RESEARCH ARTICLE OPEN ACCESS

Early Prediction of Brain Stroke Using Machine Learning

Kalaiselvi.P [1], Vasanth.G [2], Aravinth.P [3], Elamugilan.A [4], Prasanth.S [5]

Department of Artificial Intelligence and Data Science, Sri Sairam Engineering College - Chennai

ABSTRACT

Brain stroke is one of the driving causes of death and disability worldwide. Healthcare is a sector that needs continuous innovation in technologies and treatments. One of the most well-known technologies used in healthcare in recent years is Machine Learning. Machine Learning is the subset of Artificial Intelligence (AI) that involves building algorithms that is giving machines the capacity to learn and develop over time without being explicitly programmed to do so is the goal of machine learning. Brain stroke is a dangerous health disorder caused due to the interruption of blood flow. Early prediction of brain stroke and taking the necessary treatments help in reducing the mortality rate. So, the project mainly aims at predicting the chances of the occurrence of stroke using emerging machine learning techniques. There are many risk factors that cause brain stroke such as diabetes, hypertension, BMI, Blood pressure and many more. So, the Model is developed with the above parameters using the Machine learning algorithms such as Decision tree, Logistic Regression, Support Vector Machine, Naive Bayes, KNN, and Random Forest. Comparing the accuracy of the above algorithm, a model is developed for the prediction of Brain Stroke.

Keywords:- Machine Learning (ML), Support Vector Machine (SVM), Random Forest, Naïve Bayes, Logistic regression, K Nearest Neighbour (KNN)

I. INTRODUCTION

According to National Institute of Neurological Disorders and Stroke (NINDS), stroke is the second leading cause of death worldwide, with an annual death rate of around 5.5 million. Stroke is a brain disease. It is a sudden interruption of continuous blood flow to the brain and is a medical emergency. A stroke occurs when a blood vessel in the brain becomes blocked or narrowed, or when a blood vessel ruptures and spills blood into the brain. Like a heart attack, a stroke requires immediate medical attention. There are two types of strokes: ischemic stroke and haemorrhagic stroke. A transient ischemic attack (TIA) is sometimes called a "mini-stroke". This differs from the main type of stroke because blood flow to the brain is only blocked for a short time, usually no longer than 5 minutes. Several factors increase the risk of stroke, including high blood pressure, high cholesterol, diabetes, smoking, and obesity. Additionally, family history, age, and gender can also be risk factors. Stroke is preventable and treatable. By identifying those at high risk and intervening early, healthcare professionals can help prevent strokes and improve patient outcomes. Therefore, the main objective of this project is to use machine learning techniques to predict stroke risk. By analysing data from medical records, diagnostic tests, and other sources, machine learning algorithms can develop predictive models that help healthcare professionals make informed decisions about preventative measures. Machine learning algorithms analyse electronic health records (EHRs) to identify risk factors associated with stroke. By analysing large amounts of data from EHRs, machine learning algorithms can identify patterns and risk factors that medical professionals might not immediately detect. This can help healthcare professionals identify high-risk patients and take preventative action, such as prescribing medications or recommending lifestyle changes. Machine learning algorithms can also be used to develop predictive models that incorporate various risk factors such as age, gender, medical history, and lifestyle factors. Therefore, machine learning is developed using machine learning algorithms such as decision trees, logistic regression, support vector machines, naive arrays, KNN, and parameter-based random forests above. By comparing the accuracy of the above algorithms, a stroke prediction model was developed. A dataset with various physiological parameters was selected from Kaggle. The dataset is cleansed and ready to be understood by machine learning models. This process is called data pre-processing. The dataset contains 5111 records. After data pre-processing, algorithms are applied and accuracy is achieved.

II. RELATED WORKS

There have been enormous studies on stroke prediction. For example, “Stroke prediction using machine learning classifiers in the general population” by M.D. Dorr et al. (2019), In this study author used aa data from a population-based cohort to develop machine learning models for stroke prediction. The authors achieved an accuracy of 92.3% using a KNN algorithm. Several authors got high efficiency for different algorithms. For example, “Machine learning-based prediction of stroke in patients with atrial fibrillation" by Y. Ma

et al. (2020), This study used machine learning algorithms to predict the risk of stroke in patients with atrial fibrillation. The authors achieved an accuracy of 83.2% using a gradient boosting machine algorithm. Another example, "An explainable machine learning approach for stroke prediction in patients with atrial fibrillation" by W. Huang et al. (2021), This study used an explainable machine learning approach to predict the risk of stroke in patients with atrial fibrillation. The authors achieved an accuracy of 74.1% using a decision tree algorithm. These studies and works clearly shows us the importance for detecting the stroke in its early state.

III. METHODS AND MATERIALS

SVM: Support vector machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression analysis. The primary goal of SVM is to find the best decision boundary (hyperplane) that can separate the different classes in the dataset with the maximum possible margin. However, one of the limitations of SVM is that it can be sensitive to the choice of kernel function and its hyperparameters. The performance of the SVM algorithm depends heavily on the selection of kernel function and tuning its parameters, which can be time-consuming and requires domain expertise.

Logistic regression: Logistic Regression is a statistical algorithm used for binary classification tasks, which predicts the probability of an event occurring (such as a customer buying a product or not) based on one or more input variables or features. In logistic regression, the output variable is a binary variable, which means it can take only two possible values (0 or 1). The algorithm models the relationship between the input features and the output variable using a logistic function (also known as a sigmoid function), which transforms the input values into probabilities. The logistic function outputs a value between 0 and 1, which represents the probability of the event occurring. This is a popular algorithm used in various fields, such as healthcare, finance, marketing, and social sciences. It is often used for risk prediction, fraud detection, customer segmentation, and many other binary classification tasks.

KNN: K-Nearest Neighbors (KNN) is a simple and popular algorithm for supervised learning tasks such as classification and regression. It is a non-parametric algorithm, meaning that it doesn't assume any underlying probability distribution for the input data. Instead, it makes predictions based on the similarity of the input data to the data in its training set. The KNN algorithm works by finding the K nearest data points in the training set to the input data point. The distance metric used to measure the similarity between the input data point and the training data points can vary depending on the problem domain. Common distance metrics include Euclidean distance, Manhattan distance, and cosine similarity

Random Forest: Random Forest is a popular ensemble learning algorithm used for classification, regression, and other machine learning tasks. The algorithm builds a collection of decision trees, and each tree is constructed by taking a random subset of features and data samples. Random forest can be applied to a wide range of tasks, such as image classification, object detection, text classification, and predictive analytics. Additionally, the algorithm is relatively easy to use and provides interpretable results, making it a popular choice for both beginner and experienced machine learning practitioners.

Decision tree: Decision tree algorithm is a machine learning algorithm used for both classification and regression tasks. It builds a decision tree model from the training data, which is a tree-like model where each node represents a feature and each branch represents a decision rule. The decision tree algorithm starts by selecting the most important feature from the training data and creating a root node based on it. However, decision trees may suffer from overfitting if the tree is too complex or if the training data is noisy. To address this, various techniques such as pruning and ensemble methods like Random Forest and Boosting have been developed.

TOOLS USED:

WEKA: Weka is a famous machine learning tool used to build a model. It also provides us with various features like data preprocessing, model building, evaluating the model etc.

GOOGLE COLAB: It provides us with create and execute python code from the browser. It provides us with all the python functions and modules.

DATASET:

Dataset sourced from the Kaggle platform has been acquired. The dataset contains various attributes like id, bmi, average glucose level, Average blood pressure level etc.

IV. PROPOSED METHOD

It contains 5111 data in which 249 were positive and 4862 were negative and then is Data has been preprocessed from the dataset where null values or missing values are filled and data

encoding is performed to transform categorical variables into numerical values. The preprocessed data [before resampling] is imported to WEKA application[v-3.8.4] and using the Ranker search method, the attributes are selected and the machine learning algorithms are applied to the dataset and the accuracy for the respective algorithms are obtained but the accuracy are very low due the imbalanced dataset. Dataset is to be balanced for high accuracy. So, python code used to obtain the accuracy for the respective algorithms with the same dataset. The code is implemented in Google collab by using modules like NumPy, seaborn, pandas, and matplotlib

V. EXPERIMENTAL SETUPS AND RESULTS

The experimental system has been tested and trained in 30% and 70% of data respectively. Various algorithms like SVM, Logistic Regression, KNN, random forest and Decision tree were checked for the accuracy. The results were discussed in the table below.

Before resampling: The following result is obtained using WEKA software (8.3.4)

S.NO	MACHINE LEARNING ALGORITHM	ACCURACY
1	SVM	48.0626 %
2	LOGISTIC REGRESSION	48.1409 %
3	KNN	36.9276 %
4	RANDOM FOREST	42.6223 %
5	DECISION TREE	42.1918 %

Table 1: Accuracy before resampling

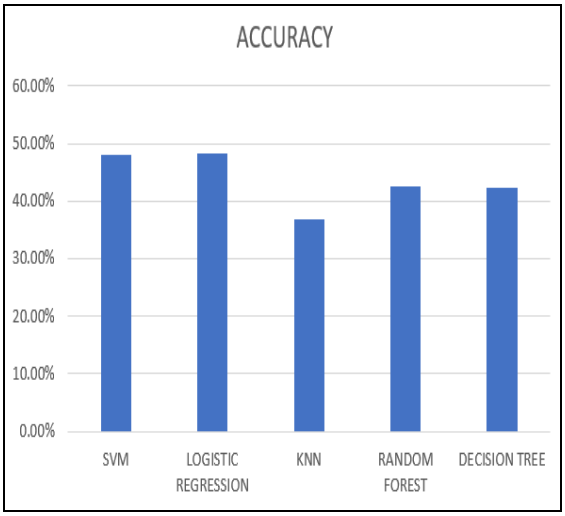

Figure 1: Graph for accuracy before resampling

After resampling: The following result is obtained using python on Google collab.

S.NO	MACHINE LEARNING ALGORITHM	ACCURACY
1	SVM	94.30%
2	LOGISTIC REGRESSION	94.36%
3	KNN	94.07%
4	RANDOM FOREST	94.25%
5	DECISION TREE	90.63%

Table 1: Accuracy after resampling

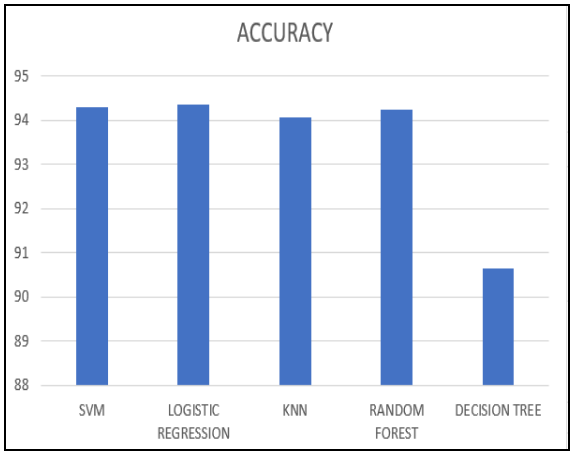

Figure 2: Graph for accuracy after resampling

VI. CONCLUSION

It can be concluded that the study provides a comprehensive analysis of the existing research work on stroke prediction. The paper has presented a systematic review of the literature, including a detailed analysis of the research methodology, findings, and limitations of the previous studies. Overall, the paper has provided valuable insights into the existing work on stroke prediction using machine learning algorithms and has identified several avenues for future research in this area. In future neural networks can be used to enhance the model's performance.

REFERENCES

[1] Nitish Biswas, Khandaker Mohammad Mohi Uddin, Sarreha Tasmin Rikta, Samrat Kumar Dey, "A Comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach", ScienceDirect, 2022.

[2] Lihi Schwartz, Roi Anteby, Eyal Klang, Shelly Soffer, "Stroke mortality prediction using machine learning". ScienceDirect,2023.

[3] Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John, "A predictive analytics approach for stroke prediction using machine learning and neural networks". ScienceDirect,2022.

[4] Rishi Raj, Jimson Mathew, Santhosh Kumar Kannath, Jeny Rajan. "Stroke it with AutoML for brain stroke classification". ScienceDirect,2023.

[5] Xinyi Zhao, Xingmei Chen, Xulong Wu, Lulu Zhu, Jianxiong Long, Li Su, Lian Gu. "Machine Learning Analysis of MicroRNA Expression Data Reveals Novel Diagnostic Biomarker for Ischemic Stroke". ScienceDirect,2021.

[6] Sang Min Sung, Yoon Jung Kang, Han Jin Cho, Nae Ri Kim, Suk Min Lee, Byung Kwan Choi, Giphil Cho. "Prediction of early neurological deterioration in acute minor ischemic stroke by machine learning algorithms". ScienceDirect,2020.

[7] Prashanth Kunwar, Prakash Choudhary. "A Stacked ensemble model for automatic stroke prediction using only raw electrocardiogram". ScienceDirect,2023.

[8] Wei Bo, Lora Cavuoto, Jeanne Langan, Heamchand Subryan, Sutanuka Bhattacharjya, Ming-Chun Huang, Wenyao Xu. "A Progressive prediction model towards home-based stroke rehabilitation programs". ScienceDirect,2022.

[9] Gangavarapu Sailasya, Gorli L Aruna Kumari. "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms". IJACSA,2021.

[10] Laercio Ives Santos, Murilo Osorio Camargos, Marcs Flavio Silveria Vasconcelos D'Angelo, Joao Baista Mendes, Egydio Emiliano Camargos de Medeiros, Andre Luiz Sena Guimraes, Reinaldo Martinez Palhares. "Decision tree and Artificial Immune Systems for stroke prediction in imbalance data". ScienceDirect,2022