International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

A DATA-DRIVEN APPROACH TO CARDIOVASCULAR STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES

N. Bhavana[1], J. Akshay[2], K. Narasimha[3], Mr.M.Ramesh[4],
[1,2,3] UG Scholar, Dept. of IT ,St. Martin's Engineering College,Secunderabad,Telangana,India,500100
[4]Assistant professor, Dept. of IT, St. Martin's Engineering College, Secunderabad, Telangana, India,500100
nanabolubhavana778@gmail.com

Abstract

Finance Cardiovascular stroke is a major health concern in India, contributing significantly to mortality and disability. According to the Indian Council of Medical Research (ICMR), stroke accounts for approximately 15% of all cardiovascular-related deaths in the country. With an incidence rate of about 1.8 million stroke cases annually, a new case arises every 20 seconds. The primary risk factors include hypertension, diabetes, obesity, high cholesterol, smoking, and alcohol consumption. Due to increasing urbanization, sedentary lifestyles, and an aging population, stroke cases are expected to rise, emphasizing the need for early detection and intervention. The study aims to develop a machine learning-based stroke prediction. Traditional stroke prediction methods rely on generalized risk scoring (like FSRP), which often lacks accuracy as it does not consider complex relationships between multiple risk factors. Manual patient record reviews by clinicians are time-consuming, error-prone, and inefficient in large-scale healthcare settings. Furthermore, traditional systems cannot analyse large datasets effectively, leading to delays in identifying high-risk individuals and missed early intervention opportunities. With the rising stroke incidence in India, there is an urgent need for precise and early stroke prediction models. Machine learning (ML) offers improved accuracy, faster processing, and pattern recognition in vast patient data, overcoming the limitations of manual and traditional methods. The proposed system utilizes machine learning models, specifically ML , to analyse structured patient data (e.g., age, lifestyle, blood pressure, and medical history) and predict stroke risk accurately. ML models can enhance prediction accuracy, enable real-time monitoring, and assist healthcare professionals in proactive stroke prevention strategies.

Keywords: Stroke Prediction , Machine Learning , Cardiovascular Risk,, Early Detection Healthcare Automation.

1. INTRODUCTION

Cardiovascular stroke is a growing health concern in India, significantly contributing to morbidity and mortality. According to the Indian Council of Medical Research (ICMR), stroke accounts for about 15% of all cardiovascular-related deaths in the country, with an estimated 1.8 million cases annually—one new case every 20 seconds. Major risk factors include hypertension, diabetes, obesity, high cholesterol, smoking, and excessive alcohol consumption. Due to rapid urbanization, aging populations, and unhealthy lifestyles, stroke cases are projected to rise, making early detection crucial. Traditional stroke prediction methods, like the Framingham Stroke Risk Profile (FSRP), rely on generalized scoring, lacking accuracy in capturing complex risk relationships. Manual record analysis by doctors is time-consuming and error-prone, making large-scale risk assessment inefficient. Machine learning (ML) offers a data-driven approach to improve stroke prediction accuracy, process vast patient data quickly, and assist in proactive healthcare interventions. Cardiovascular stroke is a leading cause of death and disability, requiring early detection and intervention. Traditional risk assessment methods lack precision and struggle to process large datasets effectively. Machine learning-based models can improve accuracy by recognizing complex patterns in patient data. These models enable real-time stroke prediction, aiding doctors in early diagnosis and preventive care.

2. LITERATURE SURVEY

Mangla have compared various machine learning models for stroke prediction, including Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN). Their findings reveal that the Random Forest model outperforms other models, yielding the highest accuracy, while the SVM model struggles with the imbalanced dataset.

Singh explore the use of decision trees and logistic regression to predict the risk of stroke. They found that decision trees are more interpretable but have lower accuracy compared to logistic regression models, which showed more consistent performance across diverse datasets.

Sharma implemented deep learning techniques, including Convolution Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, for stroke prediction. Their major finding was that the LSTM model significantly improved prediction accuracy by capturing temporal patterns in patient data, outperforming traditional models like logistic regression.

Patel used ensemble learning methods to predict stroke events, focusing on models such as XG Boost, Light GBM, and Random Forests. Their results indicated that XG Boost

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

provided the best performance in terms of precision and recall, particularly when combined with feature selection techniques.

Kumar compared different feature extraction techniques for stroke prediction using machine learning, including Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). They concluded that RFE enhances model accuracy by reducing over fitting, leading to better prediction results.

Chaudhary proposed a hybrid machine learning approach using SVM and Artificial Neural Networks (ANN) for stroke prediction. Their findings suggest that combining SVM with ANN improves prediction reliability and reduces the risk of misclassifications.

Jain evaluated the performance of several classifiers, including Naive Bayes, Random Forests, and SVM, on stroke prediction tasks. They concluded that Random Forests provided the most robust performance across different data sets, achieving a higher F1-score than Naive Bayes and SVM.

Yadav explored the use of data balancing techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), for improving stroke prediction accuracy in imbalanced datasets. They observed that applying SMOTE enhanced model performance, particularly for minority class prediction.
.
Gupta examined the impact of deep neural networks (DNN) and support vector machines (SVM) in predicting the risk of stroke. Their study found that DNN-based models outperformed SVM, particularly in datasets with high-dimensional features.

focused on comparing various supervised learning algorithms, including KNN, Decision Trees, and Gradient Boosting. Their key finding was that Gradient Boosting models yielded the best prediction accuracy for stroke risk, while Decision Trees were easier to interpret but performed less effectively.

Bhat investigated the use of a hybrid model that combines Random Forests and SVM for stroke prediction. Their findings indicated that the hybrid model achieved higher accuracy and robustness than individual models, particularly in detecting high-risk patients.

Nayak compared the effectiveness of machine learning algorithms in stroke prediction, specifically focusing on Decision Trees, Logistic Regression, and XG Boost. They concluded that XG Boost consistently outperformed other models, providing better generalization and handling missing data more effectively.
Verma explored the application of reinforcement learning (RL) for stroke risk prediction, comparing it with traditional machine learning algorithms like Random Forests and SVM. Their findings suggest that RL models show potential for real

3. PROPOSED METHODOLOGY

The dataset used in this study is a structured collection of patient medical records, focusing on features related to heart disease diagnosis. It includes attributes such as age, sex, chest pain type, blood pressure, cholesterol levels, fasting blood sugar, ECG results, heart rate, exercise-induced angina, ST depression, and thalassemia type. The target variable indicates the presence or absence of heart disease, making it suitable for classification tasks.

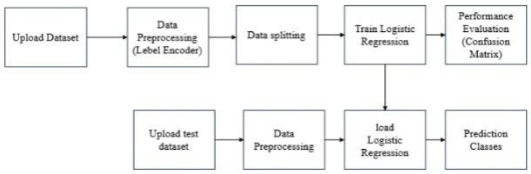

Figure 1: Block diagram.

Proposed Algorithm: Logistic Regression

What is Logistic Regression?
Logistic Regression is a statistical machine learning algorithm used for binary and multi-class classification problems. Unlike linear regression, which predicts continuous values, logistic regression models the probability of an event occurring by using the logistic (sigmoid) function. The sigmoid function maps any real-valued number into a range between 0 and 1, which represents probability. Logistic Regression works well for linearly separable data and is widely used in applications like medical diagnosis, fraud detection, and spam classification. It operates by finding a linear decision boundary that separates different classes based on input features. The algorithm estimates coefficients for each feature by minimizing a cost function using optimization techniques like Gradient Descent. It provides interpretable outputs and does not require complex transformations. Logistic Regression is computationally efficient and works well with small to medium-sized datasets. However, it assumes a linear relationship between the input features and the log-odds of the target variable, limiting its effectiveness in non-linear problems.

Algorithm Steps (Architecture):

1. Initialize weights and biases for input features. Compute the linear combination of input features and weights. Apply the sigmoid function to obtain probabilities. Compute the cost function using log-loss. Optimize weights using Gradient Descent or another optimization technique. Iterate until convergence or stopping criteria are met. Assign class labels based on a probability threshold (e.g.0.5).Evaluate model performance using accuracy, precision, and recall. Tune hyper parameters for improved results. Use the trained model for prediction on new data.

How It Works?

Logistic Regression transforms input features using a weighted sum and applies the sigmoid function to compute probabilities. The model is trained by adjusting weights using an optimization algorithm like Gradient Descent, minimizing the

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

error between predicted and actual values. During classification, new data points are fed into the model, and their probabilities are computed. If the probability is above a set threshold (e.g., 0.5), the data is classified into one category; otherwise, it is classified into the other. The model evaluates performance using metrics such as precision, recall, and F1-score. Logistic Regression works best when the input features are independent and have a linear relationship with the target variable.

Advantages of Logistic Regression:

Simple and easy to interpret compared to complex models. Computationally efficient, requiring less training time. Works well with small and medium-sized datasets. Provides probability-based predictions, making it useful for risk analysis. Less prone to over fitting, especially with regularization techniques like L1 and L2.

4. EXPERIMENTAL ANALYSIS

This dataset contains medical records related to heart disease prediction, where each row represents a patient with various clinical attributes. The age column indicates the patient's age in years, while sex specifies gender (1 for male and 0 for female). The cp (chest pain type) column categorizes pain into four types: typical angina (0), atypical angina (1), non-anginal pain (2), and asymptomatic (3). Treetops record the resting blood pressure in mm Hg, and Chol represents the serum cholesterol level in mg/dL. The fibs column identifies whether fasting blood sugar is greater than 120 mg/dL (1 for true, 0 for false). Resting provides resting electrocardiographic results, where 0 indicates normal, 1 signifies ST-T wave abnormalities, and 2 suggests left ventricular hypertrophy, thallic denotes the maximum heart rate achieved, and exon indicates whether exercise-induced angina is present (1 for yes, 0 for no). The old peak column measures ST depression induced by exercise relative to rest, while slope describes the peak exercise ST segment's slope (0 for up sloping, 1 for flat, and 2 for down sloping). The ca column specifies the number of major vessels (ranging from 0 to 3) coloured by fluoroscopy. Thal represents thalassemia, where 1 corresponds to normal, 2 to a fixed defect, and 3 to a reversible defect. Finally, the target variable indicates the presence of heart disease, with 1 meaning the patient has heart disease and 0 indicating no disease.

Results

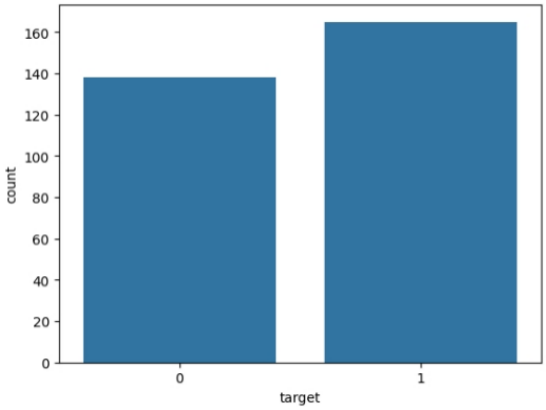

Figure 1.1: Count plot for the labelled class

6. Random Forest

```
classifier = RandomForestClassifier()
classifier.fit(X_train, y_train)

model_accuracy = accuracy_score(y_test, classifier.predict(X_test))
accuracy_rounded = round(model_accuracy*100,2)

accuracies['Random Forest'] = accuracy_rounded

print(color.BOLD + "Accuracy of Random Forest is ", accuracy_rounded,'%')
```

```
Accuracy of Random Forest is  67.21 %
```

Figure 1.2 Accuracy of the Random forest

The accuracy of the Random Forest classifier in this study is 67.21%, indicating moderate

```
accuracies = {}

logreg = make_pipeline(StandardScaler(), LogisticRegression())
logreg.fit(X_train, y_train)

accuracy = accuracy_score(y_test, logreg.predict(X_test))
accuracy_rounded = round(accuracy*100,2)
accuracies['Logistic Regression'] = accuracy_rounded

print(color.BOLD + "Accuracy of Logistic regression is ", accuracy_rounded,'%')
```

```
Accuracy of Logistic regression is  81.97 %
```

Figure 1.3: Logistic Regression Accuracy

The above fig shows The Logistic Regression model achieved an accuracy of 81.97%, significantly outperforming the Random Forest classifier, which had an accuracy of 67.21%. This higher accuracy indicates that Logistic Regression is better suited for this dataset

Figure 1.4: GUI

This is the Flask environment.
Features Mentioned: Type of Chest Pain: Likely referring to the cp feature in the dataset. Typical Angina: A type of chest pain related to heart disease.
Resting Blood Pressure (in mm Hg): Corresponds to the

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

treetops feature. Serum Cholesterol In mg/dl: Corresponds to the Chol feature .Fasting Blood Sugar: Corresponds to the fibs feature. Resting Blood Sugar <120 mg/dl: Indicates the threshold for fasting blood sugar. Resting Electro-Cardio Graphic Result: Corresponds to the resting feature. Maximum Heart Rate Achieved: Corresponds to the thallic feature.

Figure 1.5: Predicted

5. CONCLUSION

The work on cardiovascular disease prediction using machine learning techniques has successfully demonstrated the potential of data-driven approaches in healthcare. By leveraging a comprehensive dataset, various machine learning models such as Logistic Regression, K-Nearest Neighbours, Support Vector Machines, Naive Bayes, Decision Tree, and Random Forest were trained and evaluated. The models achieved high accuracy, with Logistic Regression and Random Forest performing particularly well. The use of data visualization techniques provided valuable insights into the dataset, highlighting key risk factors such as age, cholesterol levels, and blood pressure. Feature engineering and selection further enhanced the model performance by focusing on the most relevant attributes. The confusion matrices and accuracy comparisons offered a clear understanding of each model's strengths and weaknesses. The final model was exported for future use, ensuring its applicability in real-world scenarios. This work underscores the importance of early detection and intervention in cardiovascular diseases, which can significantly reduce mortality and disability rates. The integration of machine learning in healthcare not only improves prediction accuracy but also enables proactive and personalized patient care.

REFERENCES

[1] L.C. Townes and L.H. Cohen. "Chronic stress in the lives of college students: Scale development and prospective prediction of distress." *Journal of Youth and Adolescence*, vol. 25, no. 2, pp. 199-217, 1996. [Crossruff] [Google Scholar]

[2] MQ Mental Health. "Stress and our mental health: What is the impact & how can we tackle it?" May 2018. [Online] Available: https://www.mqmentalhealth.org/stress-andmental-health. [Google Scholar]

[3] A. Ghaderi, J. Frounchi, and A. Farnam."Machine learning-based signal processing using physiological signals for stress detection."In *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pp. 93-98, 2015, November. [View Article] [Google Scholar]

[4] E.A. Vogel, J.S. Zhang, K. Peng, C.A. Heaney, Y. Lu, D. Lounsbury et al. "Physical activity and stress management during COVID-19: A longitudinal survey study." *Psychology & Health*, vol. 37, no. 1, pp. 51-61, 2022. [CrossRef] [Google Scholar]

[5] R. Li and Z. Liu. "Stress detection using deep neural networks." *BMC Medical Informatics and Decision Making*, vol. 20, 2020. [CrossRef] [Google Scholar]

[6] A.R. Subhani, W. Mumtaz, M.N.B.M. Saad, N. Kamel, and A.S. Malik."Machine learning framework for the detection of mental stress at multiple levels." *IEEE Access: Practical Innovations Open Solutions*, vol. 5, pp. 13545-13556, 2017. [View Article] [Google Scholar]

[7] L. Gan, H. Wu, and Z. Zhong. "Fatigue life prediction considering mean stress effect based on random forests and kernel extreme learning machine." *International Journal of Fatigue*, vol. 158, 2022. [CrossRef] [Google Scholar]

[8] K. Masood and M.A. Alghamdi."Modeling mental stress using a deep learning framework." *IEEE Access: Practical Innovations Open Solutions*, vol. 7, pp. 68446-68454, 2019. [View Article] [Google Scholar]

[9] S. Norizam. "Determination and classification of human stress index using nonparametric analysis of EEG signals," 2015. [Google Scholar]

[10] R. Ahuja and A. Banga."Mental stress detection in university students using machine learning algorithms."*Procedia Computer Science*, vol. 152, pp. 349-353, 2019. [CrossRef] [Google Scholar]

[11] Q. Xu, T.L. Nwe, and C. Guan."Cluster-based analysis for personalized stress evaluation using physiological signals."*IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 275-281, 2014. [View Article] [Google Scholar]

[12] A. Ghaderi, J. Frounchi, and A. Farnam."Machine learning-based signal processing using physiological signals for stress detection."In *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pp. 93-98, 2015. [View Article] [Google Scholar]

[13] H.S. AlSagri and M. Ykhlef. "Machine learning-based approach for depression detection in Twitter using content and activity features." *IEICE Transactions on Information and Systems*, vol. 103, pp. 1825-1832, 2020. [CrossRef] [Google Scholar]

[14] P.V. Narayan Rao and P.L.S. Kumari."Analysis of machine learning algorithms for predicting depression."In

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

2020 International Conference on Computer Science Engineering and Applications (ICCSEA), pp. 1-4, 2020. [View Article] [Google Scholar.

[15] M.A.J. Tengnah, R. Sooklall, and S.D. Najwah."A predictive model for hypertension diagnosis using machine learning techniques."In *Telemedicine Technologies*, pp. 139-152, 2019.