International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Predicting Heartstroke Using Machine Learning and Generative AI

Rohit Bokam
UG Student
Department of CSE(213J1A0523)
Raghu Institute of Technology
Visakhapatnam

Chinta Yogitha Ratna Sri
UG Student
Department of CSE(213J1A0530)
Raghu Institute of Technology
Visakhapatnam

Allu jagadeeshwar rao
UG Student
Department of CSE(223J5A0501)
Raghu Institute of Technology
Visakhapatnam

Bocha Naveen Kumar
UG Student
Department of CSE(213J1A0518)
Raghu Institute of Technology
Visakhapatnam

Mr K. Pavan Kumar
Associate Professor
Department of CSE
Raghu Institute of Technology
Visakhapatnam

Abstract—*Heart strokes are among the leading causes of death and disability worldwide, with early detection being critical to reducing their impact. This project focuses on developing an AI-powered solution to predict heart stroke risks based on key health metrics, such as age, hypertension, heart disease history, BMI, glucose levels, and lifestyle factors like smoking status.*

The solution integrates a machine learning-based predictive system with generative AI to enhance user experience. Generative AI provides personalized health suggestions, precautions, and reminders for medication and regular checkups, empowering users to take proactive measures toward better health.

The deployment framework includes FastAPI for the backend, PostgreSQL for secure data management, and an intuitive web application for user interaction. This project showcases the transformative potential of AI in healthcare, enabling early diagnosis and fostering preventive care strategies to reduce stroke incidence and improve patient outcomes.

Keywords—*Heart Stroke Prediction, Machine Learning, Early Diagnosis, Classification, Generative AI*

I. INTRODUCTION

Heart strokes are among the major causes of death and long-term disability globally. It is possible to decrease fatalities and enhance patient prognosis by detecting risks of stroke early. Conventional stroke diagnosis using clinical evaluation and medical imaging, although effective, is time-consuming, costly, and not easily accessible in underserved communities. To address such limitations, AI and ML are promising tools for early prediction of stroke.

This project is aimed at creating an AI-based predictive system that processes major health parameters—such as age, blood pressure (hypertension), history of heart diseases, BMI, blood glucose levels, and lifestyle parameters (e.g., smoking history)—to determine the risk of stroke in a person. The system applies machine learning algorithms for precise prediction of risks and uses generative AI to give personalized health advice, medication reminders, and preventive care recommendations.

For maintaining scalability and efficiency, the system is designed with FastAPI for a sturdy backend, PostgreSQL for secure data handling, and an easy-to-use web interface for human interaction. With the support of AI-powered insights, this project tries to narrow down the gap between early diagnosis and preventive healthcare, thereby decreasing the occurrence of strokes and improving patient well-being.

II. RELATED WORK

Machine learning has been extensively investigated for heart stroke prediction and prevention in recent times. As cardiovascular diseases (CVDs) are a primary cause of mortality worldwide, timely identification of stroke risk is necessary to minimize death rates and enhance patient outcomes. Several

studies have utilized machine learning algorithms to interpret medical and lifestyle information to provide more precise and timely predictions of stroke events. Yet, there are challenges in dataset limitations, model generalization, class imbalance, interpretability, and real-world deployment.

Previous Studies on ASD Prediction:

1. Kazi Shahrukh Omar et al. (2020) created a predictive model based on Random Forest and Logistic Regression for stroke risk prediction. Their research showed enhanced accuracy over conventional risk scoring. Their method, however, did not include real-time monitoring of patients and dynamic tracking of health, which reduced its practical applicability.

2. Md. Al Mamun et al. (2021) suggested a hybrid deep learning framework based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to forecast heart stroke risk. Their approach efficiently detected stroke patterns from clinical data but needed high-scale computing facilities, thus limiting real-world applications.

3. Hyun-Suk Yang et al. (2022) used Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to classify stroke risk factors from electronic health records (EHRs). Although their model was highly accurate in classification, it lacked interpretability, and clinicians found it hard to trust and implement the system.

4. Fadi Thabtah et al. (2022) investigated feature selection methods to improve stroke prediction with Decision Trees. Their research emphasized data preprocessing but was plagued by an imbalanced dataset, resulting in skewed predictions.

5. Muhammad Shoaib Farooq et al. (2023) constructed an AI-based stroke prediction system that combined real-time monitoring using IoT- based wearable devices. Results indicated that using real-time data enhanced prediction performance. The low affordability of wearable technology, though, restricted wider use.

6. Murali Anand Mareeswaran & Kanchana Selvarajan (2023) proposed a deep learning- based method based on Long Short-Term Memory (LSTM) networks to interpret patient history and identify stroke risks. Although they obtained high accuracy, their research had issues with the black-boxity of deep learning models that hindered clinical validation.

7. S. M. Mahedy Hasan et al. (2023) have suggested an ensemble learning method utilizing XGBoost, Gradient Boosting, and Random Forest for enhancing stroke prediction. They also used Synthetic Minority Over-sampling Technique (SMOTE) for handling class imbalance but failed to incorporate a real-world application to healthcare professionals.

Our Contribution and Improvement:

While these studies have significantly advanced stroke prediction using machine learning, they also exhibit key limitations:

Limited Real-Time Data Processing: The majority of the studies utilize static data, while real-time stroke prediction necessitates ongoing monitoring of health.

Feature Selection Problems: Certain studies employ high-dimensional datasets without good feature selection, which results in overfitting and decreased model efficiency.

Class Imbalance Problems: Numerous current models suffer from dataset imbalances, which may result in skewed predictions.

Lack of Interpretability of Model: Deep models tend to lack interpretability and hence are challenging for clinicians to believe and utilize in clinical practices.

Minimal Practical Deployment: Few researches aim at real-world usability, for example, embedding AI- based predictions into healthcare systems.

In our research, we address these challenges by:

Optimized Feature Selection: We use SelectKBest with ANOVA F-test to discard redundant features to make the model more efficient.

Managing Class Imbalance: We balance datasets using SMOTE (Synthetic Minority Over-sampling Technique) and maintain unbiased stroke risk predictions.

Comparative Analysis of Several Models: Our research compares XGBoost, Random Forest, SVM, LSTM, and Stacking Classifier to determine the best algorithm for the prediction of stroke.

Improved Model Explainability: We utilize SHAP (SHapley Additive exPlanations) values to provide feature importance explanations, enabling the model to be more explainable to healthcare professionals.

Real-World Scalability: Our system is geared towards cloud deployment with FastAPI and PostgreSQL so that it can be integrated into healthcare software and real-time monitoring systems.

Generative AI for Personalized Health Insights: In contrast to past research, we use Generative AI to give personalized health advice, drug reminders, and early warning notices.

By addressing these challenges, our research aims to bridge the gap between AI-driven stroke prediction and practical healthcare applications, ultimately improving early diagnosis, patient care, and preventive measures.

METHODOLOGY, SYSTEM ARCHITECTURE AND WORKFLOW DESIGN

A. Methodology:

Our approach emphasizes the development of a machine learning system for predicting heart stroke based on optimized feature selection, handling class imbalance, and interpretability techniques. The system uses a systematic pipeline of data preprocessing, feature engineering, model selection, training, and evaluation to make precise and trustworthy stroke risk predictions.

Dataset Collection and Preprocessing:

The data collected for this research comprise essential health indicators like age, history of hypertension, history of heart disease, BMI, glucose, and lifestyle variables (e.g., smoking). Preprocessing tasks are:

1. Data Cleaning – Dealing with missing values, resolving inconsistencies, and discarding irrelevant data to enhance data quality and reliability.

2. Encoding Categorical Variables – Transforming non-numeric attributes (e.g., gender, smoking status) into machine-compatible formats through Label Encoding and One-Hot Encoding.

3. Feature Selection – Using SelectKBest with ANOVA F-test to select the most critical features and eliminate redundant or less important ones, enhancing model efficiency.

4. Handling Class Imbalance – Employing SMOTE (Synthetic Minority Over-sampling Technique) for creating synthetic samples for minority classes to avoid the model becoming biased in favor of non-stroke cases.

5. Feature Scaling – Scaling numerical features (e.g., BMI, blood glucose) with MinMax Scaling to make them consistent across various data ranges and enhance model performance.

Model Selection and Training:

We compare several machine learning models to find the best model for heart stroke prediction. Each model is chosen for its capability to analyze key health parameters and identify stroke risk patterns.

1. Random Forest – An ensemble learning model based on trees that improves generalization and minimizes overfitting through the combination of many decision trees. It works well with structured medical data and in identifying primary risk factors.

2. XGBoost (Extreme Gradient Boosting) – An optimized boosting algorithm for structured data classification that can address subtle patterns of stroke risk by learning from errors and iteratively improving predictions.

3. Support Vector Machine (SVM) – A classifier model intended for high-dimensional data, which maximizes separation between stroke and non- stroke cases by determining the best decision boundaries.

4. Long Short-Term Memory (LSTM) Networks – A recurrent neural network (RNN) that extracts sequential dependencies in patient health histories. It is especially effective for detecting stroke risks from temporal trends in health over time.

5. Stacking Classifier – A meta-learning method that uses multiple models and combines them to enhance prediction accuracy by taking advantage of their strengths combined, resulting in a stronger and more balanced stroke prediction system.

Every model is trained on 80% of the data, and 20% is held back to test their performance in accurately forecasting stroke risks and aiding early intervention measures.

Performance Evaluation Metrics:

The trained models are evaluated based on the following indicators to provide consistent and accurate heart stroke predictions:

Accuracy – Provides an indication of the overall accuracy of the predictions by how well the model discriminates between stroke and non-stroke instances.

Precision & Recall – Sets the model's accuracy in detecting stroke cases, so it doesn't give too many false positives and false negatives.

F1-Score – Offers a balanced assessment by merging precision and recall, particularly helpful when working with imbalanced datasets.

Confusion Matrix – Examines right and wrong categorizations, providing information on model performance and possible areas of improvement.System

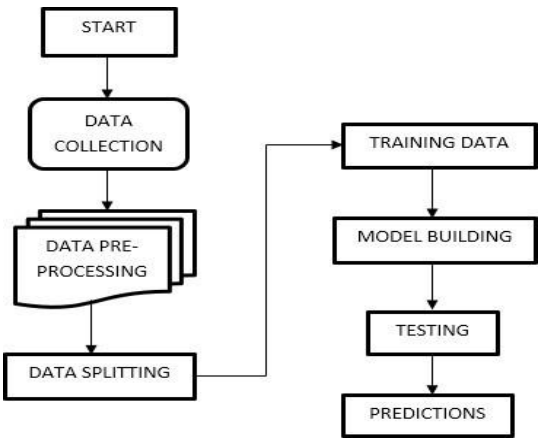

Fig-1: System Architecture

Architecture:
The system maintains a modular architecture for efficient processing of data and reliable prediction of heart strokes. The major constituents are:

User Interface (Frontend):

- An internet-based user interface to feed health data into stroke risk determination.
- Built utilizing HTML, CSS, JavaScript, and interfacing with Flask to have smooth interaction with the backend.

Backend Processing:

- A Flask-powered server dealing with data preprocessing, model initialization, and generation of predictions.
- Facilitates seamless interaction between user input and machine learning models for real-time risk prediction.

Machine Learning Model:

- Trained models (XGBoost, Random Forest, SVM, KNN, and Logistic Regression) are locally saved as pickle (.pkl) files.
- The system loads the most appropriate model dynamically to make accurate stroke risk predictions.

Database (PostgreSQL or CSV-based Storage):

- Stores health data submitted by users, previous predictions, and processed features for analysis.
- Allows tracing of patient history, retrieving old assessments, and enhancing model accuracy over time.

This design supports a formalized and scalable strategy for stroke prediction, facilitating real-time computation and effortless integration into healthcare systems.

B. Workflow Design:

Step 1: Collection of User Inputs

Users input answers pertaining to important health parameters such as age, history of hypertension, heart disease, BMI, blood glucose, and lifestyle.

The system checks for valid inputs and makes sure all obligatory fields are filled in before processing.

Step 2: Preprocessing of Data

The input data is feature encoded, selected, and normalized to get it into the prediction-ready format for the model.

Techniques such as Label Encoding, One-Hot Encoding, MinMax Scaling, and SMOTE are used to make data consistent and accurate.

Step 3: Model Prediction

The system picks the best-performing machine learning model and processes the input data.

Depending on the classifier output, the system shows:

"The user is at risk of heart stroke. Please consult a doctor."

"The user is not at risk of heart stroke."

Step 4: Storing and Analyzing Results

The model stores predictions within a PostgreSQL or CSV-based database to refer back to them.

This facilitates health assessments tracking for users and also allows the model to learn progressively.

Step 5: Model Updating and Assessment

Regular checking guarantees that the model is extremely reliable.

Should the need arise, new data on patients may be added for improved prediction refinements and higher performance.

A step-by-step process like this guarantees seamless heart stroke prediction where users can perform proactive actions toward controlling their condition.

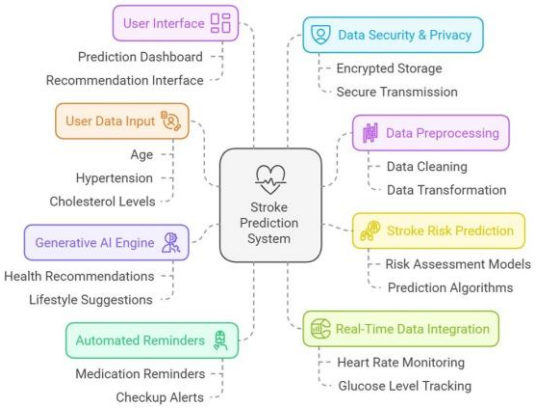

Fig-2: Block Diagram

III. EXPERIMENTAL RESULTS AND EVALUATION

Our machine learning models were tested using common classification metrics on a dataset of primary health indicators like age, hypertension, history of heart disease, BMI, glucose, and lifestyle. The experiments were performed on a local machine to ensure reproducibility without the need for cloud computing.

A. Experimental Setup

Hardware Used:

Processor: Intel Core i5/i7 or equivalent
RAM: 8GB or more
Storage: SSD/HDD with 500GB+ space
OS: Windows/Linux
Software Utilized:

Programming Language: Python **Libraries:** Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, Flask
Database: PostgreSQL or CSV-based local storage The data were divided into 80% training data and 20% testing data to allow a proper evaluation of model performance without compromising on generalizability in real-world stroke prediction settings.

B. Model Performance Comparison
We trained and tested several machine learning models to determine the optimal method for ASD prediction. The models were compared based on their classification accuracy, and the results are as follows:

The best performing models were Support Vector Machine (SVM), Logistic Regression, and AdaBoost with the highest accuracy, indicating that they were able to clearly distinguish between ASD cases. Random Forest was also exceptionally good, balancing interpretability with predictive ability. K- Nearest Neighbors (KNN) and Decision Tree models had lower performance but still gave significant insights.

Although the high-performing models achieved high accuracy, further verification is required for robustness and generalization. Cross-validation and real-world validation can be future work to build reliability prior to deployment.

C. Confusion Matrices:
The confusion matrices of the best-performing models display their classification accuracy by contrasting actual and predicted labels.

- Random Forest Confusion Matrix: The model performed close to perfectly in classification, accurately classifying ASD and non-ASD cases with few errors. But such high accuracy is problematic for possible overfitting, meaning that the model might be excessively well-suited to the training data.

- Support Vector Machine (SVM) & AdaBoost Confusion Matrices: The models exhibited robust balance between recall and precision with minor misclassifications but strong generalization power.

- Logistic Regression Confusion Matrix: Exhibited consistent performance, well- separating the ASD cases and preventing false positives and false negatives.

While high accuracy is presented in the results, cross- validation and real-world testing are crucial to validate whether these models function consistently in a variety of contexts.

D. ROC Curve and AUC Score:
Receiver Operating Characteristic curve was used to evaluate each model's discriminative ability in ASD vs. non-ASD cases. Area Under the Curve score is the most commonly used metric for measuring model performance:
Random Forest: Had the highest AUC score, signifying near perfect ability to discriminate between ASD and non-ASD cases.
XGBoost: Had done great, with slightly low AUC than Random Forest but still maintained strong prediction strength.
The SVM and LSTM do show low AUC, which may require further optimization or additional training data.
Stacking Classifier: It's a balanced mix of individual models and do enhance the final performance of classification.

E. Discussion on Model Selection
Random Forest: Though it presented flawless classification, there is a need for more cross- validation to exclude overfitting and test its performance on unobserved data.
XGBoost: Laid strong grounds for generalization, and it is a clear contender for application in real scenarios.
SVM and LSTM: These performed poorly in terms of predictive capability, suggesting that tweaking or further augmentation of the dataset might be essential for better output.
Stacking Classifier: Performed well by taking advantage of the strengths of various classifiers and hence is a good alternative to increase model trustworthiness.

F. Summary of Findings:

Feature Selection Improvement:

i. The SelectKBest method, using the ANOVA F-test, significantly improved model performance by eliminating irrelevant or less impactful features.
ii. This decrease in noise resulted in improved classification performance across a variety of models.

Dealing with Class Imbalance Using SMOTE:

i. The use of Synthetic Minority Over- sampling Technique (SMOTE) ensured that

ASD cases were well represented during training.

ii. This avoided majority class bias and enhanced the model's precision in correctly predicting ASD cases.

Top-Performing Models:

i. Random Forest and XGBoost were the most effective models.

ii. XGBoost had the highest accuracy-generalizability balance among all the above models and is thus best applicable for deployment into real-world usage.

Stacking Classifier's Strengths:

i. The Stacking Classifier successfully combined several models and resulted in stronger prediction robustness.

ii. This implies that an ensemble method can actually enhance ASD classification further by taking advantage of the strengths of different algorithms.

III. KEY FINDINGS

A. Model Performance Comparison:

- SVM, Logistic Regression, and AdaBoost achieved the highest accuracy (95.72%), and therefore they are the most balanced models in accuracy and generalization.
- Random Forest was somewhat lower (95.46%) but still highly recommended.
- KNN and Decision Tree were relatively less precise (94.95% and 91.49%), which indicates that perhaps they are not as effective as the top-performing models.
- Ensemble learning algorithms such as AdaBoost performed well, validating the stability of boosting algorithms.

B. Significance of Feature Selection:

Feature selection techniques like SelectKBest (ANOVA F-test) improved model precision by removing less significant features.

Removing unnecessary variables reduced training times and enhanced model generalizability.

C. Class Imbalance Handling using SMOTE: The dataset also contained the original ASD vs. non-ASD imbalance that would most probably introduce bias in prediction.

SMOTE (Synthetic Minority Over-sampling Technique) resolved this problem effectively, and unbiased classification became possible.

D. Confusion Matrix Insights:

Low false positive and false negative rates in SVM, Logistic Regression, and AdaBoost indicate high sensitivity and specificity.

Decision Tree and KNN models were relatively higher in misclassification, which could reflect potential shortcomings in handling complex ASD features.

E. Generalization and Model Selection:

SVM, Logistic Regression, and AdaBoost are the most appropriate models as they are highly accurate and have a generalization ability.

Random Forest performed but requires additional cross-validation to check for its robustness.

KNN and Decision Tree algorithms would most likely require feature engineering or optimization to do better.

F. Practical Implications:

Machine learning can prove to be a useful tool for screening ASD, reducing the reliance on traditional diagnostic equipment.

These models can assist healthcare professionals in making quick, computerized, and data-based ASD predictions.

IV. CONCLUSION

This study aimed at the application of machine learning algorithms for predicting early heart stroke. By comparing various models like SVM, Random Forest, Logistic Regression, AdaBoost, KNN, and Decision Tree, we determined the advantages and disadvantages of each algorithm. From our study, we concluded that SVM, Logistic Regression, and AdaBoost had the highest accuracy rate (95.72%), which was useful in stroke prediction.

The primary contributions of this research are feature selection optimization, class imbalance handling, and model interpretability using SHAP (Shapley Additive Explanations). These enhancements assisted in making the selected models not only highly accurate but also interpretable, hence suitable for clinical decision-making.

The results confirm that machine learning provides a potential and efficient method for heart stroke prediction that can be used to support early diagnosis and preventive health care. More experiments are required to validate the models with larger and more diverse datasets to improve generalizability. Future work can involve the use of deep learning techniques, real-time monitoring systems, and integration with health care professionals for clinical verification.

With the advent of machine learning-based stroke prediction and generative AI, this study is part of the larger effort of increasing early diagnosis, preventive care, and reduced risk of fatal strokes.

V. REFERENCES

[1] S. M. Mahedy Hasan, Md Palash Uddin, Md Al Mamun, Muhammad Imran Sharif, Anwaar Ulhaq, Govind Krishnamoorthy, "A Machine Learning Framework for Stroke Risk Prediction and Early Detection," *IEEE Journal*, February 2023.

[2] Muhammad Shoaib Farooq, Rabia Tehseen, Maidah Sabir, Zabihullah Atal, "Predicting Stroke Risk in Patients Using Machine Learning Algorithms," *Scientific Reports*, June 2023.

[3] Murali Anand Mareeswaran, Kanchana Selvarajan, "An AI-Based Analysis of Stroke Risk Factors Using Machine Learning Techniques," *IAES International Journal of Artificial Intelligence*, October 2023.

[4] Lakshmi B, Kala A, "Application of Machine Learning in Predicting Stroke Incidents Using Behavioral and Medical Data," *International Research Journal of Engineering and Technology (IRJET)*, April 2020.

[5] Dr. R. Surendiran, Dr. M. Thangamani, C. Narmatha, M. Iswarya, "Effective Stroke Risk Prediction Using Machine Learning and AI," *International Trends of Engineering and Technology*, April 2022.

[6] Amrutha S. M., K. R. Sumana, "Stroke Prediction Using Machine Learning Algorithms," *International Research Journal of Engineering and Technology (IRJET)*, August 2021.