e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

STROKE PREDICTION USING MACHINE LEARNING MODELS

P. Naresh[*1], S. Shreeya Reddy[*2], T. Ebenezar[*3], CH. Rajesh[*4]

[*1,2,3]Student Of B. Tech Computer Science And Engineering, Department Of Computer Science And Engineering, Malla Reddy College Of Engineering & Technology, Hyderabad, Telangana, India.

[*4]Assistant Professor, Department Of Computer Science And Engineering, Malla Reddy College Of Engineering & Technology, Hyderabad, Telangana, India.

ABSTRACT

Cardiovascular diseases (CVDs), including stroke and heart disease, remain leading causes of morbidity and mortality worldwide. Early identification of individuals at high risk of developing these conditions is crucial for preventive interventions and improving patient outcomes. In this study, we propose a machine learning-based approach for the prediction of stroke and heart disease risk.

The dataset utilized comprises a comprehensive set of demographic, clinical, and lifestyle factors collected from a diverse population sample. Various machine learning algorithms, including Decision Trees, Support Vector Machines (SVM), and Random Forest (RF), are employed to develop predictive models. Among these algorithms, RF stands out as it combines the strength of Random Forest with an iterative process enhancing model performance to 90% accuracy and interpret-ability.

This research contributes to advancing the field of cardiovascular risk assessment by leveraging machine learning techniques to develop accurate and interpretable predictive model. The proposed framework holds promise for enhancing early detection, risk stratification, and prevention of stroke and heart disease, ultimately leading to improved patient outcomes and reduced healthcare burden.

Keywords: Cardiovascular Disease, Stroke, Heart Disease, Machine Learning, Predictive Modeling, Random Forest, Risk Assessment, Healthcare.

I. INTRODUCTION

Cardiovascular diseases (CVD's) are the leading cause of global deaths, accounting for 17.9 million deaths in 2019, with heart attacks and strokes being responsible for 85% of them. Sadly, over three quarters of these deaths occur in low- and middle-income countries. The good news is that most CVD's can be prevented by addressing behavioral risk factors such as smoking, unhealthy diet, physical inactivity, and excessive alcohol consumption. Detecting CVD's early is crucial for timely management with counseling and medication.

To avoid heart attacks or strokes, it's essential to adopt a healthy lifestyle. This includes maintaining a balanced diet, engaging in regular physical activity, and avoiding tobacco products. Additionally, it's crucial to monitor and control risk factors like high blood pressure, high cholesterol, and diabetes.

Depression and anxiety can accelerate the development of risk factors for heart attacks and strokes, such as high blood pressure and high cholesterol. These mental health conditions also increase the risk of major cardiovascular events by about 35%. Cumulative stress can worsen heart and brain health by directly affecting physical well-being and promoting unhealthy behaviors like smoking and sedentary lifestyles.

A recent study conducted in rural America found that transgender individuals have a higher prevalence of cardiovascular disease risk factors compared to individuals. Transgender males, in particular, were at a significantly higher risk, with increased rates of tobacco use, obesity, and high blood pressure.

The American Heart Association and the National Institutes of Health provide annual updates on heart disease, stroke, and cardiovascular risk factors. These updates include data on key health behaviors (such as smoking, physical activity, diet, and weight) and health factors (like cholesterol, blood pressure, and glucose control). The reports also cover various clinical heart and circulatory conditions, along with associated outcomes and economic cost.

II. LITERATURE REVIEW

1. Using an optimized random forest model and random search algorithm, Ashir Javeed, Shijie Zhou, and colleagues (2017) created **"An Intelligent Learning System for Improved Heart Disease Detection.**" In order to select factors and diagnose cardiovascular illness, this article use the random forest model and the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

random search algorithm (RSA). The main purpose of this model's optimization is to use grid search algorithmic programming. Predicting cardiovascular disease involves two types of experiments. Only the random forest model is created in the first form; in the second experiment, a random forest model based on the Random Search Algorithm is created. Comparing this methodology to the traditional random forest model, it is less complex and more efficient. It generates 3.3% greater accuracy than traditional random forests. The suggested learning system can aid in the doctors' improvement.the accuracy of heart failure identification

2. **"Prediction and Diagnosis of Heart Disease by Data Mining Techniques**" was created by Mirsaeid Hosseini Shirvani and Boshra Bahrami. In order to diagnose cardiovascular disease, this paper employs multiple classification methodologies. The datasets are divided using classifiers such as KNN, SVO classifier, and Decision Tree. Following classification and performance assessment, the decision tree that performs the best at predicting cardiovascular disease from the dataset is analyzed.
3. Using a data mining technique, Mamatha Alex P and Shaicy P Shaji (2019) created **"Prediction and Diagnosis of Heart Disease Patients.**" This paper employs support vector machines, KNNs, random forests, and artificial neural networks.Neural Network Artificial
4. **"Predictive Model and Effective Analysis of Stroke Disease Using Classification Method**s"-Prayathri, A. Sudha, N. Jayasankar In order to reduce the number of dimensions and identify the characteristics that are more important for the prediction of stroke disease, the principal component analysis algorithm is employed in this research to determine whether or not the patient is experiencing a stroke.
5. "**A Study Using the National Health Insurance Database to Develop an Algorithm for Stroke Prediction**" -Min SN, Lee KS, Subramaniyam M, Park SJ, Kim DJ In this study, the model equation for creating an algorithm for pre-diagnosing strokes using possibly adjustable risk factors was derived.
6. **"Stroke Focus: Predicting and Preventing Stroke**" Regnier, Michael Modern stroke prevention

III. METHODOLOGY

Machine learning and predictive modeling techniques offer a more sophisticated approach to predicting stroke and heart disease risk by leveraging advanced algorithms to analyze data and identify patterns. Here's how machine learning and predictive modeling can be applied in this context:

Data Collection: Machine learning models require extensive datasets containing information on traditional risk factors (age, gender, hypertension, etc.) as well as additional variables

Stroke Prediction DataSet Attributes

Features	Description
Age	Age
Gender	Male and Female
Hypertension	Hypertension
Heart Disease	1 Has heart disease 0 Does not have heart disease
Ever_married	1 means Married 0 means Not married
Work_type	Children Private Never worked Govt job Self employed
Residence_type	Rural Urban
Avg_glucose_level	Average glucose level
bmi	Body mass index
smoking_status	Never smoked Formerly smoked

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

Heart Disease DataSet Attributes

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

Feature Selection: Machine learning algorithms can automatically identify relevant features (predictors) from the dataset that contribute most to the prediction of stroke and heart disease risk. This process helps in selecting the most informative variables for the model.

Model Development: Various machine learning algorithms can be employed for building predictive models, including logistic regression, decision trees, random forests, support vector machines, gradient boosting,These models learn from the data and develop relationships between risk factors and the likelihood of developing cardiovascular events.

Model Training: The predictive model is trained using historical data, where known outcomes (e.g., occurrence of strokes or heart disease) are used to teach the algorithm to recognize patterns and make accurate predictions.

Model Evaluation: Once trained, the model is evaluated using separate validation datasets to assess its performance. Common evaluation metrics include accuracy.

Optimization: Models may undergo optimization processes to improve their performance, such as hyperparameter tuning, feature engineering, and ensemble methods to combine multiple models for better predictive accuracy.

IV. EXISTING SYSTEM

The current landscape the Existing systems is only predicting the occurrence of heart disease using Support Vector machine,Logistic Regression, KNN, neural Networks. These Techniques is not giving the accurate measures because it is not considered the proper factors that is effecting. The algorithms used here are with accuracy less than 85% only. And the existing systems is not predict the future occurrence of any stroke to the patient due to blood clots because of heart disease irregular behavior. There were other existing system which were using deep learning but the dataset is not so large to apply the deep learning techniques for better result.

Problems with Existing Systems:

Limited Risk factors and bias which are not accurate of result. It may not predict the future occurrence of stroke and heart diseases.

V. PROPOSED SYSTEM

Even though there are many ways to identify Heart Disease and Stroke commonly rely on traditional machine learning algorithms like Naive Bayes, Support Vector Machines (SVM),Logistic Regressions. This proposal introduces Random Forest, Decision Tree, XGBoost, a more advanced ensemble learning algorithms. The proposed system aims to demonstrate the superior performance of those algorithms comprehensive analysis of its capabilities.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

Goal & Objectives of Proposed System:

The main goals of this model are to:

- **Provide High predictive accuracy**, it leads to better generalization and improved performance compared to existing systems.
- It will consider the **accurate risk factors** of the heart disease and stroke occurrence.

VI. TECHNOLOGY USED

Python: Python is a programming language that is easy to use and read. It has a vast library of libraries covering a wide range of topics, such as web development, data science, and machine learning. Because of its adaptability and simplicity, Python is highly preferred by developers.

Pandas: An vital tool for data scientists and analysts, Pandas is a Python library for manipulating and analyzing data. It provides simple-to-use data structures and tools for activities including importing, cleaning, transforming, and analyzing structured data.

A variety of supervised and unsupervised learning algorithms, as well as tools for model selection, evaluation, and preprocessing, are available in the **Scikit-learn** machine learning package, which is based on Python.

Based on **Matplotlib, Seaborn** is a Python visualization toolkit that offers a high-level interface for making visually appealing statistical visualizations with integrated color and themes. **Flask** for user interface

VII. ALGORITHMS

Logistic Regression:

Tailored for forecasting categorical outcomes, logistic regression distinguishes itself from linear regression by centering on classification issues.

Support Vector Machine (SVM):

Extensively employed in both classification and regression tasks, SVM strives to construct optimal decision boundaries in n-dimensional space to separate data points into discrete classes.

Decision Trees:

Suited for both classification and regression, decision trees delineate hierarchical decision rules based on input features, leading to easily understandable interpretations.

Random Forest:

A method of ensemble learning, random forest combines multiple decision trees to improve accuracy and combat overfitting by amalgamating predictions.

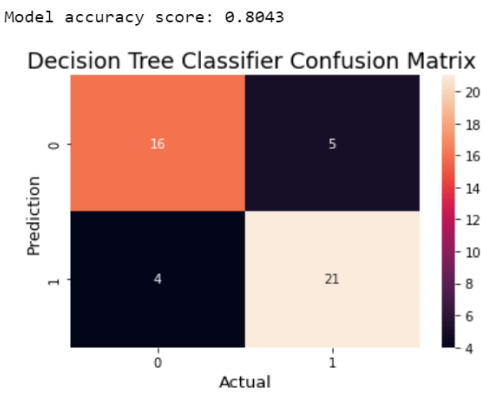
Gradient Boosting:

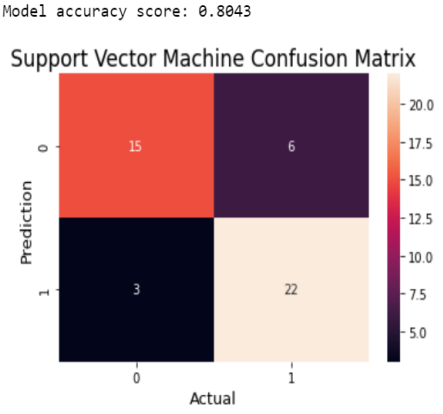
Another ensemble technique, gradient boosting sequentially constructs decision trees to rectify errors made by preceding trees, thereby enhancing predictive performance with each iteration.

XGBoost:

Recognized for its speed and efficacy, XGBoost implements gradient boosted decision trees, establishing its dominance in competitive machine learning

VIII. RESULT AND ANALYSIS





e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

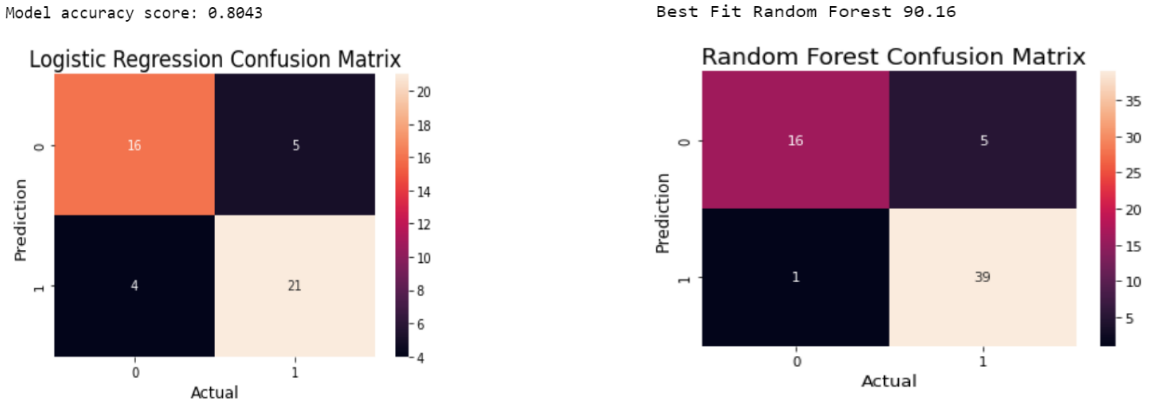


Fig 1: Models Performances

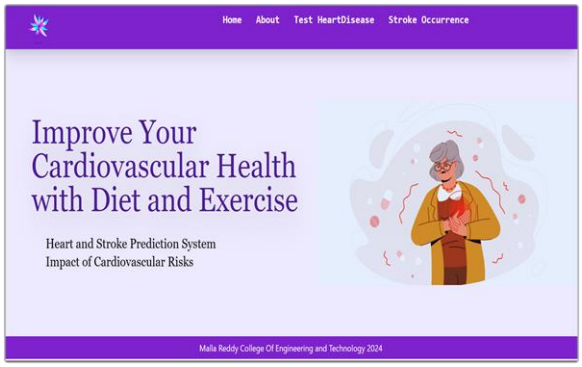


Fig 2: Home Page

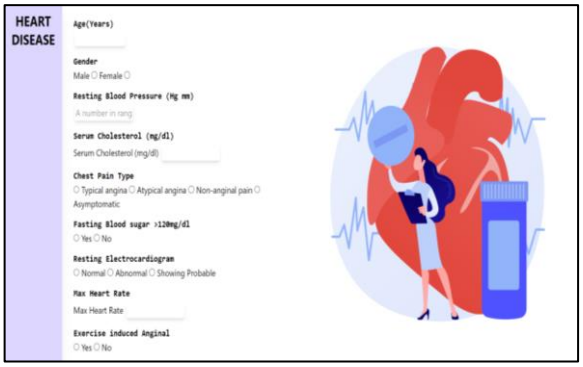


Fig 3: Heart Disease Prediction Page

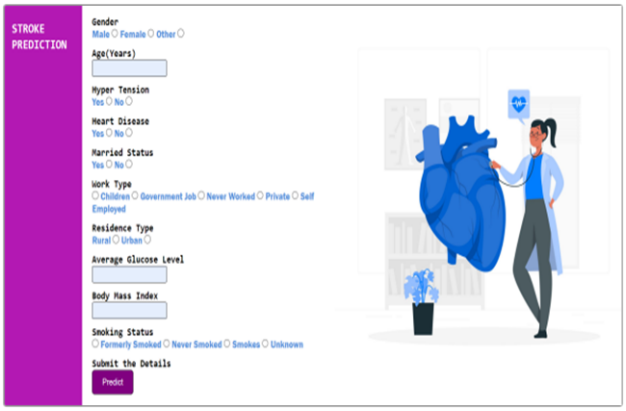


Fig 4: Stroke Prediction Page

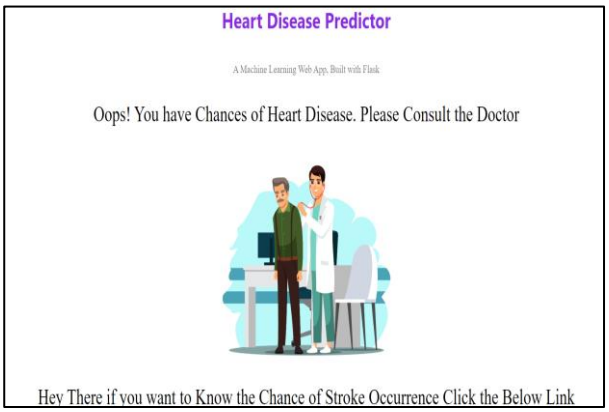


Fig 5: Heart and Stroke Positive Result Page

Fig 6: Heart and Stroke Prediction Negative Result Page

After comparison of models we came to know that random forest algorithm performs very well , so we trained the model using random forest algorithm which gives 90% accuracy. We initially provide the system with a patient details as input , then the system predicts it whether he or she has the heart disease and stroke occurrence.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

IX. CONCLUSION

In conclusion, Heart Disease and Stroke is a life-threatening medical illness that should be treated as soon as possible to avoid further complications. The development of an ML model could aid in the early detection of stroke and the subsequent mitigation of its severe consequences. The effectiveness of several ML algorithms in properly predicting stroke and heart disease based on a number of physiological variables is investigated in this study. Random forest outperforms the other methods tested with a classification accuracy of above 90 percent.

X. REFERENCES

[1] C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques",
https://www.researchgate.net/publication/323277772_Survey_on_prediction_and_analysis_the_occurrence_of_heart_disease_using_data_mining_techniques, 118(8):165-173 · January 2018

[2] Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDUL GHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK",
https://www.researchgate.net/publication/328630348_PREDICTION_OF_HEART_DISEASE_USING_ARTIFICIAL_NEURAL_NETWORK. October 2018

[3] Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for HeartDiseases",https://www.researchgate.net/publication/274718934_Data_Mining_in_Healthcare_for_Heart_Diseases. March 2015.

[4] Komal Kumar Napa, G.Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers",
https://www.researchgate.net/publication/340885231_Analysis_and_Prediction_of_Cardio_Vascular_Disease_using_Machine_Learning_Classifiers, April 2020.

[5] Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach",
https://www.researchgate.net/publication/338428682_Cardiovascular_Disease_Diagnosis_A_Machine_Learning_Interpretation_Approach, January 2019.

[6] In October 2016, Jabbar Akhil and Shirina Samreen published "Heart disease prediction system based on hidden naïve Bayes classifier," which can be found at
https://www.researchgate.net/publication/309735105_Heart_disease_prediction_system_based_on_hidden_naive_Bayes_classifier.

[7] S. Gupta and S. Raheja, "Stroke Prediction using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.

[8] N. S. Adi, R. Farhany, R. Ghina and H. Napitupulu, "Stroke Risk Prediction Model Using Machine Learning," 2021 International Conferenceon Artificial Intelligence and Big Data Analytics, 2021, pp. 56-60, doi:10.1109/ICAIBDA53487.2021.9689740.

[9] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469,
doi: 10.1109/ICECA49313.2020.9297525.

[10] R. Islam, S. Debnath and T. I. Palash, "Predictive Analysis forRisk of Stroke Using Machine Learning Techniques," 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2021, pp. 1-4, doi:10.1109/IC4ME253898.2021.9768524.

[11] A. Devaki and C. V. G. Rao, "An Ensemble Framework for Improving Brain Stroke Prediction Performance," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-7,doi: 10.1109/ICEEICT53079.2022.9768579.

[12] V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

[13] M. Sheetal singh, Prakash choudhary, " Stroke Prediction using Artificial Intelligence ", 8th Annual Industrial Automation and Electromechanical Engineering conference (IEMECON) 2017 DOI: 10.1109/IEMECON.2017.8079581.

[14] Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat, Sharmin Akter Tanna, Taslima Mostafa Alif, Romana Rahman Ema. " Detection of Stroke disease using Machine Learning Algorithms " 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT DOI: 10.1109/ICCCNT45670.2019.8944689