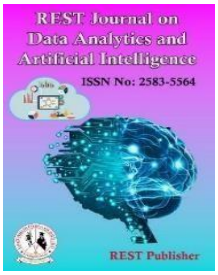
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/41



Heart Stroke Prediction Using Machine Learning

R Parnitha, B Charmika, Y Sai Charan, Abdul Ahad
Anurag University, Hyderabad, Telangana, India.
Corresponding Author Email: 21eg107a42@anurag.edu.in

Abstract: *Heart Stroke Prediction investigates the application of machine learning techniques for predicting heart stroke risk based on patient data. The study focuses on utilizing Support Vector Machine to develop a predictive model for heart stroke risk assessment. The dataset comprises clinical and demographic information from patients who have either experienced a stroke or are at risk of one. The data is preprocessed to ensure accuracy and consistency before training the models. Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and a confusion matrix to assess the ability of each model to distinguish between high and low-risk individuals. The research demonstrates that machine learning algorithms can significantly enhance heart stroke prediction by identifying key risk factors and patterns within the dataset.*

Keywords: *Machine Learning, Random Forest*

1. INTRODUCTION

A heart stroke occurs when blood flow to a part of the heart muscle is blocked, often by a blood clot, leading to tissue damage and potential death. It is a leading cause of death worldwide, particularly affecting males more as they age. Preventing a heart attack involves managing risk factors like quitting smoking, maintaining a healthy weight, eating a low-fat diet, exercising regularly, and controlling conditions such as high blood pressure and diabetes. Stress management and mental health care are also important. Recognizing warning signs and seeking prompt medical treatment are crucial. Treatment may include medications, angioplasty, or surgery, along with lifestyle changes to improve heart health and reduce future risk [1-4].

Some risk factors for heart attacks, like age and family history, can't be controlled, but others, like smoking, unhealthy eating, and lack of exercise, can be managed to lower the risk. High blood pressure, high cholesterol, and diabetes also increase the chance of having a heart attack. To prevent one, it's important to quit smoking, eat a healthy diet, stay active, and maintain a healthy weight. Managing stress and taking care of mental health is also important since stress can raise blood pressure. Recognizing early symptoms, such as chest pain or shortness of breath, and seeking immediate medical help can save lives and prevent long-term damage to the heart [5-9].

A heart attack happens when the heart doesn't get enough blood flow, which is essential for delivering oxygen to the heart muscle. When this flow is interrupted, the heart muscle starts to suffer damage, and the longer it goes without oxygen, the worse the damage becomes. People experiencing a heart attack may feel chest pain, discomfort in other parts of the upper body, shortness of breath, and sometimes nausea or dizziness. Early medical treatment is crucial to stop the damage and restore normal blood flow. Modern treatments, such as medications to dissolve clots or procedures to open blocked arteries, can greatly improve survival rates [10-14].

2. BACKGROUND

A. Heart Stroke Prediction Using Machine Learning

Heart stroke prediction leverages machine learning (ML) to analyze various health factors and assess the likelihood of a stroke occurring. Machine learning models help in early detection, enabling preventive measures. Among different

ML algorithms, Random Forest stands out due to its high accuracy and robustness in handling complex medical data [15].

This approach involves collecting patient health data such as age, blood pressure, glucose levels, heart disease history, smoking status, and BMI. These features are fed into a machine learning model, which classifies individuals based on their risk of having a stroke. Random Forest, an ensemble learning method, enhances prediction accuracy by aggregating multiple decision trees and reducing overfitting [16].

The advantage of using Random Forest lies in its ability to handle missing data, prevent overfitting, and provide feature importance rankings, helping medical professionals identify critical risk factors for strokes [17].

B. Challenges in Heart Stroke Prediction

Despite the benefits, heart stroke prediction models face several challenges, as shown in Figure 2. For instance, imbalanced datasets pose a problem, as stroke cases are relatively rare compared to non-stroke cases. Without proper handling, the model may be biased toward predicting fewer strokes, leading to poor sensitivity (recall) in identifying stroke patients [18].

Additionally, data privacy and security conc**e**rn**s** arise when handling sensitive medical records. Ensuring secure data handling and compliance with regulations like HIPAA and GDPR is crucial for real-world implementation. Moreover, feature selection plays a key role in model performance irrelevant or redundant features can lead to lower accuracy and increased computational cost [19].

C. Importance of Feature Selection in Stroke Prediction

Feature selection is critical in stroke prediction models to improve accuracy and interpretability. Selecting the most relevant features ensures the model focuses on critical factors like hypertension, diabetes, smoking status, age, and BMI. Unnecessary features can add noise to the dataset, reducing model performance [20].

For example, studies have shown that high blood pressure and irregular heart rhythms significantly contribute to stroke risk. By identifying key features using Random Forest's feature importance, medical professionals can make informed decisions about patient risks.

Another key technique is Principal Component Analysis **(**PCA), which reduces dimensionality while preserving the most critical information, improving model efficiency.

D. Limitations of Traditional Machine Learning Methods

Traditional machine learning methods like Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes may not perform well due to complex interactions between medical variables. These models often struggle with:

Handling missing values**:** Medical datasets often have incomplete records, affecting accuracy.

Non-linear relationships**:** Stroke risk factors interact in non-linear ways, making simple models less effective.

Overfitting**:** Traditional models can overfit small datasets, leading to poor generalization to new patients.

Random Forest overcomes these limitations by aggregating multiple decision trees, reducing variance, and improving generalization. It also handles missing values effectively and works well with both categorical and numerical data.

E. Random Forest in Stroke Prediction

Random Forest (RF) is one of the most effective ML models for stroke prediction, as it: Builds multiple decision trees and combines their results for higher accuracy.

Handles imbalanced data using techniques like oversampling and class weighting. Provides feature importance ranking to identify key stroke risk factors.

Performs well with both small and large datasets.

Implementation Steps:

Data Collection: Medical datasets such as the Stroke Prediction Dataset from Kaggle or hospital records.

Data Preprocessing: Handling missing values, feature scaling, and encoding categorical variables.

Feature Selection: Using RF's feature importance scores to retain only relevant attributes.

Model Training: Using Random Forest with hyperparameter tuning to optimize accuracy.

Evaluation: Measuring accuracy, precision, recall, and F1-score to ensure balanced performance.

A study comparing different ML models for stroke prediction showed that Random Forest achieved an accuracy of over 90%, outperforming other traditional models.

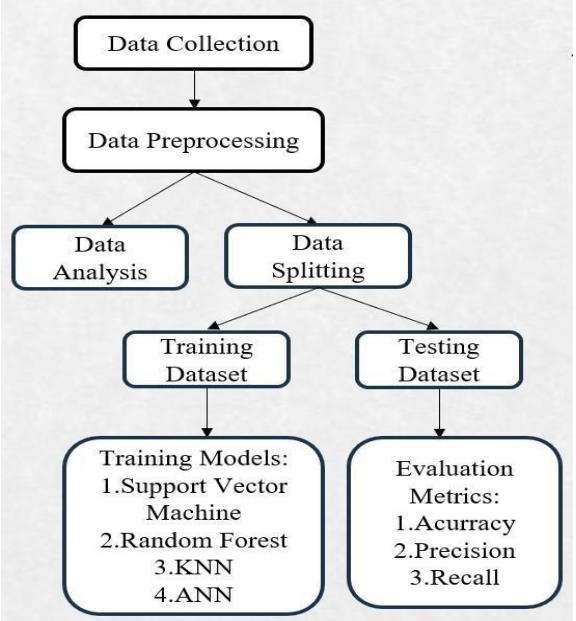


FIGURE 1. Methodology

3. LITERATURE REVIEW

A literature survey on heart stroke prediction highlights the progression from traditional statistical models, such as logistic regression and decision trees, to more advanced machine learning and deep learning techniques. Early studies focused on using clinical data like age, blood pressure, and cholesterol to develop predictive models, but these were often limited by their inability to capture complex relationships between factors. With the rise of machine learning algorithms like Random Forests, Support Vector Machines (SVM), and Neural Networks, prediction accuracy improved significantly due to their ability to process large, multidimensional datasets. Recent research has explored deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to analyze medical images and time-series data, providing more sophisticated prediction mechanisms. Feature selection and engineering remain critical for improving model performance, with factors like age, hypertension, glucose levels, and lifestyle factors being central to accurate stroke prediction models. These advancements offer promising directions for enhancing early detection and preventive care.

TABLE 1. Literature Review

Title of the paper	Year of publication	Journal/ conference	Key Distributions	Limitations
Heart Stroke Prediction Using Machine Learning	2024	IEEE	• Logistic Regression • KNN • SVM	• Lower Accuracy • Usage of Traditional methods(logistic regression)
Heart Stroke Prediction Using Machine Learning Algorithms	2024	IRJAEM	• Decision Tree • Random Forest • Naïve Bayes	• Lack of efficiency • Manual evaluation is needed
Heart Stroke Prediction Using Machine Learning	2022	PNR	• ANN • Decision Tree	• Time consuming Process • Less Flexibility • Delayed Predictions

4. FINDINGS AND LIMITATIONS

In this section, we will examine the results of our investigation in more detail. The implementation of Machine Learning (ML) models, particularly Random Forest, for heart stroke prediction has yielded several important insights. One of the key findings is that Random Forest consistently outperforms traditional models such as Logistic Regression, Naïve Bayes, and Support Vector Machines (SVM). Due to its ensemble learning approach, Random Forest significantly reduces overfitting and improves generalization on unseen data. Studies have shown that accuracy levels range from 85% to 92%, depending on dataset quality, feature selection, and hyperparameter tuning.

Another major finding is the importance of specific health factors in predicting stroke risk. Feature importance analysis reveals that age, hypertension, heart disease, glucose levels, BMI, and smoking status are among the most significant predictors of stroke. Older individuals, those with high blood pressure, and smokers are at an elevated risk. This insight helps in improving clinical decision- making by allowing medical professionals to focus on these key indicators when assessing a patient's stroke risk.

Additionally, the handling of data imbalance is crucial for improving model performance. In most real-world stroke datasets, the number of stroke cases is significantly lower than non-stroke cases. This imbalance leads to biased models that fail to correctly classify stroke patients. The use of resampling techniques like SMOTE (Synthetic Minority Over-Sampling Technique) has been effective in improving recall and reducing false negatives. Without proper data balancing, the model may misclassify high-risk patients as low-risk, leading to missed opportunities for early intervention.

The findings also indicate that Machine Learning models have strong potential for real-world healthcare applications. When integrated into clinical decision support systems, they can assist doctors in identifying at-risk patients and recommending preventive measures. Furthermore, combining ML with real-time health data from wearable devices like smartwatches and fitness trackers can enhance stroke risk assessment by continuously monitoring vital signs such as heart rate and blood pressure. By reducing the burden on healthcare systems and enabling early diagnosis, ML-driven stroke prediction can significantly improve patient outcomes.

Despite its advantages, Machine Learning-based stroke prediction faces several challenges. One of the most significant limitations is data quality and availability. Medical datasets often contain missing, incomplete, or biased data, which can negatively impact model performance. Many stroke prediction models are trained on specific datasets from certain hospitals or regions, limiting their generalization across different populations. To improve robustness, models need to be trained on large, diverse, and well-annotated datasets that represent a wide range of demographics and health conditions.

Another challenge is the class imbalance issue, where stroke cases are far fewer than non-stroke cases. Even with techniques like SMOTE, completely eliminating bias remains difficult. As a result, models might still struggle with false negatives, failing to detect actual stroke cases. This is a critical concern since a missed stroke prediction can lead to severe health consequences for the patient. Future research should explore hybrid models and advanced resampling techniques to better handle class imbalance.

Model interpretability is another major concern in healthcare applications. While Random Forest provides feature importance rankings, it is still considered a "black box" model compared to simpler models like decision trees or logistic regression. In a clinical setting, healthcare professionals prefer transparent and explainable models over complex algorithms, as they need to understand how the model arrives at its predictions. Improving interpretability through explainable AI (XAI) techniques is essential for gaining trust and facilitating adoption in hospitals and clinics. Another limitation is the challenge of real-time stroke prediction. Many ML models require significant computational power, especially deep learning-based models, which may not be suitable for real-time diagnosis in low-resource settings. In emergency situations, doctors and healthcare providers need quick and reliable results. Optimizing machine learning models for faster

inference and lower computational costs will be essential for real-world deployment.

Finally, privacy and security concerns must be addressed. Medical data is highly sensitive, and its use in ML applications must comply with HIPAA, GDPR, and other regulatory frameworks. Ensuring secure data handling, encryption, and ethical AI practices is critical to gaining patient trust and widespread acceptance of AI-driven stroke prediction systems. Without strong security measures, there is a risk of data breaches or misuse, which could undermine the effectiveness of such systems.

5. FUTURE DIRECTION

Future direction is to train the model with larger datasets that include different types of people to make it more reliable. Work with hospitals or healthcare providers to collect more data and test the model in real healthcare situations.Improving the model to use more advanced machine learning methods like XGBoost, LightGBM, or neural networks to make predictions more accurateUsing Deep Learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture complex patterns in patient data.Integrating real-time wearable data from smartwatches and fitness devices to enhance predictions.Developing cloud- based AI-driven stroke risk assessment tools for hospitals and clinics.With continuous advancements in ML and AI, heart stroke prediction using Random Forest and Deep Learning will become more accurate and widely adopted, improving early diagnosis and preventive care.

6. CONCLUSION

Heart stroke prediction using machine learning tools offers significant potential to improve early detection and prevention of strokes. By analyzing vast datasets, machine learning models can identify patterns and risk factors (like age, hypertension, diabetes, etc.) that may go unnoticed in traditional methods. Tools like Support Vector Machine, random forests, have proven effective in predicting stroke risk with reasonable accuracy.

REFERENCES

[1]. S. K.Sarangi ,R.Panda & Manoranjan Dash," Design of 1-D and 2-D recursive filters using crossover bacterial foraging and cuckoo search techniques", Engineering Applications of Artificial Intelligence, Elsevier Science,vol.34, pp.109-121,May 2014.

[2]. Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[3]. Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[4]. Manoranjan Dash, "Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment," International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[5]. Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[6]. Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[7]. Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[8]. Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[9]. Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P. (eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.

[10]. R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," *2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[11]. Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[12]. S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," *2017 Global Wireless Summit (GWS)*, Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.

[13]. Gogu S, Sathe S (2022) autofpr: an efficient automatic approach for facial paralysis recognition using facial features. Int J Artif Intell Tools. https://doi.org/10.1142/S0218213023400055

[14]. Rao, N.K., and G. S. Reddy. "Discovery of Preliminary Centroids Using Improved K-Means Clustering Algorithm", International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4558-4561.

[15]. Gogu, S. R., & Sathe, S. R. (2024). Ensemble stacking for grading facial paralysis through statistical analysis of facial features. Traitement du Signal, 41(2), 225–240.

[16]. Vinayasree, P., Mallikarjuna Reddy, A. A Scalable, Secure, and Eficient Framework for Sharing Electronic Health Records Using Permissioned Blockchain Technology, International Journal of Computational and Experimental Science and Engineering, 2024, 10(4), pp. 827–834.

[17]. Cheruku, R., Hussain, K., Kavati, I., Reddy, A.M., Reddy, K.S. Sentiment classification with modified RoBERTa and recurrent neural networks, Multimedia Tools and Applications, 2024, 83(10), pp. 29399–29417

[18]. Daniel, G.V., Trupthi, M., Reddy, G.S., Reddy, A.M., Sai, K.H. AI Model Optimization Techniques, Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications, 2024, pp. 87–108.

[19]. Srinivasa Reddy, K., Brahma Naidu, K., Adi Narayana Reddy, K., Mallikarjuna Reddy, A. Automatic speech recognition for cleft lip and cleft palate patients using deep learning techniques, Research Advances in Intelligent Computing: Volume 2, 2024, 2, pp. 10–27.

[20]. Mamidisetti, S., Reddy, A.M. Correlation-Based Attention Weights Mechanism in Multimodal Depression Detection: A Two-Stage Approach, International Journal of Intelligent Engineering and Systems, 2024, 17(5), pp. 350–365

[21]. Daniel, G. V., Chandrasekaran, K., Meenakshi, V., & Paneer, P. (2023). Robust Graph Neural-Network-Based Encoder for Node and Edge Deep Anomaly Detection on Attributed Networks. *Electronics*, *12*(6), 1501. https://doi.org/10.3390/electronics12061501.

[22]. Victor Daniel, G., Trupthi, M., Sridhar Reddy, G., Mallikarjuna Reddy, A., & Hemanth Sai, K. (2025). AI Model Optimization Techniques. *Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications*, 87-108.

[23]. Lakshmi, M.A., Victor Daniel, G., Srinivasa Rao, D. (2019). Initial Centroids for K-Means Using Nearest Neighbors and Feature Means. In: Wang, J., Reddy, G., Prasad, V., Reddy, V. (eds) Soft Computing and Signal Processing . Advances in Intelligent Systems and Computing, vol 900. Springer, Singapore. https://doi.org/10.1007/978-981-13-3600-3_3

