2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)

Brain Stroke Prediction using Decision Tree Algorithm

1st Viswanatha V
Department of ECE
Nitte Meenakshi Institute of Technology
Bengalulru, India
viswanatha.v@ieee.org

2nd Ramachandra A C
Department of ECE
Nitte Meenakshi Institute of Technology
Bengaulru, India
ramachandra.ac@nmit.ac.in

3rd Parameshachari B.D
Department of ECE
Nitte Meenakshi Institute of Technology
Bengaulru, India
paramesh@nmit.ac.in

4th Aditya Kumar Sharma
Department of ECE
Nitte Meenakshi Institute of Technology
Bengaluru, India
1nt21ec006.adityakumar@nmit.ac.in

***Abstract*—Stroke is a condition that occurs due to interruption or reduction in the blood supply to the brain. Due to this, the neurons in the brain suffer hypoxia which leads to cell injury or even death of the brain cells if not treated in time. This causes permanent or long-term injury and affects the person's life forever. Hence, early detection and prevention of stroke are essential as it is one amongst of the top causes of mortality and morbidity worldwide. Decision tree methods have been developed as a useful tool for this and machine learning (ML) models have shown promise in stroke prediction. Decision trees remain suitable for use in healthcare applications because they offer interpretability and can handle both categorical and numerical data. A dataset containing demographic figures, medical record, lifestyle features, and other pertinent variables are used by this ML model. The data set is analyzed using decision tree method, which identifies risk factors for strokes and captures intricate interactions between predictors. The handling of missing values and outliers eliminates the need of intensive preparation. By using different parameters in the decision tree algorithm this model has successfully given about 93.4% accuracy on the validation dataset. This accuracy was given when the depth of the tree was set to 4. Depths of 9 and 16 were also implemented. The depth of 9 gave us 91.4% accuracy on the validation data whereas the depth of 16 gave us 91% accuracy. It is to conclude that adopting a simple model by applying the decision tree approach provides us the best results after training this model on various parameters.**

***Keywords*—Stroke, machine learning, decision tree classification**

I. Introduction

The overall prevalence of stroke in the United States is 2.5%, with more than 7 million people aged 20 and older having had a stroke. This situation has a negative impact on the health and quality of life of affected patients. In addition, it puts a burden on hospital care, resulting in a shortage of available beds. In 2014-2015, the economic cost of stroke in the United States was approximately $351.2 billion. [1]. The dataset used here is from kaggle.com [2]. This dataset contains 5110 rows of data. It contains patient gender, age, hypertension level, heart disease conditions, marital status, job type, residence type, average glucose level, BMI, smoking status and weather they had a stroke or not.

According to India today, Stroke is the second most leading cause of deaths in India and about 1,85,000 strokes are reported every year in India with nearly one stroke every 40 seconds [3].

Decision trees are algorithms of supervised ML used for classification and regression. It is described as a hierarchical, tree structure, which consists of a root node n number of branches, internal nodes, and leaf nodes [4]. The decision tree as shown in fig.1 follows a method where it uses a tree like chart to make predictions. The dataset is split into different sets based on importance and significant feature at each tree node. Decision trees can be easily understood and are very versatile. In this study the decision tree classifier is used to classify the data based on whether a stroke will occur or not. In this section some of the documents regarding decision tree algorithm and stroke are reviewed.

The authors tried using natural language processing on social media posts. This method is called the DRFS method. This method helps in finding different symptoms of stroke. This model is undesirable [5].

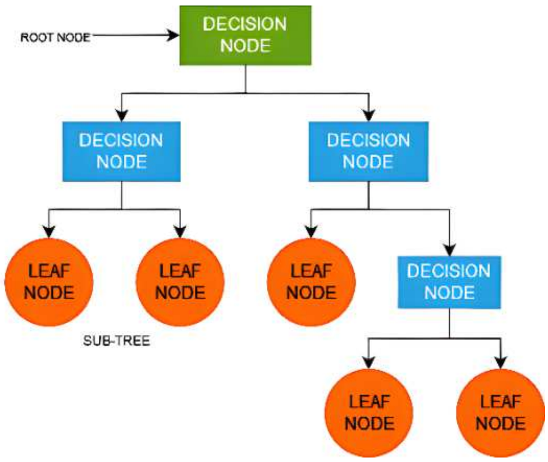

Fig. 1. Decision tree algorithm

II. Literature Survey

Research documents show three different algorithm types used to detect stroke. The algorithms used are decision tree, Naïve Bayes and neural networks. Conclusion from that paper is that the decision tree algorithm had the highest accuracy compared to the other two, but that doesn't work well with validation dataset [6].

The Cardiovascular Health Study (CHS) dataset was used in stroke prediction. The authors combined the Decision Tree classifier with C4.5 algorithm, support vector machine, principal component analysis and artificial neural networks provide the best result [7]. Since the dataset used in our paper

979-8-3503-1341-3/23/$31.00 ©2023 IEEE

is way smaller, Neural networks would not be a feasible option.

However, the CHS dataset used here has a smaller input parameter. The model had been trained on decision tree to predict stroke as demonstrated research publication [8]. The calculated accuracy of the decision tree was 74.31%.

Documentation says that both classification and regression problems are solved by decision tree classification. The input variables for this supervised learning approach algorithm already have matching.

output variable. The structure resembles a tree. With this technique, the data is continuous divided based on a specific parameter. Decision and leaf nodes are two components decision tree [9].

The first node separates the data and the second node provides the result. The accurateness of decision tree classification technique in this stroke prediction situation was 66%, which is lower than logistic regression. Precision and recall scores are equal and equivalent to 77.6%, as in logistic regression. This algorithm yielded an F1 score of 77.6%.

The classification algorithm of back-propagation neural networks is used together with Decision Tree Algorithm, PCA Algorithm and Dimension Reduction Algorithm to create classification model [10].

According to Chutim Jalayondeja, when using demographic data for prediction, decision tree, naive Bayes, and neural network are three models that have been taken into account. Decision tree was found to have the highest accuracy and the lowest false positivity rate (FP).

However, the neural network (NN) was chosen for safety because it has high False Positives and less false negatives values [11]. The popular C4.5 decision tree method analyzes NIHSS scores and classifies stroke based on severity, which is divided into four categories.

This information helps to predict the potential timing of the stroke and its associated handicap, enabling submission other medications and basic safety precautions. This information is very helpful when it is learnt about it stroke and its patterns in humans [12]-[13].

III. Methodology

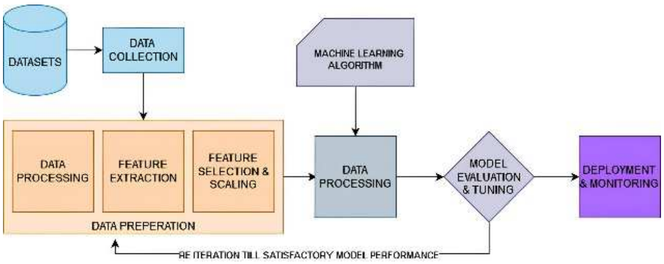
DATASETS
DATA COLLECTION
MACHINE LEARNING ALGORITHM
DATA PROCESSING
FEATURE EXTRACTION
FEATURE SELECTION & SCALING
DATA PREPERATION
DATA PROCESSING
MODEL EVALUATION & TUNING
DEPLOYMENT & MONITORING
RE ITERATION TILL SATISFACTORY MODEL PERFORMANCE

Fig. 2. Methodology workflow

Following are the brief steps followed to build the model as shown in fig.2.

A. Data collection

Identification of datasets with necessary and relevant features for a given problem statement.

B. Data preprocessing

Once data is collected, it needs to be processed before being used. This includes removing missing values and normalizing or standardizing the dataset.

C. Feature Selection

Feature selection involves identifying the most applicable features that contribute to the predictive power of the model.

D. Model Selection

Choosing a suitable ML algorithm based on nature of the problem, data size and other requirements.

E. Model training

In this step, the selected model gets trained using the prepared data set. The data is then divided into data for training and validation sets, and the model learns patterns and relationships in the train data. Various optimization techniques, such as gradient descent, are often used to iteratively update model parameters to minimize prediction error [14].

F. Model evaluation

After training, the model performance is evaluated using a validation or test set. Evaluation metrics such as accuracy, precision, recall, F1-score, or root mean square error are used to quantify model performance [15]. This step helps to understand how well the model generalizes to unseen data. A machine learning workflow is a reiterative process with reaction from evaluation and monitoring to inform next steps. This helps in continuously improving the model performance and addressing any challenges or limitations that arise during the development and deployment phase.

The following section is about the step-by-step implementation of the previously discussed methodologies [16].

1) Data set: The data set for stroke prediction is from Kaggle. This dataset has 5110 rows and 12 columns. The columns have 'id', 'gender', 'age', ‘hypertension', heart_disease','ever_married’, ‘work_type', 'Residence_type', avg_glucose_level', 'bmi', 'smoking_status' and 'stroke' as the main features. The output column 'stroke' has the value as either '1' or '0' [2].

The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. This dataset is highly imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column.

TABLE I. Dataset Description

Attribute	Description
Id	Patient ID number
Gender	Gender of patient
Age	Age of patient
Hypertension	Prescence or absence of hypertension
Heart_Disease	Prescence or absence of heart disease
Ever_Married	Marital status
Work_Type	Different job categories
Residence_Type	Urban or rural residence type of patient
Glucose Level	Average bull glucose in patient
Bmi	Body mass index of patient
Smoking_Status	Smoker or non-smoker
Stroke	Shows stroke status

Table-I consists of the list of attributes in the dataset and a small description for them. All of these are the physiological conditions that could be potential risk factors.

2) Data processing: Data processing is done to remove the unwanted data from the data set. By processing the data, it is made sure that the data used has the same data type. This helps in the making of a proper detailed analysis on the data.

First, the relevant columns for the model are chosen. Then they are added to a list. From that list the columns that have no values in them are removed.

After doing this data columns that have non numerical values are left. To convert these non-numerical values into dummy numbers an encoding known as one hot encoding is used. By using this method, dummy values can be easily given in the dataset wherever there is non-numerical data.

Further, the 'train_test_split' function is used to split the entire data frame into training data and validation data in an 80:20 ratio.

3) Model: A decision tree classifier is a popular ML algorithm that is used for both classification and regression functions. The decision tree forms a hierarchical flowchart-like structure where each internal node represents a decision based on a specific task and each leaf node represents a class label or predicted value. Decision trees are easy to interpret as they have a flowchart-like structure. Each path from root to leaf node signifies a set of conditions that lead to a particular outcome.

Decision trees can be combined into models such as random forests and gradient boosting, which often provide better performance and generalization.

At first, the algorithm with the maximum depth of 9 branches is applied as this will give a moderate level of complexity.

Along with decision tree classifier grid search cv has been used. The grid search function is from the sklearn library. This helps in achieving the best parameters from a set of parameters in a grid.

Thorough this research the CV value has been set to 10. In grid search, CV stands for cross validation. If CV is not given, then by default it cross validates it 5 times. The purpose of cross validation is to recognize if there is any overfitting or a failure to recognize a pattern [13].

Tabulating the accuracy of the training and validation data from this part as shown in Table-II.

TABLE II. THE ACCURACY WITH DEPTH=9

Data	Accuracy
Test data	93.751%
Validation data	91.44%

To improve the model by decreasing its complexity, the depth of the tree is reduced to '4' from the previous value of 9. This change requires making some adjustments in the code. The new depth is set as '4'. Using a matplotlib library, the decision tree can be visualized as shown in fig.3.

Here, it is clear after decreasing the complexity to four a better accuracy is obtained as shown in Table-III.

TABLE III. ACCURACY OBTAINED BY DEPTH=4

Data	Accuracy
Test data	95.65%
Validation data	93.40%

To maximize accuracy, an attempt to increase the depth of the decision tree to sixteen has been made. Increasing the depth of the decision tree helps in recognizing more minor patterns and relationships but may lead to overfitting. To prevent overfitting in decision tree, cross validation is used.

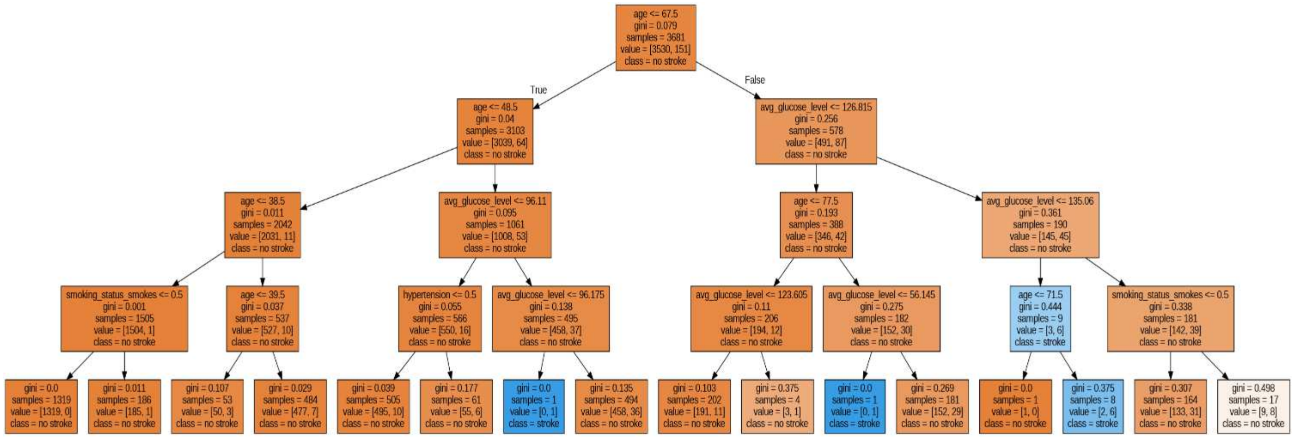

Fig. 3. Decision Tree with Depth = 4

By seeing the accuracy in Table-IV, it can be confirmed that the accuracy is decreasing with the increase in complexity of the model. Figure-3 shows a representation of the model with a depth of 4.

TABLE IV. ACCURACY OBTAINED BY DEPTH=16

Data	Accuracy
Test data	93.697 %
Validation data	91.38 %

Hence, it can be said that the best results obtained out of the three models is the one with depth of four included in table III.

4) Equations: Entropy is a metric for determining how random or unpredictable a dataset is. When classifying data, the dataset's distribution of class labels is used to calculate the randomness. For a subset of the original dataset with K classes for the ith node, the entropy can be defined as:

$$H_i = -\sum^{n} p(i,k) log_2\, p(i,k) \tag{1}$$

In equation (1), S is a sample of the dataset. K is a specific class out of K classes. The amount of data points in class k as a percentage of all the data points in dataset sample S is known as p(k) and P (i, k) shouldn't be equal to 0 in this case.

The Gini Impurity evaluates a score between 0 and 1, where 0 is the case when all observations fall into one class and 1 is the case when the elements within classes are distributed randomly. In this situation, we wish to have a low Gini index score. The expression for gini impurity is defined in (2). Where pi is elements of of the ith category.

$$Gini\ Impurity = \ 1 - \sum p_i^2 \qquad (2)$$

IV. Results And Discussion

A. Stroke and BMI relationship

A higher BMI increases risk of stroke. Studies have shown that people with a high BMI have a higher risk of stroke than those with a low BMI. Obesity is a known risk factor for many diseases, including stroke. Excess weight, especially visceral fat, contributes to the development of cardiovascular risk factors such as high blood pressure, diabetes and high cholesterol, which can increase the risk of stroke.

A high BMI can also affect recovery after a stroke. Obese individuals who have a stroke may face greater challenges in the recovery process, such as reduced mobility, increased risk of complications, and difficulty in rehabilitation.

In the Figure-4, BMI has been plotted on x-axis and the respective count has been plotted on y-axis.

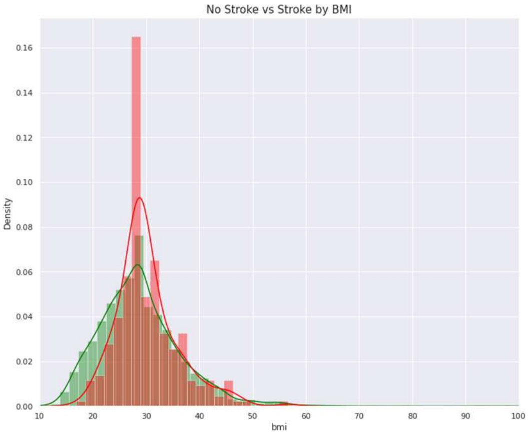

Fig. 4. Stroke & BMI Relation

Patients with higher BMI have more chances of getting stroke.

B. Average Glucose level relation with stroke

In the below graph, Average glucose level has been plotted on x-axis and Density has been plotted on y-axis. Figure 5 displays the patients' glucose levels and how it affects their chances of getting a stroke. Diabetes is a risk factor for stroke. People with diabetes have higher blood sugar levels because insulin is not produced and used. Chronically elevated blood sugar levels increase the risk of atherosclerosis (hardening and narrowing of the arteries), which can damage blood vessels and result in stroke. Hyperglycemia (high blood sugar) is common in the acute phase of stroke and is associated with a worse outcome. Studies have shown that hyperglycemia during stroke is associated with improved mortality, disability and function compared to stroke patients with normal blood glucose levels.

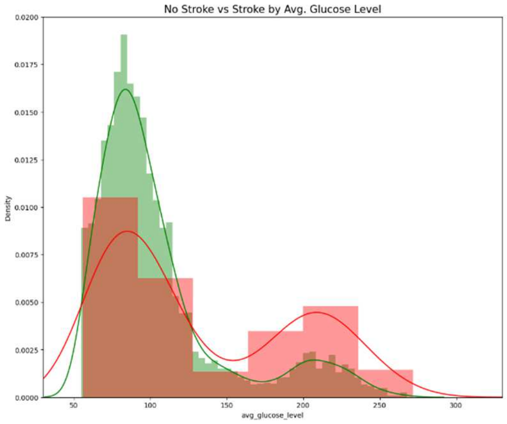

Fig. 5. Glucose Level and Stroke relationship

Patients with higher glucose level had more risk of stroke.

C. Age and stroke relationship

In figure 6, age has been plotted on x-axis and density has been plotted on y-axis. With increase in age the chances of the patients getting stroke also increases.

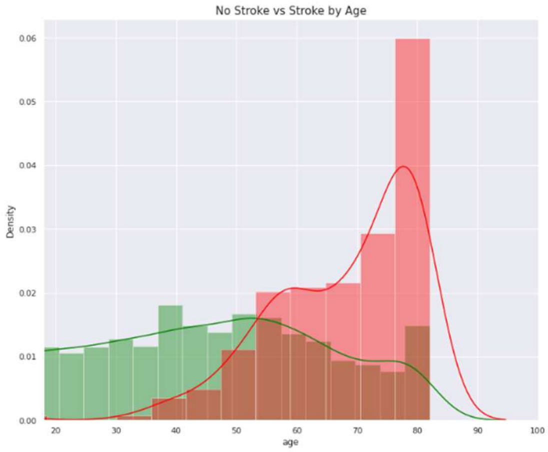

Fig. 6. Stroke & Age Relation

Patients between 70-85 had the most strokes compared to other age groups.

By looking at this plot one can confidently say that the chances of a person getting a stroke gradually increases with age.

D. Heatmap of dataset

A heatmap is a visual representation of data that uses a color gradient to represent values in a matrix.

It is commonly used to identify relationships or patterns between features and can provide insights into the strengths and direction of those relationships. The heatmap of this dataset is represented in figure 7.

A decision tree classifier with depth 9 has nine levels, including the root node and eight other levels.

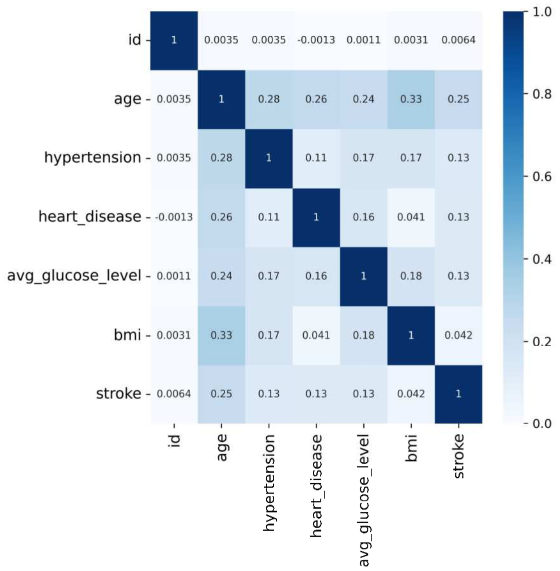

Fig. 7. Heatmap of all the data values used in the dataset.

Deep trees can represent more complex relationships and interactions between elements. However, there is a risk of overfitting the training data, for example with deep trees. This means that the model cannot generalize to unseen data.

Decision tree classification with a depth of 4 consists of four levels, including the root node and three other levels. The root node will make the first decision based on the most significant feature. Subsequent levels are decided based on remaining features.With a depth of '4' , the decision tree can capture moderately complex relationships between elements and the target variable.

It strikes a balance between simplicity and capturing more nuanced patterns in the data.

Using this decision tree algorithm gives us the best accuracy on training and validation dataset.

E. Comparative Analysis

A comparison of the decision tree and other models is shown in Table V.

TABLE V. COMPARATIVE ANALYSIS

ML Model	Accuracy
Decision Tree	95.6%
Logistic Regression	77.201%
SVM	79.843%
KNeighbous	83.65%
GaussianNB	19.27%
BernoulliNB	60.56%

This table makes it quite evident that the decision tree model performs better for this dataset. Table V, which compares our model of choice with various levels of complexity, is shown below. Table VI delve into the best depth of decision tree that the decision tree method performs best with a low complexity of 4, as rather than 9 or 16. Hence, we can say that simple decision tree works best for this dataset for its parameters.

F. Accuracy

The metric used for validating our model is accuracy score.

$$Accuracy = \frac{True\ Positives(TP)+True\ Negatives(TN)}{TP+TN+False\ Positives+False\ Negatives} \qquad (3)$$

Equation 3 represents the accuracy used to evaluate performance of the model.

TABLE VI. ACCURACY AND DEPTH

Depth	**Accuracy**	
	Train data	***Val data***
4	95.6%	93.4%
9	93.75%	91.4%
16	93.6%	91.0%

V. CONCLUSION AND FUTURE SCOPE

Decision trees are simple to use and comprehend, it can be a beneficial model for machine learning. It is especially effective in healthcare settings where clarity is necessary. To analyse patient data and identify relevant risk factors for strokes, ML models might leverage the potent tool of decision trees. Decision trees can handle both category and numerical data, making them flexible for stroke prediction. Again, Decision trees require extensive data preprocessing because they can handle outliers and missing values. Decision trees are adaptable for stroke prediction because they can handle both category and numerical data. They eliminate the need for intensive data preprocessing because they can deal with missing values and outliers.

The management of imbalanced datasets by decision trees can also be a problem, as is the case with stroke prediction where the prevalence of strokes is quite low in comparison to non-stroke cases. This problem can be solved using approaches like resampling methods (under sampling, oversampling), or by using the right assessment metrics, such AUC-ROC.

Decision trees provide a useful algorithmic strategy for stroke prediction in machine learning models, to sum up. They are a good option for healthcare applications due to interpretability, capacity to manage heterogeneous data types, and capacity to record complicated relationships. To enhance the efficacy in stroke prediction, due consideration should be given to potential issues such overfitting and unbalanced data.

Decision trees will work better with the help of larger datasets. Deep learning: Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two examples of deep learning algorithms that have shown great effectiveness in a range of healthcare applications. Stroke prediction algorithms that combine deep learning and decision trees can be more effective and precise.

Overall, improving accuracy, interpretability, and clinical value will be the focus of future machine learning (ML) models on stroke prediction utilizing decision trees method. We can enhance stroke risk assessment, early detection, and individualized therapies, thereby lessening the burden of strokes on people and healthcare systems, by embracing a variety of data sources, utilizing hybrid methodologies, and implementing these models in real-world situations.

REFERENCES

[1] Alanazi, Eman M., Aalaa Abdou, and Jake Luo. "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models." JMIR Formative Research 5.12 (2021): e23440

[2] Pradeepa, S., et al. "DRFS: detecting risk factor of stroke disease from social media using machine learning techniques." Neural Processing Letters (2020): 1-19.

[3] Kansadub, Teerapat, et al. "Stroke risk prediction model based on demographic data." 2015 8th Biomedical Engineering International Conference (BMEiCON). IEEE, 2015.

[4] Alanazi, Eman M., Aalaa Abdou, and Jake Luo. "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models." JMIR Formative Research 5.12 (2021): e23440.

[5] Singh, M. Sheetal, Prakash Choudhary, and Khelchandra Thongam. "A comparative analysis for various stroke prediction techniques." Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part II 4. Springer Singapore, 2020.

[6] Nwosu, Chidozie Shamrock, et al. "Predicting stroke from electronic health records." 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019.

[7] Joshi A, Vishnu C, Mohan C K. Early detection of earthquake magnitude based on stacked ensemble model. Journal of Asian Earth Sciences: X. 1;8:100122, Dec 2022.

[8] Singh, M. Sheetal, and Prakash Choudhary. "Stroke prediction using artificial intelligence." 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON). IEEE, 2017.

[9] Kansadub, Teerapat, et al. "Stroke risk prediction model based on demographic data." 2015 8th Biomedical Engineering International Conference (BMEiCON). IEEE, 2015.

[10] Yu, Jaehak, et al. "Semantic Analysis of NIH stroke scale using machine learning techniques." 2019 International Conference on Platform Technology and Service (PlatCon). IEEE, 2019.

[11] V. V, R. A. C, S. B. M, A. Kumari P, V. S. Reddy R and S. Murthy R, "Custom Hardware and Software Integration: Bluetooth Based Wireless Thermal Printer for Restaurant and Hospital Management," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-5, doi: 10.1109/MysuruCon55714.2022.9972714

[12] V. V, R. A. C, V. S. R. R, A. K. P, S. M. R and S. B. M, "Implementation of IoT in Agriculture: A Scientific Approach for Smart Irrigation," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972734.

[13] Li, G., Wang, J., Jia, X. and Yang, Z., 2021, December. A new piecewise linear representation method based on the R-squared statistic. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 515-519). IEEE.

[14] Khaled Mohamed Almustafa. Prediction of heart disease and sensitivity analysis of classifiers. Almustafa BMC Bioinfirmatives (2020) 21: 278. https://doi.org/10.1186/s12859-020-03626-y.

[15] V. Viswanatha, A. C. Ramachandra, P. T. Hegde, M. V. Raghunatha Reddy, V. Hegde and V. Sabhahit, "Implementation of Smart Security System in Agriculture fields Using Embedded Machine Learning," 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 2023, pp. 1-6, doi: 10.1109/ICAISC58445.2023.10200240.

[16] Mohawesh, Rami, et al. "Fake or genuine? contextualised text representation for fake review detection." arXiv preprint arXiv:2112.14343 (2021).