International Journal of Innovative Research in Engineering
Volume 5, Issue 5 (September-October 2024), PP: 37-39.
www.theijire.com
ISSN No: 2582-8746

Predicting Stroke Risk Using Random Forest Algorithm

D. Pravin kumar[1], S. Alagarsamy Raja[2], J. Arunsunai Gowtham[3], B. Kalanithi[4]

[1]*Associate Professor, Department of Computer Science and Engineering, K.L.N. College of Engineering, Sivagangai, Tamilnadu, India.*

[2,3,4] *Final Year Students, Department of Computer Science and Engineering, K.L.N. College of Engineering, Sivagangai, Tamilnadu, India.*

How to cite this paper:
D. Pravin kumar[1], S. Alagarsamy Raja[2], J. Arunsunai Gowtham[3], B. Kalanithi[4], "Predicting Stroke Risk Using Random Forest Algorithm ", IJIRE-V5I05-37-39.

Abstract: *This model aims to predict stroke risk using a dataset containing features like age, BMI, glucose levels, and lifestyle factors. The goal is to build a predictive tool that identifies individuals at high risk of stroke using machine learning models. The primary challenge of class imbalance is addressed using SMOTE, which enhances model performance on the minority class.Both Logistic Regression and Random Forest models were trained, with Random Forest outperforming Logistic Regression. Random Forest algorithm achieves high accuracy, precision and an AUC score, when comparing with Logistic Regression. Overall, Random Forest is a more accurate and reliable tool for stroke prediction.*

Key Word : *Stroke Risk, Class Imbalance, SMOTE, Random Forest, Logistic Regression, Predictive Tool*

I.INTRODUCTION

Machine learning helps estimate the risk of stroke by analysing health and lifestyle data, making early detection and prevention possible. The goal is to improve stroke prediction through advanced technology. Traditional methods are often slow and struggle with complex data, making them less reliable. The Random Forest algorithm, which processes a wide range of patient information, offers greater accuracy and speed in predicting stroke risk. This leads to better understanding of stroke risks, earlier detection, and more personalized treatment for patients.

II. OBJECTIVE

The mission focuses on to develop a predictive model for stroke risk using the Random Forest algorithm to improve early detection and intervention. It aims to analyse patient data, including age, gender, medical history, and lifestyle factors, to accurately assess stroke risk. This approach will automate data cleaning, fixing issues like missing values and class imbalance using techniques like SMOTE. Additionally, exploratory data analysis will help identify important risk factors for stroke. The Random Forest algorithm will improve prediction accuracy, while Logistic Regression will be used for comparison. Ultimately, the aim is to provide healthcare professionals with a reliable tool to identify high-risk individuals, allowing for timely interventions and better patient outcomes in stroke prevention.

III.LITERATURE SURVEY

1. An Improved Concatenation of Deep Learning Models for Predicting and Interpreting Ischemic Stroke:

Predicting ischemic stroke using a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The dataset includes 5110 health records with imbalanced stroke cases, balanced using SMOTE. The CNN+LSTM model achieved high accuracy, outperforming other models like Logistic Regression, Random Forest, and kNN. Performance metrics used include accuracy, precision, recall, F1-score, and AUC. SHAP was employed to explain the model's predictions, ensuring interpretability for healthcare professionals and promoting personalized decision-making. Further testing is recommended.

2. An Imbalanced Data Preprocessing Algorithm for the Prediction of Heart Attack in Stroke Patient:

Predicting heart attacks in stroke patients using clinical data and machine learning techniques. It addresses the challenge of imbalanced data, as most stroke patients do not experience heart attacks. A combination of undersampling and oversampling techniques is applied to balance the dataset. The study utilizes Logistic Regression and Random Forest algorithms, evaluating them based on accuracy, precision, recall, and F1-score. Data visualization tools, such as heatmaps and ROC curves, provide actionable insights for healthcare professionals to identify at-risk patients early.

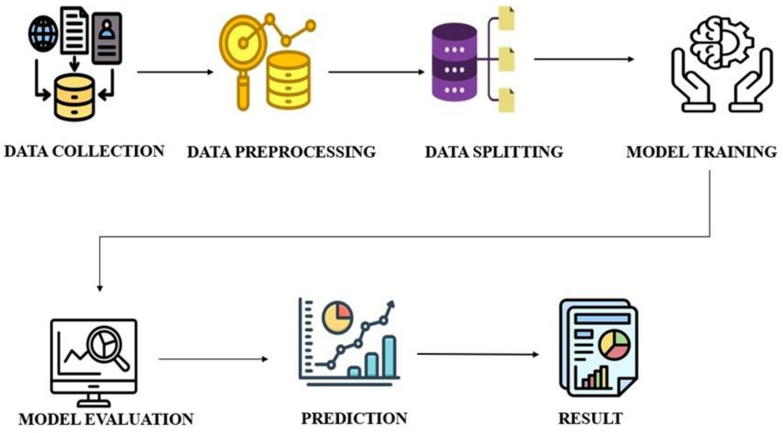
IV.EXISTING SYSTEM

This system identifies ischemic strokes using a hybrid deep learning model that combines CNN and LSTM. It effectively addresses class imbalance through the SMOTE technique, enhancing accuracy with imbalanced datasets. Trained on a healthcare dataset, it utilizes key features like age, glucose levels, and heart disease. However, it faces challenges such as high computational costs and the need for user training.

V.PROPOSED SYSTEM

The system utilizes the Random Forest algorithm for accurate stroke predictions. It preprocesses data by handling missing values and converting lifestyle habits into numeric form. Key risk factors such as age, BMI, and smoking status are emphasized. Real-time predictions provide healthcare providers with quick insights, while patient details are collected for a comprehensive analysis and better decision-making.

VI.ARCHITECTURE DIAGRAM



VII.SYSTEM OVERVIEW

1. Data Collection

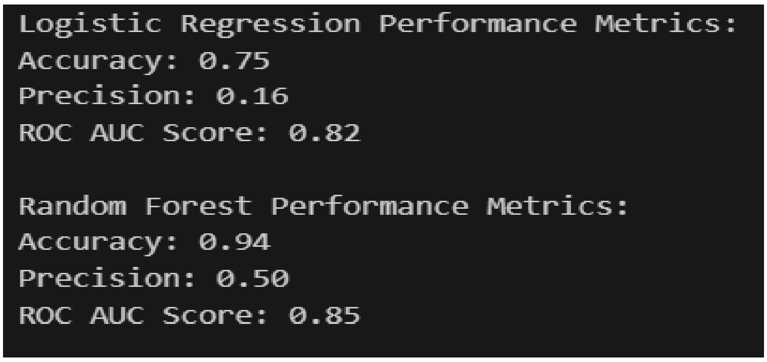
The Data is gathered from a CSV file and explored to understand its structure. Missing values, such as those in the BMI column, are filled with the mean. Unnecessary columns like the ID are dropped to reduce noise and focus on meaningful data.

2. Data Preprocessing

The Categorical variables like gender and work type are encoded using Label Encoder and OneHotEncoder, respectively, to convert them into numerical format. SMOTE (Synthetic Minority Over-sampling Technique) is applied to handle class imbalance, ensuring the model is trained on a balanced dataset.

3. Data Splitting

The pre-processed dataset is divided into training and test sets using an 80-20 split. The training set is further resampled using SMOTE to ensure balanced class distribution. Standard scaling is applied to normalize feature values, improving model performance and comparison.



```
Logistic Regression Performance Metrics:
Accuracy: 0.75
Precision: 0.16
ROC AUC Score: 0.82

Random Forest Performance Metrics:
Accuracy: 0.94
Precision: 0.50
ROC AUC Score: 0.85
```

Fig 7.1 Comparision of Performance Metrics

4. Model Evaluation

Two models, logistic regression and random forest, are trained on the scaled data. Each model’s performance is evaluated based on accuracy, precision and ROC-AUC score. ROC curves are plotted to visually compare model performance, and key metrics are displayed for analysis.

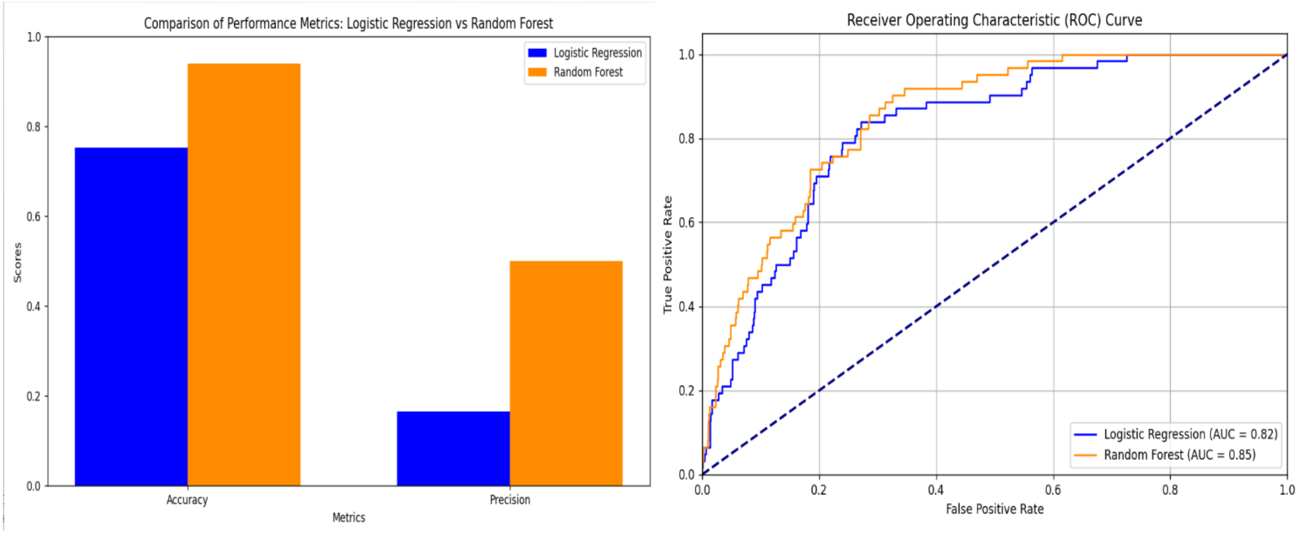


Fig 7.2 Pictorial Representation of Performance Metrics

VIII.CONCLUSION

In comparing the Logistic Regression and Random Forest model, Random Forest algorithm performed better, achieving about 94% accuracy, 50% precision and 85% AUC score. In Comparison, Logistic Regression had around 74% accuracy, 16% precision and 82% AUC score.The Enhancements in performance, supported by SMOTE for balancing the data, show that Random Forest algorithm is more effective in prediction, making it a valuable tool for healthcare decision making.

References

1. *K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," in IEEE Access, vol. 11, pp. 52288-52308, 2023*
2. *S. Sakri et al., "An Improved Concatenation of Deep Learning Models for Predicting and Interpreting Ischemic Stroke," in IEEE Access, vol. 12, pp. 53189-53204, 2024*
3. *S. Peñafiel, N. Baloian, H. Sanson and J. A. Pino, "Predicting Stroke Risk With an Interpretable Classifier," in IEEE Access, vol. 9, pp. 1154-1166, 2021*
4. *M. Wang, X. Yao and Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," in IEEE Access, vol. 9, pp. 25394-25404, 2021*
5. *A. N. V. K. Swarupa, V. H. Sree, S. Nookambika, Y. K. S. Kishore and U. R. Teja, "Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm," 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, 2021*
6. *K. Kamata et al., "Development of stroke detection method by heart rate variability analysis and support vector machine," 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, China, 2015, pp. 1257-1261, doi: 10.1109/APSIPA.2015.7415475.*
7. *K. S. R. S, B. Chandra, K. Kausalya, C. RM and G. R. V, "Prognosis of Stroke using Machine Learning Algorithms," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 1-6, doi: 10.1109/ICCMC56507.2023.10084158.*
8. *M. Dahiya, N. Mishra, S. Agarwal and Z. Parveen, "Predicting the occurrence of Ischemic stroke by Gradient Boost Approaches," 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2023, pp. 1-4, doi: 10.1109/ICIEM59379.2023.10166287.*
9. *P. Gahiwad, N. Deshmane, S. Karnakar, S. Mali and R. Pise, "Brain Stroke Detection Using CNN Algorithm," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-4, doi: 10.1109/I2CT57861.2023.10126125.[10].R. Kuksal, M. Vaqur, A. Bhatt, H. Chander and K. Joshi, "Stroke Disease Detection and Prediction using Extreme Gradient Boosting," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 187-191, doi: 10.1109/AISC56616.2023.10085514.*
10. *R. Kuksal, M. Vaqur, A. Bhatt, H. Chander and K. Joshi, "Stroke Disease Detection and Prediction using Extreme Gradient Boosting," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 187-191, doi: 10.1109/AISC56616.2023.10085514.*