Original Article

Heart Stroke Prediction Using Machine Learning

Syed Shareefunnisa[1*], S N Lakshmi Malluvalasa[2],T R Rajesh[3], Maridu Bhargavi[4]

[1*]Assistant Professor,Department of CSE ,Vignan's foundation for Science Technology and Research, Vadlamudi-522213, Guntur(dist), Andhra Pradesh(state), India, E-Mail:syedshareefa@gmail.com
[2]Assistant Professor,Artificial intelligence and Datascience,Koneru Lakshmaiah Education Foundation,Vaddeswaram, Andhra Pradesh, India
[3]Assistant Professor,Department of CSE ,Vignan's foundation for Science Technology and Research, Vadlamudi-522213, Guntur(dist), Andhra Pradesh(state), India
[4]Assistant Professor,Department of CSE ,Vignan's foundation for Science Technology and Research, Vadlamudi-522213, Guntur(dist), Andhra Pradesh(state), India, E-Mail: bhargaviformal@gmail.com

***Corresponding Authors:** Syed Shareefunnisa

*Assistant Professor,Department of CSE ,Vignan's foundation for Science Technology and Research, Vadlamudi-522213, Guntur(dist), Andhra Pradesh(state), India, E-Mail:syedshareefa@gmail.com
DOI: 10.47750/pnr.2022.13.S05.395

Abstract

Present day, Deaths due to heart strokes are increasing greatly day by day. Unfortunately, detecting such conditions in humans is a complex task. Handling such complex tasks can be done by using data sets. Heart strokes occurrence prediction can be done only by automation because we need to keep monitoring the heart rate. MITBIH arrhythmia is one of the datasets which help us, so it is used in this paper. Automation for the mentioned task can be obtained by using various data mining techniques. Some of the techniques used in this paper are decision trees, naïve bayes, ANN algorithm & Random Forest algorithm. Therefore, the motto of this paper is to compare the above-mentioned algorithms & find out which is more accurate in accomplishing the task. At the end, after all the assessments we can say that the algorithm Random Forest has got 99% of accuracy which is recorded as highest among all the algorihms. But in the case of ECG images ANN algorithm has achieved an accuracy of 94%.

Keywords: involved in this research— KNN, Naïve bayes, Decision Tree, Random Forest, Prediction of Heart Stroke, ECG images, ANN, Arrhythmia.

I. Introduction

This paper speaks about the prediction of heart strokes in human beings using different data mining techniques and comparing them.Heart of a human body is one of the chief part.So, Any irregularity in the functioning of the heart affects the whole body. In Present world, cases of deaths by heart strokes are increasing day to day. Some of the habits that cause heart issues are smoking, alcohol & high content of cholesterol presence. According to WHO, few millions are losing their lives every year. Reducing Stress, tension & having habits that promote heart health can prevent heart issues.

When the cause is known then it is easy to treat it i.e., diagnosis plays an important role in the health care system. Detecting the heart strokes at right time accurately can save lives which is done by this paper. Data mining technique is nothing but the study of huge chunks of data & using the stats obtained from it in a progressive way.Machine Learning helps in using this data by converting vague to formatted datasets.So, ML can be used in the diagnosis phase of medicine & thus treatment is made easy. IN this heart stroke scenario, the data is ECG readings & graphs.ML helps us in analyzing this data to predict a heart stroke in near future. All the above-mentioned techniques help further more in terms of accuracy.

Comparison of various ML data mining techniques like Random Forest, ANN & naïve bayes for accuracy & performance. So that it becomes easy for us to predict a heart stroke at primitive stage. There are many datasets in the online community but the dataset i.e., used in this paper is MITBIH Arrhythmia. Cardiac Arrhythmia refers to irregular pounding of heart. Class 'S' in the taken dataset tells us about cardiac arrhythmia. It is one of the primal warnings for a heart stroke.[1]–[6]

II. Literature Survey

Heart stroke kills a lot of people, and persons with a smoking status of 1 are more likely to have a heart attack. If the ElectroCardioGram (ECG-image) type is 'S,' meaning super ventricular, the person may be at risk for a heart attack.

A machine learning-based approach has previously been utilized by many academics to predict cardiac attacks. Chen et al. provided a report on heart attack prediction. They employed 82 patients' data who are affected by stroke in the analysis done by them, two Artificial neural network (ANN) models to predict the accuracy, the accuracy achieved in this process is 79% and 95% respectively. Chung et al. conducted the research on heart stroke to predict stroke patient fatality rates. A total of 15,099 patients were employed in the study to detect the onset of a heart attack. To detect heartbeats, they used a deep neural network technique. PCA was applied to extract a medical history and forecast a heart attack by the authors. Govindarajan et al had collected data from 507 patients for a study to classify heart attack disorders using a combination of text extraction classification and machine learning. Various machine learning (ML) algorithms were used for ANN

training purposes in their study, the SGD algorithm yielded the highest result of 95%. Amini et al. showed a study to forecast the manifestation of stroke in 807 healthy and sick participants, 50 of stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia and alcohol consumption in the study. Risk factors were classified. They used the c4.5 decision tree algorithm, which was 95% accurate, and the nearest neighbor approach, which was 94% accurate. Shin et al conducted research on AI-assisted heart rate prediction. The Cardiovascular Health Research (CHS) dataset was used in their study to predict stroke. For principal component analysis, the feature extraction decision tree approach was used. They have used a neural network classification method to create a model. The model achieved 97% accuracy. Chin et al. showed a study to see if an automatic early stroke could be sensed. The main goal of his study was to generate a scheme that could systematize a chief stroke using CNN. To train and test the CNN model, they collected 256 images. They used the data extension method to augment the captured image in their imaging scheme by prearranging to eradicate the incredible, non-existent field of the stroke. CNN's approach has a 90% accurateness rate. After reviewing the prev articles, the fundamental idea behind the proposed scheme was to build an input-based stroke prediction system. Based on their accuracy, precision, recall, and f-score measurement, we studied the KNN, Decision Tree, and Random Forest classification algorithms and selected the best classification method for heart disease prediction. Murat et al. offered an assessment of knowledge, and an increased focus on a deep understanding of modes has become commonplace for categorizing ECG data. They used ECG data from five organizations totalling 100,022 strokes from the MIT-BIH arrhythmia dataset to evaluate the most widely used deep learning algorithms in this works. Hong and colleagues The review author provides a methodical analysis of deep learning methods for ECG data, either from a modelling or post point of view. From January 1, 2010 to February 29, 2020, the author used the ECG (DL) Deep Learning model issued by the Google Scholar, PubMed, and the Digital Appendix & Library Mission. CNN, DBN, recurrent neural community, quick period memory (LSTM) and closed repeated machine are among the various DL approaches proposed by Ebrahimi et al. they were examined. To extract and classify features, most of the reviewed studies used at least some kind of DL approach. With thirteen deep layers of a fully folded neuronal community, Acharya et al. Popular and Predictive Doctrines identified (CNN). This method is time overwhelming, inadequate by the use of a realistic object, and is in control for variable grades on the student level. The overall performance of the concluding version of CNN, like ANN, will be determined by the shape weights of the community and the settings of the prev layers. The pooling method reduces the size of output neurons of the convolutional layer, reduces the computation width, and prevents overfitting. The accuracy and specificity, and sensitivity of the recommended technique were 88.67%, 90.00%, and 95.00%, respectively. Acharya et al. identified. The conclusion is that fuzzy base summation is successful in evaluating automated ECGs, but more research is needed. The accuracy of the meaning is 86.67 percent and the specificity is 93.75 percent. The constructed version has been well verified on a variety of general performance indicators, but can be altered for more realistic applications. According to Limaye and Deshmukh, ECG drawing thought includes very low occurrence alarms from 0.5 to 100 Hz. Devices that allow filtering of the ECG recording are called cardiac video display units. The Low Pass Filter (LPF) is used to filter out unsolicited high frequency noise. Mbachu et al. supplied LPF, HPF and BSF structures by the Kaiser window in which these three filters are connected in series, processing the signal in the range from zero to one hundred Hz and a signal of thirteen dB for each gap of two hundred and with interference mute warnings. days et al. used the Hamming window to create a virtual clean notch design with a 50Hz interference effect resulting in 13.4dB of diminution. Litjen et al. reviewed more than eighty cardiac MRI, computed tomography, and single photon computed tomography studies covering intravascular visible cohesion imaging and echocardiography. Teaches the fundamentals underlying the most advanced deep research algorithms. Kaplan Berkaya et al. published the review of ECG signs in which they analyzed 1,538 recordings, including measurements of heart rate, cardiovascular function, diagnosis of heart problems, emotion recognition, and biometric identification. They also summarized the most outstanding work in the pre-processing part of the ECG alarm evaluation. Studying its effects under different pre-processing methods is very important. Haroon explained the importance of deep learning techniques such as B. Convolution Neural Network (CNN) created algorithms that can save you the manual ECG signal function. Python PTB besides MIT-BIN Dataset ECG (wfdb) database fjle libraries were used for testing and several functions and versions of information were created.[1]–[6]

We took the arrhythmia dataset, which has roughly 1 lakh photos, and identified different types of heartbeats using all of these images. Following the categorization, we used the ANN algorithm to forecast the heart stroke rate and concluded that the S type of heart beat could result in a heart attack. Dataset and Methodology Used in this research.

We have taken the arrhythmia dataset which contains around 1 lakh images and by using the all these images we have classified types of heartbeats. After the classification is done, we have applied the ANN algorithm and predicted the heart stroke rate and predicted that the S type of heart beat may lead to heart stroke.

III. DATASET AND METHODOLOGY USED IN THIS RESEARCH

Datasets used:

.csv dataset: we have used heart stroke prediction.csv

To perform this study we have used the heart stroke prediction dataset from Kaggle. The dataset consists of 11 parameters and that parameters are id, age, gender, hypertension, worktype, residence type, heart disease, avg level of glucose, body mass index (bmi), marital status, smoking status, and stroke.

Description of the dataset:

The dataset contains 11 attributes and each attribute describes that the data is categorical or numerical data.

id: This element indicates a person's unique identifier. Data that can be calculated.

age: The age of the person is indicated by this characteristic. Information about the categories.

gender: The gender of the person is indicated by this characteristic. Data that can be calculated.
Hypertension: This feature indicates whether or not this person has high blood pressure. Information about the categories.
Work type: This property indicates a person's job situation. Information about the categories.
Residence type: The person's life situation is represented by this trait.
heart disease: If this person has heart disease, this trait indicates it. Data that can be calculated.
Avg glucose level: This feature reflects how high a person's blood sugar was on average. Data that can be calculated.
Bmi: Bmi stands for "numerical data." This property refers to a person's BMI (body mass index).
ever married: information from the category. The marital status of a person is defined by this property.
Smoking status: Categorical statistics on smoking status. This property indicates whether or not a person smoke.
stroke: numerical data. This trait reveals whether a person had a stroke or not. The decision class is the full attribute dash, and the rest of the attributes are the answer class.
Our dataset has approximately 62000 records. The input dataset is categorized into train and test datasets, with dataset the training model accounting for 80% of the total. A training dataset is the set of data, used to train a machine learning model. The test datasets are used to demonstrate the trained model's performance.

Sl.No	Attribute Name	Type
1	Gender	Numerical type
2	Age	Category type
3	Hypertension is there or not	Category type
4	Heart disease stroke	Numerical type
5	Marital status	Category type
6	Work type	Category type
7	Residence type	
8	Average_glucoselevel	Numerical type
9	Bmi (body mass index)	Numerical type
10	Smoking status	Category type
11	Stroke	Numerical type

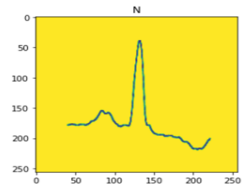
Image dataset:

We have taken the dataset from Kaggle i.e., ECG heartbeat classification (MITBIH Arrhythmia dataset) in which it contains 5 types of signals. The dataset contains around 6000 ECG records referring to the five super classes of cardiac arrhythmia. The dataset is divided into 2 splits i.e., training split and test split. The train split consists of 80% of data and test split consists of 20% of data.
The results for grouping of 5 classes from the MITBIH arrhythmia dataset are shown in the table below

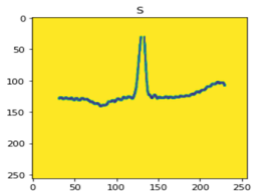
Category	Description	Annotations
Class N	Normal or non ectopic beat class	Normal type of heart beat and atrial escape
Class S	Supra-ventricular ectopic beat class	Fusion of both paced beat class and normal beat class nodal premature
Class V	Ventricular ectopic beat class	Premature ventricular contraction class
Class F	Class of Fusion beat	Combination of both ventricular class (V) and normal (N) class
Class Q	Class of Unknown beat	Combination of both paced class and normal (N) beat class Non classifiable

This is a combination of two well-known heart rate categorization datasets. We have used a dataset with some sampling frequency and a total of 6000 ECG beat images to test the deep learning model. The MITBIH arrhythmia dataset includes the main F, Q, N, V and S. This dataset was used to evaluate the particular transmission learning functions and to research

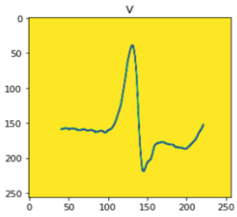
the heartbeat distribution using a deep neural network architecture. The signal simulates to an electrocardiogram (ECG)-style heartbeat in distinctive cases impacted by different arrhythmias and myocardial infarctions.
Each component of these signals represents a heartbeat and is segmented and pre-processed.
i) Normally, the "N" group includes bundle branch of right block or bundle branch of left block, nodular prolapse, atrioventricular prolapse.



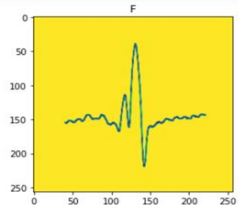
ii) The "S" category includes atrial-premature class, aberrant atrial-premature class, nodal-premature class, and the supra-ventricular premature class.



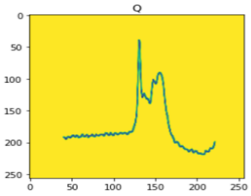
iii) The group “V” in computational intelligence and neuroscience includes ventricular prolapse class and pre-mature ventricular contraction class.

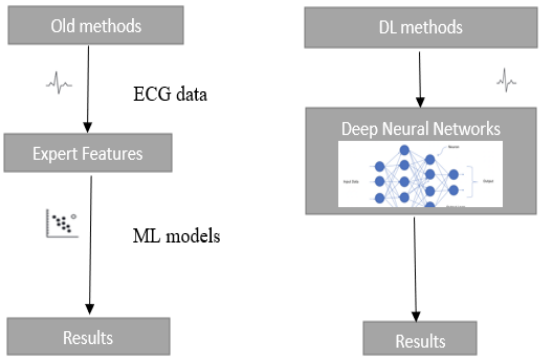


iv) The "F" class stands for fusion, which is combination of ventricular (V) class and normal (N) class.



v) The “Q” class stands for Supra-ventricular class it is the combination of the paced and Normal (Normal) class





METHODOLOGY AND MODELS USED:

There are some steps involved in the methodology and every machine learning model should follow this methodology.

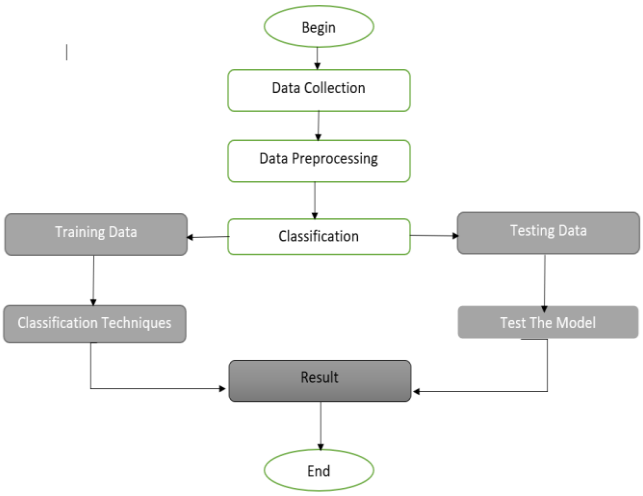


Fig Steps involved in the methodology

Techniques used for .csv format dataset:

The attributes utilized in the dataset are fed into the various machine learning (ML) algorithms as input.
In order to divide this dataset into train and test samples we need to preprocess the data. There are some techniques used for preprocessing
Finding mean of a column, finding median of a column, finding mode of a column, Checking missing values, Checking unique values in the dataset, Finding no.of affected persons by stroke, Finding the gender count, Grouping the gender based on stroke, Finding the smoking status count, Grouping the smoking status based on stroke, Splitting the data into two parts, Label encoder, Defining features and label, Splitting dataset into train and test, Building the models

We have used 4 different types of ML algorithms to find the performance metrics for this project. They are

I. Decision tree algorithm
II. Naïve Bayes algorithm
III. Random Forest algorithm
IV. K- Nearest neighbor algorithm

Algorithm name: Decision tree
Decision Tree algorithm comes under Managed learning that can be used to solve both classification and regression problems. It is mainly based upon selecting the order of the nodes (attributes). We can follow specific parameters such as gini index, impurity, entropy, information gain.

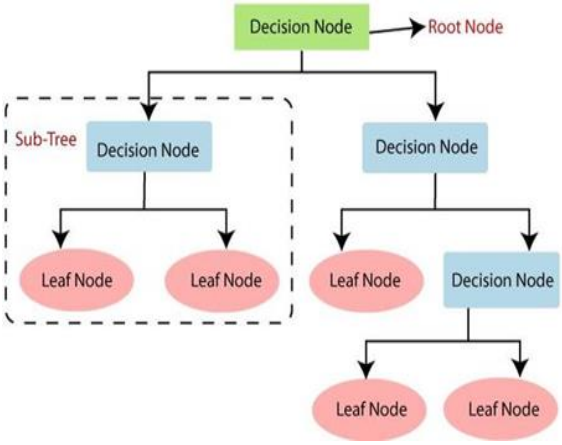


Fig., Decision-Tree model

Algorithm name: Naïve Bayes

The Naïve Bayes Classifier takes the reference of the bayes theorem, Bayesian classifier are the statistical classifiers. It contains evidence, hypothesis, likelihood.
A – Hypothesis and B – Evidence and P(B|A) – Likelihood
P(A|B) – it is a posterior probability and P(H) – Prior probability

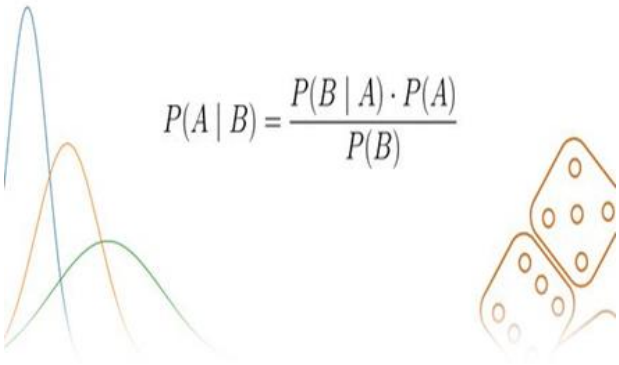


Fig Naïve bayes model

Algorithm name: Random Forest Algorithm

A random forest algorithm is a classifier that combines a large no.of decision trees from different subsets of a dataset and averages the predictability of the dataset. It is a controlled machine learning algorithm commonly used to resolve classification and regression problems.

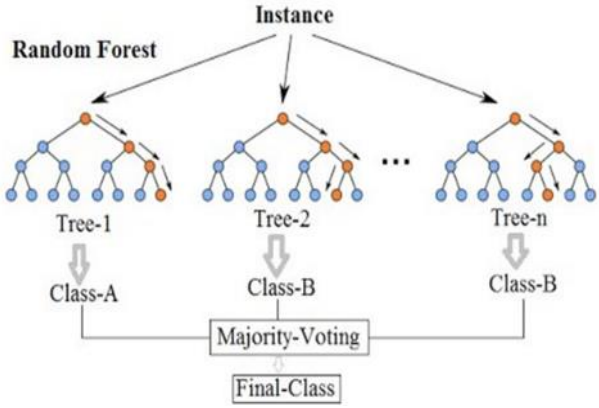


Fig., Random-Forest classifier

Algorithm name: K-nearest neighbour

The K-classifier nearest neighbor in the pattern searches for space training patterns and selects the pattern closest to the unknown pattern. The proximity between samples, defined by the Euclidean distance. The K-Nearest Neighbor method is a non-parametric algorithm that takes all existing data and organizes a new data point based on likeness. This means that new data can be straightforwardly classified into the appropriate group using the K-NN algorithm.

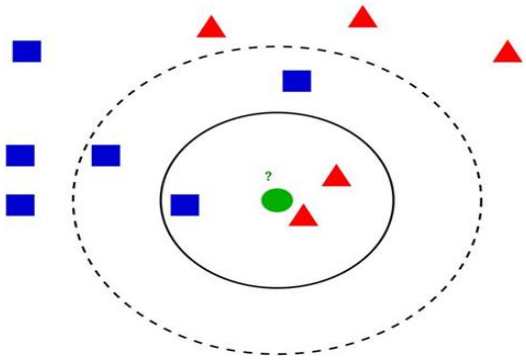
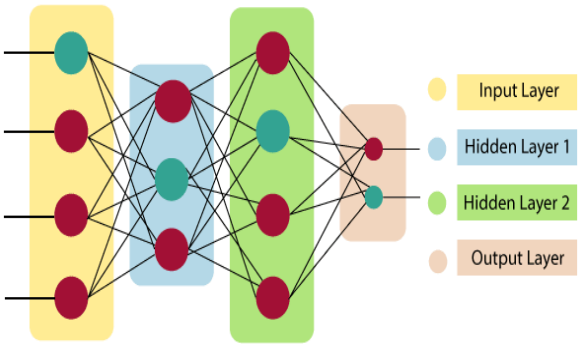


Fig., K- Nearest-Neighbour

Techniques used in ECG dataset:

Algorithm name: Artificial Neural Network (ANN)

The term "artificial neural network" comes from the biotic neural networks that develop the structure of the human brain. Like the human brain, which connects neurons, artificial neural networks have interconnected neurons at different layers of the network. The neurons are knows as nodes.



This algorithm contains different layers such as Input, Hidden, Output layers.
Name of the layer: Input
By the name, it accepts several types of input given by the programmer.
Name of the layer: Hidden
The layer hidden is located in between layer input and the layer output. It executes all computations to search all hidden structures and designs.
Name of the layer: Output
The layer input given in the input layer goes through a series of transformations using a hidden layer to reach the output delivered using this layer.
An artificial neural network (ANN/ takes an input, calculates the weighted sum of the input, and includes the bias. This calculation is expressed in the form of transmission function.

IV. RESULTS ANALYSIS

This part contains the explanation of how experiments are been carried out and how findings are obtained throughout the research. This experiment was carried out with the aid of the python programming language, and machine learning and deep learning techniques were used to carry out this research. To start this project, we have used csv data and ecg images data, we have loaded the dataset in CSV format by using the methods in python and applied preprocess technique and applied all the machine learning classifier models and predicted the performance metrics for that dataset. The train, test split function is divided such that the train split contains 80% and test split contains 20% of data from the dataset.
We have executed this project on csv file as well as on electrocardiogram (ECG) images. For the csv file we have used various algorithms in the machine learning classification such as the Decision tree algorithm, Naïve Bayes algorithm, k-Nearest Neighbor algorithm, Random Forest algorithm. Based on the accuracy metrics the scores are as follows.

Model used	Accuracy naïve
Naïve bayes	78%
Decision tree	98.5%
Knn	98.7%
Random forest	99%

For the ecg images we have used the artificial neural networks and built three models. For the three models we have used the hidden layers for the three models i.e.,
Model 1: We have used one hidden layer ran 50 epochs and the activation function is “sigmoid”. The accuracy obtained is 91%.
Model 2: We have used two hidden layers ran 100 epochs and the activation functions used are “relu” and “sigmoid”. The accuracy obtained is 94%.
Model 3: We have used two hidden layers ran 40 epochs and the activation functions used are “relu” and “sigmoid”. The accuracy obtained is 82%.

V. CONCLUSION

Hence this research focuses on the prediction of heart stroke, these techniques can be used to predict the heart stroke us acquire how the heart stroke is predicted based on csv format and ecg images.
Some of the algorithms in machine learning (ML) and deep learning (DL) undergoes training to predict heart stroke. The study results show that Random forest and for the ecg images model 2 of ANN algorithm gives the best accuracy among the remaining models
As per the results we can conclude that for csv data we recorded the best accuracy for Random Forest algorithm (accuracy). For the ecg images we recorded the best results for model 2 (no.of epochs=100, number of hidden hidden=2, activation functions= “Relu”, “sigmoid”).

REFERENCES

1. M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation," Apr. 2018, doi: 10.1109/ICHI.2018.00092.
2. D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, and M. Chen, "An ECG Heartbeat Classification Method Based on Deep Convolutional Neural Network," *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/7167891.
3. T. M. Geethanjali, D. M. D, M. S. K, S. M. K, and A. Professor, "STROKE PREDICTION USING MACHINE LEARNING," JETIR, 2021. [Online]. Available: www.jetir.orgd710
4. B. Deepak Kumar, S. Yellaram, S. kothamasu, S. Puchakayala, and A. Professor, "Heart Stroke Prediction using Machine Learning," 2021. [Online]. Available: www.ijcrt.org
5. W. Ullah, I. Siddique, R. M. Zulqarnain, M. M. Alam, I. Ahmad, and U. A. Raza, "Classification of Arrhythmia in Heartbeat Detection Using Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, 2021, doi: 10.1155/2021/2195922.
6. G. Sailasya and G. L. Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms." [Online]. Available: www.ijacsa.thesai.org