ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

A Machine Learning Framework for Early Prediction of Brain Stroke Using Clinical Attributes

[1]B NANDANA KUMAR, [2]P. PARVATHI, [3]L. VINOD, [4]M. KRISHNA NAGA SAI, [5]M. S RIDEVI

[1]Dept of CSE, D.N.R. College of Engineering & Technology, Balusumudi, Bhimavaram - 534 202, W.G. Dist., A.P., India.

[2,3,4,5]Student, Dept of CSE, D.N.R. College of Engineering & Technology, Balusumudi, Bhimavaram - 534 202, W.G. Dist., A.P., India.

ABSTARCT

Stroke remains one of the leading causes of mortality and long-term disability worldwide, necessitating early detection for timely clinical intervention and improved patient outcomes. Inspired by methodologies used in ADMETbased drug side-effect prediction—where functional group patterns and engineered descriptors facilitate early risk assessment—this study proposes an explainable machine learning framework for predicting stroke occurrence based on clinical, demographic, and lifestyle factors. The curated dataset incorporates key attributes such as age, hypertension status, history of heart disease, average glucose level, smoking behavior, and body mass index (BMI). Comprehensive data preprocessing techniques, including exploratory data analysis (EDA), feature correlation analysis, and class imbalance handling using the Synthetic Minority Over-sampling Technique (SMOTE), were employed to enhance data quality and model robustness. Multiple machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), were developed and evaluated. Among these, the Gradient Boosting model demonstrated superior and consistent generalization performance across both balanced and imbalanced datasets. Furthermore, SHapley Additive exPlanations (SHAP) analysis was utilized to interpret model predictions, revealing that age, average glucose level, hypertension, and BMI are the most influential features—aligning with established clinical risk factors for stroke. The proposed framework provides a reliable, interpretable, and data-driven approach for early stroke risk prediction. By leveraging structured feature engineering concepts analogous to functional group-based modeling in drug discovery, this work highlights the potential of machine learning in enhancing predictive accuracy in healthcare diagnostics. The system can assist clinicians in identifying high-risk individuals and lays the foundation for future integration of multimodal biomedical data to further improve stroke prediction models.

Keywords: Stroke Prediction, Machine Learning, Gradient Boosting, SHAP Analysis, Healthcare Analytics, Risk Assessment, SMOTE, Explainable AI (XAI), Clinical Decision Support, Feature Engineering

1. INTRODUCTION

1.1 Background and Overview

Stroke is one of the leading causes of mortality and long-term disability worldwide, posing a major burden on global healthcare systems. Hypertension, which is often referred to as a "silent killer," is one of the most significant risk factors contributing to stroke incidence and related complications [1]. The global prevalence of hypertension has been increasing steadily, affecting a large proportion of the adult population and significantly elevating the risk of cardiovascular diseases, including stroke [2].

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

A substantial proportion of stroke cases can be prevented through early detection and timely intervention. However, accurate prediction of stroke risk remains complex due to the involvement of multiple interrelated factors. These include clinical conditions such as hypertension and heart disease, demographic attributes such as age and gender, and lifestyle behaviors such as smoking, diet, and physical inactivity [2]. The interaction among these factors is often non-linear and dynamic, making traditional prediction approaches less effective.

Conventional stroke risk assessment models, such as the Framingham Stroke Risk Profile, are primarily based on statistical techniques that assume linear relationships among variables [3]. While these models provide useful population-level insights, they often lack the ability to capture complex dependencies and interactions among risk factors. As a result, their performance in personalized risk prediction is limited, particularly in diverse and real-world clinical environments.

1.2 1.2 Machine Learning for Stroke Prediction

In recent years, machine learning (ML) has emerged as a powerful tool in healthcare analytics, enabling more accurate and data-driven disease prediction. ML algorithms are capable of analyzing large volumes of data, identifying hidden patterns, and modeling complex non-linear relationships among variables. These capabilities make ML particularly suitable for predicting stroke risk, where multiple heterogeneous factors contribute to the outcome. Several machine learning models, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), have been successfully applied in medical prediction tasks [4]. Ensemble methods, particularly boosting algorithms, have demonstrated superior performance due to their ability to combine multiple weak learners into a strong predictive model.

Despite these advantages, ML-based systems face several challenges in healthcare applications. One of the primary issues is **class** imbalance, where the number of non-stroke cases significantly exceeds stroke cases. This imbalance can lead to biased models that perform well on the majority class but fail to accurately identify high-risk patients. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been proposed to address this issue by generating synthetic samples for the minority class, thereby improving model performance [5]. Another critical challenge is the lack of interpretability in machine learning models. In clinical settings, it is essential for healthcare professionals to understand the reasoning behind model predictions. Explainable Artificial Intelligence (XAI) techniques, such as SHapley Additive exPlanations (SHAP), provide a solution by quantifying the contribution of each feature to the model's output [6]. This enhances transparency and facilitates trust in ML-based decision support systems.

1.3 1.3 Motivation

Stroke remains a major global health concern due to its high mortality rate and long-term complications. Many stroke cases are associated with modifiable risk factors such as hypertension, poor medication adherence, and unhealthy lifestyle behaviors [4]. Despite the availability of clinical data, existing systems often fail to provide accurate and personalized risk assessments.

The motivation behind this work is to leverage machine learning techniques to develop an intelligent and interpretable stroke prediction system. By integrating advanced ML algorithms with data preprocessing and explainability methods, the proposed framework aims to improve prediction accuracy and support early intervention.

1.4 1.4 Problem Statement

Traditional stroke prediction methods are limited in their ability to model complex and non-linear relationships among multiple risk factors. These methods often rely on simplified statistical

assumptions, which restrict their effectiveness in real-world clinical scenarios. Furthermore, the presence of class imbalance in medical datasets results in biased model performance, particularly in identifying stroke cases, which belong to the minority class.

In addition, many existing machine learning models lack interpretability, making it difficult for healthcare professionals to trust and adopt these systems in clinical practice. Without clear explanations of predictions, the practical deployment of such models remains a challenge.

To address these limitations, this study proposes a comprehensive machine learning framework for early stroke prediction. The framework incorporates data preprocessing techniques such as handling missing values and class imbalance using SMOTE, evaluation of multiple machine learning models to determine optimal performance, and integration of explainable AI methods such as SHAP to enhance interpretability.

The primary objective is to develop an accurate, reliable, and interpretable system for early stroke risk detection, enabling timely medical intervention and improved decision-making in healthcare environments.

2. LITERATURE REVIEW

A literature review provides a detailed understanding of existing research, methodologies, and gaps in a particular domain. In the field of stroke prediction, significant research has been conducted using both traditional statistical methods and modern machine learning approaches. This section reviews key contributions and highlights their limitations.

Early research in cardiovascular risk assessment emphasized the role of hypertension and related factors. The World Health Organization identified hypertension as a major contributor to stroke and cardiovascular diseases, highlighting the need for early detection and management [1]. Studies on global hypertension trends further confirmed its strong association with stroke risk, emphasizing the importance of predictive modeling in healthcare [2].

One of the foundational works in this domain is the Framingham Stroke Risk Profile, developed by Kannel et al. as part of the Framingham Heart Study [1]. This model utilized clinical parameters such as age, blood pressure, diabetes, and smoking status to estimate stroke risk. Despite its importance, the model was based on linear regression techniques and lacked the ability to capture non-linear relationships and interactions among multiple risk factors.

With the advancement of computational capabilities, machine learning (ML) techniques have emerged as a powerful alternative to traditional methods. Khosla et al. [2] explored the use of machine learning algorithms for stroke prediction using clinical datasets. Their study demonstrated that ML models could process multiple input features simultaneously and uncover hidden patterns that are not evident in conventional statistical models.

Interpretability is a critical requirement in healthcare applications, as clinicians must understand the reasoning behind model predictions. To address this challenge, Rudin and Letham et al. [3] proposed an interpretable machine learning approach using Bayesian Rule Lists. This model generates simple and humanreadable decision rules, making it easier for healthcare professionals to interpret the results.

Another significant challenge in stroke prediction is the issue of class imbalance in medical datasets. Ghanipour et al. [4] addressed this problem by applying the Synthetic Minority Over-sampling Technique (SMOTE). This technique generates synthetic samples of the minority class, thereby balancing the dataset and improving the model's ability to detect stroke cases.

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

Ensemble learning techniques have shown remarkable success in handling complex datasets and improving prediction accuracy. Zhang et al. [5] proposed a stroke prediction model using the Extreme Gradient Boosting (XGBoost) algorithm. XGBoost is a powerful ensemble method that combines multiple weak learners to form a strong predictive model.

Recent research has also focused on integrating multiple data sources to enhance prediction performance. Chen et al. [6] developed a multimodal machine learning framework that combines clinical data with medical imaging. By incorporating diverse data types, the model achieved improved accuracy and robustness.

The growing demand for transparency in AI systems has led to the development of Explainable Artificial Intelligence (XAI) techniques. Lee et al. [7] applied SHapley Additive exPlanations (SHAP) to interpret machine learning models used for stroke prediction. SHAP provides a quantitative measure of the contribution of each feature to the model's prediction.

A comprehensive review conducted by Chang et al. [8] analyzed various machine learning approaches for stroke prediction. The study compared different models and discussed their advantages and limitations. It highlighted key challenges such as data imbalance, lack of interpretability, and limited real-world deployment.

2.1 Research Gap

Although significant progress has been made in the field of stroke prediction, several limitations still exist. Traditional statistical models are unable to effectively capture complex, non-linear relationships among multiple risk factors. While machine learning models improve prediction accuracy, many of them lack interpretability, which limits their adoption in clinical settings. Additionally, class imbalance remains a critical issue, affecting the model's ability to accurately identify high-risk patients.

Moreover, most existing studies focus on individual aspects such as model development, data preprocessing, or interpretability, rather than providing a unified framework that integrates all these components. There is a need for a comprehensive system that combines efficient data preprocessing, robust machine learning algorithms, and explainable AI techniques to deliver accurate and reliable stroke predictions.

2.2 Conclusion of Literature Review

In summary, the literature indicates that machine learning techniques have significantly improved the performance of stroke prediction systems. Ensemble methods such as XGBoost and Gradient Boosting have demonstrated high accuracy and robustness, while preprocessing techniques like SMOTE effectively address class imbalance. Furthermore, explainable AI methods such as SHAP enhance model transparency and support clinical decision-making. However, there is still a need for an integrated and interpretable framework that can address existing challenges and provide reliable predictions in real-world healthcare environments.

3. PROPOSED SYSTEM

The proposed system is a machine learning-based framework designed to predict brain stroke risk using clinical and lifestyle attributes. It utilizes multiple input features such as age, hypertension, heart disease, glucose level, BMI, and smoking status to analyze patient health conditions. The system applies data preprocessing techniques including handling missing values, encoding categorical variables, and normalization to improve data quality.

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

To address the issue of class imbalance, SMOTE (Synthetic Minority Oversampling Technique) is used to generate synthetic samples and enhance model performance. Various machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting are implemented and evaluated to identify the most accurate model. Additionally, SHAP (Explainable AI) is incorporated to provide clear insights into feature importance and model predictions. The final model is deployed through a webbased application that enables real-time stroke risk prediction, allowing healthcare professionals to make timely and informed decisions.

Advantages of Proposed System:

- Provides higher prediction accuracy compared to traditional methods.
- Captures complex and non-linear relationships between risk factors.
- Handles imbalanced datasets effectively using SMOTE.
- Enables early detection of stroke risk for timely medical intervention.
- Offers personalized predictions for individual patients.
- Provides real-time prediction through web-based deployment.

3.1 System Architecture

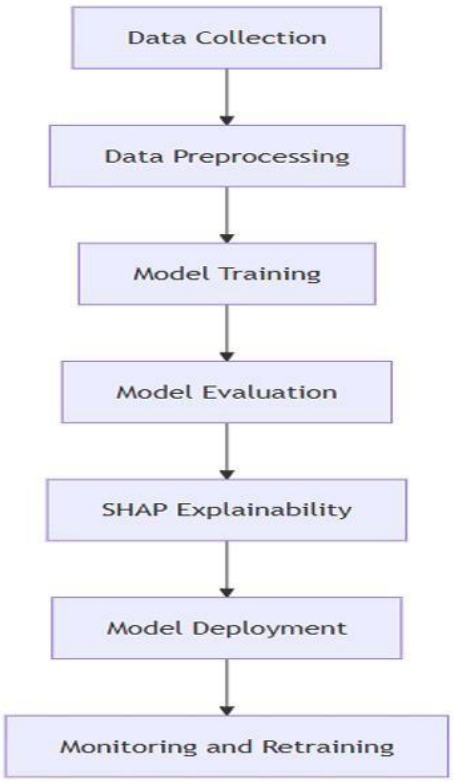


Fig 1: System Architecture

3.2 Use Case Diagram

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

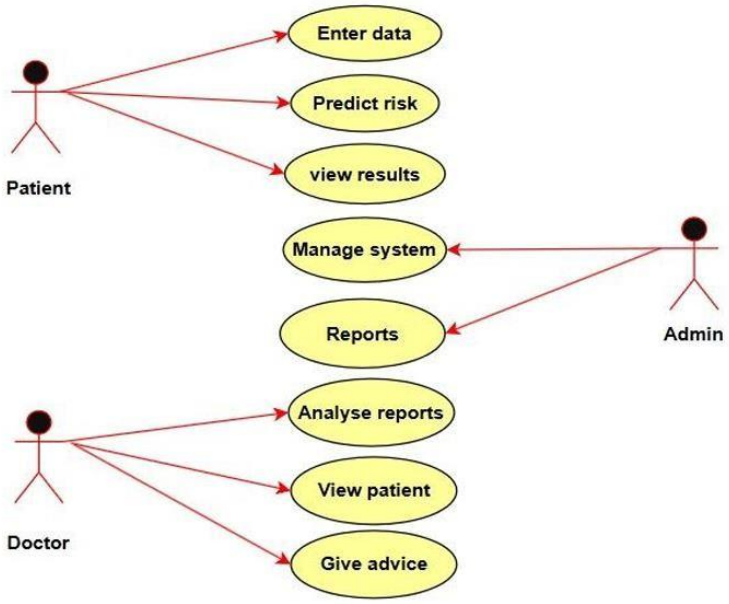


Fig 2: Use Case Diagram

3.3 Class Diagram

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram may be capable of providing certain functionalities. These functionalities provided by the class are termed "methods" of the class. Apart from this, each class may have certain "attributes" that uniquely identify the class.

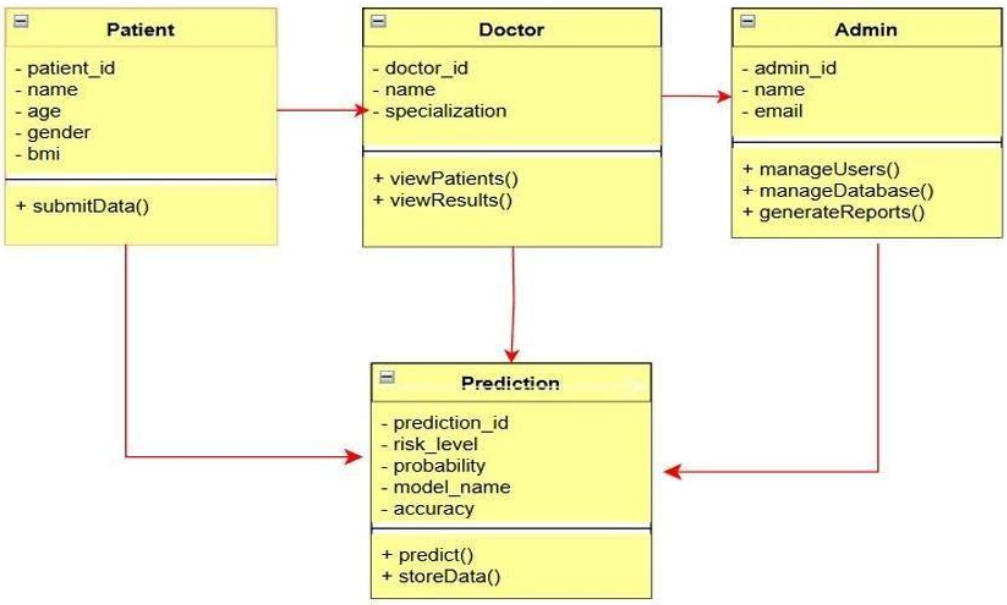


Fig 3: Class Diagram

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

3.4 Sequence Diagram

A sequence diagram represents the interaction between different objects in the system. The important aspect of a sequence diagram is that it is time-ordered. This means that the exact sequence of the interactions between the objects is represented step by step. Different objects in the sequence diagram interact with each other by passing "messages".

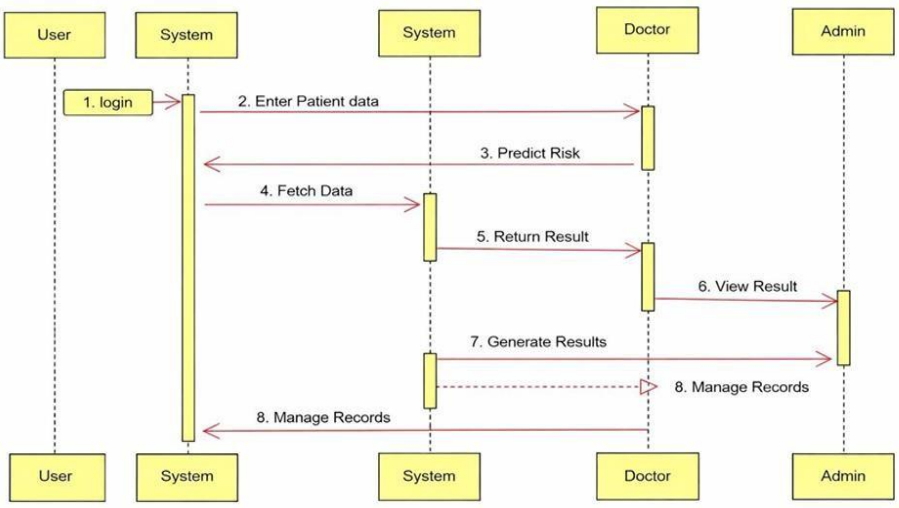


Fig 4: Sequence Diagram

3.5 Activity Diagram

In UML, the activity diagram is used to demonstrate the flow of control within the system rather than the implementation. It models the concurrent and sequential activities. The activity diagram helps in envisioning the workflow from one activity to another. It puts emphasis on the condition of flow and the order in which it occurs. The flow can be sequential, branched, or concurrent, and to deal with such kinds of flows, the activity diagram has come up with a fork, join, etc. It is also termed as an object-oriented flowchart.

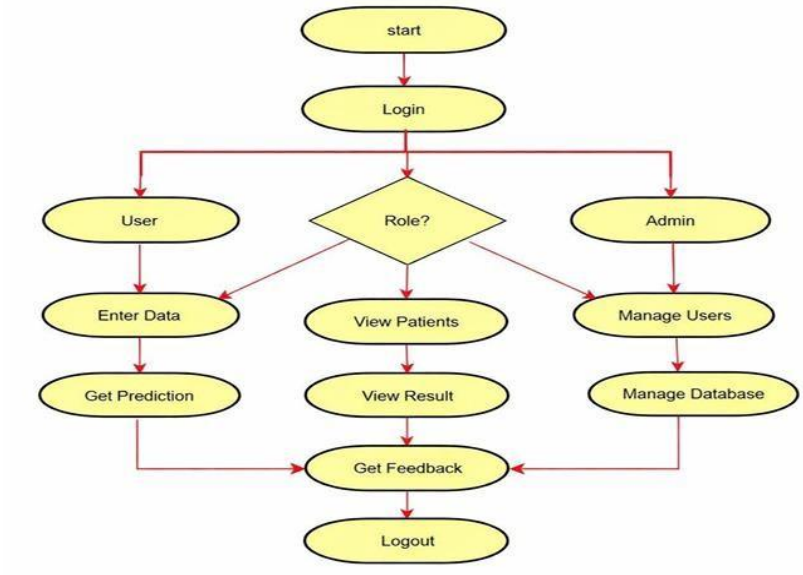


Fig 5: Activity Diagram

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

4. RESULTS

The proposed Machine Learning Framework for Early Prediction of Brain Stroke was successfully implemented and tested. The system processes patient data in real-time, applying preprocessing, model evaluation, and SHAP-based explainability to generate accurate and interpretable stroke risk predictions.

Home Page

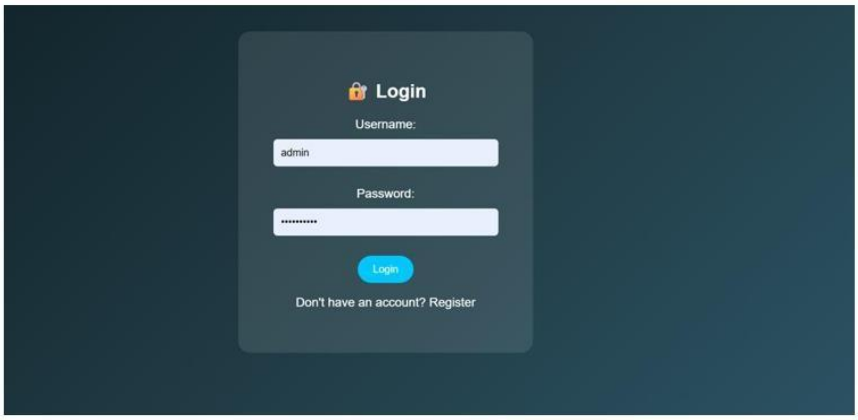


Fig 6: Home Page

New User Registration

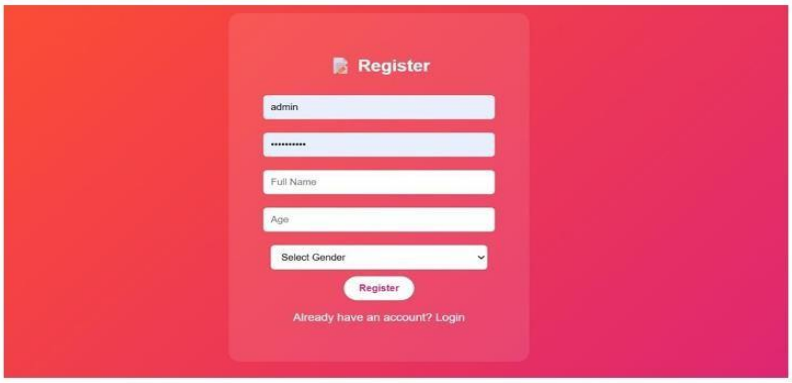


Fig 7: New User Registration

Dashboard

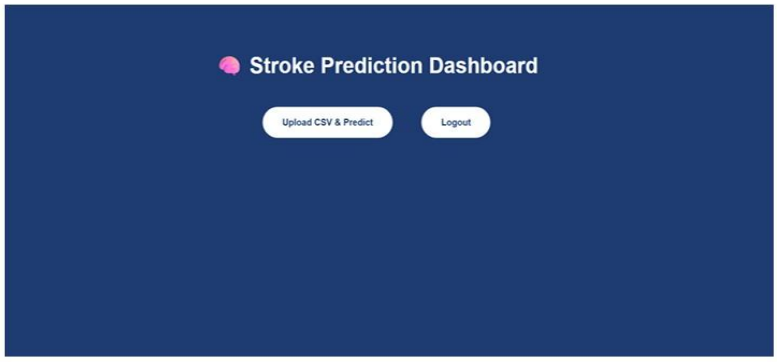


Fig 8: Dashboard Dataset

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

Dataset

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	prediction
0	1	67.0	0	1	1	2	1	228.69	36.6	1	1	0
1	0	61.0	0	0	1	3	0	202.21	0.0	2	1	0
2	1	80.0	0	1	1	2	0	105.92	32.5	2	1	0
3	0	49.0	0	0	1	2	1	171.23	34.4	3	1	0
4	0	79.0	1	0	1	3	0	174.12	24.0	2	1	0
5	1	81.0	0	0	1	2	1	186.21	29.0	1	1	0
6	1	74.0	1	1	1	2	0	70.09	27.4	2	1	0
7	0	69.0	0	0	0	2	1	94.39	22.8	2	1	0
8	0	59.0	0	0	1	2	0	76.15	0.0	0	1	0
9	0	78.0	0	0	1	2	1	58.57	24.2	0	1	0

Fig 9: Dataset

Prediction Results

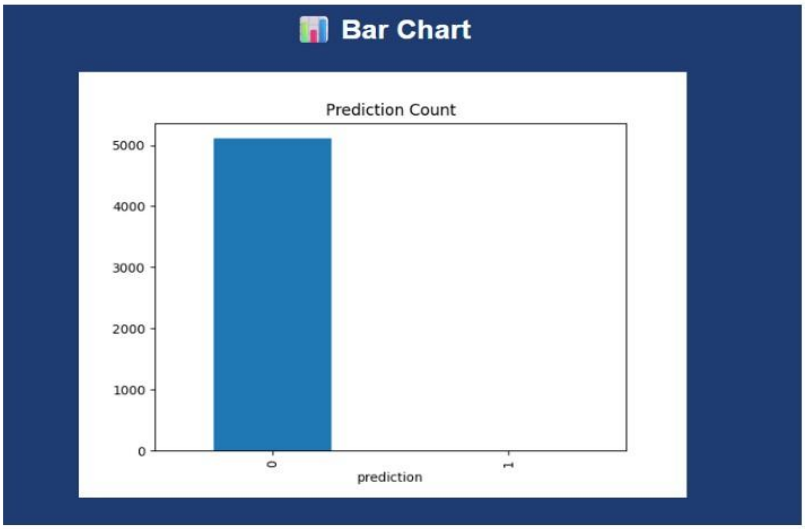


Fig 10: Prediction Results ROC Curve

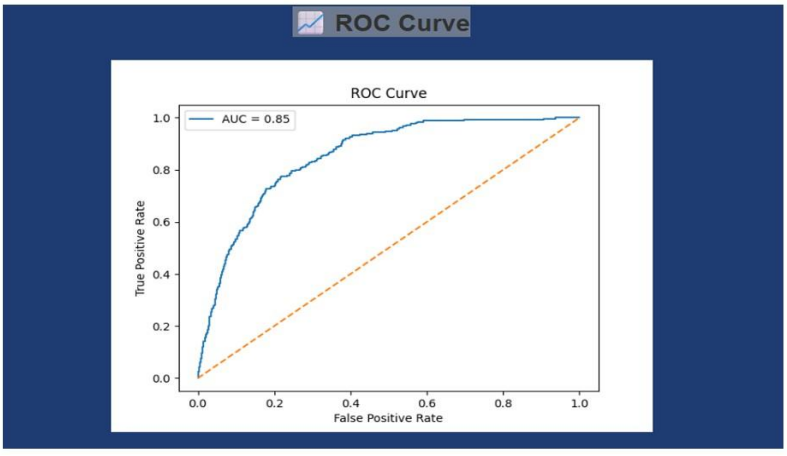


Fig 11: ROC Curve Pair Plot

ISSN: 2582 - 6379
IJISEA Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2026, Volume 7 Issue 2
www.ijisea.org

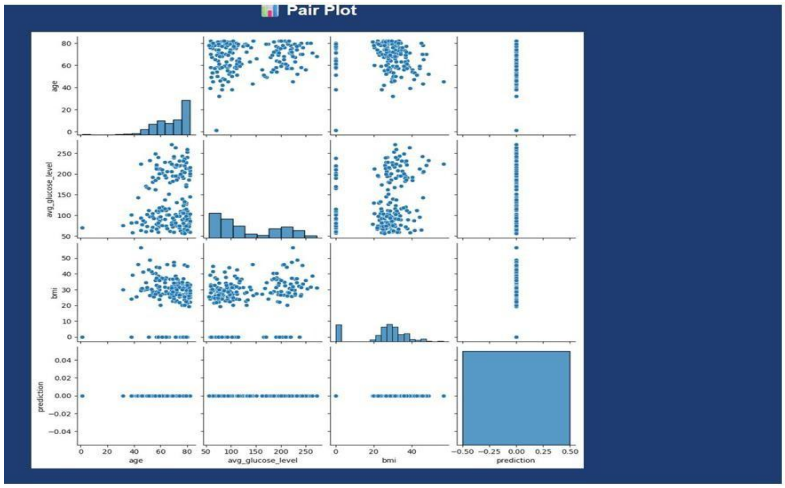


Fig 12: Pair Plot

Model Performance

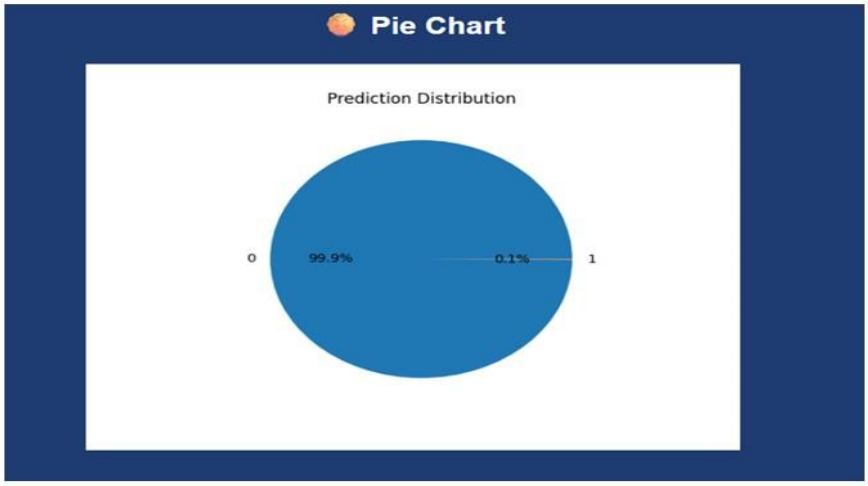


Fig 13: Model Performance

5. CONCLUSION

This study presented a machine learning-based framework for the early prediction of brain stroke using clinical, demographic, and lifestyle-related features. The results obtained through comprehensive exploratory data analysis (EDA) and multi-model evaluation highlight the critical challenges and opportunities associated with stroke prediction in healthcare systems. One of the key findings of this work is that class imbalance represents a major obstacle in developing reliable stroke prediction models. Initial experiments conducted on the original imbalanced dataset revealed that models such as Support Vector Machine (SVM) and Random Forest achieved high overall accuracy; however, they failed to identify stroke cases effectively. In addition, the feature analysis conducted in this work confirms that attributes such as age, average glucose level, hypertension status, and body mass index (BMI) are among the most influential predictors of stroke occurrence. These results are consistent with established clinical knowledge, thereby validating the reliability of the proposed framework. Overall, the proposed system demonstrates that machine learning can serve as a powerful tool for early stroke risk assessment when combined with proper data handling and evaluation strategies. Ensemble learning methods, particularly Random Forest, have shown the ability to capture

complex and non-linear relationships within clinical datasets. The integration of such models into healthcare systems can assist clinicians in prioritizing high-risk patients, enabling timely intervention and preventive care. Future work can focus on incorporating advanced imbalance-handling techniques, optimizing decision thresholds, and integrating multimodal data sources such as medical imaging and genetic information. These enhancements have the potential to further improve prediction accuracy and support the development of more comprehensive and reliable stroke prediction systems for real-world clinical applications.

6. FUTURE SCOPE

While the proposed machine learning-based stroke prediction framework demonstrates promising performance, several opportunities exist to further enhance its accuracy, robustness, and clinical applicability.

One of the primary areas for improvement is the handling of class imbalance. Advanced hybrid methods such as SMOTEENN and ADASYN (Adaptive Synthetic Sampling) can be explored to generate more representative minority class samples. Additionally, cost-sensitive learning approaches can further enhance the model's ability to identify high-risk patients.

Another important direction is the exploration of more advanced algorithms. State-of-the-art boosting techniques like XG Boost, Light GBM, and Cat Boost offer improved computational efficiency and predictive accuracy. Furthermore, the integration of deep learning models, including Artificial Neural Networks (ANNs) and hybrid architectures, can enable the system to learn high-level feature representations. The incorporation of multimodal biomedical data represents another significant area for future enhancement. Integrating additional data sources such as electronic health records (EHRs), medical imaging (CT/MRI scans), genetic information, and real-time data from wearable sensors can provide a more comprehensive understanding of patient health.

Future work can focus on the integration of advanced Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME (Local Interpretable Model-Agnostic Explanations). These methods can provide detailed insights into model predictions, thereby helping clinicians make informed decisions and increasing trust in the system.

Finally, developing a user-friendly web or mobile application that integrates the prediction model can make the system accessible to healthcare providers and patients. Such deployment would enable continuous data collection, real-time prediction, and timely intervention, ultimately contributing to improved patient outcomes and reduced healthcare burden.

REFERENCES

[1] World Health Organization. A Global Brief on Hypertension: Silent Killer, Global Public Health Crisis: World Health Day 2013; WHO: Geneva, Switzerland, 2013.
[2] Mills, K.T.; Stefanescu, A.; He, J. The global epidemiology of hypertension. Nat. Rev. Nephrol. 2020, 16, 223–237.
[3] Lehasa, O.M.-E.; Chude-Okonkwo, U.A.K. Dataset for discovering new hypertension small molecules using machine learning-aided computational fragment-based design. Data Brief 2024, 55, 110677.
[4] Olowofela, A.O.; Isah, A.O. Profile and predictors of antihypertensive adherence among patients in a tertiary care setting in Southwestern Nigeria. Am. J. Hypertens. 2017, 30, 919–927.
[5] Takuathung, M.N. et al. Adverse Effects of Angiotensin Converting Enzyme Inhibitors in Humans: A Systematic Review and Meta-Analysis. Int. J. Environ. Res. Public Health 2022, 19, 8373.
[6] Mao, F. et al. Chemical Structure-Related Drug-Like Criteria of Global Approved Drugs. Molecules 2016, 21, 75.

[7] Zanders, E.D. Preclinical Development. In The Science and Business of Drug Discovery; Springer: Cham, Switzerland, 2020.
[8] Lehasa, O.M.-E.; Chude-Okonkwo, U.A.K. Machine Learning-aided Computational Fragment-based Design of Small Molecules for Hypertension Treatment. Intell.-Based Med. 2024, 10, 100171.
[9] Blanco-González, A. et al. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. Pharmaceuticals 2023, 16, 891.