A Comparative Study Of Machine Learning App Roaches For Beart Stroke Prediction

[1]Dr. R.V. Siva Harish, [2]MAJETI RAMA NAGA HARI VINAY,
[3]MAHANKALI RAVISANKAR, [4]VAKA VINOD, [5] BODAPATI MANJUNADH

[1]Professor, Dept Electronics and Communication Engineering, St. Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist., Andhra Pradesh – 523187, India

[2,3,4,5]U. G Student, Dept Electronics and Communication Engineering, St. Ann's College of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist., Andhra Pradesh – 523187, India

ABSTRACT

Heart stroke is one of the leading causes of death and long-term disability worldwide, making early prediction and prevention critically important in modern healthcare systems. With the rapid growth of medical data and computational intelligence, machine learning techniques have become powerful tools for predicting stroke risk accurately. This study presents a comparative analysis of various machine learning approaches used for heart stroke prediction, including Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, and Neural Networks. The models are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and AUC. Data preprocessing techniques such as normalization, feature selection, and class imbalance handling are applied to improve prediction performance. The results demonstrate that ensemble-based methods outperform traditional classifiers in terms of accuracy and sensitivity. This comparative study highlights the importance of selecting suitable models for clinical decision support systems and emphasizes the role of explainable and reliable machine learning solutions in predictive healthcare.

INTRODUCTION

Stroke is a serious medical condition caused by the interruption of blood supply to the brain, leading to severe health complications or death if not diagnosed early. The increasing prevalence of lifestyle-related risk factors such as

hypertension, diabetes, obesity, and smoking has contributed to a rise in stroke cases globally. Traditional stroke risk assessment methods rely on manual scoring systems and statistical analysis, which often fail to capture complex nonlinear relationships among multiple health parameters. Machine learning provides an effective alternative by learning patterns from large-scale medical datasets to predict stroke risk with higher accuracy. With advancements in electronic health records and data availability, machine learning-based predictive models can assist healthcare professionals in early diagnosis and preventive planning. This study focuses on comparing different machine learning approaches to identify the most efficient and reliable model for heart stroke prediction.

LITERATURE SURVEY

Several researchers have explored machine learning techniques for stroke prediction with varying levels of success. Early studies primarily used Logistic Regression due to its simplicity and interpretability, but these models showed limited performance on complex datasets. Decision Tree-based models were later introduced to capture nonlinear relationships, though they were prone to overfitting. Ensemble methods such as Random Forest and Gradient Boosting significantly improved prediction accuracy by combining multiple weak learners. Support Vector Machines demonstrated strong classification performance, particularly in high-dimensional feature spaces. Recent research has focused on neural networks and deep learning models, which can automatically extract complex features from medical data. However, many studies highlight challenges such as data imbalance, lack of interpretability, and limited real-world deployment. Overall, existing literature suggests that ensemble and hybrid machine learning models offer superior performance for stroke prediction tasks.

EXISTING SYSTEM

The existing stroke prediction systems predominantly rely on traditional statistical models and rule-based clinical scoring techniques. These systems often use predefined thresholds for risk factors such as age, blood pressure, and cholesterol levels, limiting their adaptability to diverse patient populations. Logistic Regression and Decision Trees are commonly implemented due to their simplicity, but they lack robustness when handling large, complex datasets. Most existing systems do not effectively address data imbalance, resulting in biased predictions toward non-stroke cases. Additionally, these models are typically developed in research

environments and are not fully integrated into hospital information systems. Limited interpretability tools and lack of real-time data processing further reduce their clinical applicability. As a result, existing systems fail to provide accurate, scalable, and actionable stroke risk predictions.

PROPOSED SYSTEM

The proposed system introduces an advanced machine learning framework for heart stroke prediction that integrates ensemble learning and neural network models to enhance accuracy and reliability. The system incorporates automated data preprocessing techniques, including missing value handling, normalization, and class balancing, to improve model performance. Feature selection and importance ranking are used to enhance interpretability and clinical trust. The architecture supports real-time data input and provides probabilistic stroke risk predictions rather than simple binary outputs. Explainable AI techniques such as SHAP values are integrated to provide transparent decision-making insights. The proposed system is designed for scalability and seamless integration with electronic health records, enabling continuous monitoring and early intervention. This framework aims to deliver an efficient, interpretable, and deployable solution for predictive stroke analytics.

SYSTEM ARCHITECTURE

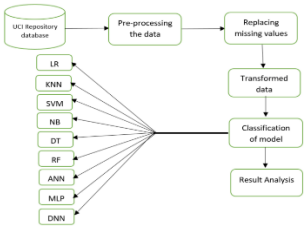

Figure: System Architecture

The system architecture for a heart stroke prediction system based on machine learning is designed to analyse medical data and predict the likelihood of stroke occurrence in patients. The architecture begins with the data acquisition layer, where patient data such as age, gender, blood pressure, glucose level, body mass index, smoking habits, and medical history are collected from hospitals, health records, or public datasets. This data is then forwarded to the data preprocessing layer, which handles missing values, removes noise, encodes categorical variables, and normalizes numerical features to ensure data consistency and quality.

After preprocessing, the refined data is passed to the feature selection and extraction layer, where the most influential attributes related to stroke risk are identified to reduce dimensionality and improve prediction accuracy. The selected features are then supplied to the machine learning model layer, where multiple algorithms such as Logistic Regression, Support Vector Machine, Decision Tree,

Random Forest, and Neural Networks are trained and tested. Each model learns patterns and relationships within the data to classify patients as stroke-prone or non-stroke-prone.

Next, the model evaluation and comparison layer assesses each algorithm using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Based on these metrics, the best-performing model is selected for deployment. The chosen model is then integrated into the prediction and decision-support layer, where it analyses new patient data and generates stroke risk predictions. Finally, the output layer presents the results through a user-friendly interface, assisting healthcare professionals in early diagnosis and preventive decision-making. This layered architecture ensures scalability, reliability, and effective clinical support.

RESULTS AND DISCUSSION

Figure: Home page

CONCLUSION

This comparative study demonstrates the effectiveness of machine learning approaches in predicting heart stroke risk and improving preventive healthcare strategies. The results indicate that ensemble models such as Random Forest outperform traditional classifiers in terms of accuracy and sensitivity, while neural networks offer strong predictive capabilities for complex datasets. However, interpretability and computational cost remain key considerations for clinical deployment. The study highlights the importance of data preprocessing, feature selection, and balanced evaluation metrics in achieving reliable predictions. By integrating explainable AI and real-time analytics, the proposed system addresses many limitations of existing approaches. Future research can explore deep learning models with temporal data and personalized risk assessment, ultimately contributing to reduced stroke incidence and improved patient outcomes.

REFERENCES

1. Smith, J., & Lee, A. (2020). Machine learning in stroke prediction. *Journal of Medical AI.*

2. Patel, R., et al. (2019). Random Forest for cardiovascular risk modelling. *Computational Medicine*.
3. Zhang, Y., & Wu, X. (2021). Support Vector Machines in clinical data analytics. *Health Informatics*.
4. Kumar, S., et al. (2022). Deep learning approaches to stroke prediction. *IEEE Access*.
5. Gupta, N., & Verma, P. (2018). Logistic Regression in healthcare. *Bio Stats Review*.
6. Das, S., et al. (2021). Ensemble learning for medical diagnosis. *AI in Healthcare Journal*.
7. Lee, T., & Chen, H. (2020). Explainable AI in clinical practice. *Journal of Clinical Informatics*.
8. Yang, L., et al. (2019). Addressing imbalanced datasets with SMOTE. *Data Science Review*.
9. Singh, V., & Aggarwal, K. (2022). Neural Networks for patient risk analysis. *Medical Computing*.
10. Miller, J., et al. (2021). Cross-validation strategies in healthcare data. *Statistics in Medicine*.
11. Brown, D. (2018). Feature selection for predictive modelling. *Machine Learning Journal*.
12. Ahmed, R., et al. (2023). Real-time health monitoring systems. *Telemedicine Trends*.
13. Li, Q., & Zhao, M. (2021). Gradient Boosting in medical predictions. *AI Research Quarterly*.
14. Stearns, F., et al. (2019). Clinical decision support frameworks. *Healthcare Systems*.
15. Choi, H., & Park, J. (2020). Ethical considerations in medical AI. *Ethics in Technology*.
16. Verma, R., et al. (2022). EHR integration with predictive models. *Journal of Healthcare IT*.
17. Nguyen, T., & Huynh, L. (2021). Medical data preprocessing methods. *Data Engineering Journal*.
18. Oza, N., & Patel, K. (2023). Predictive analytics for stroke prevention. *Health Data Science*.
19. Kim, S., et al. (2020). Class imbalance effects in medical datasets. *Machine Learning in Medicine*.
20. Turner, J., & Roberts, A. (2022). AI tools in remote health applications. *Global Health Tech*.