STM JOURNALS

Journal of
Microwave Engineering & Technologies

ISSN: 2349-9001
Volume 11, Issue 2, 2024
May–August
DOI (Journal): 10.37591/JoMET

https://journals.stmjournals.com/jomet

Review JoMET

Identification of Brain StrokeUsing Artificial Intelligence

Anuradha T.[1,*], Abhishek J.[2], Maheboob Patel[3], Mohammad Akbar Ali[4]

Abstract

Globally, strokes are the primary cause of disability and mortality. Recently, machine learning (ML) and deep learning (DL) have been employed by artificial intelligence algorithms as effective stroke diagnosing techniques. These days, machine learning and data mining technologies are used in the construction of the main models. We have used fivemachine learning algorithms to determine if a stroke has occurred or is likely to occur based on a patient's physical state and information from medical reports. We are utilizing the numerous hospitals records we have gathered to address our issue. The categorization outcome demonstrates that the outcome is appropriate and useful for a real-time medical report. We discovered that machine learning algorithms can help us comprehend diseases and be a helpful tool in the healthcare industry.

Keywords: Random forest, k-NN, artificialintelligence, decision tree, logistic regression, SVM

INTRODUCTION

Everybody believes that their health is vital, and to keep track of information regarding illnesses and the connections between them, a registration system is required. Most disease-related data was in patient reports located in clinics, patient case summaries, and other manually maintained data. They could be used to interpret sentences using a variety of text mining and machine learning (ML) techniques.

Machine learning is a method that can be used ininformation retrieval to distribute content where the syntactic and semantic components of the content are predominant. Various machine learning methods are proposed and applied for classification and feature extraction. Most medical professionals refer to injuries caused by an interruption in blood flow to the brain and spinal cord as strokes. A stroke always causes a visceral reaction, even though its meaning can vary depending on the situation.

In domains like information management, surveillance, and medical, machine learning can be regarded as a valuable auxiliary when properly trained and applied with algorithms. Using data mining technologies, this study provides an overview of data monitoring from a syntactic and semantic perspective. To train the system, the disease forms will be taken, the patient's symptoms will be extracted, and the data will be used. The suggested Stemmer extracts common and unique qualities for the detection of stroke disease, after the labelling and maximum entropy methods' mining of case sheets. Following processing, a variety of machine learning methods were used to the data, including decision trees, random forests, logistic regression, support vector machines, and K-nearest neighbors [9–15].

***Author for Correspondence**
Anuradha T.
E-mail: anuradhat@pdaengg.com

[1]Student, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering Kalaburagi, Karnataka-585102, India
[2]Student, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering Kalaburagi, Karnataka-585102, India
[3]Student, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering Kalaburagi, Karnataka-585102, India
[4]Student, Department of Computer Science and Engineering Poojya, Doddappa Appa College of Engineering Kalaburagi, Karnataka-585102, India

Received Date: July 06, 2024
Accepted Date: July 19, 2024
Published Date: July 30, 2024

Citation: Anuradha T., Abhishek J., Maheboob Patel, Mohammad Akbar Ali. Identification of Brain StrokeUsing Artificial Intelligence. Journal of Microwave Engineering & Technologies.2024;11(2):15–22p.

Among the methods that achieve good accuracy is Support Vector Machine. Since everyonebelieves that their health is essential to their survival, a method for recording illnesses and their relationships is needed. With a classification accuracy of 96%, random forest classification performs better than the other techniques examined. The study shows that for brain stroke predicting, the random forest method performs better than alternative procedures when cross-validation metrics are applied.

Patient cases, clinical patient records, and other manually maintained records include the major disease information. It is possible to understand their utterances with a variety of machine learning techniques (ML). Machine learning is a method that may separate content forinformation retrieval when the syntactic and semantic components of the content are given priority. To extract and classify features, a plethora of machine learning has been put out and undertaken. Most medical practitioners define stroke as abnormal blood flow causing damage to the brain and spinal cord. Thoughts and viewpoints vary in how a stroke is experienced, a clear visceral reaction is typically produced.

Every memory is encoded and preserved in a network within the brain, which is made up of a trillion glia, 100 billion neurons, and weighs more than three kilos. Every person's breathing and movement are supported by brain activity. Ten times as many people have died by stroke indeveloping nations as in affluent ones since 1970, or more than 50 years ago, and by 2030, the global total is predicted to double. There are three main categories of strokes: transient ischemic attacks (TIA), hemorrhagic strokes, and ischemic strokes. The most prevalent kind ofstroke is an ischemic stroke. 87 percent of strokes are ischemic strokes, according to theAmerican Heart Association (AHA), which occur when a blood clot or other blockage is stillpresent in a brain blood vessel [15-20].

RELATEDWORK

The section discusses several Brain Stroke Identification Systems that have been proposed with each having their own unique features.

The authors in [1], have developed a deep learning model for predicting the final stroke infarct CNN. The highest Dice similarity coefficient (DSC) was achieved in both reperfused and non-reperfused patients.

The authors used CNN in final infarct volume prediction from CTA 2021 CNN. Compared CNN-derived ischaemic lesion volumes to final infarct volumes that were manually segmented from follow-up CT and to CTPRAPID ischaemic core volumes in [2].

ANNs to optimally predict the ischemic core in acutestroke patients, using advanced imaging 2019 ANN. Automated detection of acute ischemic stroke regionshas been observed in [3].

The authors had classified segment brain strokes. 2021 UNet with CNN.U-Net, one of the encoders-decoder deep learning-based CNNs, has been developed and proposed for classification and segmentation of brain stroke is discussed in [4].

The authors have accurately predicted the intracranial hemorrhage on NCCT brain images 2022 Transfer DLmethod an automated transfer deep learning method that combines ResNet-50 and dense layers for accurate prediction of intracranial hemorrhage on NCCT brain images in [5].

In [6] the authors predicted strokes and emergency 2022 Hybrid LSTM Improved accuracy and efficiencyin diagnosing brain stroke.

The authors in [7] did Polynomial kernel discriminant analysis for 2d visualization of classificational problems. Neural Computing and Applications.

The authors in [8] developed framework for ai driven neurorehabilitation training-the profiling challenge. InHEALTHINF.

The authors got Outcomes and complications afterendovascular treatment for the brain arteriovenous malformations: a prognostication attempted usingartificial intelligence in [10].

SYSTEMARCHITECTURE

Dataset

A dataset consists of several data files. In the context of tabular data, a dataset is comparable to one or more database tables, where each table's column represents a specific variable, and each row represents a specific dataset record see below in Figure 1. The dataset includes a list of each variable's values, including an individual's age, sex, and BMI. Documents or files can also be gathered into databases. Data from stroke prediction studieswere used in the research. There were 12 columns and 5110 rows in this material. In the result column, the stroke value is either 1 or 0. A value of 1 denotes the presence of a detected stroke risk, whereas a value of0 shows the absence of a detected risk. In this data set,the chance of 0 in the result column (bet) is greater than the likelihood of 1 in the same column. In the column, only 249 rows have a value of 1, whereas 4861 rows have a value of 0. Pre-processing of the data is done to balance it and increase accuracy.

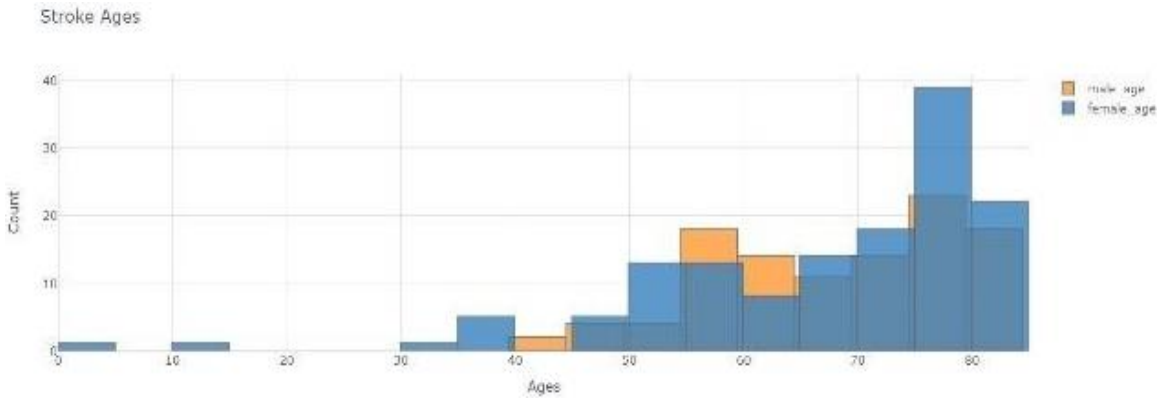

Figure 1. Dataset with gender and age categories.

Pre-processing

To prevent the model from deviating from the training strategy, data must be treated to remove undesirable noise and outliers from the dataset. In this step, everything that is keeping the model from performing better is fixed. After obtaining the necessary dataset, the information needs to becleaned up and made ready for model construction. The data collection has twelve features, as was already indicated. Since the ID column's existence has no bearing on the model's structure, it is first left out. Then, the dataset is filled in if any null values are found.

Here, in this case, the BMI column's zero values are filled in by the average value of the data column. Tag encoding transforms string values in a data set into machine-understandable integer values. Strings need to be translated to integers because the computer is frequently educated on numbers. Thegathered data set has five strings of each data type. Every string in the data collection is encoded during tag encoding, converting it all into a series of numbers. There is a severe imbalance in the stroke prediction data set. There are 5110 rows in the data set overall; 249 rows indicate that a row may exist, and 4861 rows indicate that a row does not. While training a machine-level model with such data can result in accuracy, other metrics of accuracy, such precision and recall, are insufficient. Proper handling of such uneven data is necessary to prevent false observations and poor forecasts. Therefore, these uneven data must be handled first to develop an effective model.

Algorithms used

The accuracy of each algorithm used in the proposed system is shown in Figure 2.

a. Decision tree
b. Random Forest

c. K-Nearest Neighbor (KNN)
d. Logistic regression
e. Support Vector Machine

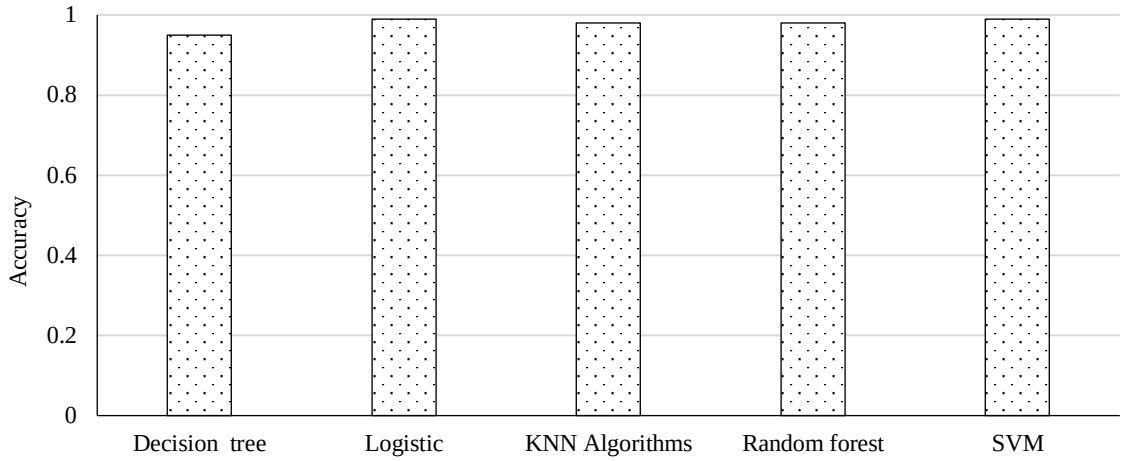

Figure 2. Accuracy of each algorithm.

Decision Tree: Although it is mostly suggested for classification tasks, a decision tree is a supervised learning technique that may be used for regression and classification tasks alike. This classifier is tree-structured, with internal nodes standing in for dataset features, branches for decision rules, and leaf nodes for each output.

K-Nearest Neighbors: Supervised learning techniquesare the foundation of K-Nearest Neighbors, one of the most fundamental machine learning algorithms. Assuming that the new case and the data are comparable to the cases that are already available, the KNN algorithm allocates the new case to the class thatis most like the present classes. The K-NN method categorizes each new data point according to similarity after storing all of the existing data. This implies that the K-NN algorithm can quickly classify newly discovered information into a series of wells. Regression analysis can be done using the K-NN technique, even though classification problems are themain application for it. During the training phase, the KNN algorithm classifies newly received data into a class that is quite close to the previously stored data set.

Logistic regression: A machine learning classification method called logistic regression is usedto ascertain the likelihood of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data that is coded as either 1 (yes, success, etc.) or 0 (no, failure, etc.). Put another way, the logistic regression model predicts P(Y=1) as a function of X.

Support Vector Machine: One of the most popular supervised learning techniques for regression and classification issues is support vector machines, or SVMs. However, the solution of machine learning classification problems is its main use. To make it simple to classify fresh data points in the future, the SVM technique seeks to determine the optimal line or decision boundary that might split the n-dimensional space into groups. This ideal decision boundary is known as the hyperplane. Extreme vectors or points are selected by SVM to help create the hyperplane. Since these extreme circumstances are referred to as support vectors, the approach is called a support vector machine.

WORKING OF ALGORITHMS

Algorithm: Decision Tree

Partitioning: Starting at the root node, the process divides the data into subsets according to the values of its attributes. Making data subsets that are more homogeneous (cleaner) in relation to the target variable is the aim.

Choosing the best splits: To decide where to split, the algorithm evaluates several criteria, such as:

Gini Additivity: measures the additivity of a node.A lower Gini coefficient indicates a cleaner node.

Information Gain: measures the decrease in entropy(disorder) after division.

Chi-square: evaluates the statistical significance of differences between the expected and observed distributions.

Variance Reduction: Used for regression trees to minimize variance in subsets.

Recursion: Until one of the stopping criteria is satisfied, the splitting procedure is applied iteratively to each subgroup, creating branches of the tree.

Stopping criteria: The recursion stops when a node reaches a certain depth, can no longer be split, or when a node reaches a predefined purity threshold.

Algorithm: Support Vector Machine(SVM)

For linearly separable data, SVM finds a hyperplane that maximizes the margin between classes.

Mathematical formulation: the hyperplane can be represented as$w \cdot x+b$=0w·x+b=0, where$w$w is the weight vector, x is the feature vector and$b$b is the apparent term.

Optimization objective: minimize$\|w\|2$‖w‖2

(to maximize margin) provided all data points are correctly classified:' n i+b)≥1whereyiyiis the class identifier of the$i$ith data point.

Non-linear SVM

SVM employs kernel functions to translate non- linearly separable data into a higher dimensional space where it can be separable linearly.

Core trick: The core function computes the inner product of two points in a higher dimensional space rather than directly transferring the data to that space.Common kernels are:

Linear kernel: $(x,) = x \cdot y$K(x,y)=x·y

Polynomial kernel: $K\ (x, y) = (x \cdot y+1)$) n K(x,y)=(x·y+1)d.

Radial Basis Function (RBF) kernel: $(x,) =$ \ n exp$f_0$$(-\gamma\|x-y\|2)$ K(x,y)=exp(− γ‖ x − y‖2)

Sigmoid kernel: $(x,) =$tanh$f_0$$(x$\n·$y+c)$ K(x,y)=tanh(αx·y+c)

Soft Margin SVM: In practice, complete separation of classes is often not possible. Soft margin SVM allows some errors to be classified for better generalization.

Regularization parameter (C): controls the trade- off between maximizing the margin and minimizing the classification error. A smaller$C$C allows for more classification errors, which increases the margin. Larger$C$C tends to classify less, which can lead to a narrower margin.

Algorithm: Logistic Regression

The training process involves finding the optimal weights$w$w and bias$b$b that minimize the cost function. This is usually done by calculating the gradient:

1. Initialize the weights and offsets.

2. Calculate the predicted probabilities from thetraining data.
3. Calculate the cost function.
4. Update the weights and biases using thegradients of the cost function.
   - Weight update: $w=w-\alpha\partial J\partial w\mathbf{w}=\mathbf{w}-\alpha\partial\mathbf{w}\partial J$
   - Bias update: $b=b-\alpha\partial J\partial bb=b-\alpha\partial b\partial J$
     where$\alpha\alpha$ is the learning rate.

A logistic regression model's performance can beassessed using a variety of indicators, including *Accuracy:* the proportion of correctly classifiedinstances.

Precision: The proportion of true positive predictionsout of all positive predictions.

Recall (Sensitivity): Proportion of true positivepredictions out of all true positives.

F1 score: harmonic means between recall and precision.

ROC-AUC: The genuine positive rate as a percentage of false positives, as represented by the area under the receiver operating characteristic curve.

Algorithm: k-Nearest Neighbors (k-NN)

Select the quantity (k) of neighbors: Select how many of your closest neighbors will be considered for regression or categorization. K is a user-defined constant with a value.

Small k can be noisy and lead to overfitting, while large k can over-smooth the decision boundary.

Calculate the distance

Calculate the distance between the input data point and all training data points. Common distance measures are:

Euclidean distance: distance=$\sum i=1(xi-y$n) n 2distance=$\sum$i=1n(xi−yi)2

Manhattan distance: distance=$\sum i=1n|xi-y\square$ n distance n= $\sum$ i = 1n|xi−yi|

Minkowski distance: Generalization of Euclidean and Manhattan distances.

Identify neighbors: Sort the distance and select the k nearest neighbors for the input data point.

Make a prediction:
Classification: Determine the class label that is most common among k nearest neighbors.

Regression: Calculate the mean (or weighted mean) of k nearest neighbor values.

Algorithm: Random Forest:Data preparation

- *Input data:* Data set with properties$X$X andtarget variable$y$y.
- *Data partitioning:* The data is partitioned into training and testing sets.

Forest Construction

- *Bootstrap Sampling:* Create n training data subsets using sampling and replacement. The size of each subset is the same as that ofthe original training set, however examples may overlap.
- *Decision tree training:* Train a decision treefor each subset.

Random feature selection: Only a random subset of characteristics is taken into consideration for sharing during tree construction, as opposed to considering every feature that could be present at every

node. As a result, there is less correlation between trees and more randomness. Without trimming, every tree reaches its full depth, producing low bias but possibly significant variance.

Making predictions

- *Classification:* Each tree in the forest gives aclass prediction. The final category is determined by a majority vote of all trees.
  Final Prediction=mode (predictions fromaltrees).
- *Regression:* Each tree gives a numerical prediction. The final prediction is the average of all three predictions.
  Final prediction=1n∑i=1$n$prediction from tree$i$.

METHODOLOGY

The Block diagram of Identification of Brain Stroke Using Artificial Intelligence is shown in Figure 3.

1. Load the datasets.
2. Pre-process the data.
3. Divide the dataset into test and train sets.
4. The machine learning models are trained using the traindataset.
5. Test the model for prediction and accuracygeneration using the test data.

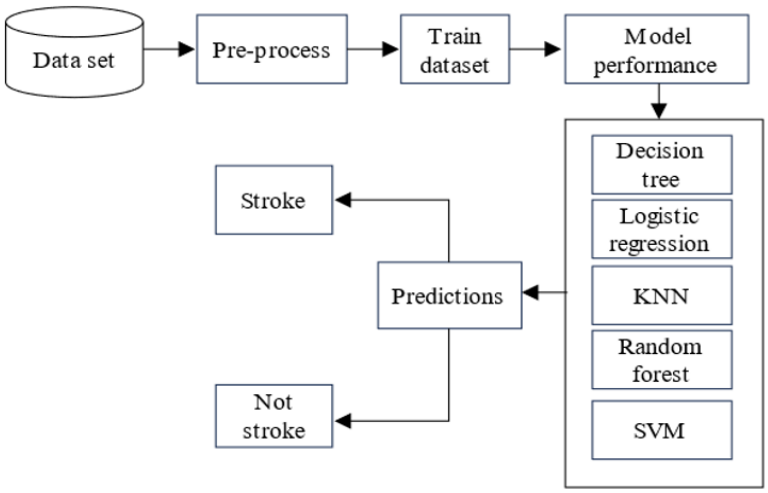

Figure 3. Block diagram of Identification of Brain Stroke Using Artificial Intelligence.

CONCLUSION

Finally, this work explores and evaluates different AI- based stroke diagnosis techniques, including ML and DL models. This offers helpful details regarding their functionality, scalability, computing efficiency, and advantages and disadvantages. This will helpresearchers and medical professionals choose the best strategy for accurate and successful stroke detection. In addition, we include performance evaluation tables,advantages and disadvantages of ML and DL approaches in our study. In this paper, we provide various open data sets in tabular format to help researchers find data to improve their models. The "Stroke detection using AI and ML" project aims to use the power of artificial intelligence and machine learning to change the way stroke is detected and diagnosed. Although the anticipated outcomes would help specific patients, they will also have a significant impact on the healthcare system. This initiative is a promising first step toward improved patient care and outcomes through more effective and precise health diagnostics.

REFERENCES

1. Karthik R, Menaka R, Johnson A, Anand S. Neuroimaging and deep learning for brain stroke detection-A review of recent advancements and prospects. Computer Methods and Programs in Biomedicine. 2020 Dec 1; 197:105728.

2. Donkor ES. Stroke in the century: A Snapshot of the burden, epidemiology, and quality of life. Stroke Res. Treat. 2018. Article ID. 2018;3238165(10).
3. Feigin VL, Stark BA, Johnson CO, Roth GA, Bisignano C, Abady GG, Abbasifard M, Abbasi-Kangevari M, Abd-Allah F, Abedi V, Abualhasan A. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology. 2021 Oct 1;20(10):795–820.
4. Andersen KK, Olsen TS, Dehlendorff C, Kammersgaard LP. Hemorrhagic and ischemic strokes compared: stroke severity, mortality, and risk factors. Stroke. 2009 Jun 1;40(6):2068–72.
5. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016 Nov 10.
6. Shoily TI, Islam T, Jannat S, Tanna SA, Alif TM, Ema RR. Detection of stroke disease using machine learning algorithms. In2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2019 Jul 6 (pp. 1–6). IEEE.
7. Li X, Bian D, Yu J, Li M, Zhao D. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. BMC medical informatics and decision making. 2019 Dec;19:1–7.
8. Sailasya G, Kumari GL. Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications. 2021;12(6).
9. P. Govindarajan, et al, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, pp. 817– 828, 2020.
10. Yu J, Park S, Kwon SH, Ho CM, Pyo CS, Lee H. AI-based stroke disease prediction system using real-time electromyography signals. Applied Sciences. 2020 Sep 28;10(19):6791.
11. Stroke Prediction Dataset [accessed on May 25, 2022] Available online: https://www.kaggle.com/datasets/fedesoriano/stroke- prediction-dataset.
12. Ruban S, Dcosta F, Saldanha R, Jabeer MM. Stroke prediction using machine learning techniques. Sacaim-2022. 2023 Jul 2:166.
13. Duda RO, Hart PE. Pattern classification. John Wiley & Sons; 2006.
14. Bentley P, Ganesalingam J, Jones AL, Mahady K, Epton S, Rinne P, Sharma P, Halse O, Mehta A, Rueckert D. Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage: Clinical. 2014 Jan 1;4:635–40.
15. Jayachitra S, Prasanth A. Multi-feature analysis for automated brain stroke classification using weighted Gaussian naïve Bayes classifier. journal of circuits, systems and computers. 2021 Aug 18;30(10):2150178.
16. Feng R, Badgeley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. Journal of neurointerventional surgery. 2018 Apr 1;10(4):358–62.
17. Zhu G, Bialkowski A, Guo L, Mohammed B, Abbosh A. Stroke classification in simulated electromagnetic imaging using graph approaches. IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology. 2020 May 18;5(1):46–53.
18. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, Estany-Gestal A, Sobrino T, Campos F, Castillo J, Rodríguez-Yáñez S. Random forest-based prediction of stroke outcome. Scientific reports. 2021 May 12;11(1):10071.
19. Kokkotis C, Moustakidis S, Giarmatzis G, Giannakou E, Makri E, Sakellari P, Tsiptsios D, Karatzetzou S, Christidi F, Vadikolias K, Aggelousis N. Machine Learning Techniques for the Prediction of Functional Outcomes in the Rehabilitation of Post-Stroke Patients: A Scoping Review. BioMed. 2022 Dec 27;3(1):1–20.
20. Nusinovici S, Tham YC, Yan MY, Ting DS, Li J, Sabanayagam C, Wong TY, Cheng CY. Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology. 2020 Jun 1;122:56–69.

