2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-3368-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOSEC51865.2021.9591840

Early Detection of Brain Stroke using Machine Learning Techniques

Dr.Vempati Krishna[1], Professor, Dept of CSE, Teegala Krisha Reddy College of Engineering and Technology(A),Hyderabad, Telangana
krishna.vempati2015@gmail.com

Dr. PVRD Prasada Rao[3], Professor, Dept of Computer Science and Engineering, Koneru Lakshmaih Education Foundation, Vaddeswaram, AP,India
pvrdprasad@kluniversity.in

Dr.G.John Babu[5], Professor, Dept of CSE, Vice Principal, Vijaya Engineering College, Khammam, Telangana
johnbabug@gmail.com

Dr. J.Sasi Kiran[2], Professor in CSE & Dean, Lords Institute of Engineering and Technology(A), Hyderabad, Telangana
sasikiranjangala@gmail.com

Dr.G.Charles Babu[4], Professor, Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana
charlesbabu26@gmail.com

Abstract- The brain is the most complex organ in the human body. Brain Stroke is a long-term disability disease that occurs all over the world and is the leading cause of death. A stroke occurs when the brain's blood supply is cut off and it ceases to function. There are two primary causes of brain stroke: a blocked conduit (ischemic stroke) or blood vessel spilling or blasting (hemorrhagic stroke). Early brain stroke prediction yields a higher amount that is profitable for the initiating time. Brain stroke is caused primarily by people's lifestyle decisions, particularly in the current scenario by evolving elements such as high blood sugar, heart disease, obesity, diabetes, and hypertension. This research study has used various machine learning (ML) algorithms like K nearest neighbour, logistic regression, random forest (RF) classifier and SVC. This research work designs a model using one among the following

algorithms with high accuracy to predict the stroke for newly given inputs.

Keywords: Brain Stroke, Machine Learning, Diseases, Algorithms, Decision tree (DT), Random forest classifier, svc

I. INTRODUCTION

Now-a-days brain stroke has become a major disease that is leading to death. Prediction of brain stroke in the early stage has become very difficult and it is time taking tasks as a result many people are losing their lives. To overcome this we use machine learning approach and build a model to predict whether a person is suffering from brain stroke or not. We can do this by considering various attributes of the patient and predict the output. A stroke is a medical condition in which poor bloodstream to the brain causes cell death. There are 2 significant types of stroke: ischemic, due to nonappearance of the bloodstream, and hemorrhagic, due to bleeding. Both reasons of Brain gets damaged if symptoms last lower than 1 or 2 hours, the stroke is transient ischemic assault (TIA), also named a mini-stroke. A hemorrhagic stroke might also relate to an extreme migraine. The symptoms of a stroke might be stable. The long-term problems might contain bladder control loss and pneumonia. The primary danger factor for stroke is hypertension. Various danger factors contain tobacco smoking, diabetes mellitus, heftiness, a past TIA, high blood cholesterol, end-stage kidney sickness, and preliminary fibrillation. An ischemic stroke is ordinarily caused by blood vessel blockage however there are additionally more uncommon causes. A hemorrhagic stroke is caused either because of space among the membranes of the brain or bleeding directly from the brain. The bleeding might happen because of a burst brain aneurysm. An analysis is usually founded on physical tests and helped by clinical imaging like MRI or CT scan. A CT scan might rule out bleeding, however, might not essentially rule out ischemia that immediately regularly doesn't appear on a CT scan. The numerous tests like blood tests and an electrocardiogram (ECG) are never really dangerous factors and preclude other potential causes [8]. Low glucose might cause comparable symptoms. The prevention incorporates diminishing danger factors, medical procedures to open up the veins to the brain in those with tricky carotid narrowing, and warfarin in individuals with trial fibrillation. The aspirin might be suggested by doctors for anticipation. A stroke or TIA frequently needs crisis care. An ischemic stroke, whenever recognized within 3 to four and half hours, might be treatable with a medicine, which could separate the clot. Few hemorrhagic strokes profit with a medical procedure. Treatment to attempt recuperation of lost capacity is named as stroke recovery, and preferably occurs in a stroke unit; in any case, these are not accessible in a significant part of the world.

II. LITERATURE SURVEY

In order to get essential information about different models associated existing literature were studied. Few significant conclusions were made through those are listed below. The Purpose of the paper was Calculation of 10-year stroke prediction probability and classifying the customer's discrete probability of stroke into 5groups. In the paper, a health risk appraisal function is established for stroke prediction with the use of Framingham Study cohort. In the survey, this manuscript aimed to develop a formula for increasing a stroke pre-diagnosis method with potentially modifiable risk factors. In this paper, decision tree (DT) method is utilized for feature selection procedure; principle component analysis method is utilized for decreasing the dimension and adopted back propagation neural network classification method, to create a classification method. The paper suggests the application interface design for related medical data visualization and management for neurologists in stroke clustering and prediction framework named as a Stroke MD. The paper focuses on cutting-edge prevention methodologies of stroke. The paper, main segment examination method is utilized for diminishing the measurements and it defines the qualities including more towards the expectation of stroke illness and predicts if the patient is experiencing stroke sickness or not. In the paper, Non-contrast head CT scan will be the present standard for first imaging of patients with head injury or stroke side effects. This paper planned to create and approve a bunch of deep learning methods for computerized identification.

III. PROPOSED MODEL

Stroke is the subsequent leading reason for death worldwide and stays significant health trouble both for national healthcare frameworks and people. Possibly modifiable danger factors for stroke incorporate hypertension, heart sickness, diabetes, and liberation of glucose digestion, preliminary fibrillation, and way of life factors. Subsequently, the target of our work is to apply standards of ML over enormous current data sets to adequately calculate the stroke dependent on conceivably modifiable danger factors. Then, it planned to improve the application to give a customized warning based on every client's degree of stroke risk and a way of life correction message about stroke hazard factors. To reduce the risk of death we had come up with a model which was built using KNN algorithm of Machine Learning. Using this algorithm we had built a model and later we had built a Graphical User Interface (GUI) for this model using FLASK framework. And so by using this GUI any person can open our website and enter certain details of their health and can predict whether they are suffering from brain stroke or not.

IV. OBJECTIVES

The significant objective of this study is to predict whether a person is suffering from brain stroke or not. This work helps to reduce the risk of patient because in the early stages itself we can predict the stroke. To establish a KNN model that is highly efficient in terms of accuracy. Stroke prediction can be done very efficiently and effectively. This model can be used in medical fields to reduce the death rate of stroke patients.

V. IMPLEMENTATION

The classification methods are a fundamental piece of data mining and ML applications. Around 70% of issues in data science are characterization issues. There are many classification issues that are accessible, yet the strategic relapse is normal and is a helpful relapse technique for solving binary classification issues. Another class of classification is Multinomial classification that controls the problems whereas numerous classes are available in the target variable.

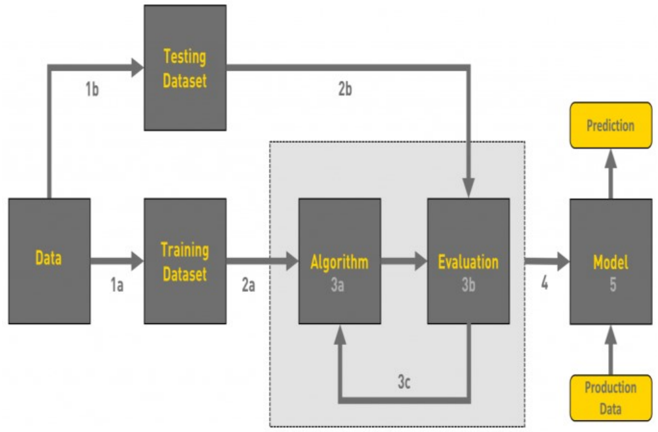


Fig.1. Proposed System Block Diagram

Implementation of work includes the below steps:

a. Data collection

b. Data pre-processing

c. Training model

d. Testing model

e. Deployment

Data Collection: Dataset required for our work is collected from Google website. Our dataset consists of 5110 rows and 10 columns.

Data Pre-Processing: This step has performed various visualizations on the data and also converted all the categorical variables into continuous variables for our work requirement.

Training Model: For training the model, we had used various ML algorithms such as KNN, logistic regression, SVM, and DT classifier.

A. *Logistic Regression*

Logistic regression may be a managed taking in arrangement calculation used to anticipate the likelihood of a focus variable. The nature of the target or subordinate variable is dichotomous that intends that there might be a chance to be main two conceivable classes. Hosting information coded concerning illustration Possibly 1 (stands to success/yes) alternately 0 (stands to failure/no). Mathematically, a logistic relapse method predicts P(Y=1) Likewise a work for X. It may be a standout amongst those simplest ml calculations that might a chance to be utilized for Different arrangement issues, for example, spam detection, diabetes prediction, growth identification and so forth throughout this way, observing and stock arrangement of all instrumentation may be enhanced.

B. *SVM*

The SVMs are incredible yet adaptable supervised ML methods that are utilized both for characterization and regression. However, usually they are utilized in classification issues. In 1960's, SVMs were initially presented, however, later they got refined in 1990. SVMs have their method of execution when contrasted with other ML methods. Recently, they are very famous due to their capacity to deal with numerous categorical and continuous factors.

An SVM method will be fundamentally a portrayal of numerous categories in hyper-plane in multidimensional space. The hyper-plane is created in an iterative method by SVM so the mistake might be reduced. SVM will aim to separate the datasets into classes to discover a maximum marginal hyper plane (MMH).

The below points are significant conceptions in SVM –

a. Support Vectors – Information focuses that need the aid of the closest hyper-plane may be known as backing vectors. Dividing lines will have a chance to be characterized with the help of this information focuses.

b. Hyper plane – concerning illustration should be obvious in over diagram, it will be a choice plane alternately space that will be partitioned between a situated of Questions Hosting distinctive classes.

c. Margin – It might be characterized similarly with the middle of two lines on the storeroom information focuses about diverse classes. It might a chance to be computed similarly as the peroxide blonde separation from that accordance of the help vectors. Expansive edge will be acknowledged concerning illustration a great edge Also little edge may be viewed as similarly as an awful edge.

The primary objective from claiming SVM may be to partition those datasets underclasses with figure and maximum marginal hyper plane (MMH).

C. *Random Forest (RF) Classifier*

Random forest will be used for both classification and also regression. Be that however, it is mostly utilized to order issues. As we think that a wood is committed up for trees What's more, all the more trees methods a greater amount strong woods. Similarly, an irregular wood algorithm selects trees once information specimens and so on to obtain that prediction beginning with each for them. At long last, select those best results by voting method. It will be a group technique that may be superior to a single choice tree on account of it lessens those over-fitting by averaging those result.

RF Algorithm function

We might know RF method functioning with the support of below steps –

a. Initially, start with random samples selection from the provided dataset.

b. Further, this method will create a DT for each sample. Then it will become the prediction result from each DT.

c. In this step, voting is executed for each predicted result.

d. Lastly, choose the most voted prediction result as last prediction result.

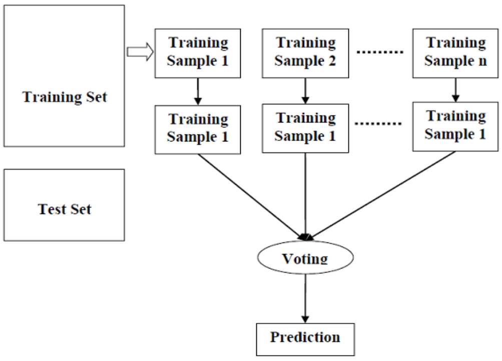


Figure.2. Random Forest Block Diagram

D. *Decision Tree Classifier*

For general, decision tree examination will be a predictive displaying device that might make connected crosswise over numerous zones. Choice trees can make constructed toward an algorithmic approach that could part those datasets in distinctive approaches dependent upon distinctive states. The DT needs majority capable calculations that fall under the class from claiming regulated calculations. They could make utilized for both order and regression assignments. The two fundamental substances of the tree would decision nodes, the place the information may be part also leaves, and the place we got the result.

E. *DT Algorithm Implementation*

Gini Index (GI)

It will be the term of cost function, which will be utilized to estimate the binary splits in the dataset and operates with the categorical target variable "Success" or "Failure".

Higher the GI value, higher the homogeneity. A perfect value of GI will be 0 and worst is 0.5 (for 2 class issues). The GI for a split might be estimated with support of subsequent steps –

Initially, estimate GI for sub-nodes by utilizing the equation p^2+q^2, which will be the sum of squares of probability for failure and success.

Then, estimate GI for split utilizing the weighted Gini score of every node of that split.

The Classification and Regression Tree (CART) method utilizes Gini technique to produce binary splits.

Split Creation

A split is essentially containing an attribute in value and dataset. We might create a split in dataset with support of next three parts –

Part1: Computing Gini Score − we have just deliberated this part in past segment.

Part2: Splitting a dataset − It might make characterized Likewise dividing An dataset under two schedules for rows Hosting list of a trait Also An part esteem from claiming that quality. Following getting those two gatherings - straight Furthermore left, from those dataset, we can ascertain the worth for part Eventually Tom's perusing utilizing Gini score computed over initially a feature. The part-worth will choose for which gathering the quality will live.

Part3: Evaluating all splits − Next major aspect following discovering Gini score Furthermore Part dataset will be that assessment about constantly on parts. To this purpose, we must weigh each worth connected with each quality likewise a hopeful part. After that we need should find the best workable part toward assessing the cosset of the part. The best chance to be utilized similarly will have a hub in the choice tree.

F. K Nearest Neighbour (KNN)

The KNN method is a kind of regulated ML method that might be utilized for both classification and also regression predictive issues. However, it is principally utilized to classify things in the industry.

The subsequent 2 properties would describe KNN well –

G. Lazy learning algorithm

It does not have a particular training phase and utilizes whole data for training whereas classification.

H. Non-parametric learning algorithm

it doesn't assume anything about underlying data.

The KNN method utilizes feature similarity to calculate the values of novel data points that mean the novel data point is allocated a value built on how closely it matches the points in the training set. We might know it's working with the support of subsequent steps

Step 1 – For executing any method, we require a dataset. So through the initial step of KNN, we should load the training and test data.

Step 2 − Then, we require to select a value for K i.e. the nearest data points. K might be any integer.

Step 3 – For every point in test data do the following –

1 − Compute those separations among test information Furthermore every column about preparing information for those help about whatever of technique specifically: Euclidean, maul alternately hamming distance. The majority regularly utilized strategy to ascertain separation maybe not easy.

2 − Now, built on distance value, sort them in ascending sequence.

3 − Then, it will select top K rows from the sorted array.

4 − Now, it will relegate the test side population of point reliant on the majority incessant population about these rows.

Step 4 – End

I. Advantages of KNN

a. It is a very easy method to understand and interpret.
b. It is very beneficial for nonlinear data due to there will be no assumption about data in this method.
c. It is a versatile method as we might utilize it for classification and regression.

It needs moderately more accuracy yet the entire there need aid substantially exceptional managed Taking in models over KNN.

J. Summary of all algorithms used

From the below table we can clearly understand what algorithms are used in the work and what are accuracies generated for each algorithm used.

Algorithms used	**Accuracy**
Logistic Regression	94.32

Support Vector Machine	94.57
Random Forest Classifier	98.57
Decision Tree Classifier	98.98
K Nearest Neighbour	99.35

Table.1. Algorithm used with Accuracy values

VI. RESULTS & ANALYSIS

Let us see the bar chart representation of our experimental results.

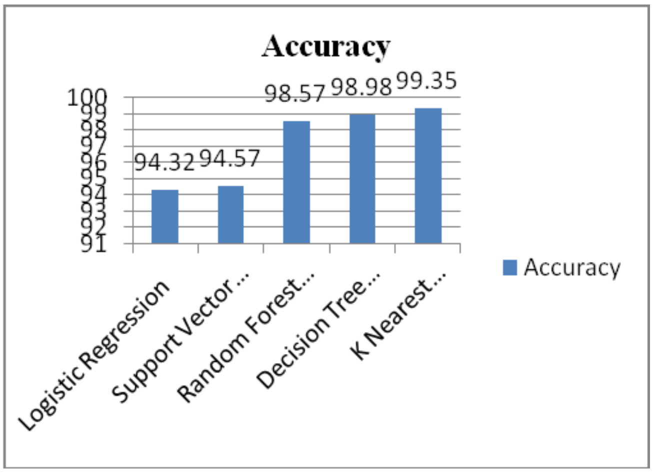
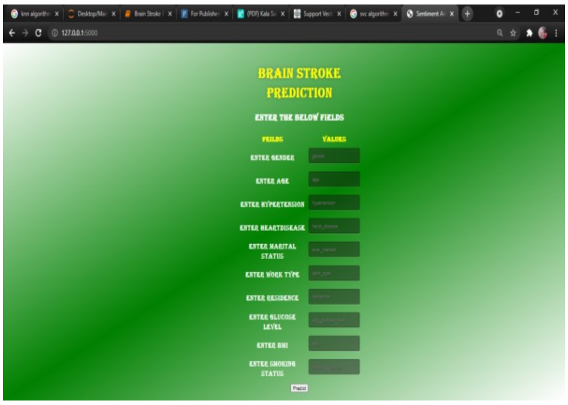


Figure.3.Graphical Representation

So, finally from the above graph mentioned we had got the accuracy of 94.32% for the algorithm Logistic Regression, accuracy of 94.57% for Support Vector Machines, accuracy of RF Classifier is 98.57%, accuracy of 98.98% for the algorithm Decision tree Classifier and the accuracy of 99.35% for the K Nearest Neighbour algorithm. Among all the algorithms used we had got the best and accurate accuracy of 99.35%. It is clear that we had first trained the model by using KNN and we had predicted the output for the newly given inputs. The predicted output would be either 1 or 0. Here 0 represents that the person is not suffering from Stroke and 1 represents that the person is suffering from stroke. So, here using KNN algorithm we had built the model and using this model we had developed a Graphical User Interface through which a person can provide his health details and predict whether he is suffering from brain stroke or not. After the prediction, if the result comes as he is suffering from brain stroke then he needs to consult the doctor for further treatment.

Let us see the results obtained.

The below image is the User Interface through which a person can check whether he is suffering from stroke or not. You should enter the fields to predict.



BRAIN STROKE PREDICTION

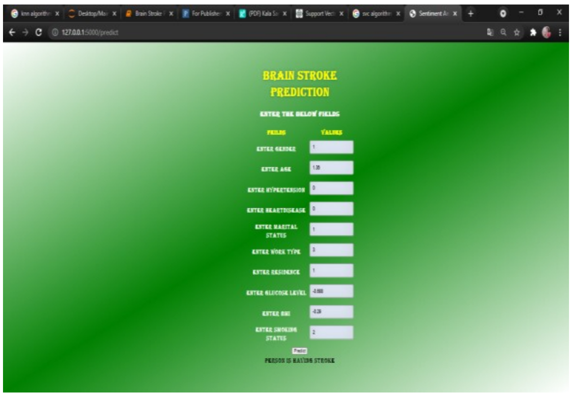
ENTER THE BELOW FIELDS

FEILDS	VALUES
ENTER GENDER	gender
ENTER AGE	age
ENTER HYPERTENSION	hypertension
ENTER HEARTDISEASE	heart_disease
ENTER MARITAL STATUS	ever_married
ENTER WORK TYPE	work_type
ENTER RESIDENCE	residence
ENTER GLUCOSE LEVEL	avg_glucose_level
ENTER BMI	bmi
ENTER SMOKING STATUS	[illegible]

Predict

Figure.4. Output 1

Later on entering the fields the output obtained that the person is suffering from stroke is shown below.



BRAIN STROKE PREDICTION

ENTER THE BELOW FIELDS

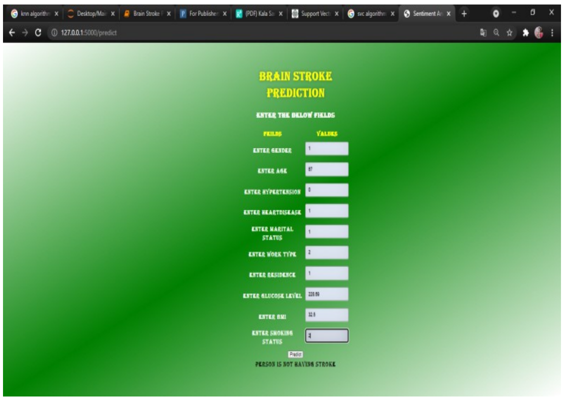
FEILDS	VALUES
ENTER GENDER	1
ENTER AGE	1.30
ENTER HYPERTENSION	0
ENTER HEARTDISEASE	0
ENTER MARITAL STATUS	1
ENTER WORK TYPE	3
ENTER RESIDENCE	1
ENTER GLUCOSE LEVEL	-0.500
ENTER BMI	-0.26
ENTER SMOKING STATUS	2

Predict

PERSON IS HAVING STROKE

Figure.5. Output 2

The different values are given repeatedly and then the output obtained as the person is not suffering from stroke as shown.



BRAIN STROKE PREDICTION

ENTER THE BELOW FIELDS

FIELDS	VALUES
ENTER GENDER	1
ENTER AGE	67
ENTER HYPERTENSION	0
ENTER HEARTDISEASE	1
ENTER MARITAL STATUS	1
ENTER WORK TYPE	2
ENTER RESIDENCE	1
ENTER GLUCOSE LEVEL	228.69
ENTER BMI	32.5
ENTER SMOKING STATUS	2

PERSON IS NOT HAVING STROKE

Figure.6. Output 3

VII. CONCLUSION

It is concluded that by using the proposed work we can predict whether a person is suffering from brain stroke or not very easily in the earlier stage predicting the disease might decrease the causes of death. So, with the help of our work, there will be a great benefit in the medical field. Lots of deaths caused due to brain stroke will also be reduced to a great extent by using our work.

REFERENCES

[1] Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; Guizani, M. Deep learning for IoT Big data and streaming analytics: A survey. IEEE Commun. Surv. Tutor. 2018, 20, 2923–2960. [CrossRef]

[2] Ajayi, O.O.; Bagula, A.B.; Ma, K. Fourth industrial revolution for development: The relevance of Cloud federation in healthcare support. IEEE Access 2019, 7, 185322–185337. [CrossRef]

[3] Morrar, R.; Arman, H.; Mousa, S. The fourth industrial revolution (Industry 4.0): A social innovation perspective. Technol. Innov. Manag. Rev. 2017, 7, 12–20. [CrossRef]

[4] Garcia, A.R. AI, IoT, Big data, and technologies in digital economy with blockchain at sustainable work satisfaction to smart mankind: Access to 6th dimension of human rights. In Smart Governance for Cities: Perspectives and Experiences, 2nd ed.; Lopes, N.V.M., Ed.; Springer: Gewerbestrasse, Switzerland, 2020; pp. 83–131.

[5] Johnson, C.O.; Nguyen, M.; Roth, G.A.; Nichols, E.; Alam, T. Global, regional, and national burden of stroke, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019, 18, 439–458. [CrossRef]

[6] Subudhi, A.; Dash, M.; Sabut, S. Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. Biocybern.Biomed. Eng. 2020, 40, 277–289. [CrossRef]

[7] Lee, H.J.; Lee, J.S.; Choi, J.C.; Cho, Y.J.; Kim, B.J.; Bae, H.J.; Kim, D.E.; Ryu, W.S.; Cha, J.K.; Kim, D.H.; et al. Simple estimates of symptomatic intracranial hemorrhage risk and outcome after intravenous thrombolysis using age and stroke severity. J. Stroke 2017, 19, 229–231. [CrossRef]

[8] Kim, Y.D.; Jung, Y.H.; Saposnik, G. Traditional risk factors for stroke in East Asia. J. Stroke 2016, 18, 273–285. [CrossRef]

[9] Poorthuis, M.H.; Algra, A.M.; Algra, A.; Kappelle, L.J.; Klijn, C.J. Female-and male-specific risk factors for stroke: A systematic review and meta-analysis. JAMA Neurol. 2017, 29, 86–93. [CrossRef]

[10] Malik, V.; Ganesan, A.N.; Selvanayagam, J.B.; Chew, D.P.; McGavigan, A.D. Is atrial fibrillation a stroke risk factor or risk marker? An appraisal using the bradford hill framework for causality. J. Heart Lung Circ. 2020, 29, 86–93. [CrossRef]

[11] Karuppusamy, Dr P. "Hybrid Manta Ray Foraging Optimization for Novel Brain Tumor Detection." Journal of Soft Computing Paradigm (JSCP) 2, no. 03 (2020): 175- 185.

[12] b. Vijayakumar, T. “Classification of brain cancer type using machine learning.” Journal of Artificial Intelligence 1, no. 02 (2019): 105-113

[13] Alawadi, S., Fern´andez-Delgado, M., Mera, D., and Barro, S. (2019). Polynomial kernel discriminant analysis for 2d visualization of classification problems. Neural Computing and Applications, 31(8):3515–3531.

[14] Almeida, Y., Sirsat, M. S., i Badia, S. B., and Ferm´e, E. (2020). Airehab: A framework for ai driven neurorehabilitation training-the profiling challenge. In HEALTHINF, pages 845–853.

[15] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., et al. (2019). An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201):861–867

[16] Bento, M., Souza, R., Salluzzi, M., Rittner, L., Zhang, Y., and Frayne, R. (2019). Automatic identification of atherosclerosis subjects in a heterogeneous mr brain imaging data set. Magnetic resonance imaging