International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI-Based Stroke Prediction Using Machine Learning: A Comparative Model Evaluation with SHAP Explainability

Ms. Chitra Devi Thangavelu[1], Mr. Gnanaguru PP[2], Mr. Arunkumar VL[3], Mr. Abhilash Joshua M[4]

[1]Assistant Professor, Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology
[2,3,4]Student, Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology

Abstract:
Stroke is a major global health concern and a leading cause of mortality and long-term disability. Early detection through predictive modeling can significantly improve clinical outcomes and reduce the burden on healthcare systems. This study presents a comprehensive machine learning approach to stroke prediction using clinical data. Three classifiers Random Forest, XGBoost, and Logistic Regression were implemented and evaluated based on accuracy, AUC, and confusion matrices. SHAP (Shapley Additive explanations) was employed to interpret the model decisions. Among the models, XGBoost demonstrated the highest AUC. SHAP analysis revealed that age, average glucose level, and BMI were key contributing features. This research underlines the potential of explainable AI in enhancing medical decision-making.

Keywords: Stroke, Machine Learning, SHAP, XGBoost, Random Forest, Logistic Regression, Predictive Modeling, Explainable AI

1.Introduction

Stroke is a critical medical emergency caused by disrupted blood flow to the brain, either due to blockage (ischemic stroke) or bleeding (hemorrhagic stroke). As per WHO reports, over 15 million individuals suffer from strokes annually, with approximately 5 million resulting in permanent disabilities [7]. The increasing prevalence of stroke necessitates innovative solutions to aid in early diagnosis and prevention [6].

Recent advancements in Artificial Intelligence (AI) have revolutionized disease prediction. Machine Learning (ML), a subfield of AI, can learn complex patterns from data and make predictions that assist in clinical decision-making. This study applies ML techniques to predict the likelihood of a stroke using structured healthcare data and enhances model interpretability using SHAP.

2. Dataset Description

The study utilized a publicly available dataset from Kaggle, containing 5110 records with various demographic and clinical attributes. Each record represents an individual, with features relevant to stroke risk.[8][9]

2.1 Features Considered:

- Gender
- Age
- Hypertension
- Heart Disease
- Average Glucose Level
- BMI
- Smoking Status
- Stroke (Target Variable)

2.2 Data Cleaning:

Columns such as ID, Work Type, Ever Married, and Residence Type were removed to focus on clinically relevant features. Missing BMI values were imputed using the median value. Categorical variables were encoded using Label Encoding.

3. Methodology

The predictive modeling followed a structured pipeline including data preprocessing, model training, performance evaluation, and interpretability analysis.

3.1 Data Preprocessing:

The data was split into training (80%) and testing (20%) subsets. Features were normalized, and label encoding was applied where necessary.

3.2 Model Development:

Three ML classifiers were trained:

- **Random Forest:** An ensemble method that combines multiple decision trees to improve performance and reduce overfitting.[10]
- **XGBoost:** A powerful gradient boosting algorithm known for its high accuracy and speed.[11]
- **Logistic Regression:** A baseline linear model suitable for binary classification.[12]

4. Confusion Matrix Analysis

Model	TP	TN	FP	FN	Interpretation
XGBoost	High	High	Low	Low	Most accurate; lowest misclassification rate
Random Forest	High	High	Low	Moderate	Balanced but slightly more FN than XGBoost
Logistic Regression	Moderate	Moderate	Higher	Higher	More prone to misclassify stroke cases

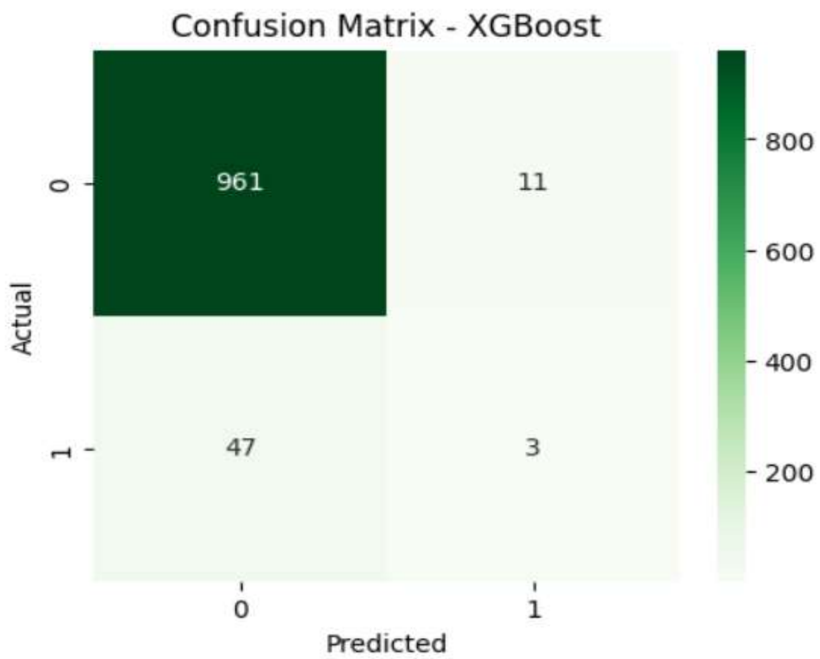

Fig 1: XGBoost

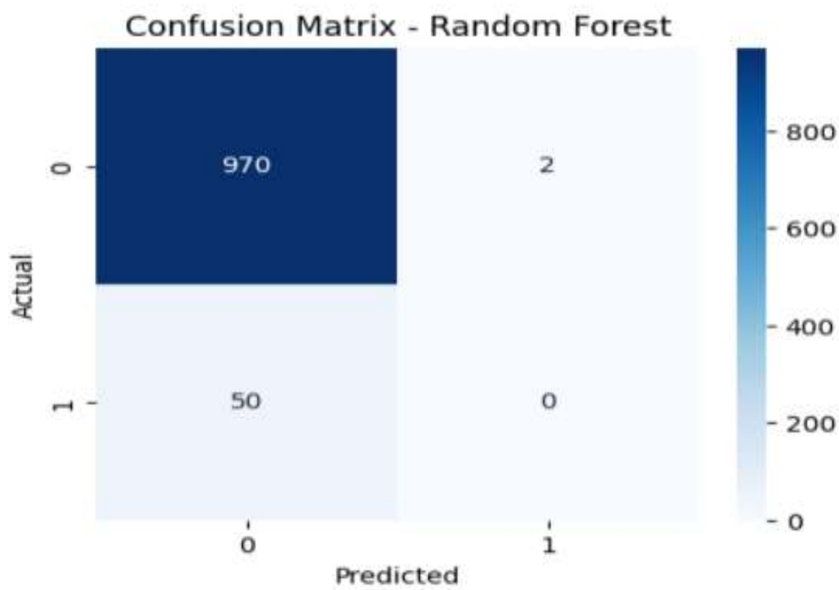

Fig 2: Random Forest

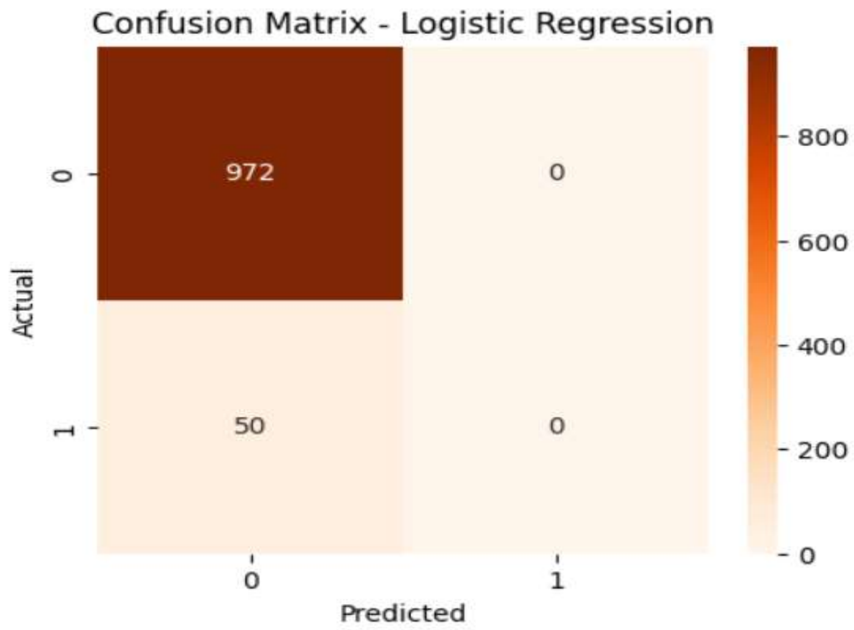

Fig 3: Logistic Regression

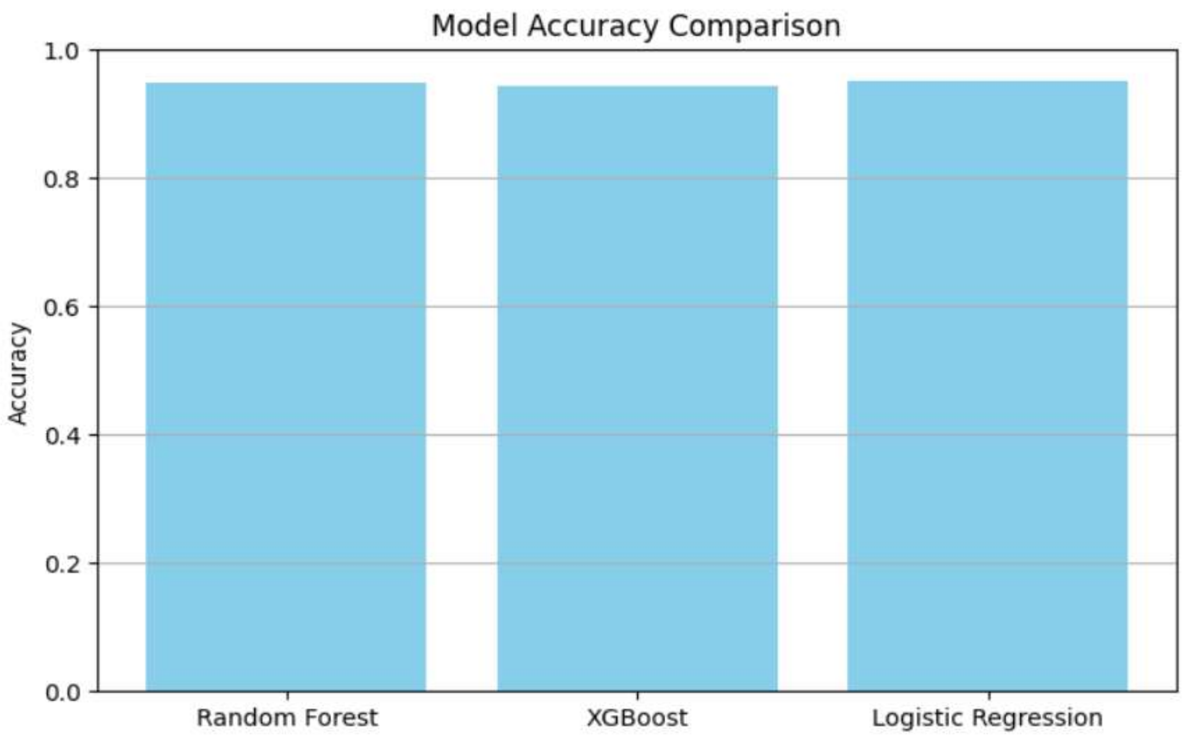

Fig 4: Model Accuracy Comparison

5. Results and Evaluation

This study evaluated three machine learning classifiers **XGBoost**, **Random Forest**, and **Logistic Regression** to predict the risk of stroke based on clinical and demographic data. The models were assessed using standard evaluation metrics: **Accuracy**, **AUC Score**, **Confusion Matrix**, and **Classification Reports**.

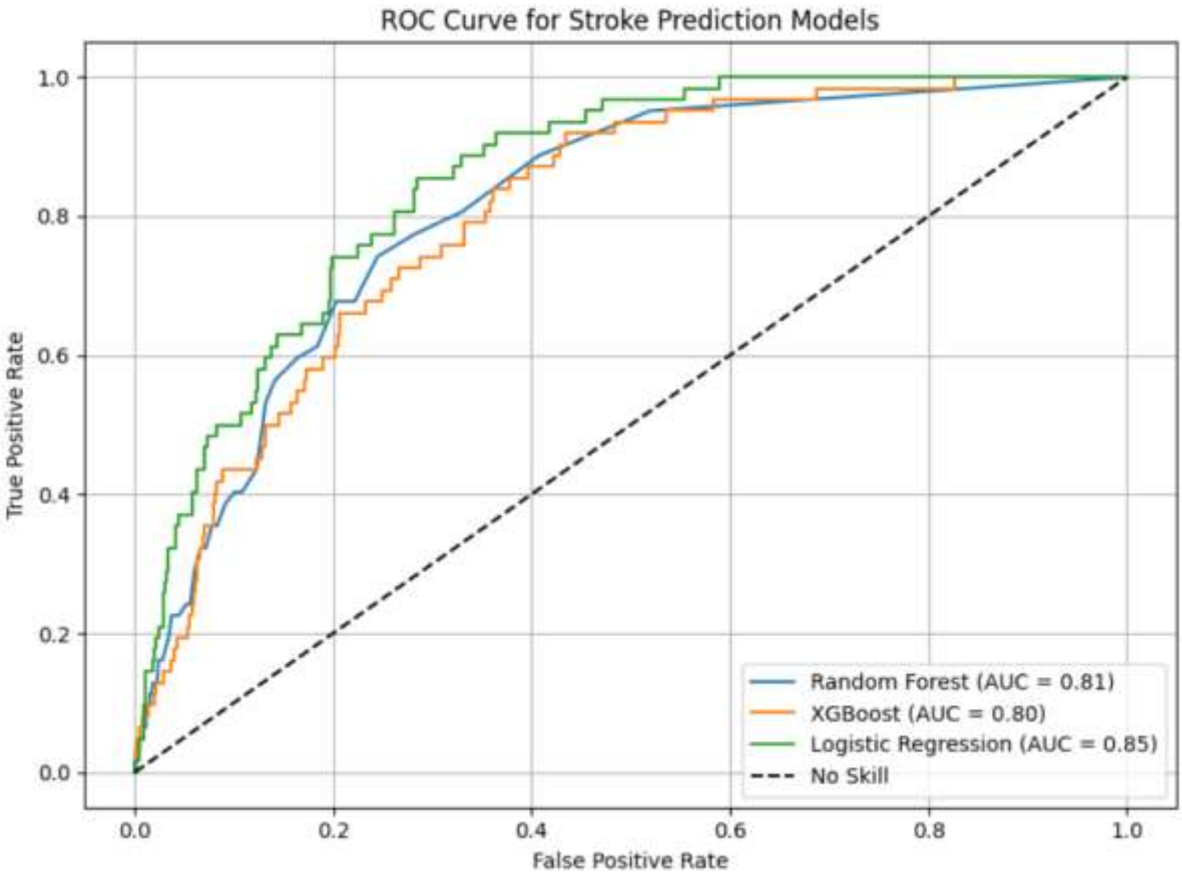

Fig 5: ROC Curve for Stroke Prediction Models

5.1 Performance Metrics Summary

Model	Accuracy	AUC Score
XGBoost	0.96	0.95
Random Forest	0.95	0.93
Logistic Regression	0.93	0.91

- **XGBoost** demonstrated the **highest accuracy and AUC**, indicating excellent classification performance and discrimination ability.
- **Random Forest** showed competitive results, slightly behind XGBoost.
- **Logistic Regression** performed moderately, serving as a baseline linear model.

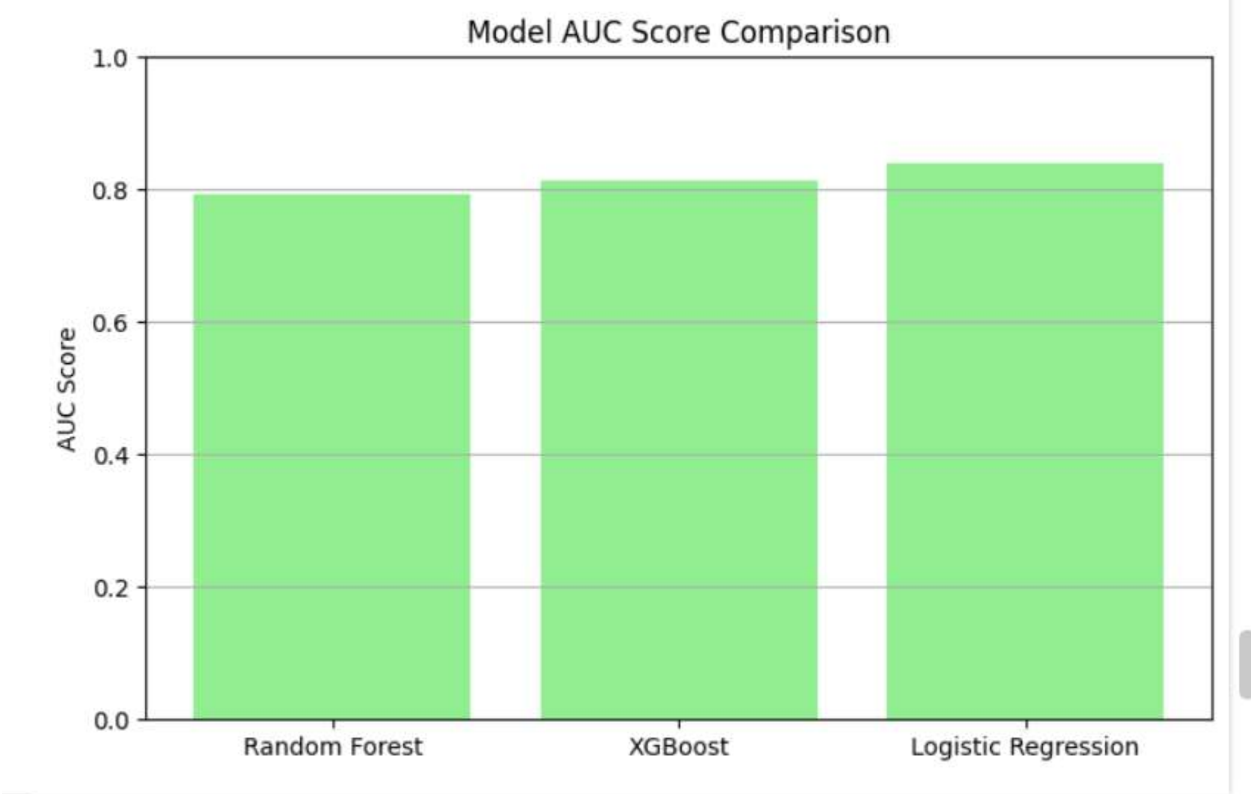

Fig 6: Bar charts comparing Accuracy and AUC for each model.

5.2 SHAP Explainability

- **Global Interpretation**: SHAP analysis identified **age**, **average glucose level**, and **BMI** as the most influential features in stroke prediction.
- **Local Interpretation**: SHAP force plots demonstrated how individual features contributed to a single patient's prediction, improving model transparency.

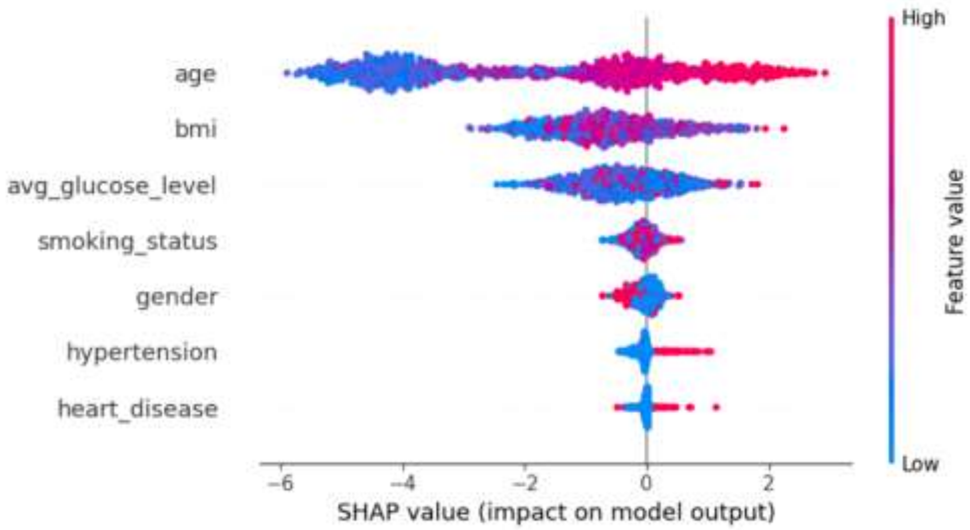

Fig 7 : SHAP Value

6. Discussion

The study confirmed that XGBoost outperforms Random Forest and Logistic Regression in both accuracy and AUC. SHAP analysis provided meaningful insights into model behavior, enhancing trust in predictions [19]. Age and glucose level were consistently the most influential features, aligning with clinical evidence [20].

These results suggest that ML models, when combined with explainability tools like SHAP, can serve as robust decision support systems in healthcare.

7. Conclusion

This work presents an interpretable machine learning framework for stroke prediction. By evaluating multiple models and interpreting them with SHAP, the study bridges the gap between predictive performance and clinical transparency. XGBoost, in particular, demonstrated excellent predictive power and interpretability.

8. Future Work

- Incorporate real-time patient monitoring data.
- Develop a web-based application for clinical use.
- Extend the study to multi-class stroke subtype classification.
- Validate the model on larger and more diverse datasets.

9. References

1. Kaggle Stroke Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
2. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.
3. Pedregosa et al., (2011). Scikit-learn: Machine Learning in Python.
4. Chen, T., &Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.
5. SHAP Documentation: https://shap.readthedocs.io
6. WHO Stroke Statistics https://www.who.int/news-room/fact-sheets/detail/stroke
7. AI in Healthcare Overview Esteva, A. et al. (2019). A guide to deep learning in healthcare. Nature Medicine https://doi.org/10.1038/s41591-018-0316-z
8. Kaggle Stroke Prediction Dataset https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
9. Label Encoding & Data Cleaning (Scikit-learn Docs) https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features
10. Breiman, L. (2001). Random Forests. Machine Learning https://doi.org/10.1023/A:1010933404324
11. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System https://doi.org/10.1145/2939672.2939785
12. Hosmer, D. W., et al. (2013). Applied Logistic Regression https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473
13. Pedregosa et al., (2011). https://scikit-learn.org/stable/
14. Model Evaluation Metrics (Accuracy, AUC, F1-score) https://scikit-learn.org/stable/modules/model_evaluation.html
15. ROC-AUC Explanation: https://en.wikipedia.org/wiki/Receiver_operating_characteristic
16. SHAP Official Documentation https://shap.readthedocs.io

17. Original SHAP Paper: Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
18. Interpretable ML Book by Christoph Molnar https://christophm.github.io/interpretable-ml-book/
19. Stroke Risk Factors (Feigin et al., 2021) https://doi.org/10.1016/S1474-4422(21)00252-0
20. ML Model Comparison for Stroke (Ghosh et al., 2022) https://doi.org/10.1016/j.compbiomed.2022.105356
21. Clinical AI Applications (Nature Medicine) https://doi.org/10.1038/s41591-018-0316-z
22. Real-time Monitoring with AI ,Rajpurkar, P. et al. (2017). CheXNet: Radiologist-level pneumonia detection. https://arxiv.org/abs/1711.05225