ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

BRAIN STROKE DETECTION USING MACHINE LEARNING ALGORITHMS

G. Kowshiki, PG Scholar, Department of CSE, K.S.R.M. College of Engineering, Kadapa
Kowshiki836@gmail.com
Dr. M. Sreenivasulu Professor, Department of CSE, K.S.R.M. College of Engineering, Kadapa
ms@ksrmce.ac.in

ABSTRACT
A stroke, a detrimental medical condition, can occur as a result of the rupture of blood vessels within the brain or due to a disruption in the flow of essential nutrients. The World Health Organization recognizes stroke as a significant global cause of both mortality and disability. While significant research has been conducted in the domain of heart stroke prediction, relatively fewer studies have addressed the risk of brain stroke. The rapid and accurate identification of stroke occurrence plays a pivotal role in providing timely medical assistance to patients. Our model achieves a high accuracy rate in distinguishing between stroke and non-stroke cases, with a sensitivity and specificity that exceed those of existing methods. The study encompasses a comprehensive analysis of multiple factors, employing machine learning (ML) algorithms such as Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) to achieve precise prognostications. By delving into these predictive models, the research seeks to advance the understanding of stroke risk assessment and empower healthcare practitioners with effective tools for early intervention and prevention strategies.

Keywords:
Machine learning; Logistic Regression; Decision Tree Classification; Random Forest Classification; K-nearest Neighbor; Support Vector Machine.

I. INTRODUCTION
In recent days, due to a variety of physical and environmental conditions, the occurrence rate of infectious and chronic diseases is gradually rising and appropriate detection and treatment implementation is essential to reduce the disease impact. In the current era, modern disease detection and treatment implementation methods are existing to reduce the medical stress. Stroke is an ailment that impacts vessels that supply blood to the thoughts. mind stroke takes region which list blood glide to the mind is each reduced or interrupted. While this occurs, the mind no longer gets sufficient oxygen or other crucial components, and the brain cells start to die. A stroke effects important lengthy-time period incapacity or demise. Here, mind stroke is one of the leading causes of death all around the world. There are 3 kinds of brain strokes: ischemic strokes, hemorrhagic strokes, and transient ischemic assault, which is also referred to as a caution or mini-stroke [1] [14]. A stroke victim's chances of making a full recovery are improved the earlier they receive medical care. It is very crucial for any individual who has experienced a stroke before to seek immediate medical attention. Understanding the multifaceted landscape of strokes is not only pivotal for medical professionals but also for individuals at large, as knowledge empowers swift action and potentially life-saving interventions. The machine learning algorithm can improve patients' health through early detection and treatment. We have used several machine learning algorithms to detect the type of stroke that can occur in a patient or already occurred from their clinical report and statistical data People lose their lives in large numbers due to stroke and it is increasing in developing countries [2]. There are several stroke risk factors that regulate different types of strokes. Predictive algorithms help to understand the relationship between these risk factors and types of strokes. We have built a stroke dataset by collecting data from various sources validated by medical experts. Then the dataset can be processed and to be used with the machine learning algorithms. We have built several models of classification. These models can be used to classify stroke patients in real time. Machine learning can be depicted as a valuable tool in domains

ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

such as surveillance, medicine, and data management, leveraging appropriately trained machine learning algorithms [3]. The proposed concept involves mainly extracting patient symptoms from medical case sheets and training the system with this collected data. Subsequently, case sheets are analysed through maximum entropy techniques and tagging, and a proposed stemmer is utilized to identify common and distinctive attributes for stroke disease detection. The utilization of data mining techniques in this study provides a comprehensive assessment of information tracking, encompassing both semantic and syntactic standpoints. Then, the processed data were fed into various machine learning algorithms such as K-Nearest Neighbors, Random Forest, Decision tree, Logistic Regression, Support vector machine [10]. Among these algorithms, Support vector Machine gives high accuracy [4].

II. LITERATURE SURVEY

Govindarajan [5] managed the data assembled from Sugam Multi-speciality Hospital. The dataset contained more than 500 records of patients and many fascinating class names of two huge Stroke types[14]. They applied Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Bagging, and Boosting. Among the above Machine Learning Algorithm, they got the highest accuracy using ANN Algorithm with ~95%. Badriyah, Tessy et al.[6] Data can be analyzed and used as consideration for the decision making. It can be carried out with a variety of approaches such as using the Deep Learning method which is increasingly being used today because it is proven to be powerful in solving various problems. The forerunner of Deep Learning itself began in 1980 when Kunihiko Fukushima made Neocognition, the first model of the Convolutional Neural Network before being refined by Yann LeCun, Leon Bottou, Joshua Bengio. G. A. P. Singh et al.,[7] Lung cancer is one of the most common causes of death among all cancer-related diseases (Cancer Research UK in Cancer mortality for common cancers. Automated classification of lung cancer is one of the difficult tasks, attributing to the varying mechanisms used for imaging patient's lungs. C. L. Chin et al.,[8] [9]. Over the past few years, stroke has been among the top ten causes of death in Taiwan. Stroke symptoms belong to an emergency condition, the sooner the patient is treated, the more chance the patient recovers. The purpose of this paper is to develop an automated early ischemic stroke detection system using CNN deep learning algorithm.

III. PROPOSED METHODOLOGY

In this proposed system, we are using different machine learning algorithms are K-Nearest Neighbors, Random Forest, Decision tree, Logistic Regression, Support vector machine[11]. In our proposed system, we evaluate multiple algorithms in comparison with one another, ultimately choosing the model with the highest accuracy. The selection of the best model for our dataset is contingent upon the accuracy score. This section will provide a comprehensive overview of the work conducted in the detection of Brain Stroke.

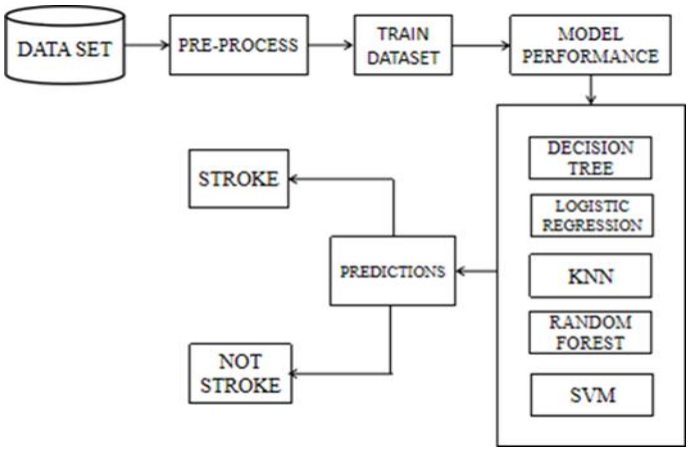

Fig 1: Block Diagram of Proposed System

ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

The architecture of proposed methodology is shown in fig.1 contains the following steps:

1. Data Set

A data set is a collection of data. A dataset is associated with one or more database tables when dealing with tabular data. In this context, each column within a table represents a specific variable, while each row corresponds to a particular record within the dataset. These records contain information about various attributes, such as gender, age, and BMI of individuals. It's important to note that datasets can take other forms, including collections of documents or files. The data set for the detection of Brain stroke is collected from the Kaggle website. For this purpose, the Brain_Stroke.csv has been collected from the Kaggle website. The study utilized the Stroke Prediction dataset for its analysis. There were 4982 rows and 11 columns in this dataset. In the output column "stroke," the values are binary, represented as either 1 or 0. A value of 0 signifies the absence of a stroke risk, while a value of 1 signifies the presence of a stroke risk. Notably, the dataset contains a higher number of instances with a value of 0 in the "stroke" column, with 4861 rows indicating no stroke risk, compared to 249 rows where a stroke risk is identified (with a value of 1). Data pre-processing techniques are employed to balance the dataset to enhance the accuracy of the analysis.

2. Data Pre-processing

This stage mainly focuses on addressing all factors hindering the model at peak efficiency. It commences with the collection of datasets, followed by the essential steps of data cleaning and preparation for model development. Prior to model construction, data pre-processing is essential, as it involves the removal of undesirable noise and outliers from the dataset that could otherwise cause the model to deviate from its intended training path. As stated before, the dataset has characteristics. This can involve either imputing missing values with estimated values for removing rows or columns with missing data, depending on the nature of the dataset and the extent of missingness. The dataset comprises 4982 rows, with 249 rows suggesting the potential occurrence of a stroke, while 4861 rows affirm the absence of a stroke. While training a machine learning model on such data might yield high accuracy, it can lead to shortcomings in other critical performance metrics like precision and recall. If this imbalance in the dataset is not appropriately addressed, the resulting conclusions will be inaccurate, rendering the predictive capabilities of the model ineffective. Therefore, the initial step toward achieving an effective model is to address this data imbalance.

3. Data Splitting

Portraying Features and Target: Separate the dataset into features (x) and the objective variable (y), where X contains with or without area from the “target” section and y contains only the “target” portion. Train-Test split: Partition the data into Planning and testing Subsets. The planning set is used to set up the man-made intelligence models, while the testing set is put something aside for surveying their show. The training set is used for the training the machine learning algorithms whereas the test set is for the testing the model accuracy.

4. Machine Learning Algorithms

The following machine learning algorithms were considered for prediction of brain stroke.

4.1 Logistic Regression

Logistic regression is a classification algorithm in Machine Learning utilized for predicting the probability associated with a categorical dependent variable. In logistic regression, this dependent variable is typically binary, with data coded as 1 (representing a positive outcome, success, etc.) or 0 (representing a negative outcome, failure, etc.). Logistic regression is a supervised machine learning algorithm primarily employed for classification tasks. Its primary objective is to predict the probability of an instance belonging to a specific class or category.

ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

4.2 Random Forest Classifier

Random Forest is a popular machine learning algorithm used for both classification and regression tasks, with a primary focus on classification problems. It is considered an ensemble method because it leverages the strength of multiple decision trees to provide more robust and accurate predictions.

One of its significant advantages is its ability to reduce overfitting, a common issue in single decision trees. Random Forest is built on the principle of ensemble learning, a technique that involves the amalgamation of multiple classifiers to address intricate problems and enhance model performance significantly. This is achieved by averaging the results from multiple trees, resulting in a more reliable and generalized prediction model.

4.3 Decision Tree

A Decision Tree is a supervised learning method capable of addressing both regression and classification tasks. This tree-like classifier employs nodes to signify dataset features, branches to convey decision rules, and each leaf node to denote the outcome.

Decision Trees find applications in various domains, including finance, healthcare, marketing, and beyond, making them a fundamental tool in the realm of supervised machine learning. They also serve as foundational components in ensemble methods like Random Forests and Gradient Boosting, enhancing predictive accuracy and robustness.

4.4 K-Nearest Neighbor(KNN)

K-NN stands out as a fundamental but crucial classification algorithm within the field of Machine Learning. This approach relies on assessing the likeness between incoming data and existing instances, subsequently assigning the new case to the category that closely resembles the ones already present.

It is a part of supervised learning and finds significant utility in domains such as intrusion detection, pattern recognition, and data mining. The K-NN algorithm retains all the existing data and employs similarity metrics to categorize a new data point.

The "K" in K-NN represents the number of nearest neighbors to consider, and a distance metric, commonly Euclidean distance, quantifies the similarity. In classification, the algorithm assigns the majority class among the K nearest neighbors to the new data point, while in regression, it calculates predictions by averaging or weighting the target values of these neighbors. Selecting an appropriate K-value is crucial, as it influences the model's performance. However, K-NN can be computationally expensive, particularly with large datasets, and benefits from feature scaling.

4.5 Support Vector machine(SVM)

The Support Vector Machine (SVM) stands out as one of the most widely employed Supervised Learning algorithms, serving both Classification and Regression tasks [12]. Nonetheless, its primary application lies in the realm of Classification within the field of Machine Learning.

The objective of the SVM algorithm is to craft the optimal line or decision boundary, capable of partitioning an n-dimensional space into distinct classes, ensuring that future data points can be accurately assigned to their respective categories. This paramount decision boundary is referred to as a hyperplane. SVM identifies the pivotal points or vectors crucial for constructing the hyperplane. These pivotal instances are denoted as support vectors, hence lending the algorithm its name, Support Vector Machine.

5. Model evaluation

The following measures can be used for evaluating the machine learning model for stroke prediction.

5.1 Accuracy:

The Accuracy Rate stands out as the most used measure, representing a straight forward ratio of correctly predicted instances to the total number of instances observed. In simpler terms, accuracy provides the percentage of instances predicted correctly.

ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

5.2 Precision:
Precision is defined as the proportion of correctly classified positive samples (True Positives) out of the total number of samples classified as positive (whether classified correctly or incorrectly).

Precision = T P/ T P+FP

Here TP is True positive tuples and FP is False Positive tuples.

5.3 Recall:
Recall is calculated as the ratio of positive samples correctly classified as positive to the total number of positive samples.

Recall = TP/TP + FN

Here TP is True positive and
FN is False Negative Tuples.

IV. RESULTS AND ANALYSIS

The Machine Learning algorithms were implemented using Python language. The Brain stroke.csv data set has been used and that can be split into Train set and Test set. The accuracy has been measured for different machine learning algorithms is tabulated in Table 1. Among all algorithms Support Vector Machine gives satisfactory results and can be used in a real time patient's stroke classification.

Table 1. Accuracy Measure for machine learning algorithms

ALGORITHM	PERCENTAGE ACCURACY
Logistic Regression	94.98%
Decision Tree	90.96%
KNN	94.78%
SVM	95.71%
Random Forest	94.38%

V. CONCLUSION

The primary objective of the proposed methodology is to reduce the death rate in the population attributed to brain strokes. By employing ML algorithms, the Support Vector Machine consistently provides reliable results and stands out as an effective choice for real-time stroke classification in patients when compared to other algorithms. The Brain_Stroke.csv data set has been used for training the machine learning algorithms and evaluating the model's performance. So, the Support Vector Machine Classifier can be used to detect brain stroke detection early to decrease the death rate.

VI. REFERENCES

[1] M. S. Raja, M. Anurag, C. P. Reddy, and N. R. Sirisala, "Machine Learning Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics
[2] Dataset named 'Stroke Prediction Dataset' from Kaggle: https://www.kaggle.com/fedesoriano/stroke- prediction-dataset.
[3] Stroke Prediction Using Machine LearningAlgorithms: https://ijirem.org/DOC/2-stroke-prediction-using-machine-learning- algorithms.pdf
[4] Stroke prediction using SVM R S Jeena; Sukesh Kumar https://ieeexplore.ieee.org/document/7988020
[5] P. Govindarajan, R. K. Soundara Pandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, stroke disease classification using machine learning algorithms," Neural Computing and Applications in 2019.
[6] Sung, S.M., Kang, Y.J., Cho, H.J., Kim, N.R., Lee, S.M., Choi, B.K., Cho, G. (2020). early neurological prediction deterioration by machinelearning algorithms. CN and Neurosurgery.
[7] Pahus S. H, Hansen A. T, Hvas A. M (2020), "Thrombophilia testing in young patients with

ISSN: 2278-4632
Vol-13, Issue-09, No.04, September: 2023

Ischemic stroke". Thromb Res 137:108–112.

[8] Dupont S. A, Wijdicks E. F, Lanzino G, Rabinstein A. A (2020), "Aneurysmal subarachnoid hemorrhage: an overview for the practicing neurologist." 30(5):45–54.

[9] Santos E. M. M, Yoo A. J, Beenen L. F, Majoie C. B, Marquering H. A (2019), "Detection of stroke using Neurology methods"

[10] S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," Int J Comput Appl, vol. 149, no. 10, pp. 26–31, 2019.

[11] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.

[12] R. Jeena and S. Kumar, "Stroke prediction using svm," in 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600–602, IEEE, 2016.

[13] [M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in 20178th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 158–161, IEEE, 2017.

[14] L. T. Kohn, J. Corrigan, M. S. Donaldson, et al., To err is human: building a safer health system, vol. 6. National academy press Washington, DC, 2000.