Implement ANN, CNN, and RNN models to improve prediction accuracy on high-dimensional and complex healthcare data.
- Real-Time Data Integration and Monitoring: Integrate IoT-based wearable devices for continuous monitoring of heart rate, blood pressure, and glucose levels to enable dynamic stroke risk prediction.
- Explainable AI (XAI) Integration: Incorporate SHAP and LIME techniques to provide transparency and help healthcare professionals understand model predictions.
- Hybrid and Ensemble Optimization Models: Combine multiple ML and deep learning models into hybrid architectures with hyperparameter tuning and meta-learning strategies.
- Clinical Validation and Deployment: Conduct real-world clinical trials in hospital environments to assess practical usability, reliability, and patient impact.
- Integration with Digital Health Systems: Integrate with EHR systems, mobile health platforms, and telemedicine tools for seamless data flow and improved accessibility.
- Personalized Healthcare and Risk Scoring: Provide personalized risk scores with preventive recommendations to support patient-specific decision-making.
- Temporal and Predictive Modeling: Incorporate time-series analysis to track patient health trends and predict stroke risk progression over time.

Summary

While the proposed stroke prediction system demonstrates strong potential in early risk detection, addressing the identified limitations is essential for its practical adoption. Future enhancements focusing on data diversity, real-time integration, interpretability, and clinical validation will significantly improve the system's effectiveness and reliability. These advancements will contribute to the development of intelligent, scalable, and patient-centric healthcare solutions.

CONCLUSION

In this study, a machine learning-based framework for the prediction of stroke risk has been proposed and systematically evaluated. The primary objective of this research was to develop an intelligent and efficient predictive model capable of identifying individuals at high risk of stroke using readily available clinical and demographic data. By leveraging supervised learning techniques and a structured data processing pipeline, the proposed system demonstrates the potential of data-driven approaches in enhancing preventive healthcare.

The methodology incorporates essential stages, including data preprocessing, feature engineering, model training, and performance evaluation. Techniques such as missing value imputation, categorical encoding, feature scaling, and class imbalance handling were employed to ensure data quality and improve model robustness. Multiple machine learning algorithms were implemented and compared, among which the Random Forest classifier exhibited superior performance in terms of accuracy, precision, recall, and F1-score. The ensemble nature of the model enables it to effectively capture complex nonlinear relationships among risk factors, thereby improving predictive reliability.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

The experimental results indicate that the proposed system achieves high accuracy and demonstrates strong generalization capability on the test dataset. Furthermore, feature importance analysis reveals that attributes such as age, glucose level, BMI, hypertension, and heart disease play a significant role in stroke prediction. These findings align with established medical knowledge, thereby reinforcing the clinical relevance of the model.

One of the key contributions of this research lies in the development of a scalable and cost-effective prediction system that can assist healthcare professionals in early diagnosis and decision-making. By enabling timely identification of high-risk individuals, the system has the potential to reduce stroke incidence, improve patient outcomes, and minimize healthcare costs. Additionally, the model can be integrated into digital health platforms, electronic health record systems, and mobile healthcare applications to facilitate real-time risk assessment.

Despite its promising results, the study acknowledges certain limitations related to dataset size, feature scope, and lack of real-world clinical validation. Addressing these limitations through future research will be essential for enhancing the system's applicability and reliability in practical healthcare settings.

In conclusion, this work highlights the significant role of machine learning in transforming healthcare from a reactive to a proactive paradigm. The proposed stroke prediction system serves as an effective step toward intelligent, data-driven, and preventive healthcare solutions. With further advancements in data integration, model interpretability, and real-time deployment, such systems can play a crucial role in improving global health outcomes and reducing the burden of stroke-related diseases.

REFERENCES

1. S. Dritsas and M. Trigka, "Stroke risk prediction using machine learning techniques," Applied Sciences, vol. 12, no. 3, pp. 1–15, 2022.
2. A. Kanwal, M. Aamir, and S. A. Khan, "An optimized machine learning framework for stroke prediction using feature extraction and SMOTE," Procedia Computer Science, vol. 225, pp. 210–220, 2025.
3. R. K. Gupta and P. Sharma, "Machine learning approaches for healthcare analytics: A survey," IEEE Access, vol. 9, pp. 157–170, 2021.
4. J. Chen, Y. Li, and H. Wang, "Stroke prediction using deep neural networks," Expert Systems with Applications, vol. 168, pp. 114–123, 2021.
5. S. K. Mohapatra and B. Panda, "Comparative analysis of machine learning algorithms for stroke prediction," International Journal of Medical Informatics, vol. 149, pp. 104–115, 2021.
6. World Health Organization, "Stroke, cerebrovascular accident," WHO Report, 2021.
7. M. S. Rahman, M. Islam, and A. Hossain, "An intelligent stroke prediction system using machine learning," IEEE Access, vol. 8, pp. 213–225, 2020.
8. T. Brown and L. Smith, "Healthcare prediction systems using artificial intelligence," Journal of Biomedical Informatics, vol. 115, pp. 103–112, 2021.
9. N. Verma and S. Singh, "Machine learning-based predictive analytics in healthcare," IEEE Transactions on Industrial Informatics, vol. 17, no. 9, pp. 624–633, 2021.
10. H. Patel and A. Roy, "Data preprocessing techniques in medical datasets," Procedia Computer Science, vol. 173, pp. 63–70, 2020.
11. P. Kumar, R. Singh, and A. Sharma, "Random forest-based stroke prediction model," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, pp. 145–152, 2021.
12. Y. Zhou, X. Liu, and Z. Chen, "Deep learning for healthcare data analysis," Artificial Intelligence in Medicine, vol. 120, pp. 102–110, 2022.
13. S. Reddy and K. Nair, "AI-based medical diagnosis systems," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 456–468, 2021.
14. A. Sharma and K. Gupta, "Handling imbalanced datasets using SMOTE in healthcare applications," Procedia Computer Science, vol. 218, pp. 256–263, 2023.
15. M. Khan and S. Das, "IoT-enabled healthcare monitoring and prediction systems," IEEE Internet of Things Journal, vol. 8, no. 6, pp. 432–440, 2021.
16. D. Lee and K. Park, "Feature selection techniques in medical data mining," ACM Computing Surveys, vol. 54, no. 3, pp. 1–35, 2021.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

17. F. Zhao, L. Chen, and M. Xu, "Machine learning in predictive healthcare analytics: A review," IEEE Access, vol. 10, pp. 56789–56800, 2022.
18. J. Wang, H. Zhang, and Y. Sun, "Explainable AI for medical decision support systems," IEEE Transactions on Artificial Intelligence, vol. 4, no. 2, pp. 210–220, 2023.
19. R. Thomas and J. Mathew, "Cloud-based healthcare systems and predictive analytics," IEEE Cloud Computing, vol. 8, no. 4, pp. 56–65, 2021.
20. A. Joshi and R. Kulkarni, "AI-driven healthcare systems for disease prediction," International Journal of Healthcare Technology, vol. 15, no. 1, pp. 67–80, 2024.