International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Machine Learning Model for Stroke Disease Classification

**Chaitanya Sairam Naidu**

Department of Computer Science Engineering, Chirala Engineering College, Jawaharlal Nehru Technological University Kakinada (JNTUK), Chirala, India
*sairamchaitanya[at]yahoo.com*

**Abstract:** ***In many nations, stroke is the primary factor contributing to mortality and obesity. This study improves the quality of the picture to strengthen image results and to diminish noise to increase the image quality of CT scans of stroke patients, as well as using machine learning algorithms to categorize the scans of the patients into the two subtypes those are ischemic stroke and stroke hemorrhage. This research utilized eight machine learning algorithms to classify stroke diseases, those are Naive Bayes, K-Nearest Neighbors, Decision Tree, Logistic Regression, Multi-layer Perceptron (MLP-NN), Random Forest, Support Vector Machine and Deep Learning. Our study showed that Random Forest delivers the most accurate outcomes (95.97%), together with recall values (96.12%), f1- Measures(95.39%), and precision values (94.39%).***

**Keywords:** Machine learning algorithms, CT Scan image, stroke hemorrhage, stroke ischemic

## 1. Introduction

In many countries, stroke is a leading cause of death and has a high morbidity rate that results in disability. Before stroke treatment can begin, a proper diagnosis must be made, because the sort of stroke a person has will determine how they are treated. Based on CT scan image data, this study divides stroke patients into hemorrhagic stroke and ischemic stroke groups. Ischemic stroke is often brought on by a blood vascular obstruction. While hemorrhage stroke is brough on by a leaking in brain tissue.

Chiun-Li Chin, et al. [4] conducted research on the diagnosis and prediction of stroke and created an early ischemic stroke detection system that automatically employs the CNN Deep Learning algorithm. The fully connected layer and pooling layer are the two convolutional layers used in the CNN architecture. The main use of this pooling layer is down-sampling, which is nothing but in order to lessen the issue of overfitting, the layer will limit the quantity of data and parameters.

The categorization outcomes have a 90% accuracy rate. Using an open stroke dataset obtained from www.radiopaedia.org, Marbun, JT. et al. [5] classified patient data into three types using CNN's Deep Learning architecture through CT scan images. Those three types are ischemic stroke, normal and hemorrhagic stroke. The obtained accuracy is 90%. While other researchers were utilizing the same dataset, Badriyah, Tessy et al. [6] improved the stroke diagnostic accuracy of the Deep Learning method via hyperparameter optimization.

Accuracy may be increased to 100% using Deep Learning's random search optimization technique and Bayesian search for hyperparameter tuning. While Jenna R.S. and Dr. Sukesh Kumar [9] used the effectiveness of several kernel functions in the Support Vector Machine approach to predict stroke, their research yielded satisfying findings as well. The Kernel Linear Function and polynomial obtained the best experimental results, 91.7% & 87% respectively.

In addition to studies that used specific techniques for categorizing medical datasets, the research by Gur Amrit Pal Singh and P.K. Gupta [3] identified and categorized lung cancer by using a number of algorithms in machine learning , such as Super Vector Machine (SVM), K-Nearest Neighbor (KNN), Nave Bayes, Decision Tree, Random Forest, Stochastic Gradient Descent (SGD), and Multilayer perceptron (MLP), which is a type of Deep Learning architecture. The accuracy rate was 88.55% with the MLP strategy from the results of the classification of a medical image dataset of 15,752 with a distribution of 6,912 for the benign class and 8.84% for the malignant class.

In studies that contrasted the three ways, Hima Haridas and Aswathy Wilson combined different methodologies. The first strategy makes use of the Neural Network technique, whereas the second strategy combines two algorithms—Principal Component Analysis (PCA) for decomposition and classification using Neural Networks. While the third strategy includes three algorithms those are PCA for decomposition reduction, Neural Network for stroke prediction and Decision Tree for feature selection. The comparisons done using the three ways yielded the following results: 95.0%, 95.2%, and 97.7%, with the third approach producing the best results.

## 2. Analysis Methodology

### 1) System Design

Data collection, pre-processing, and performance analysis method of classification are the three primary steps of the system design for this study shown in Figure 1.

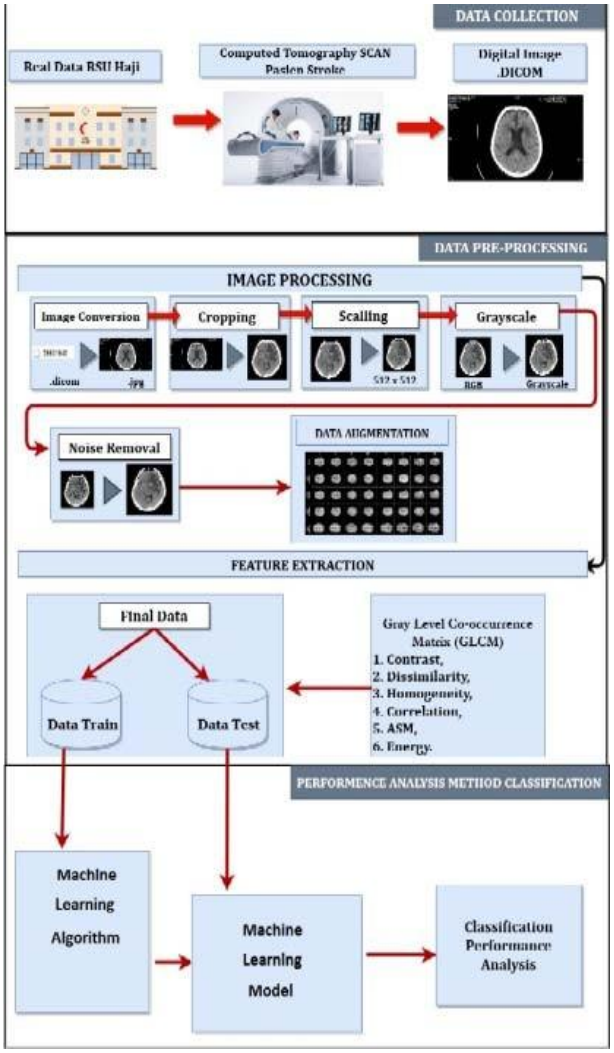



**Figure 1:** Design System

### 2) Data Collection

The CT scan dataset of the patient's brain that contains the image data from both hemorrhagic and ischemic strokes was used as the data collection in this investigation. The dataset was gathered from numerous hospitals in our region of India.

### 3) Initialization Data

Data Pre-processing is a procedure used to enhance the quality of image data, and it includes conversion, cropping, scaling, grayscale, which regulates the size of the pixels utilized, noise removal, which reduces noise and adds blur, as well as feature extraction. Using the Gray Level Co-occurrence Matrix (GLCM) approach, the retrieved feature at the feature extraction stage is a texture feature. The six characteristics of GLCM are contrast, dissimilarity, homogeneity, correlation, ASM, and energy.

### 4) Performance Analysis Method of Grouping.

The performance analysis stage of the classification, which will be applied straight from the process between training data and testing data, comes after the feature extraction stage. The grouping in this research is used to identifystrokes in the brain.

The grouping in this research produces Hemorrhagicstroke, Ischemic stroke, with the help of machine learning algorithm.

The primary components of the system design—data collecting, data preprocessing, which involves feature extraction and image processing and, classification procedures—will be described in the subsections that follow.

### 5) Data Collection

The study's primary source of data was a CT scan study of stroke patients from the various Hospital's in our area, India where the details of 103 patients whose scans were used in the study, 98 had ischemic strokes and three had hemorrhagic strokes. We can obtain 1 to 5 images of stroke objects from each patient by using the 5.0 mm Slice Thickness option in DICOM, and only high-quality image data will be used. There are 7 image data obtained for hemorrhagic stroke and 226 image data obtained for ischemic stroke because of some patients having numerous images.

### 6) Image Procedure

Image processing on CT scans is finished by following the below six steps: (1) data altering (2) cropping (3) scaling (4) greyscale and (5) data augmentation. Each process will be explained below.

### a) Data altering

In this procedure, an application is used to carry out the DICOM data conversion process. The dataset that was collected from the hospital's is still in the form of a dicom data extension. Only a Personal Computer that is integrated with the CT scan tool can see image data with the dicom extension. Dicom converter, Micro Dicom, CoolUtils.com, and Syngofast View are some of the open-source applications that are used to convert dicom photos to jpg images.

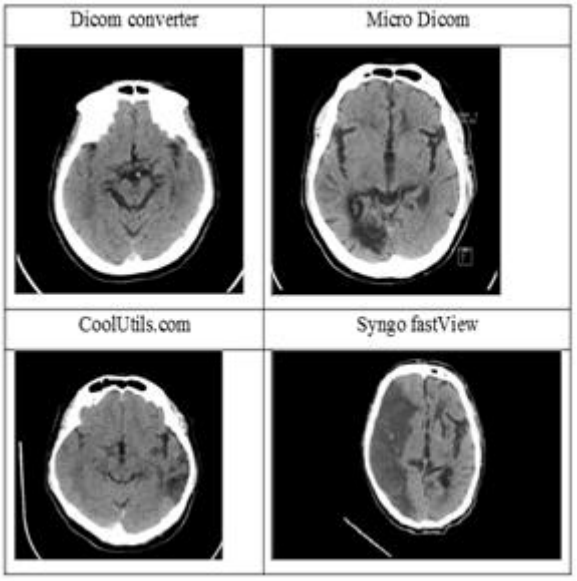



**Figure 2:** Conversion output

The Syngofast View application is utilized in this investigation. because the conversion outcomes are higher in dimensions and appear crisper and cleaner than those of other converters. In addition, the information present in the

dicom image can be deleted, leaving only the objects visible.

**b) Cropping**

The second step, data cropping, comes next. By leaving a small amount of the background black, cropping is done to draw attention to the head object in the current CT Scan image.

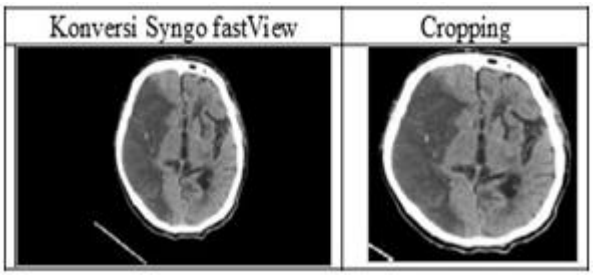
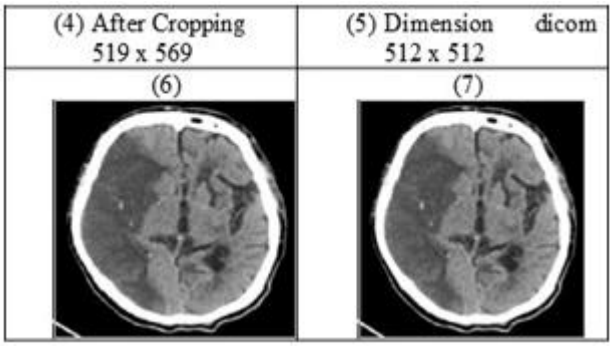



**Figure 3:** Cropping output

**c) Scaling**

Scaling is used to make the image's dimensions uniform because cropping results in varied image dimensions. Scaling is reset to the dicom image's dimensions with dimensions.




of 512 x 512

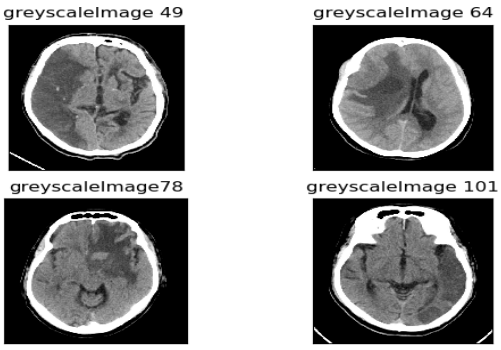



**Figure 5:** Greyscale output

**d) Data Augmentation**

Data augmentation involves recreating an image while preserving its essential qualities. Currently, there are 7 (seven) options to enhance the image, including random brightness, horizontal shift, vertical Flip, random rotation, and zoom image. The dataset utilized for hemorrhagic stroke and ischemic stroke after augmentation contains the same number of 1000 data for each type of stroke.

**7) Feature Extraction**

An image analysis method called Gray-Level Co-Occurrence Matrix (GLCM) calculates the corresponding texture features. Images made up of pixels, each with a distinct grey level intensity, are the image data from which features need to be retrieved. Contrast, Dissimilarity, Homogeneity, Energy, Correlation, and ASM are the six features employed in GLCM feature extraction.

The outcomes of feature extraction from data on hemorrhage and ischemic strokes with Class labels of 0 for hemorrhage and 1 for ischemic 1 are shown in the instances below.

**Table 1:** Outcome of Feature Extraction with GLCM

| Contrast | Dissimilarity | Homogeneity | Energy | Correlation | ASM | Class |
|---|---|---|---|---|---|---|
| 1.844102 | 0.3609838 | 0.8779072 | 0.3479091 | 0.9705421 | 0.1210407 | 0 |
| 2.029881 | 0.3434394 | 0.8914868 | 0.4156345 | 0.9726359 | 0.172752 | 1 |
| 1.8518008 | 0.3646675 | 0.873956 | 0.3622859 | 0.9722011 | 0.1312511 | 1 |
| 2.1414279 | 0.4448626 | 0.8491489 | 0.3544439 | 0.9669001 | 0.1256305 | 0 |
| 2.0834416 | 0.4300745 | 0.8509771 | 0.3381213 | 0.9673815 | 0.114326 | 1 |
| 2.073278 | 0.3800122 | 0.8751074 | 0.3798733 | 0.9691523 | 0.1443038 | 1 |
| 1.7004744 | 0.3282935 | 0.8874237 | 0.3568815 | 0.9698076 | 0.1273644 | 0 |
| 1.8116768 | 0.3223041 | 0.8947491 | 0.3560765 | 0.9642877 | 0.1267904 | 1 |
| 2.0153447 | 0.3853258 | 0.8691239 | 0.3375996 | 0.9691704 | 0.1139735 | 1 |
| 1.7087298 | 0.3274587 | 0.8884911 | 0.3617156 | 0.9698593 | 0.1308382 | 0 |

**Figure 4:** Scaling output

**8) Greyscale**

Greyscale is used to make an image's degree of grey uniform. Since the radiography digital image value will only have a pixel value between 0 and 255, the image scaling procedure will instead employ a greyscale method, where black is represented by the lowest number of colors (0), while white is the highest number (1), to create the image.

**a) Machine learning algorithm for grouping**

Naive Bayes, Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Neural Network MLP, Random Forest (RF), Deep Learning (DL), and Support Vector Machine are 8 machine learning techniques used to compare classification performance (SVM). experiments carried conducted with the Python library's default settings.

## 3. Experiment and Research

Data categorization is carried out utilizing sampling procedures, including k-fold cross validation with k = 10 and Leave-One-Out (LOO) cross validation, after data pretreatment using image processing and feature extraction. Regarding the evaluation of performance utilizing accuracy, precision, recall, and full measure. All algorithms utilized in this experiment make use of the Python library's default parameter settings.

5.39%. Even if the KNN method achieves an accuracy value of 94.54% better.

Following are the results of measurements of the seven machine learning algorithms with 10-fold cross validation.

**Table 2:** Performance Difference using 10-fold Cross Validation

| Grouping Method | Accuracy | Precision | Recall | F1- measure |
|---|---|---|---|---|
| KNN | 95.42% | 94.89% | 94.62% | 94.62% |
| Naïve Bayes | 71.29% | 70.92% | 74.47% | 72.65% |
| Logistic Regression | 83.01% | 81.28% | 77.47% | 79.33% |
| Decision Tree | 93.33% | 92.33% | 92.35% | 92.23% |
| Random Forest | 95.68% | 95.68% | 94.11% | 94.89% |
| NN-MLP | 88.44% | 87.49% | 84.87% | 86.16% |
| Deep Learning | 86.47% | 84.73% | 83.05% | 83.87% |
| SVM | 85.68% | 85.04% | 80.38% | 82.65% |

According to the findings of the trials, the classification algorithm utilizing the Random Forest approach had the best validation outcomes, with accuracy scores of 95.67%, precision scores of 95.69%, recall scores of 97.33%, and F-Measure scores of 94.88%.

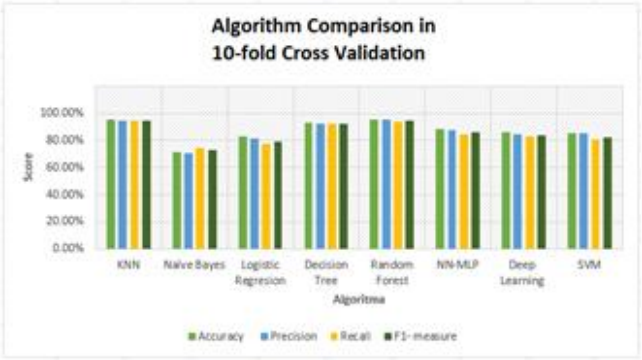



**Figure 6:** Performance Rate by 10-fold Cross

The performance results obtained while employing the Leave-One-Out (LOO) cross validation sampling approach are displayed in Table 5 and Figure 7. With LOO, the experimental findings demonstrate that the Random Forest approach yields the classification algorithm that describes the best accuracy value based on accuracy value, which is equivalent to 95.97% with a precision value of 94.39%, 96.12% recall and f1-Measure

**Table 3:** Performance Difference using Leave-One-Out (LOO) Cross Validation

| Grouping Method | Accuracy | Precision | Recall | F1- measure |
|---|---|---|---|---|
| KNN | 95.59% | **94.55%** | 95.79% | 94.80% |
| Naïve Bayes | 71.19% | 74.43% | 70.86% | 72.60% |
| Logistic Regression | 83.03% | 77.51% | 81.28% | 79.35% |
| Decision Tree | 93.57% | 92.62% | 92.34% | 92.48% |
| Random Forest | **95.98%** | 94.40% | **96.13%** | **95.40%** |
| NN-MLP | 85.33% | 86.72% | 93.16% | 89.77% |
| Deep Learning | 80.63% | 79.27% | 73.06% | 76.48% |
| SVM | 85.60% | 80.33% | 84.86% | 82.54% |

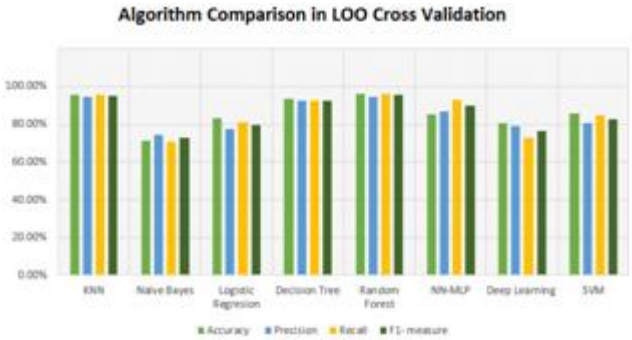



**Figure 7:** Performance Rate by LOO Cross

## 4. Conclusion

In this work, stroke data on CT scan image data is classified using machine learning methods. Picture processing and feature extraction are carried out on the image data prior to categorization. And after that the K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Multi-layer Perceptron (MLP-NN), Deep Learning, and Support Vector Machine are utilized in comparison to each other to do the classification. According to our experiment, when compared to other examined classification algorithms, the Random Forest method's classification algorithm has the best level of accuracy. The accuracy of the classification algorithm with the default optimization parameter value hasn't been tested, nevertheless. From this point forward, the categorization model may be enhanced to accomplish. To improve the accuracy of the machine learning algorithm, parameter adjustment is required.

## References

[1] L. Feigin et al., The GBD 2k13 research, vol. 45, no. 3, 2k15, provides an update on the worldwide burden of ischemic and hemorrhagic stroke in 1990-2k13.
[2] Stroke Epidemiology in South, East, and South-East Asia: A Review, B. W. Yoon, N. Venketasubramanian, J. C. Navarro and J. Pandian volume no. 20, no. 1, 2k18.
[3] P. K. Gupta and G. A. P. Singh, Performance analysis of multiple machine learning-techniques approaches for diagnosing and grading lung cancer in humans, vol. 3456789, Springer London, 2k18.
[4] Highly efficient initial ischemia stroke detection approach utilizing CNN deep learning algorithm, C. L. Chin et al., vol. 2K18-Janua, number. iCAST. 2k18.
[5] U. Andayani, Seniman and J. T. Marbun Grouping of strokes hitting diseases with the help of CNN, vol. 978, number. 1.2K18.
[6] W. C. Chen, C. Y. Hung C. H. Lin, P. T. Lai, and C. C. Lee, measuring both deep neural network and other ML algorithms for stroke detection in a huge-scale population-based on database of electronic medical claims.2K17.
[7] Sukesh Kumar, Jenna R.S "Stroke detection with the help of SVM", Instrumentation, International Conference on Control, Computational Technologies, and Communication (ICCICCT),2K16.
[8] P. Kornprobst, S. Paris, F. Durand, and J. Tumblin, A nice and useful introduction to filtering of bilateral and to the applications of it.2008.
[9] M.K.S, Alsmadi, K., Omar & Noah, S. A (2009) Back Propagation Algorithm: Topmost Algorithm when compared to the other Multi-layer Perceptron Algorithms. CS and Network Security International Journal, 9 (4),PP.378-383.
[10] W. J. Powers et al., "2K15 Heart Association of America/Stroke Association of America Focused on Update of the 2K13 instructions for the initial Management of Patients Regarding Endovascular Treatment with Acute Ischemic Stroke," Stroke, volume 46, number. 10, pp. 3020–3035, 2K15.