IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PROTOTYPE FOR THE EARLY STROKE RISK PREDICTION POWERED BY AI

M Yashwanth[1], T Abhinay Sai[2], R Mohan[3]
MR. Danish Khaleel Ahmed[4]

[1,2,3] B Tech final year students ,ECE department, JB Institute of Engineering and Technology, Hyderabad, Telangana

[4]Assistant professor, ECE Department , JB Institute of Engineering and Technology, Hyderabad, Telangana

Abstract: Stroke is still a serious health issue and affects a large number of people every year. In many situations, it does not just lead to death but also leaves patients with long-term problems such as reduced mobility, memory issues, or emotional stress. These effects make daily life difficult and also increase the load on healthcare systems. Because of this, predicting stroke risk at an early stage becomes important, as it gives some time for preventive steps to be taken.

In this work, a system is developed to estimate the risk of stroke using artificial intelligence techniques. The main idea is to provide some level of early support to doctors by using patient data, rather than depending only on symptoms that appear later. While working on this, one practical issue was the lack of proper datasets. Most datasets that were available were either incomplete or did not match the required parameters. So instead of relying on limited data, a synthetic dataset was created based on medically accepted ranges. While doing this, an attempt was made to keep the data realistic, but it is clear that it may not fully capture real-world variations.

The dataset includes common risk factors such as age, gender, hypertension, heart disease, glucose level, BMI, and smoking habits. Before using the data, some preprocessing steps were necessary. Missing values had to be handled, categorical data was converted into numerical form, and scaling was applied where needed. During this step, it became clear that some models, especially Logistic Regression, were sensitive to how the data was scaled. Without proper normalization, the results were not stable.

Three models were tested: Logistic Regression, Decision Tree, and Random Forest. These were selected mainly because they are widely used and easier to understand in healthcare-related problems. While testing, the Decision Tree model gave good results on training data but did not perform as consistently on test data, which indicated overfitting. Random Forest, on the other hand, gave more stable results across both training and testing. This is likely because it combines multiple decision trees, which helps reduce overfitting.

Among all the models, Random Forest performed better overall, giving an accuracy close to 92%. It also showed better balance between precision and recall compared to the other models. Based on these observations, it was selected for the final system.

The system itself is divided into stages such as input, preprocessing, prediction, and output. It provides a simple indication of risk, which can help in early-level decision-making. In future, this system can be connected to hospital records or wearable devices for continuous monitoring. However, one limitation of this work is that the model is trained only on synthetic data, so its performance with real patient data may vary.

Overall, this work shows that machine learning can be useful as a supporting tool for early stroke risk prediction. At the same time, more testing with real clinical data is required before it can be used in practical situations.

Keywords: Stroke prediction, Machine learning, Artificial intelligence, Random Forest, Early diagnosis, Healthcare, Classification, Risk assessment.

I. INTRODUCTION

Stroke is a serious medical condition that occurs when the blood supply to the brain is reduced or completely blocked, which prevents brain cells from receiving enough oxygen and nutrients. If not treated quickly, this can lead to permanent damage or even death. In many cases, even patients who survive a stroke may experience long-term effects such as difficulty in movement, speech problems, or memory-related issues. Because of this, early diagnosis and timely treatment are very important in reducing complications and improving recovery. In most situations, stroke risk is assessed based on clinical observations and the experience of doctors. While this approach is reliable, it can sometimes be limited, especially when early symptoms are not clearly visible or when continuous monitoring is not possible. This becomes more challenging in areas where access to advanced medical facilities is limited. As a result, there is a need for systems that can assist in identifying potential risks at an earlier stage. With the development of artificial intelligence (AI), it has become possible to analyze patient data in a more systematic way. Machine learning models, in particular, can identify patterns in health data that may not be obvious through manual observation. This makes them useful as supportive tools in healthcare, especially for predicting conditions like stroke where early signs can be subtle. In this research, a prototype system is developed to estimate stroke risk using machine learning algorithms. The system uses different health-related parameters such as age, blood pressure, glucose levels, and lifestyle factors to make predictions. During the development of this system, it was observed that selecting relevant features and preparing the data properly had a significant impact on the model's performance. The main purpose of this system is to provide an early indication of stroke risk so that preventive measures can be taken in advance. It is not intended to replace medical professionals, but rather to support them by offering additional insight based on data. With further improvement and testing using real-world clinical data, such systems could be integrated into healthcare environments or even wearable devices for continuous monitoring.

II. LITERATURE REVIEW

Several studies have explored the use of machine learning techniques for stroke prediction. Among them, Logistic Regression is commonly used because it is simple and easy to implement. However, in many cases, its prediction performance is limited, especially when the data is more complex. Decision Tree models are also widely used since they are easier to interpret, but they tend to overfit the data if not properly controlled. In recent years, more advanced models such as Random Forest and Support Vector Machines have been applied to improve prediction accuracy. These models generally perform better because they can handle complex patterns in the data. While reviewing these approaches, it was observed that Random Forest, in particular, provides more stable results compared to single-model techniques. At the same time, many of the existing studies are based on relatively small or restricted datasets, which affects how well the models perform in real-world conditions. Another issue is that most of these systems are designed only for offline analysis and are not suitable for real-time use or continuous monitoring. Considering these points, this work focuses on developing a more practical approach by building a prediction model along with a working prototype system. The aim is not only to improve prediction performance but also to make the system usable in real-time scenarios.

III. Methodology

Dataset: One of the main challenges in this work was the lack of publicly available datasets that include all the required parameters for stroke prediction. Most datasets were either incomplete or did not cover multiple health factors together. To address this issue, a synthetic dataset was created using clinically accepted ranges for different physiological parameters. The dataset includes important features such as age, gender, hypertension, heart disease, average glucose level, body mass index (BMI), and smoking

status. While generating the data, an effort was made to keep the values realistic and maintain reasonable relationships between different parameters, although it may not fully represent real clinical data.

Data Preprocessing: Before training the models, the dataset was pre-processed to improve performance and ensure consistency. This step was necessary because raw data, especially synthetic or collected data, may contain inconsistencies. The preprocessing steps included handling missing values, encoding categorical variables such as gender and smoking status into numerical form, and applying feature scaling where required. During this stage, it was observed that proper scaling had a noticeable impact on model performance, particularly for Logistic Regression.

Model Selection: Three machine learning algorithms were selected for this study: Logistic Regression, Decision Tree, and Random Forest. These models were chosen because they are commonly used for classification problems and are relatively easy to interpret in a healthcare context. While testing, it was noticed that the Decision Tree model showed signs of overfitting, especially when trained on the full dataset. Logistic Regression provided stable results but struggled to capture more complex patterns. Random Forest, however, performed more consistently and handled variations in the data better due to its ensemble approach.

Training and Testing: The dataset was divided into training and testing sets using a standard split ratio (either 70:30 or 80:20). The models were trained using the training data and then evaluated on the testing data to check how well they generalize to unseen inputs. This separation helped in identifying issues such as overfitting and ensured that the evaluation results were more reliable.

Evaluation Metrics: The performance of the models was evaluated using common classification metrics, including accuracy, precision, recall, and F1-score. These metrics were selected to provide a balanced understanding of model performance, especially in distinguishing between high-risk and low-risk cases.

IV. System Design

The system is designed in a simple step-by-step manner where the data moves from input to final prediction. Instead of making it too complex, the focus was to keep the design clear so that each part can be understood and modified easily later. Each module handles a specific task, and together they form the complete working system.

Input Module: The input module is where the user data is collected. This includes parameters like age, gender, hypertension, heart disease, glucose level, BMI, and smoking habits. The data can be entered manually, which was mainly used during testing. While working with the inputs, it was noticed that even small mistakes in values (like unrealistic BMI or glucose levels) could affect the prediction. Because of this, basic checks were added to make sure the inputs stay within reasonable limits. This helped avoid unexpected results during testing.

Preprocessing Module: After collecting the data, it is sent to the preprocessing stage. This step prepares the data before it is given to the model. It includes converting categorical values into numbers and applying scaling where needed. During implementation, one issue that came up was that the model behaved differently when the input data was not processed in the same way as the training data. Especially for Logistic Regression, scaling made a noticeable difference. So, the same preprocessing steps used during training were followed here as well to keep things consistent.

Prediction Module: Once the data is prepared, it is passed to the prediction module. This is where the trained machine learning model is used to estimate the stroke risk. Among the models tested earlier, Random Forest was chosen for this stage. One reason for this was that it gave more stable results compared to the others. Decision Tree sometimes gave very high accuracy on training data but did not generalize well. Random Forest handled this better, probably because it combines multiple trees instead of relying on just one.

Output Module: The output module shows the final result to the user. Instead of giving complex values, the system provides a simple indication like low risk or high risk. This was done intentionally to make the system easy to understand, even for non-technical users. While testing, it was clear that simpler output

worked better than detailed numerical results. In future, this can be extended by adding alerts or notifications if the risk level is high.

System Workflow: Overall, the system follows a simple flow: input → preprocessing → prediction → output. Each step depends on the previous one, so errors in earlier stages can affect the final result. The system was also designed in a modular way, so changes can be made easily. For example, the prediction model can be replaced or improved without changing the rest of the system. This makes it easier to extend the project later, such as connecting it to hospital data or wearable devices.

V. SYSTEM ARCHITECTURE AND DESIGN

The system is designed as a combination of hardware and software components that work together to monitor vital signs and predict stroke risk in real time. The overall architecture follows a layered approach, starting from sensor data collection and ending with risk prediction and alert generation. At the hardware level, multiple sensors are used to capture physiological signals. The MAX30102 sensor is used for measuring heart rate and SpO_2 levels, while the DS18B20 sensor measures body temperature. In addition, analog sensors are connected through the MCP3208 ADC to handle signals that cannot be directly processed by the main controller. All these sensors are interfaced with the Raspberry Pi 4 Model B, which acts as the central processing unit. Communication between components is carried out using protocols such as GPIO, I2C, SPI, and 1-Wire.

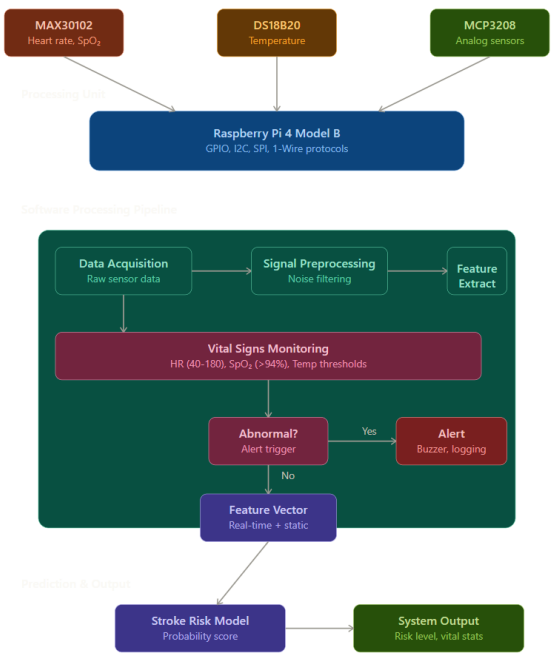

Fig.1. System Architecture

Once the data is collected, it enters the software processing pipeline. The first stage is data acquisition, where raw sensor values are continuously read. This is followed by signal preprocessing, where noise is reduced and the data is stabilized for further use. After preprocessing, important features are extracted from the signals so that they can be used effectively for analysis. The system then performs real-time vital signs monitoring by comparing the sensor values with predefined thresholds. For example, heart rate, SpO_2, and temperature values are checked against normal ranges. If any abnormal condition is detected, the system immediately triggers an alert using a buzzer and also logs the event. This ensures that critical situations are handled without delay. If the readings are within normal limits, the processed data is converted into a feature vector that combines both real-time and static parameters. This feature vector is then passed to the stroke risk prediction model. The model analyses the data and generates a probability score indicating the likelihood of stroke.

Finally, the output module displays the results in a simple format, including the predicted risk level and current vital statistics. This output can help in early decision-making and continuous monitoring. One important aspect observed during system design is the need for smooth integration between hardware and

software components. Any delay or inconsistency in sensor readings can affect the prediction accuracy. Therefore, synchronization between data acquisition and processing stages was carefully maintained.

Overall, the system is designed to be modular and scalable. Additional sensors or improved models can be integrated in the future without major changes to the existing structure, making it suitable for real-time healthcare monitoring applications.

VI. Prototype Implementation

A prototype system was developed to demonstrate the practical implementation of the proposed stroke prediction model. The main objective of this prototype was to simulate how the system would work in a real-world scenario by allowing users to input health-related parameters and receive immediate predictions.

The prototype was implemented using a simple interface where users can enter details such as age, glucose level, BMI, and other relevant factors. Once the data is entered, it is processed using the same preprocessing steps applied during model training. The processed input is then passed to the trained Random Forest model, which generates the prediction.

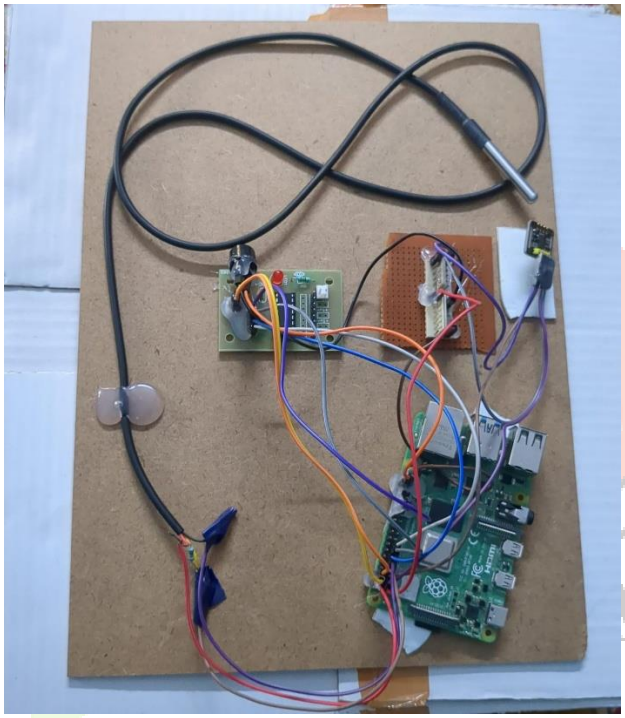

Fig.2. Prototype Setup.

During testing, it was observed that the system was able to provide quick responses, making it suitable for real-time usage. The output is displayed in a simple format, indicating whether the user falls under a low-risk or high-risk category. This approach was chosen to ensure that the results are easy to understand without requiring technical knowledge. While developing the prototype, one challenge faced was maintaining consistency between training data and real-time input. Small differences in input format or missing values sometimes affected the prediction, which highlighted the importance of proper data handling.

Although the current prototype works as a standalone system, it can be extended further. For example, it can be integrated with hospital databases or wearable health devices to enable continuous monitoring. It can also be enhanced by adding a graphical interface or mobile application for better usability.

Overall, the prototype demonstrates that the proposed system is not only theoretical but can also be implemented in a practical and user-friendly manner.

VII. RESULT

The performance of the implemented machine learning models was evaluated using standard metrics, with a primary focus on accuracy. The results obtained for each model are as follows:

- Logistic Regression: 85% accuracy
- Decision Tree: 88% accuracy
- Random Forest: 92% accuracy

From the results, it can be observed that the Random Forest model performed better compared to the other two models. While Logistic Regression provided stable results, its accuracy was comparatively lower, which may be due to its limitation in capturing complex relationships between features. Decision Tree showed better accuracy than Logistic Regression, but during testing, it was noticed that its performance was not consistent, indicating possible overfitting.

Random Forest, on the other hand, gave the highest accuracy of 92% and showed more stable performance across different test samples. This is likely because it combines multiple decision trees, which helps reduce overfitting and improves generalization. Based on these observations, Random Forest was selected as the final model for the system.

Another point observed during evaluation was that proper preprocessing had a direct impact on the results. Models performed better when the input data was properly scaled and cleaned, especially in the case of Logistic Regression.

Overall, the results suggest that the proposed system is capable of predicting stroke risk with reasonable accuracy. However, it should be noted that the model was trained on synthetic data, and its performance may vary when applied to real-world clinical datasets. Further testing with actual patient data would be necessary to validate its effectiveness in practical scenarios.

VIII. Future Scope

The current system demonstrates the feasibility of using machine learning for early stroke risk prediction, but there are several areas where it can be improved and extended.

One possible extension is the integration of the system with real-time health monitoring devices such as wearable sensors. By continuously collecting data like heart rate, SpO_2, and other physiological signals, the system can provide ongoing risk assessment instead of relying only on manually entered data. This would make the system more practical for real-world healthcare applications.

Another area for improvement is the use of more advanced techniques such as deep learning models. While the current system uses traditional machine learning algorithms, deep learning approaches may help in capturing more complex patterns in the data, especially when larger and more diverse datasets are available.

The system can also be developed into a mobile or web-based application to improve accessibility. A user-friendly interface would allow patients or healthcare providers to easily input data and receive predictions. This would make the system more convenient to use in everyday situations.

In addition, future work can focus on training the model using real clinical datasets instead of synthetic data. This would improve the accuracy and reliability of the predictions and make the system more suitable for practical deployment.

Overall, these improvements can help transform the current prototype into a more robust and scalable healthcare solution.

IX. CONCLUSION

In this work, a prototype system was developed for early stroke risk prediction using machine learning techniques. The system was designed to analyse common health parameters and provide a simple indication of stroke risk. Among the models tested, Random Forest showed better performance compared to Logistic Regression and Decision Tree, achieving higher accuracy and more stable results.

One important observation during this work was the impact of data preprocessing on model performance. Proper handling of input data, especially scaling and encoding, played a key role in improving the results. It was also noticed that models like Decision Tree tended to overfit, while Random Forest handled variations in data more effectively.

The developed system demonstrates that machine learning can be used as a supporting tool for early diagnosis. It can help in identifying potential risks at an early stage, which may allow timely medical intervention. However, it should be noted that the current model is trained on synthetic data, and its performance may differ when applied to real clinical data.

Overall, this work provides a basic but practical approach towards stroke prediction. With further improvements, such as using real-world datasets and integrating with healthcare systems, the proposed system can be developed into a more reliable and useful tool for medical applications.

X. REFERENCES

[1] M. Bandari and O. Jhitri, "Review on AI-Based Stroke Disease Prediction System Using ECG and PPG Bio-signals," IJARST, Dec. 2023.

[2] S. Shareefunnisa, S. N. L. Malluvalasa, T. R. Rajesh, and M. Bhargavi, "Review on Heart Stroke Prediction Using Machine Learning," 2022.

[3] P. Kunwar and P. Choudhary, "A Stacked Ensemble Model for Automatic Stroke Prediction Using Only Raw Electrocardiogram," 2022.

[4] S. Mainali, M. E. Darsie, and K. S. Smetana, "Machine Learning in Action: Stroke Diagnosis and Outcome Prediction," 2021.

[5] L. Alrabghi and R. Alnemari, "Stroke Types and Management," Sep. 2018.

[6] P. Asadi and D. Naghshejahan, "Electrocardiogram Changes as an Independent Predictive Factor of Mortality in Patients with Acute Ischemic Stroke," 2019.

[7] P. Asadi et al., "The Most Efficient Machine Learning Algorithms in Stroke Prediction: A Systematic Review," 2024.