International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org **Original Research Paper**

A Novel Machine Learning based Stroke Prediction System using Magnetic Resonance Imaging and Adaptive New Fuzzy Inference System

Dr. Ajanthaa Lakkshmanan*[1], Dr. Adaline Suji R.[2], Dr. Priyanka N.[3], Dr. D. Bright Anand[4], Mr. Komatigunta Nagaraju[5], Dr. U. Ganesh Naidu[6]

Submitted: 21/10/2023 **Revised**: 17/12/2023 **Accepted**: 24/12/2023

Abstract*:* Smart health analytics is a highly researched field that employs the power and intelligence of technology for efficient treatment and prevention of several diseases. We are currently living in the post COVID phase, which has seen a tremendous rise in sudden deaths caused by many neurological diseases, among which stroke is the major one. It is considered to be the second largest causative disease of death amongst human population according to the World Health Organization. Hence this paper proposes a new method for predicting the onset of stroke using the machine learning approach of Adaptive Neuro Fuzzy Inference System (ANFIS). The input data set for stroke prediction is obtained from Kaggle data repository called as the Brain Stroke prediction dataset which contains 5111 electronic health records of patients with 11 different parameters related to the stroke disease along with brain MRI images. The data obtained is preprocessed using data cleaning methods, segmented using SegNet and features relevant are extracted using CapsuleNet. Predictive analytics is done using ANFIS model and is compared with existing classifiers like Logistic Regression, Random Forest, XG Boost algorithm, Adaboost algorithm and Gated Recurrent Unit. The predictive performance of the proposed model is tested using metrics like accuracy, precision, sensitivity, specificity, F1 measure and ROC curve analysis.

Keywords: *ANFIS, Brain, CapsuleNet, Prediction, SegNet, Stroke*

1. Introduction

Brain is considered to be the prime organ that supports conscious living and stroke results in either death or unconscious living depending upon the individual’s conditions. After the attack of the novel Corona Virus Disease 2019(COVID-19), there has been an upsurge of various diseases and disorders. Among them, the most common is the cardiovascular and neurological disorders which are seen as the prominent consequence caused by COVID-19[1]. This is because during the first and second waves of COVID, which hit almost all the countries across the globe, many people succumbed, and all the survivors are continuing to experience one or the other post COVID complications. It is a well-known fact that COVID created an impact and damaged most of the vital organs of the human body, among which the heart and brain are the worst ones to be affected. There are several diseases that affect the brain such as stroke, brain tumors, dementia, epilepsy, Parkinson’s disease and Alzheimer disease [2]. Some of them are considered to be neurodegenerative in nature which happens with aging and many other factors. Here in this paper, we are concentrating on the disease of stroke.

Stroke is considered to be a neurological disorder that is developed because of the blood supply disturbance in the brain [3]. It can be divided into two types called the Ischemic stroke and hemorrhagic stroke. While ischemic stroke is caused because of the clot in blood vessel, hemorrhagic stroke is caused due to the sudden rupture or burst of one of the major brain vessels. Without proper supply of blood to brain cells, they begin to die which results in disability of the brain and hence the disease. This is also called a cerebrovascular accident. While ischemic stroke accounts for 80 to 85% of total stroke rate, hemorrhagic stroke accounts for 15 to 20% of stroke disease [4]. It can also be classified as long-term stroke and short-term stroke based on the length of suffering. For example, if stroke can be resolved and normally restored within three years, it is called a short-term stroke. Treatment and rehabilitation existing more than a period of three years would be long term stroke.

According to the World Health Organization, a recent health survey says that 15 million human beings get affected due to stroke annually out of which 5.5 million is irrecoverable and loss their lives [5]. Another report says that in every 4 to 5 minutes, people die due to stroke. It is becoming a highly prevalent and fatal disease which needs prior identification and proper treatment to

[1] *Assistant Professor, Department of Computing Technologies, SRM Institute of Science and Technology, Chennai-603 202.*
[2] *Associate Professor grade-1, Scope, Vellore Institute of Technology, Vellore campus 632 014.*
[3] *Assistant Professor senior grade-1, Scope, Vellore Institute of Technology, Vellore campus – 632*
[4] *Professor, Department of Computer Science and Engineering, Narayana Engineering College, Gudur, Tirupati District, 524101.Andhra Pradesh.*
[5] *Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist, A. P, India. Email: knagaraju@kluniversity.in.*
[6] *Assistant professor,CSBS, B V Raju Institute Of Technology, Narsapur, Medak, Telangana, India-502313,ganesh613@gmai.com*

avoid any loss of life. Stroke is considered to be the largest contributor of acquired disability and is responsible for 5% of disability adjusted life years. It accounts for 11% of total deaths amongst human population in the world [6]. There are various factors which act as a source of stroke such as low blood supply, lifestyle modifications, food and nutrition, body mass index maintenance, age, hypertension, heart dysfunction. While factors such as gender, hereditary imperfections and are non-modifiable, certain factors are considered controllable which can be managed such as obesity, good health, physical and mental wellness [7].

Predictive analysis has been applied to various medical fields such as oncology, cardiology, neurology, pancreatic diseases, hypertension etc. [8]. Stroke prediction has been given less attention because of the complexity involved in identification of the disease. Stroke prediction is now gaining importance because of its high prevalence and increased number of fatality involved. There has been a massive increase in the number of cases reported and hence any form of automated diagnosis of stroke is considered to be helpful for the physicians [9]. Human diagnosis without the support of imaging technologies such as radiography, sonography, Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) results in error prone diagnostics. While there are many medical imaging modalities available for stroke diagnosis, MRI is considered to be the most efficient one as it produces a high dimensional image that gives a detailed view of brain that is three dimensional with good resolution [10]. CT scans are also helpful in the diagnosis of stroke, but CT scans are helpful only in the case of post stroke analysis which produces multi slice scans that are recognizable only after the onset of the disease. But since we are dealing with predictive analytics, we are considering using MRI scan images.

Stroke analysis and diagnosis usually require a detailed study of the body along with neurological images of the brain to arrive at the proper conclusion. Stroke treatment is comparatively costly when compared to other diseases because of which people are now more involved in the prediction of stroke, so that it can be treated at a very early stage. It is beneficial to identify the disease at an earlier stage rather than identifying it at the later stage which adds on to the medication cost and rehabilitation cost. Physiological medications are considered to be cost effective 3 to 8 times when compared to surgical measures that are opted in the later stage of treatment, because of which physicians and patients are interested in the early prediction of stroke [11].

The ultimate aim is to forecast the attack of stroke in the considered individual and identify the symptoms and warnings produced by the human brain in a body in a very early stage in order to lessen the severity of the disease after its onset. Recovery is also based on proper treatment and early medications, good lifestyle practices such as diet, exercise, avoiding tobacco and alcohol and maintaining good health of the heart. It is also considered to affect the economy as more and more people are becoming disabled due to this disease. Accurate prediction and timely diagnosis becomes very essential. Stroke is caused by cerebral disturbance which can be identified only by means of medical imaging and scans. These radiological scans can easily adapt to machine learning models and artificial neural networks can play a very vital role in this. Computer aided diagnosis has shown good performance in the area of cognitive analysis. Machine learning has been believed to produce good results in a variety of medical fields. It can be assured that predictive analysis in the cognitive domain can aid the clinical services to a great extent and it has been proven to be a powerful tool in the realm of health.

2. Literature Survey

In paper [12], the authors describe the efficiency of Capsnet architecture of Convolutional Neural Networks (CNN). They have used these two algorithms for classification of brain hemorrhage using 3607 CT images of the brain. The images were subjected to preprocessing techniques like noise removal, skull scrapping and then given as input to UNet for segmentation and Capsnet for classification which resulted in a final accuracy of 92.1 %. This paper explains the advantages of using Capsnet over other CNN architectures.

Paper [13] analyzes the classification performance of various machine learning algorithms for predicting stroke. The algorithms studied in this review article are logistic regression, decision tree, random forest, k nearest neighbor, support vector machine and Naives bayes and the paper has been concluded stating that Naives bayes performed well than the other machine learning models. The input data set was obtained from Kaggle and preprocessed with techniques like handling imbalanced and missing values. The data was then split in the ratio of 80:20 for training and validation and various algorithms were applied on the data set for classifying. The model was implemented in simple HTML and Python.

In [14], the authors have discussed the prediction performance of CapsNet regarding Alzheimer disease which is considered to be a kind of neurodegenerative disorder. The performance of the proposed capsule net model was compared with techniques like logistic regression, random forest and multilayer perceptron. Datasets from Andy and Kaggle websites were used, and classification metrics of precision, recall and F score were calculated for both the datasets in which the Kaggle data set obtained highest prediction average of 93.54.

In this article [15], a pilot study has been conducted for prediction of ischemic stroke using different deep learning methods. Data was collected from two hospitals in Australia in real time and comparative prognostication score was calculated. Open-source Python libraries were used for preprocessing the clinical data acquired such as resizing interpolation, features scaling etc., and classification was done using convolutional neural networks and artificial neural networks. Results were statistically analyzed which included 204 patients roughly of the age 70. Metrics like sensitivity, specificity, F1 score, accuracy and AUC were calculated to estimate the performance of the proposed system.

This paper [16] presents to us a systematic review using neuro fuzzy techniques for predicting neurological disorders. In order to write this paper, a total of 303 well written articles were examined. Prediction was done for diseases including depression, heart attacks, Parkinson’s disease, brain tumors, Alzheimer, seizure disorder etc. This paper has been concluded in a way that among all the neuro fuzzy systems, the Gaussian member function and Sugeno fuzzy system are the most efficient ones which is closely followed by the Mamdani system.

Paper [17] describes the prediction process of brain stroke that has been done with the help of many machine learning and deep neural network algorithms. Random forest, Extreme gradient boosting, light gradient boosting, Adaboost, SVM linear kernel, three layer and four-layer artificial neural networks were used in this study. Random forest algorithm achieved the highest classification accuracy of 99% followed by three layer and four-layer deep neural networks. Various preprocessing techniques like data categorization, label encoding, standard scaling and the principal component analysis were done on the dataset before classification and prediction was conducted.

In this article [18], the authors have compared the performance of shallow and deep learning methods for predicting stroke from MRI images. Dataset was obtained from the Washington University

School of Medicine stroke patients and methods like Ridge regression, support vector regression, convolutional neural networks, a hybrid model combining regression with CNN was also executed and the results of all the proposed techniques were measured in terms of accuracy, centroid redundancy, location and topology redundancy and topological similarity.

Paper [19] explains the detection of brain stroke using MRI images. The proposed system uses LeNet and SegNet segmentation and classification algorithms. The input data set was obtained from anatomical tracing supplications from a stroke website containing 229 weighted MRI scans. The images were preprocessed with techniques like resizing, image denoising, normalization and contrast enhancement. Segmentation and classification was done using SegNet and LeNet. Finally stochastic gradient descent optimizer was applied towards the end of the computation after classification. Mean square error and cross entropy was calculated which resulted in a training accuracy of 97% and testing accuracy of 93%.

This paper [20] also explains the performance and efficiency of Segnet algorithm for local refinement and brain segmentation in a three-dimensional format and the model has been evaluated against the brats 2015 benchmark for brain segmentation in which it achieved very promising results. This paper uses Ischemic stroke lesion segmentation 2017 database. The proposed algorithm was implemented in Pytorch where each MRI scan had 155 slides of the size 12 * 128*128. Random dropping was used for preprocessing before segmentation. Sensitivity and specificity and dice values were calculated for arriving at the results.

In paper [21], a hybrid deep learning model combining Group Handling Method (GMDH) and Long Short-Term Memory (LSTM) was used for predicting stroke in a smart hospital based mobile platform. Many datasets including EMG lower limb data set mHealth data set well used for collecting data and preprocessing techniques like normalization and feature scaling were used before training the GMDH LSTM classifier. For predicting the occurrence of stroke, you need to know the patient details. Hyper parameter details included are batch size of 256 with a learning rate of 0.0001, training loss of 0.1890 and validation loss of 0.2364 with an epoch size of 100.

[22] explains the predictive capacity of Adaptive Neuro Fuzzy Inference System (ANFIS) for predicting long term and short-term stroke using clinical data and ultrasound carotid imaging. Data was collected from asymptotic patients with the ultrasound scan of the carotids. Patients included in this cohort study did not have previous cases of stroke or any such neurological disorders and written consent was obtained from all of them. Features extracted included statistical features, spatial gray level dependence matrices, correlation, contrast, multi scale morphology features etc. with the help of support vector machine and classification was done using ANFIS model which achieved an accuracy of 97%.

3. Proposed System

The proposed system collects input data from Kaggle repository and a well-known MRI data set for brain strokes. The acquired data is preprocessed using techniques like redundancy removal, filling in the missing data, handling imbalanced data, label encoding etc. The acquired MRI images are then segmented using Segnet model of convolutional neural networks and features are extracted from the segmented model using capsule net architecture. Using the features extracted, prediction is calculated based on ANFIS model and metrics such as accuracy, precision, sensitivity, specificity, F1 score, and Receiver Operating Characteristic (ROC) curve are calculated. The proposed system is also measured against existing techniques like logistic regression, random forest, XGboost algorithm, Adaboost algorithm and gated recurrent units based on stroke prediction. Figure 1 below explains the workflow of the proposed model.

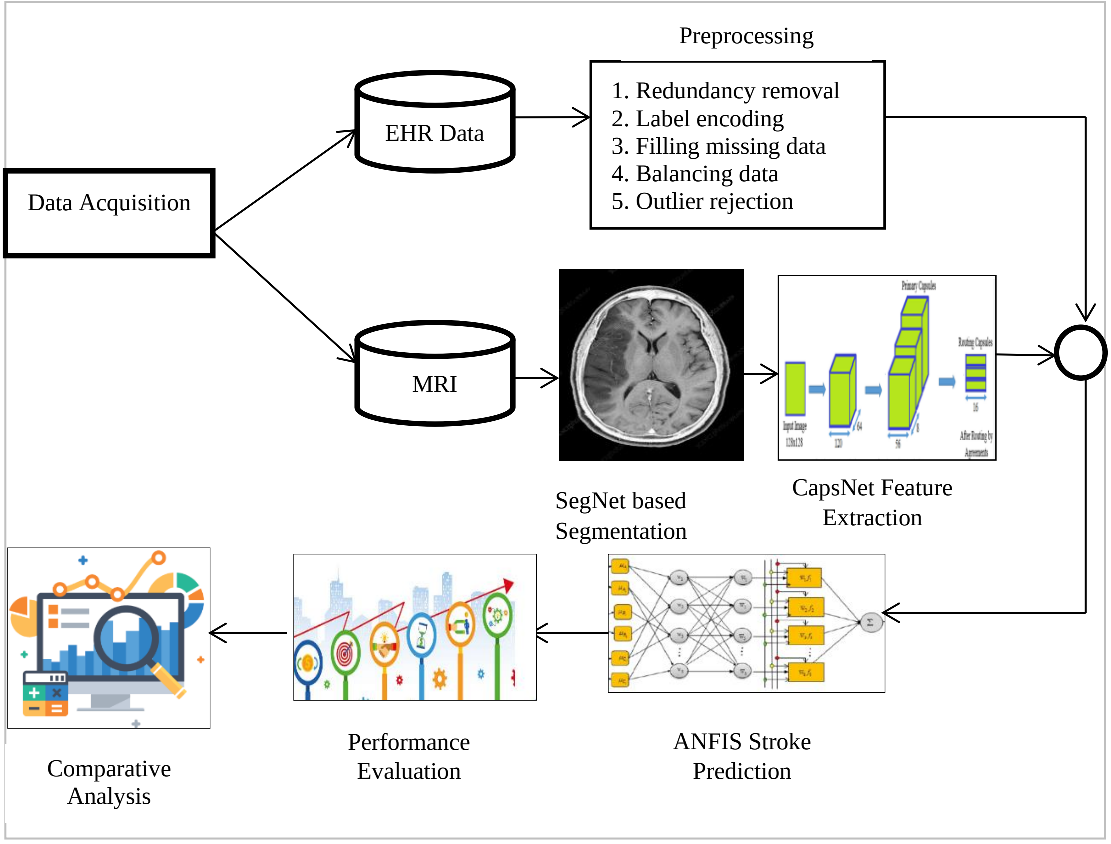


Fig. 1. Workflow of proposed system

3.1 Data acquisition

The proposed system combines clinical MRI data along with various physiological and psychological parameters of patients so that better prediction accuracy can be attained. Hence the input data used here is of two types including electronic health records of stroke and non-stroke patients collected from Brain Stroke Prediction Dataset of Kaggle repository. Another data is the MRI dataset that has been acquired from a publicly available website. Kaggle dataset contains health records of 5110 patients with 12 parameters namely patient ID, gender, age, hypertension, heart disease, marital status, work type, residence, glucose level, BMI, smoking status and previous indication of stroke attack etc.

3.2 Preprocessing

MRI scan images can be used in a raw format and no further processing is required for them as it is clinical data. But the electronic health records obtained from the Kaggle repository needs some kind of data cleaning as using them as such may degrade the predictive performance of the proposed system. Hence preprocessing techniques like data redundancy removal, handling missing data and imbalanced data, label encoding, outlier removal etc. have been carried out on the input data set [23]. Data redundancy removal is a very simple technique which removes any repetition of health records by mistake. Missing data are filled with average values of the same and label encoding is nothing but converting the strings that are present in the dataset into a numerical format so that it is interpretable by the proposed classifier. Outlier removal removes irrelevant information from the dataset such as marital status, residence, work type etc. as they are not considered to be of any importance to the stroke prediction task. It is enough that minimum preprocessing is performed in case of any medical data as removing more and more information will tend to deviate from the proposed work.

3.3 Segmentation

It is essential that the input MRI is properly segmented so that it can give good prediction accuracy of stroke disease. The MRI images are segmented using SegNet algorithm of CNN. It is a semantic based segmentation algorithm that is used mainly for the purpose of segmenting each and every pixel of the image to its target class. It is also based on the encoder decoder format [24]. The encoder is responsible for operations like convolutions and max pooling in order to produce feature maps. The structure of the encoder resembles that of the convolutional layers present in the VGG16 architecture. Batch normalization and element wise activation is performed on the generated feature maps which are then max pooled and subsampled so that the output feature map is invariant to any translation operations.

The decoder is similar to the encoder in structure but performs reverse up sampling which includes a softmax classification layer at the end. It segments the pixels of the input image and maps it to the corresponding output class label. SegNet is usually applied to images which possess a spatial relationship among them. It is considered to be efficient because of its lesser number of parameters. The decoder utilizes the feature map generated by the encoder and in turn produces sparse and dense feature maps which are then given as input to the softmax classifier for appropriate segmentation. Figure 2 depicts the architecture of SegNet based MRI segmentation.

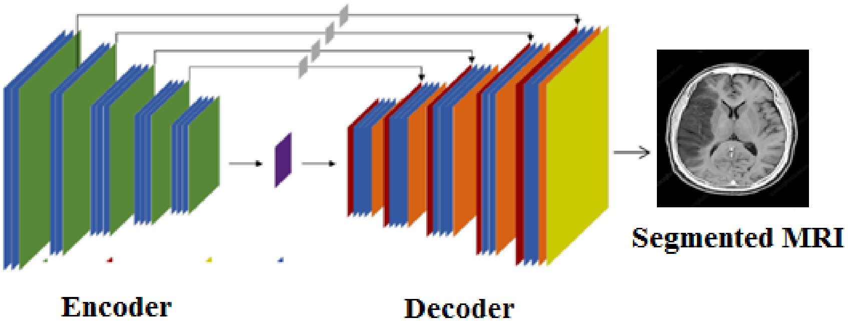


Fig. 2. SegNet Architecture

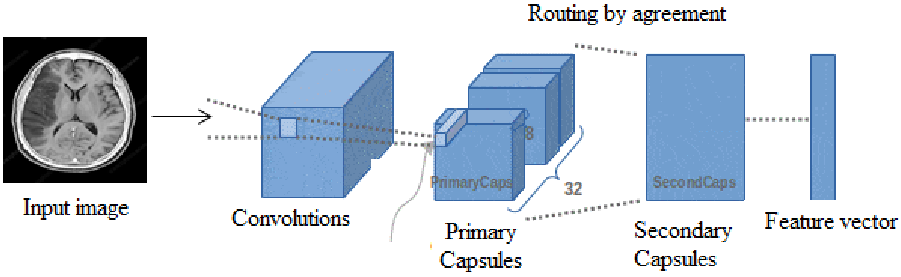


Fig. 3. CapsNet based feature extraction

3.4 Feature Extraction

From the segmented MRI images, relevant features are extracted using capsule net model of convolutional neural network. CapsNet is used in a variety of machine learning tasks but has found a recent place in medical feature extraction because of its promising results. The main idea behind Capsnet is to enable transfer of learning information amongst the capsules with the help of a dynamic routing protocol so that each capsule is aware of the activity of the other capsule in order to produce a better output [25]. This is because of which it has turned out to be a powerful feature extractor which has the ability to pull out very important information with a lesser learning rate. It is considered to be advantageous over others in terms of lesser parameters needed as nodes are grouped to form capsules and connections are less between them. CapsNet is a model that has been designed in close relationship with the organization of biological neurons called capsules. Each capsule has an individual activity vector that has been assigned to it. The job of every capsule is to predict the output of the higher capsules by using the input in hand. The capsules at the higher layers tend to use the correct predictions of lower capsules and make their own decisions. This model is considered to be more accurate when compared to others because of this property of interdependence among the capsules. A decision is taken only when it is supported by the majority of capsules which agree to each other with the help of a routing by agreement protocol. This model of CNN avoids max pooling strategy in order to reduce the processing overhead and hence it is able to extract features in a vector format rather than the standard scalar features. It produces a more generalized features vector that is invariant to viewpoints and is believed to be defensive against white box adversarial attacks. It can extract positional features, features based on texture, size, orientation, deformation etc. It is also based on the encoder decoder structure which in turn contains convolutional layers that are divided into primary capsule layers and secondary capsule layers. The primary capsule layers are eight dimensional and secondary digit capsules are 16 dimensional capsules for each class. The decoder on the other hand contains only fully connected rectified linear units and sigmoid functions. Figure 3 shows the architecture of CapsNet model.

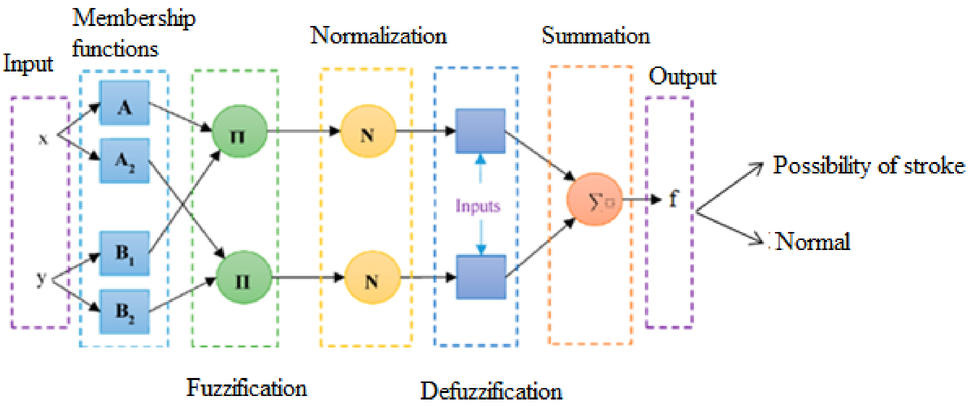


Fig. 4. ANFIS architecture

3.5 Prediction

Based on the segmented MRI images, extracted features and preprocessed EHR Kaggle dataset, the onset of stroke is predicted with the help of Adaptive Neuro Fuzzy Inference System (ANFIS) classifier. The adaptive neuro fuzzy inference system is basically a machine learning technique that employs artificial neural networks for carrying out its task. It constitutes a group of fuzzy rules based on Takagi Sugeno system of fuzzy inference [26]. It is composed of five layers called the input layer, fuzzification layer, normalization layer, defuzzification layer and output layer each of which is enabled with a specific work. The equations for each of the layers are given below in (1) to (5).

$$O_i^1 = \mu A_i(x),\ i = 1{,}2\ or\ O_i^1 = \mu B_{i-2}(y),\ i = 3{,}4 \qquad (1)$$

$$O_i^2 = w_i = \mu A_i(x) \times \mu B_i(y),\ i = 1{,}2 \qquad (2)$$

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2},\ i = 1{,}2 \qquad (3)$$

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i(p_1 x + q_1 y + r_1),\ i = 1{,}2 \qquad (4)$$

$$O_i^5 = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad (5)$$

where,

x and y - input value of a node *i* and

A_i and B_{i-2} - linguistic values of a node *i*.

w_i represents the firing strength of a rule *i*

(p_1, q_1, r_1) represents the parameter set.

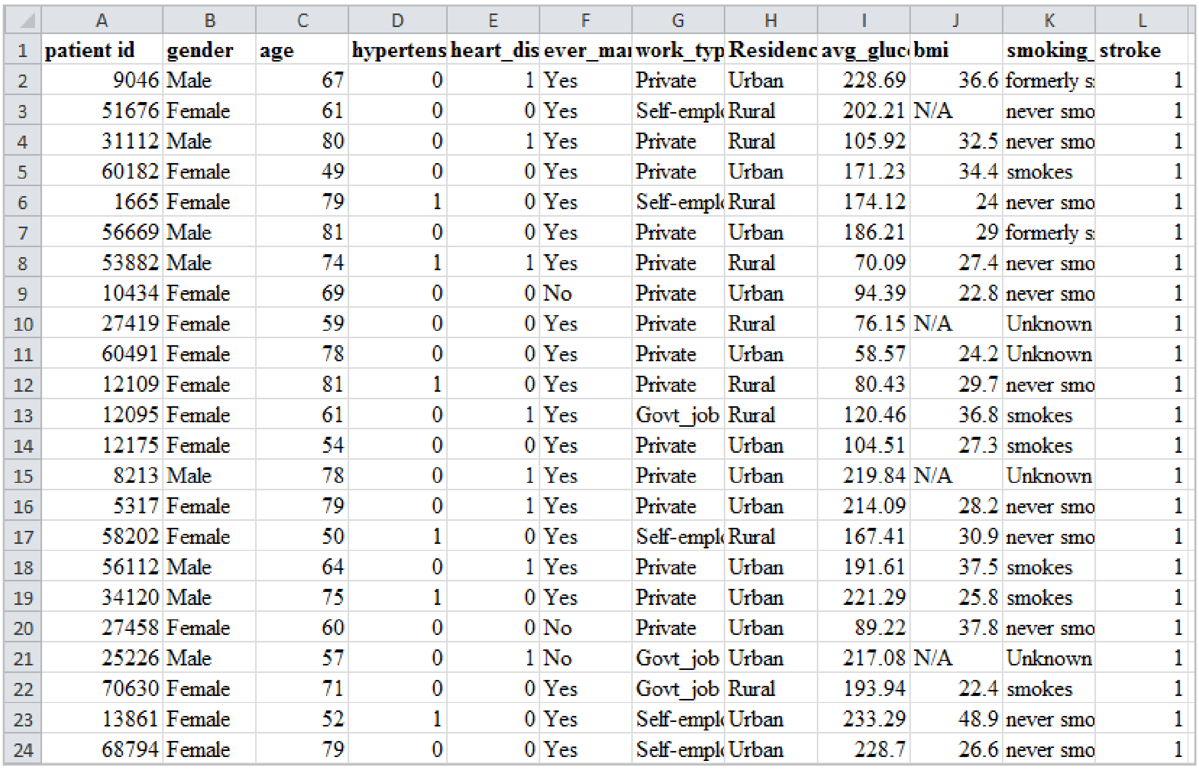
$\bar{w}_i f_i$ - consequent rule.

It is now considered a highly efficient model of prediction as it concatenates the working style of artificial neural networks and fuzzy logic. One of the major benefits of fuzzy logic is its ability to handle fuzzy data that is incomplete in nature. This may be of particular use in the medical domain where not all the information could possibly be collected or available for our prediction task. This model is based on an adaptive learning pattern, and it is dynamic in nature as it can alter the connections between the fuzzy neurons based on the requirement. This nature imparts adaptability to the underlying model because of which it is very much suitable for prediction in the medical domain [27]. It has been used in a variety of areas like control systems, big data, recognition of patterns, advanced decision making and predictive analysis to a very great extent. Figure 4 shows the architectural diagram of this model.

4. Results and Discussion

4.1 Experimental results

In order to carry out the proposed system, input data was acquired from Kaggle repository's Brain stroke prediction dataset. It contained electronic health records of 5110 patients. Each patient record shows patient ID, gender, age, hypertension, heart disease, marital status, work type, residence, glucose level, BMI, smoking status and previous indication of stroke attack. Figure 5 below shows the sample electronic health records.



	A	B	C	D	E	F	G	H	I	J	K	L
1	**patient id**	**gender**	**age**	**hypertens**	**heart_dis**	**ever_mar**	**work_typ**	**Residenc**	**avg_gluc**	**bmi**	**smoking_**	**stroke**
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s	1
3	51676	Female	61	0	0	Yes	Self-empl	Rural	202.21	N/A	never smo	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-empl	Rural	174.12	24	never smo	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
17	58202	Female	50	1	0	Yes	Self-empl	Rural	167.41	30.9	never smo	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
23	13861	Female	52	1	0	Yes	Self-empl	Urban	233.29	48.9	never smo	1
24	68794	Female	79	0	0	Yes	Self-empl	Urban	228.7	26.6	never smo	1

Fig. 5. Sample EHR data

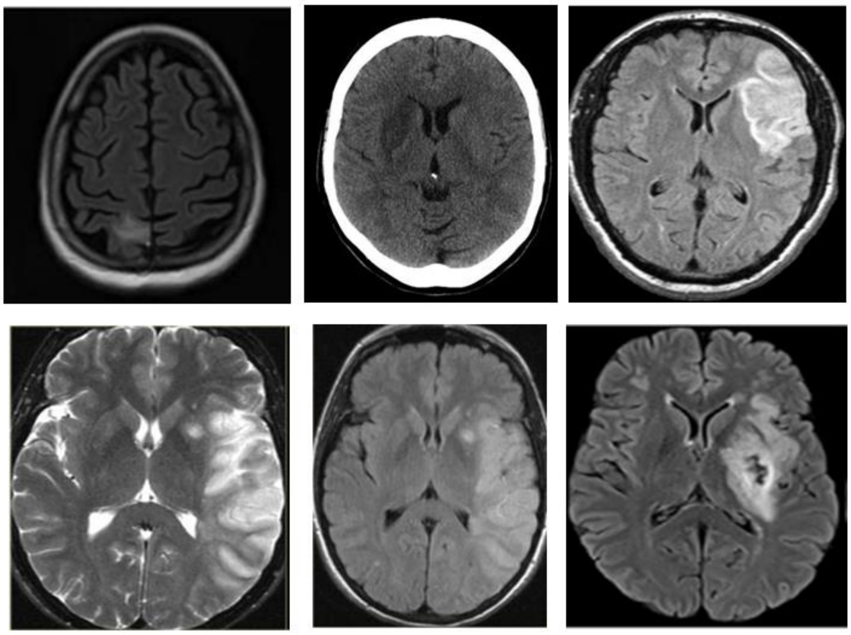
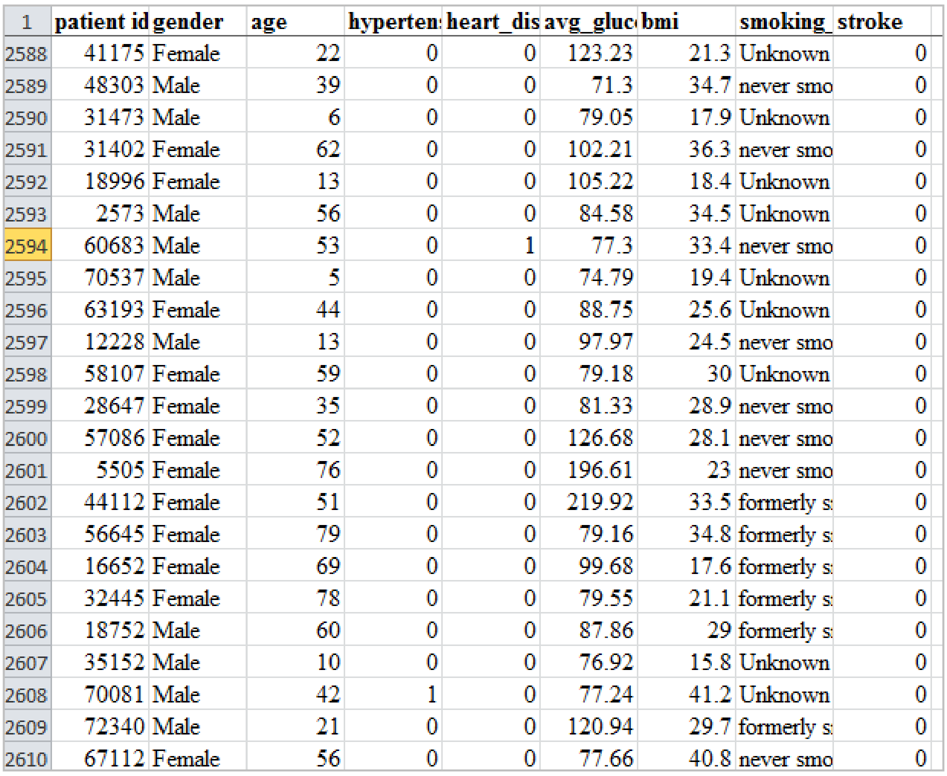


Fig. 6. Sample MRI images collected

Figure 6 shows the sample MRI images containing stroke and non-stroke attack brain segments. During preprocessing the electronic health records, 78 redundant records were identified and removed from the dataset. Redundancy could not be avoided in such large databases which contain information of more than 5000 patients. Similarly, 523 values pertaining to BMI and average glucose level were found missing in the database which were filled with average values of the corresponding rows. Figure 7 shows the output of outlier rejection where the information is not considered necessary for the proposed system.



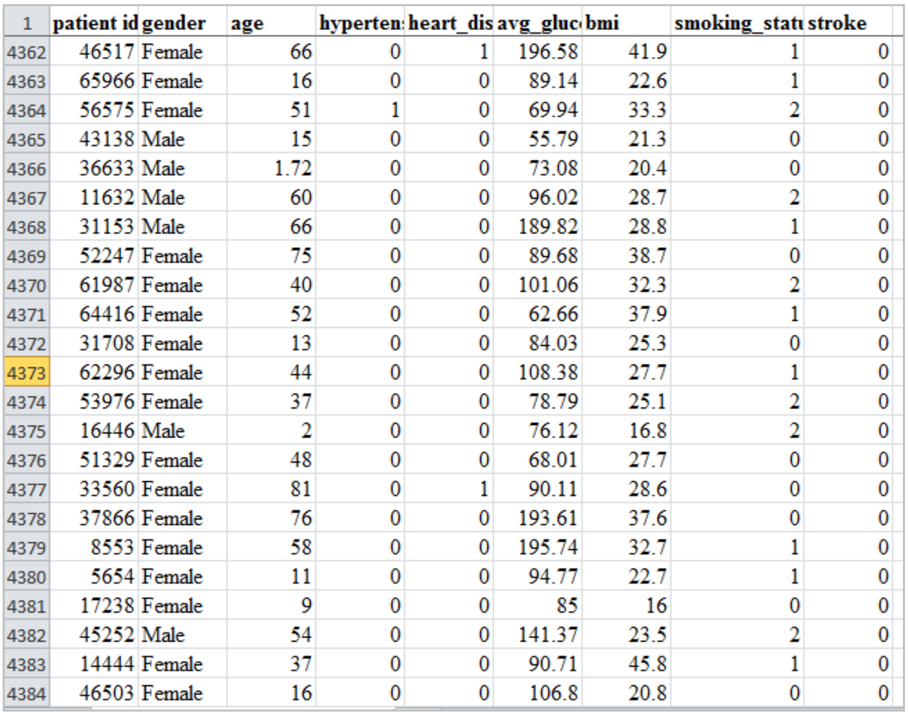
1	patient id	gender	age	hyperten	heart_dis	avg_gluc	bmi	smoking_	stroke
2588	41175	Female	22	0	0	123.23	21.3	Unknown	0
2589	48303	Male	39	0	0	71.3	34.7	never smo	0
2590	31473	Male	6	0	0	79.05	17.9	Unknown	0
2591	31402	Female	62	0	0	102.21	36.3	never smo	0
2592	18996	Female	13	0	0	105.22	18.4	Unknown	0
2593	2573	Male	56	0	0	84.58	34.5	Unknown	0
2594	60683	Male	53	0	1	77.3	33.4	never smo	0
2595	70537	Male	5	0	0	74.79	19.4	Unknown	0
2596	63193	Female	44	0	0	88.75	25.6	Unknown	0
2597	12228	Male	13	0	0	97.97	24.5	never smo	0
2598	58107	Female	59	0	0	79.18	30	Unknown	0
2599	28647	Female	35	0	0	81.33	28.9	never smo	0
2600	57086	Female	52	0	0	126.68	28.1	never smo	0
2601	5505	Female	76	0	0	196.61	23	never smo	0
2602	44112	Female	51	0	0	219.92	33.5	formerly s	0
2603	56645	Female	79	0	0	79.16	34.8	formerly s	0
2604	16652	Female	69	0	0	99.68	17.6	formerly s	0
2605	32445	Female	78	0	0	79.55	21.1	formerly s	0
2606	18752	Male	60	0	0	87.86	29	formerly s	0
2607	35152	Male	10	0	0	76.92	15.8	Unknown	0
2608	70081	Male	42	1	0	77.24	41.2	Unknown	0
2609	72340	Male	21	0	0	120.94	29.7	formerly s	0
2610	67112	Female	56	0	0	77.66	40.8	never smo	0

Fig.7. Outlier Rejection

Three columns such as marital status, residence and work type are not essential for the stroke prediction task and hence are considered as outliers and removed from the dataset. Figure 8 shows the output of label encoding task of data preprocessing. In this dataset string values were possessed only by smoking status column. It has been label encoded and its values are converted into integers as follows.

'Never smoked' = 0; 'Formerly smoked' = 1; 'Smokes' = 2

Figure 9 shows the segmented MRI images using SegNet algorithm. Figure 10 shows the workflow of ANFIS based stroke prediction.



1	patient id	gender	age	hyperten	heart_dis	avg_gluc	bmi	smoking_statu	stroke
4362	46517	Female	66	0	1	196.58	41.9	1	0
4363	65966	Female	16	0	0	89.14	22.6	1	0
4364	56575	Female	51	1	0	69.94	33.3	2	0
4365	43138	Male	15	0	0	55.79	21.3	0	0
4366	36633	Male	1.72	0	0	73.08	20.4	0	0
4367	11632	Male	60	0	0	96.02	28.7	2	0
4368	31153	Male	66	0	0	189.82	28.8	1	0
4369	52247	Female	75	0	0	89.68	38.7	0	0
4370	61987	Female	40	0	0	101.06	32.3	2	0
4371	64416	Female	52	0	0	62.66	37.9	1	0
4372	31708	Female	13	0	0	84.03	25.3	0	0
4373	62296	Female	44	0	0	108.38	27.7	1	0
4374	53976	Female	37	0	0	78.79	25.1	2	0
4375	16446	Male	2	0	0	76.12	16.8	2	0
4376	51329	Female	48	0	0	68.01	27.7	0	0
4377	33560	Female	81	0	1	90.11	28.6	0	0
4378	37866	Female	76	0	0	193.61	37.6	0	0
4379	8553	Female	58	0	0	195.74	32.7	1	0
4380	5654	Female	11	0	0	94.77	22.7	1	0
4381	17238	Female	9	0	0	85	16	0	0
4382	45252	Male	54	0	0	141.37	23.5	2	0
4383	14444	Female	37	0	0	90.71	45.8	1	0
4384	46503	Female	16	0	0	106.8	20.8	0	0

Fig. 8. Label Encoding

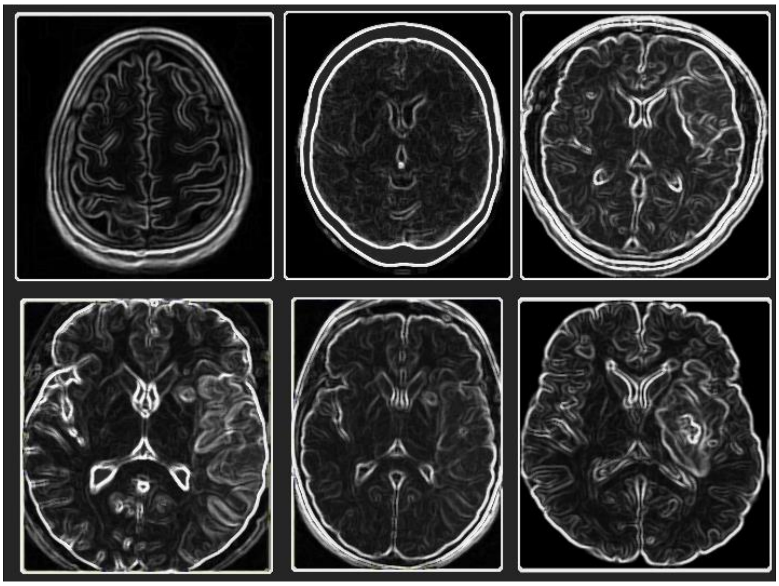


Fig. 9. SegNet based brain MRI segmentation

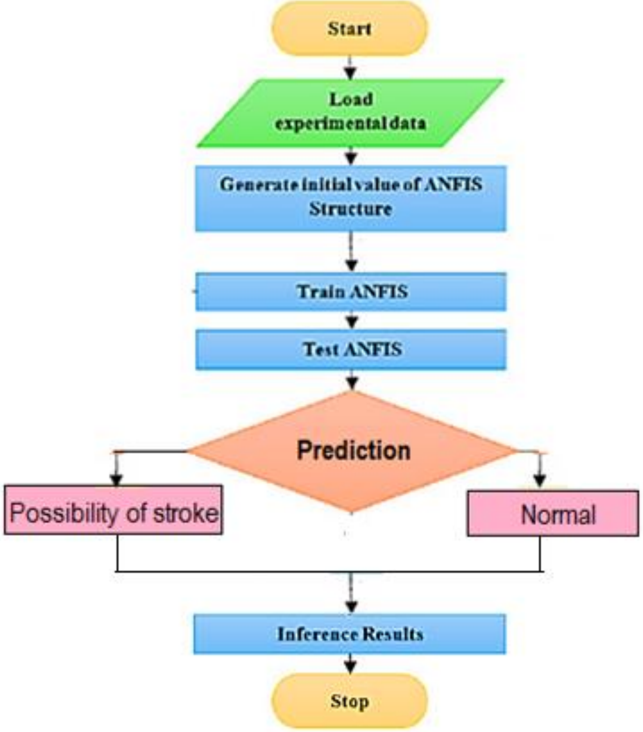


Fig.10. Stroke Prediction of ANFIS

4.2 Performance metrics

Predictive performance of the proposed model is calculated using metrics like accuracy, sensitivity, specificity, F1 score. They are calculated from True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values. Equations of all of them have been given below.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Specificity} = \frac{TP}{TP+FN} \quad (8)$$

$$F_1\ score = \frac{2*sensitivity*specificity}{sensitivity+specificity} \quad (9)$$

Table 1. Stroke Prediction Performance

S.No.	Metrics	Values
1.	Accuracy	97.46
2.	Sensitivity	94.85
3.	Specificity	95.3
4.	F1-Score	96.67

Table 1 shows the predictive performance of the proposed model. Figure 11 illustrates the prediction performance of the ANFIS model in a graphical format.

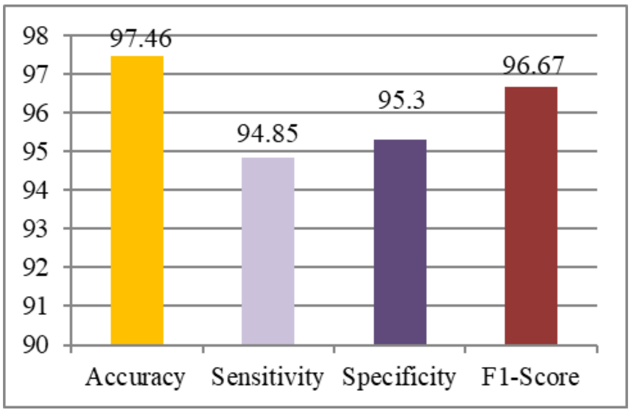


Fig. 11. Performance of ANFIS stroke prediction

4.3 Comparative Analysis with Existing Techniques

The performance of the proposed ANFIS model is compared with predictive performance of classifiers such as Logistic Regression, Random Forest, Xgboost, Adaboost and Gated Recurrent Units (GRU). Table 2 shows the comparison of the various model's accuracies produced in terms of stroke prediction.

Table 2. Performance Comparison with existing methods

S.No.	Methods	Prediction Accuracy
1.	Logistic Regression	73.4
2.	Random Forest	81.75
3.	XGboost	87.2
4.	Adaboost	91.2
5.	GRU	93.42
6.	Proposed system	97.46

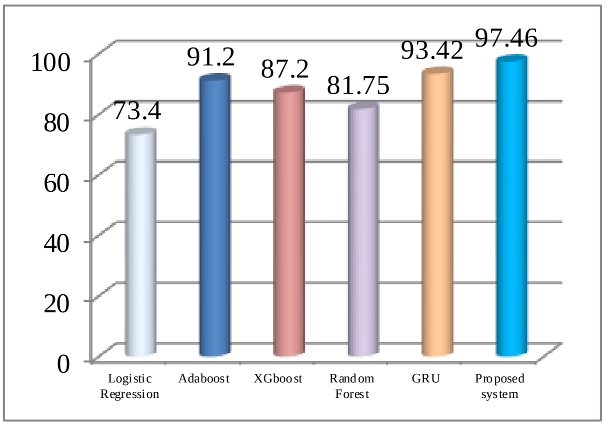


Fig. 12. Comparative analysis

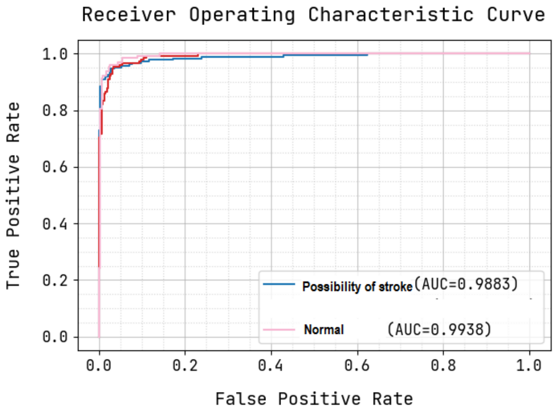


Fig. 13. ROC Curve analysis

Figure 12 portrays the comparative analysis of the proposed model along with its existing classifiers. Figure 13 shows the ROC curve analysis of ANFIS model. It is evident from the figure that both the cases of occurrence and non-occurrence of stroke have AUC values higher than 0.98 which indicate good prediction performance of the proposed ANFIS model.

5. Conclusion

Machine learning has produced enormous revolutions in the medical domain as far as predictive analysis is concerned. Disease prediction is now becoming the pioneer in medical analysis as there is a rage of new deadly diseases surging in. The disease of stroke is believed to pose a great threat to humanity in future as there is a continuous rise of neurological disorders and the psychological tension faced by human beings seems no bound nowadays. Hence in order to combat the prevalence of stroke, this paper proposes a novel method of stroke prediction using adaptive neuro fuzzy inference system. The proposed system achieves an accuracy of 97.46%, sensitivity of 94.85%, specificity of 95.3% and F1 score of 96.67%. It is evident from the results that our proposed model outperforms all the other existing techniques in terms of stroke prediction and is found to showcase good predictive analytics.

References

[1] G, T.R., Bhattacharya, S., Maddikunta, P.K.R. et al, "Antlion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset," *Multimedia Tools and Applications*, vol.81, pp. 41429–41453, 2022. 10.1007/s11042-020-09988-y.

[2] Pedersen A, Stanne TM, Nilsson S, Klasson S, Rosengren L, Holmegaard L, Jood K, Blennow K, Zetterberg H, Jern C, "Circulating neurofilament light in ischemic stroke: temporal profile and outcome prediction," *Journal of neurology*, vol.266, no.11, pp.2796-2806, 2019. 10.1007/s00415-019-09477-9

[3] Kwon HS, Lee D, Lee MH, Yu S, Lim JS, Yu KH, Oh MS, Lee JS, Hong KS, Lee EJ, Kang DW, Kwon SU; PICASSO investigators, "post-stroke cognitive impairment as an independent predictor of ischemic stroke recurrence: PICASSO sub-study," *Journal of neurology*, vol.267, pp.688-693, 2020. 10.1007/s00415-019-09630-4

[4] Wang H, Sun Y, Ge Y, Wu PY, Lin J, Zhao J, Song B (2021), "A clinical-radiomics nomogram for functional outcome predictions in ischemic stroke," *Neurology and Therapy*, vol.10, no.2, pp.819-832. 10.1007/s40120-021-00263-2.

[5] Yu, Y., Parsi, B., Speier, W., Arnold, C., Lou, M., Scalzo, F. (2019). LSTM Network for Prediction of Hemorrhagic Transformation in Acute Stroke. In: Shen, D., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11767. Springer, Cham. 10.1007/978-3-030-32251-9_20

[6] Liu, L., Chen, S., Zhang, F., Wu, F. X., Pan, Y., & Wang, J., "Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI," *Neural Computing and Applications*, vol.32, pp.6545-6558, 2020. 10.1007/s00521-019-04096-x

[7] Ananya, P.R., Pachisia, V., Ushasukhanya, S. (2022). Optimization of CNN in Capsule Networks for Alzheimer's Disease Prediction Using CT Images. In: Manogaran, G., Shanthini, A., Vadivu, G. (eds) Proceedings of International Conference on Deep Learning, Computing and Intelligence.

Advances in Intelligent Systems and Computing, vol 1396. Springer, Singapore. 10.1007/978-981-16-5652-1_49

[8] Anusha Bai, R., Sangeetha, V. (2023). A Comparative Study on Brain Intracerebral Hemorrhage Classification Using Head CT Scan for Stroke Analysis. In: Ranganathan, G., EL Allioui, Y., Piramuthu, S. (eds) Soft Computing for Security Applications. ICSCS 2023. Advances in Intelligent Systems and Computing, vol 1449. Springer, Singapore. 10.1007/978-981-99-3608-3_44

[9] Zhang, Y., Yu, M., Tong, C., Zhao, Y., & Han, J., "CA-UNet Segmentation Makes a Good Ischemic Stroke Risk Prediction," *Interdisciplinary Sciences: Computational Life Sciences*, pp.1-15. 2023. 10.1007/s12539-023-00583-x.

[10] Weng, YT., Chan, HW., Huang, TY. (2020). Automatic Segmentation of Brain Tumor from 3D MR Images Using SegNet, U-Net, and PSP-Net. In: Crimi, A., Bakas, S. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2019. Lecture Notes in Computer Science, vol 11993. Springer, Cham. 10.1007/978-3-030-46643-5_22

[11] Odeh, S., Kyriacou, E., & Pattichis, C. S. (2023). Using ANFIS to Predict the Long- and Short-Term Stroke Risk Based on Ultrasound Carotid Imaging and Clinical data of Initially Asymptomatic Patients. *Journal of Millimeterwave Communication, Optimization and Modelling*, vol.3, no.1, pp.27-31, 2023.

[12] Pathanjali, C., Monisha, G., Priya, T., Ruchita, S. K., & Bhaskar, S, "Machine learning for predicting ischemic stroke," *International Journal of Engineering Research & Technology*, vol.9, no.5, pp.1-4, 2020. 10.17577/IJERTV9IS050836

[13] Bali, B., & Garba, E. J, "Neuro-fuzzy approach for prediction of neurological disorders: a systematic review," *SN Computer Science*, vol.2, no.4, pp.307, 2021. 10.1007/s42979-021-00710-9

[14] Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol.2, pp.100032, 2022. 10.1016/j.health.2022.100032

[15] Wu, Y., & Fang, Y., "Stroke prediction with machine learning methods among older Chinese," *International journal of environmental research and public health*, vol.17, no.6, pp.1828, 2020. 10.3390/ijerph17061828

[16] Sailasya, G., & Kumari, G. L. A., "Analyzing the performance of stroke prediction using ML classification algorithms*," International Journal of Advanced Computer Science and Applications*, vol.12 no.6, pp. 539-545, 2021. 10.14569/IJACSA.2021.0120662

[17] Dritsas, E., & Trigka, M., "Stroke risk prediction with machine learning techniques," *Sensors*, vol.22, no.13, pp.1-13, 2022. 10.3390/s22134670

[18] Vasukidevi, G., Ushasukhanya, S., & Mahalakshmi, P, "Efficient image classification for alzheimer's disease prediction using capsule network," *Annals of the Romanian Society for Cell Biology*, vol.25, no.5, pp.806-815, 2021.

[19] Manoharan, J. S., Braveen M. & Ganesan Subramanian, S, "A hybrid approach to accelerate the classification accuracy of cervical cancer data with class imbalance problems," *International Journal of data mining*, vol.25, no.3-4, pp. 234 – 259, 2021. 10.1504/IJDMB.2021.122865

[20] Yu, Y., Xie, Y., Thamm, T., Gong, E., Ouyang, J., Huang, C., ... & Zaharchuk, G, "Use of deep learning to predict final ischemic stroke lesions from initial magnetic resonance imaging," *JAMA network open*, vol.3, no.3, pp. 1-13, 2020.10.1001/jamanetworkopen.2020.0772

[21] Lakkshmanan, Ajanthaa, Anbu Ananth, C., and Tiroumalmouroughane, S, "Multi-objective Metaheuristics with Intelligent Deep Learning Model for Pancreatic Tumor Diagnosis," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 5, pp. 6793-6804, 2022. 10.3233/JIFS-221171

[22] Chauhan, S., Vig, L., De Filippo De Grazia, M., Corbetta, M., Ahmad, S., & Zorzi, M, "A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images," *Frontiers in Neuro informatics*, vol.13, no.53, pp. 1-12, 2019. 10.3389/fninf.2019.00053

[23] Hu, X., Luo, W., Hu, J., Guo, S., Huang, W., Scott, M. R., ... & Reyes, M, "Brain SegNet: 3D local refinement network for brain lesion segmentation," *BMC medical imaging*, vol.20, pp.1-10, 2020. 10.1186/s12880-020-0409-2

[24] Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S., & Jannes, J, "Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study," *Academic radiology*, vol.27, no.2, pp. e19-e23, 2020. 10.1016/j.acra.2019.03.015

[25] Liu, T., Fan, W., & Wu, C, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artificial intelligence in medicine*, vol.101, pp.1-30, 2019. 10.1016/j.artmed.2019.101723

[26] Sathya, V, Mahendra Babu, G.R.; Ashok, J.; Lakkshmanan, Ajanthaa. "A Novel Predicting Students Performance Approach to Competency &amp; Hidden Risk Factor Identifier Using a Various Machine Learning Classifiers," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no.6, pp. 9565–9579, 2023. 10.3233/JIFS-224586

[27] Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H, "Machine learning–based model for prediction of outcomes in acute stroke," *Stroke*, vol.50, no.5, pp.1263-1265, 2019. 10.1161/STROKEAHA.118.024293

[28] Dehdar Karsidani, S., Farhadian, M., Mahjub, H., & Mozayanimonfared, A, "Intelligent prediction of major adverse cardiovascular events (MACCE) following percutaneous coronary intervention using ANFIS-PSO model," *BMC Cardiovascular Disorders*, vol.22, no.1, pp.1-8, 2022. 10.1186/s12872-022-02825-0

[29] Mohanty, R., Solanki, S.S., Mallick, P.K., Pani, S.K. (2021). A Classification Model Based on an Adaptive Neuro-fuzzy Inference System for Disease Prediction. In: Bhoi, A., Mallick, P., Liu, CM., Balas, V. (eds) Bio-inspired Neurocomputing. Studies in Computational Intelligence, vol 903. Springer, Singapore. 10.1007/978-981-15-5495-7_7

Declaration

[1] Funding – The author did not receive support from any organization for the submitted work.

[2] Conflicts of Interest - The author has no relevant financial or non-financial interests to disclose.

[3] Availability of Data & Material – The author hereby declares that no specific data sets are utilized in the proposed work.

[4] Code Availability – Since future works are based on the custom codes developed in this work, the code may not be available from the author.

[5] Author's contribution – The author is solely responsible for the experimental works conducted in this paper, drafting of the paper and presentation of all the sections.