Brain Stroke Prediction Using Machine Learning

Dr C V Madhusudhan Reddy, T Somashekar M.Tech, M G Usharani, M K Pavithra,
K Ramalakshmi, U Sharanya
Dept. of Computer Science and Engineering (Artificial Intelligence), St. Johns College of Engineering and Technology, Yemmiganur – 518301, India.

Abstract: **Stroke is a big deal when it comes to our health. This is one of the reasons why people die or become disabled for a long time. We need to determine who is at risk of having a stroke as early as possible. Thus, we can take steps to prevent it. Individuals will have a better chance of survival. This study is about a system that uses machine learning to predict whether a person will have a brain stroke. It examines the history of patients to determine whether they might have a stroke. This system is called a brain stroke prediction system. It uses machine learning algorithms to make predictions about stroke. The proposed system can assist doctors. It looks at things like where people are from and how old they are. It also examines how their bodies work and the kind of life they lead. All of these factors are related to the risk of stroke. The system helps doctors by looking at these things like factors, physiological factors and lifestyle factors that are related to the possibility of a person having a stroke.**

We used several different machine learning algorithms, such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and Gradient Boosting. We compared these machine learning algorithms to determine which one works best.

Before training these machine learning algorithms, we performed a lot of work to prepare the data. We have to deal with missing data, ensure that the numbers are all on the scale, and fix problems where one group of data is much larger than the others.

The results of our tests show that Random Forest and Gradient Boosting, which are based models, work better than the other machine learning algorithms. Random Forest and Gradient Boosting are more accurate and work well even when the data are not perfect. The system can help doctors make decisions. It can help doctors determine whether a patient is at risk of having a stroke before it actually occurs. This system can be trusted by doctors to obtain information about stroke risks.

Index Terms: **Brain Stroke Prediction, Machine Learning, Healthcare Analytics, Medical Data Mining, Supervised Learning**

I.INTRODUCTION

A stroke occurs when blood flow to the brain stops. This implies that the brain does not receive the required oxygen and nutrients. Brain cells are damaged rapidly. If you do not get help, you can be disabled for the rest of your life or even die. There are different types of strokes. You have strokes, hemorrhagic strokes, and transient ischemic attacks. Ischemic stroke is the most common type of stroke.

The number of stroke cases is increasing. This places significant pressure on the healthcare system. Doctors usually use imaging and blood tests. They also perform clinical examinations after someone starts showing symptoms of a stroke. These methods do not show what is happening with all the different risk factors for stroke before it actually happens. Therefore, the chance of preventing a stroke is often missed. Stroke cases are a problem, and we need to consider stroke prevention. The healthcare system is struggling with the increasing number of stroke cases.

However, with the increasing amount of electronic health data and advances in computing capabilities, machine learning has emerged as a useful method for healthcare analytics. Machine learning algorithms can process large amounts of data, identify nonlinear relationships, and produce predictions that are difficult to obtain through conventional statistical analyses.

II. LITERATURE SURVEY

Over the past ten years, researchers have been studying how machine learning can be used in the medical field. They wanted to determine whether machine learning could help predict when someone would develop heart disease or have a stroke. Initially, researchers used statistics, such as Logistic Regression, to make predictions. They liked these models because

they were easy to comprehend. The problem was that these simple models could not handle all the complicated factors that can cause heart disease and stroke. Machine learning is increasingly being used to predict cardiovascular and cerebrovascular diseases.

Some people conducted more studies and used models that are not straight-line models, such as Support Vector Machines and Decision Trees. These models were better at predicting what would happen. They were easier to comprehend. Decision Trees are an example they give you clear rules to follow and Support Vector Machines work really well even when you have a lot of features to look at. In recent studies, people have been using ensemble methods, such as Random Forest and Gradient Boosting, which are really good at what they do. Random Forest and Gradient Boosting combine what many models think will happen to obtain an answer and to be more certain that the answer is correct.

The need for proper data preprocessing, such as normalization, missing value imputation, and resampling to handle class imbalance, has also been highlighted by researchers. However, many previous studies have focused on the evaluation of only a few algorithms or metrics. This study extends the range of previous research by comprehensively comparing various models and selecting those that perform best in predicting the risk of brain stroke using healthcare datasets.

III. SYSTEM ARCHITECTURE

The new brain stroke prediction system is simple and easy to add. This means that it can work well and handle many tasks. The brain stroke prediction system has five parts:

The Data Input Module was used to collect patient information. It obtains details such as patient demographics. The Data Input Module also captures the patient's history and lifestyle variables. This means that it stores patient details, including their previous illnesses and lifestyle. The Data Input Module is important for storing information, including patient demographics, medical history, and lifestyle variables.

The Preprocessing Module is an important step. This ensures that the raw data are clean and ready for use. This handles missing values and things like that. It also performs variable encoding. The Preprocessing Module also makes sure everything is normal and balanced. This means that it performs normalization and fixes any data imbalance issues that it finds. The Preprocessing Module performs all of these tasks.

The Machine Learning Engine is really cool. This helps us to implement and train useful algorithms. These algorithms include Logistic Regression, SVM, Decision Tree, Random Forest and Gradient Boosting. The Machine Learning Engine is where we actually use these algorithms, such as Logistic Regression and Decision Tree, and others, to obtain the results we need.

We used the Machine Learning Engine to train algorithms such as Random Forest and Gradient Boosting. The Machine Learning Engine is important for implementing and training the chosen algorithms, such as SVM and Logistic Regression.

The Prediction Module performs an important task. It looks at patients. Decides whether the patient is a high-risk or low-risk patient. This decision was made based on the predictions made by the Prediction Module. The Prediction Module is helpful because it tells us which patients are high-risk and which ones are low-risk.

The Result Visualization Module is extremely helpful. It provides pictures and graphs that make sense so that doctors and nurses can make decisions. The Result Visualization Module is very useful for healthcare professionals because it shows them what is going on in a way that is easy to understand.

The proposed system ensures a systematic data flow and facilitates the easy maintenance and integration of models in the future.

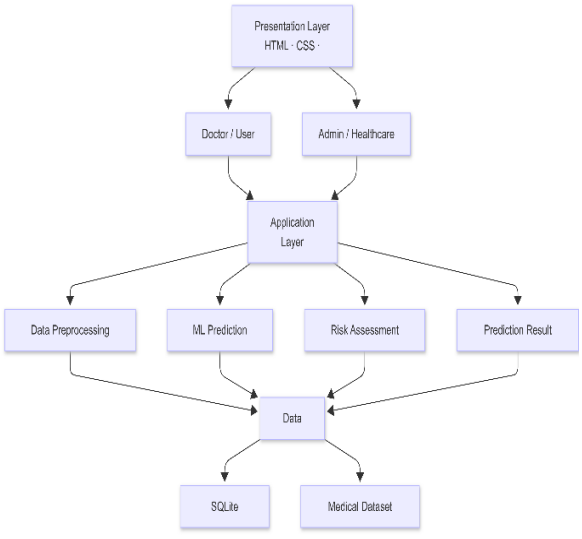

Fig: System Architecture of the Brain Stroke Prediction

IV. METHODOLOGY

The dataset has information about patients, including things like how old they're if they are a man or a woman if they have high blood pressure if they have had heart disease before if they smoke, their body mass index and their average blood sugar level. The main thing we are trying to figure out is if a patient will have a stroke. This information is what we use to teach and test the machine learning models, like the ones that try to predict if someone will have a stroke. Machine learning models and datasets on patients and stroke are very important. Data preprocessing was performed to ensure that the data were reliable and consistent. When data are missing, we use methods to handle the missing data, and we change all categorical data variables into numerical data variables. We also ensured that all continuous data variables were normalized. If the datasets are not balanced, we use methods such as oversampling or SMOTE to ensure that the model is not biased towards the majority class of the datasets. We do this to handle the datasets. Machine Learning Algorithms. We used different methods to teach computers: Logistic Regression, SVM, Decision Tree, Random Forest, and Gradient Boosting. Each of these methods was trained on the cleaned data. To see how well each method works we look at things, like how it is right how precise it is, how well it remembers things how well it does overall and something called ROC-AUC. We used these to compare the performance of each method. Model Evaluation Comparative analysis was performed to identify the best-performing algorithm. Ensemble models, specifically Random Forest and Gradient Boosting, are found to be better predictors because they have the capability to reduce variance and make the model more robust by aggregating the predictions.

V. IMPLEMENTATION

The new system was implemented using Python. It uses libraries such as Scikit-learn, Pandas, and NumPy to work with the data. The Python system has a backend that performs the following tasks: It gets the data ready, pulls out the parts, trains the models, and saves the trained models so that they can be used again. This way the Python system can work efficiently with the data and the models. The system is really easy to use. It has a simple user interface. This interface helps doctors and nurses enter information, about their patients and show the results of predictions in a way that's easy to understand. The system is also made in a way that makes it simple to add things to it later on and connect it to the systems that hospitals already use for storing information about patients. Component Detail Programming Language Python ML Category Supervised Learning Algorithms Used Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting Dataset Patient Health Records Evaluation Metrics Accuracy, Precision, Recall, F1--Output Stroke Risk (High / Low).

Component	Details
Programming Language	Python
Machine Learning Type	Supervised Machine Learning
Algorithms Used	Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting
Dataset	Patient health records
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Prediction Output	Stroke Risk (High / Low)

VI. RESULTS AND DISCUSSION

The experimental results show that ensemble learning techniques perform better than individual classifiers. The random Forest and Gradient Boosting algorithms successfully identified complex, non-linear relationships between health attributes, achieving higher accuracy and better reliability of predictions. The results demonstrate the excellent potential of ML-based approaches for the early detection of individuals at risk of stroke, thus providing a basis for preventive healthcare.

VII. BENEFITS AND APPLICATIONS

Benefits

The Early Risk Detection system is really good at finding out if someone might have a stroke before they even feel sick. This is a deal because it helps doctors catch the problem early which is important for the stroke. The Early Risk Detection is useful, for detecting strokes so people can get help before the stroke happens.

Using automated analysis is a thing. It helps to reduce mistakes that people might make. This means that the analysis is more accurate. Automated analysis is very

helpful because it reduces the possibility of error. Automated analysis also increases the accuracy of analysis.

Scalable Framework: Capable of efficiently processing a large amount of patient data.This is really good because it helps reduce the need for diagnostic imaging. The best way to do this is by allowing for care with the medical equipment. This medical equipment is cost-effective. It reduces the need for diagnostic imaging by allowing for preventive care, with the medical equipment.

Applications

Hospital Stroke Risk Screening: Quickly analyzes patients during regular check-ups. Clinical Decision-Support Systems are really helpful for doctors. They assist doctors to figure out which patients are, at a risk of getting very sick. This means doctors can use Decision-Support Systems to quickly identify high-risk patients and give them the care they need. Clinical Decision-Support Systems are useful tools for doctors to have.

Community Health Monitoring: Useful for large-scale health surveys. Medical Research is really important because it helps us understand how some things can increase the chance of having a stroke. Medical Research looks at the relationship between things that might put us at risk and the likelihood of having a Medical Research study, on stroke. This means Medical Research can tell us what things make it more likely that we will have a stroke.

Healthcare resource optimization is, about using resources in the best way possible. It does this by finding the patients who need help the most, which are the high-priority patients. This way healthcare resource optimization helps these high-priority patients get the care they need away.

VIII. CONCLUSION

The new research is about a computer system that helps figure out if someone's likely to have a brain stroke. This system looks at lots of information about the patient. Makes a good guess. It does this by cleaning up the information and then using special computer programs that are taught to make good predictions. The system found that two methods, called Random Forest and Gradient Boosting are really good at working to make the best predictions, about brain stroke risk. Brain stroke risk is what the system is trying to predict.

The framework has the potential to notify medical professionals about high-risk patients, which can lead to early clinical intervention and save lives. Future research can be extended to incorporate deep learning models, real-time monitoring using wearable technology, and cloud-based implementation for faster processing.

ACKNOWLEDGMENT

The authors acknowledge with thanks the Department of Computer Science and Engineering (Artificial Intelligence), St. John's College of Engineering and Technology, for providing the necessary guidance and infrastructure. They also thank their project guide for his invaluable mentorship and continued support in building this piece of work.

REFERENCES

[1] S. K. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors, vol. 22, no. 13, pp. 1–19, Jun. 2022.

[2] M. W. Ashraf, M. S. Ahsan, M. H. Rahman and A. M. Al Mamun, "Prediction of Brain Stroke Using Machine Learning Algorithms," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 6, pp. 539–545, Jun. 2021.

[3] S. S. Saranya, R. S. Kumar and P. M. Durai, "An Intelligent System for Stroke Prediction Using Machine Learning," Journal of Medical Systems, vol. 45, no. 3, pp. 1–11, Mar. 2021.

[4] A. Choudhury, S. Gupta and R. Kumar, "Stroke Prediction Using Supervised Machine Learning Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 8, pp. 123–128, Aug. 2020.

[5] M. R. Islam, M. R. Haque and S. A. Rahman, "A Machine Learning Approach for Stroke Prediction Based on Health Data," Procedia Computer Science, vol. 171, pp. 2079–2088, 2020.

[6] K. S. Reddy, P. S. Reddy and M. V. Rao, "Early Prediction of Brain Stroke Using Data Mining Techniques," International Journal of Computer Applications, vol. 178, no. 7, pp. 24–29, Jan. 2019.

[7] S. Sharma and A. Aggarwal, "Analysis of Stroke Prediction Using Machine Learning Models,"

International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 11, pp. 4531–4536, Sep. 2019.

[8] R. V. Phanindra and P. S. Avadhani, "Predictive Analytics for Brain Stroke Using Machine Learning," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 2, pp. 3412–3417, Dec. 2019.

[9] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.