IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134210

Stroke risk prediction using K-Nearest Neighbors algorithm

Sudhakar Avareddy[1], Chandrashekhara P[2], Pramod C[3], Harish T[4], Ayyallappa[5]

Assistant Professor, Department of Computer science and Engineering, Ballari Institute of Technology and Management, Ballari – 583104, India.[1]

Final Year Students, Department of Computer science and Engineering, Ballari Institute of Technology and Management, Ballari – 583104, India. [2,3,4,5]

Abstract: Stroke is a critical and life-threatening medical condition that necessitates early detection and intervention to mitigate its impact. This project presents a stroke prediction model using the K- Nearest Neighbors (KNN) algorithm, a popular machine learning technique known for its simplicity and effectiveness in classification tasks. In KNN algorithm is applied to classify dataset into two categories. First is at high risk of stroke and second is at low risk of stroke. The objective of this project is to develop a reliable and accurate prediction system that can assist healthcare professionals in identifying individuals at risk of stroke. The dataset used in this project comprises various demographic, clinical, and lifestyle features of a diverse group of individuals, including age, gender, hypertension status, marital status, heart disease history, work type, smoking habits, and more. The project findings indicate that the KNN-based stroke prediction model achieves promising results in terms of accuracy, sensitivity, and specificity. This suggests that KNN can be a valuable tool for identifying individuals who may be at risk of stroke, allowing for early intervention and preventive measures to be taken.

Keywords: stroke, k-nearest neighbors, logistic regression, random forest, machine learning algorithms

I. INTRODUCTION

In recent years, the application of machine learning techniques in predicting Stroke risk has gained significant attraction. This introduction provides a comprehensive overview of stroke risk prediction, with a focus on the utilization of diverse machine learning algorithms. The introduction outlines the key objectives of stroke risk prediction, highlighting its importance in medical field making informed decisions, various machine learning methodologies, such as K-nearest neighbors, linear regression, random forest, SVM, Stochastic Gradient Descent, Naïve Baye , XG Boost, and gradient boosting ,which are instrumental in forecasting stroke risk by analyzing medical data. Additionally, the introduction delves into the inherent challenges and opportunities associated with stroke risk prediction, emphasizing the critical role of accurate predictions in optimizing medical field and system that can assist healthcare professionals in identifying individuals at risk of stroke. By providing a comprehensive overview of the subject matter, this introduction lays the groundwork for a detailed exploration of stroke risk prediction using machine learning techniques.

This research outcomes underscore the efficacy of the K-nearest neighbors (KNN) algorithm in forecasting stroke occurrences, demonstrating commendable outcomes regarding precision, sensitivity, and specificity. These findings imply that KNN holds significant potential as a discerning instrument for flagging individuals predisposed to stroke, thereby facilitating timely intervention and preemptive strategies. Furthermore, the model's flexibility and ease of implementation make it a practical choice for healthcare professionals and researchers interested in predictive modelling for stroke risk assessment. High blood pressure and smoking are considerable factors contributing to stroke. The WHO states that around 15 million people worldwide are victims of this disease every year, out of which five million pass away and another five million become permanently disabled .

A. ABOUT STROKE RISK

A stroke occurs when the blood supply to part of the brain is interrupted or reduced, leading to damage or death of brain cells. This interruption can happen due to a blockage in a blood vessel (ischemic stroke) or due to a ruptured blood vessel (hemorrhagic stroke). Strokes are medical emergencies that require immediate attention as they can lead to serious long-term disabilities or even death if not treated promptly. Recognizing the signs of a stroke and acting quickly by seeking medical help is crucial in minimizing the damage and improving the chances of recovery.

DOI: 10.17148/IJARCCE.2024.134210

B. TECHNICAL ANALYSIS

Predicting stroke risk through technical analysis involves utilizing various data sources and computational methods to identify patterns and factors associated with stroke occurrence. Technical analysis relies on data collection, data preprocessing, feature selection, model selection, model training, evaluation metrics, validation, deployment and integration. By leveraging technical analysis methodologies, researchers and clinicians can develop robust predictive models for stroke risk assessment, enabling proactive interventions and personalized preventive strategies to mitigate the burden of stroke on individuals and healthcare systems

II. RELATED WORK

A.LITERATURE REVIEW

This section highlights on the literature review carried out for stoke risk prediction using various machine learning algorithms.

Jie Chen, Yingru Chen, Jianqiang Li, Jia Wang; Zijie Lin, Asoke K. Nandi(2021). Here study leverages deep learning and transfer learning to enhance stroke risk prediction, a challenge exacerbated by the strict privacy policies governing healthcare data and severe data imbalance. The proposed Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) scheme harnesses knowledge from multiple correlated sources, including external stroke data and chronic disease data like hypertension and diabetes [1].

Haifeng Xu, Mei Li, Xuemeng Li, Lei Cao, Jianfei Pang, Dongsheng Zhao(2022). Here researchers focused on the development of long-term recurrent risk prediction models for high-risk non-disabling ischemic cerebrovascular events (HR- NICE) using continuous national stroke screening cohort data from 2015 to 2018. The study identifies and selects features relevant to HR-NICE patients and employs machine learning algorithms to create stroke risk classification models. Additionally, the complexity of machine learning models may pose challenges for healthcare professionals who need to interpret the reasons behind risk predictions, potentially limiting their adoption in clinical settings [2]

M. Anand Kumar; N Abirami; M S Guru Prasad; M. Mohankumar (2022).Here study leverages Stroke has become a major cause of death, second only to cardiovascular disease, according to the World Health Organization (WHO). ECG data plays a critical role in identifying various stroke risk factors, including left ventricular hypertrophy. ECG data plays a critical role in identifying various stroke risk factors, including left ventricular hypertrophy. Additionally, the focus on ECG data and its role in identifying stroke risk factors is a commendable approach as ECG is a widely accessible and cost-effective diagnostic tool [3].

A.P.Ponselvakumar; S.Nivetha; M.Nevithaprakashi (2022).
Here researchers focused on importance of accessible healthcare and the application of machine learning techniques for stroke risk detection. The research employs data mining techniques to analyze patient medical data and explore the interrelation of different risk factors. The study focuses on the application of machine learning algorithms, particularly the Random Forest method, using electronic health data for detecting stroke risk. Moreover, the emphasis on a specific machine learning algorithm (Random Forest) may limit the exploration of alternative methods that could potentially yield better results [4].

Santosh Kumar Satapathy, Abhi Patel, Pushti Yadav, Yashvi Thacker, Dhaval Vaniya, Drashti Parmar(2023). A here researchers focused on machine learning algorithms to predict the probability of individuals experiencing a stroke. The research utilizes data spanning a wide age range, collected from the health reports of over five thousand individuals. By employing various machine learning models, including Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression, the study aims to identify factors contributing to stroke risk. Notably, higher glucose levels and advanced age in females are found to be associated with increased stroke risk[5].

Puranjay Savar Mattas, Pradeep Kumar Mishra, Himanshu Singh; Dev(2023). Here researchers used dataset that undergoes cleaning to ensure data accuracy and completeness, ensuring the model's predictions are based on reliable information. The DTC model exhibits high accuracy in predicting stroke risk, potentially providing valuable insights for both patients and medical professionals. Patients can become informed about their stroke risk and take preventive measures, while healthcare providers can use the model to enhance patient care. The study offers advantages such as the development of an accurate ML model for stroke risk prediction, which has the potential to empower patients with personalized risk assessments, enabling proactive preventive measures. Additionally, it demonstrates the effectiveness of the Decision Tree Classifier, a widely used classification algorithm [6].

DOI: 10.17148/IJARCCE.2024.134210

B. DATA SOURCES

By merging stroke data from several sources, we were able to construct a dataset. Our database has data on 5110 patients, 2115 of whom are male and 2994 of whom are female. A single word (row) corresponds to one patient, and the attributes are variables in the dataset about the health status of every patient. There are a total of 11 features in the dataset. The “Healthcare-dataset-stroke-data” is a dataset of stroke prediction from Kaggle. In the overview, the categorical variables are Marital status, age, Work type (never worked, private, self-employed, children, government job), Gender (male, female), Hypertension (yes/no), heart disease(yes/no), Residence Type (urban, rural), Smoking Status (formerly smoked, never smoked, smokes) and As for qualitative variables, we have one’s BMI, age, and average glucose(sugar) level.

III. METHODOLOGY

The approach entails the incorporation of essential libraries and the application of machine learning classification algorithms. To predict stroke occurrences, pertinent data is derived from medical test outcomes, encompassing factors such as hypertension, heart disease, glucose levels, and BMI. Users have the option to input their data directly, or we can extract the necessary information from their medical reports, which are then stored in a server-side database. Subsequently, with the dataset of respective users containing all requisite attributes for prediction, preprocessing steps are initiated to ensure performance. Figure 1 illustrates the block diagram delineating the proposed methodology for stroke prediction. Data retrieval and storage from the database are facilitated through a user input model.

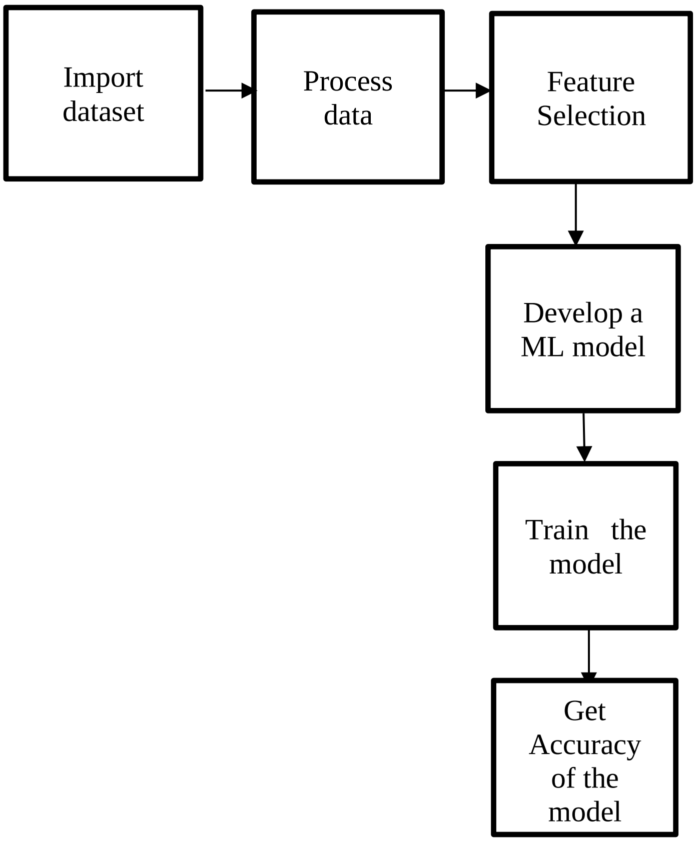

Fig. 1. Complete layout of the proposed research work

A. DATA PREPROCESSING

It encompasses several steps aimed at preparing the dataset for analysis. Once the dataset is collected, it needs to be preprocessed to ensure that it is clean and ready for analysis. This involves removing any missing or duplicate values, handling outliers, and converting categorical variables to numerical variables. Categorical variables can be converted to numerical variables using techniques such as one-hot encoding or label encoding. After preprocessing the data, the next step is to scale the features to ensure that they all have the same scale. This is important because some features may have larger values than others, which can affect the performance of the model.

DOI: 10.17148/IJARCCE.2024.134210

Feature scaling can be done using techniques such as standardization or normalization. Since our database includes string values incompatible with VS code, we utilize integer encoding via label encoding. For instance, "Male" and "Female" are represented by 0 and 1, respectively. Moreover, to handle missing attributes such as N/A, we employ null value handling techniques, substituting them with the mean value to prevent null value exceptions. Furthermore, we rescale attributes to a uniform range of 0 to 1 through data normalization, enhancing comparability and analysis.

B. FEATURE SELECTION

Feature selection plays a crucial role in stroke risk prediction using KNN (K-Nearest Neighbors) algorithm. By selecting relevant features, we can improve the efficiency and accuracy of the prediction model while reducing computational complexity After selecting the subset of features, it's important to evaluate the performance of the prediction model using these features. Finally, the selected subset of features is integrated into the KNN algorithm for stroke risk prediction. KNN classifies new instances by identifying the K nearest neighbors based on feature similarity and assigning the most common class label among them.

C. SYSTEM SETUP

System configuration is the procedure for characterizing a framework's architecture, parts, components, interfaces, and information to fulfill the specified preconditions. Configuration of frameworks could be interpreted as using the theory of frameworks to advance objects. The study and techniques situated in Article develop into the most commonly used method for the design of PC frameworks. In this way, the configuration of frameworks is the way to characterize and build frameworks to meet the client's defined needs. The UML has become the standard language in object situated investigation and structure. Structure design is an applicable model that characterizes the framework's structure and behavior. This includes the frame pieces and the relationship explaining how they work together to modify overall structure.

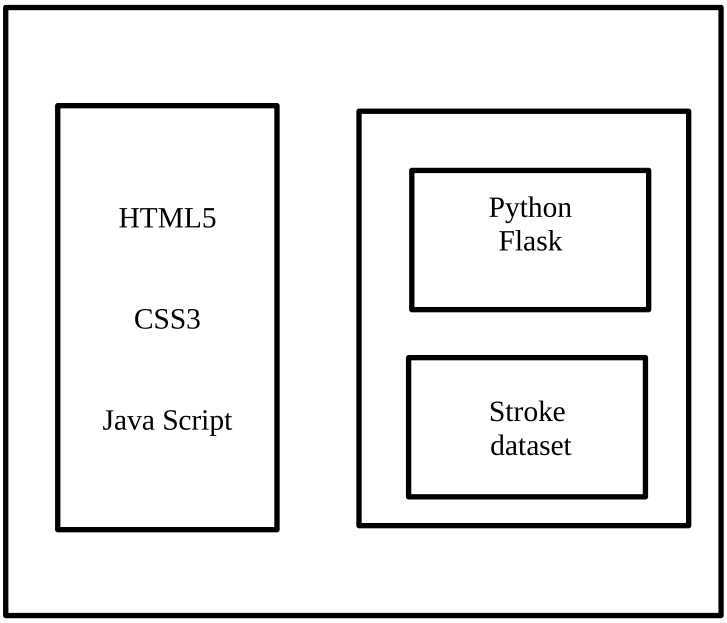

Fig. 2. System Architecture

D. STROKE PREDICTION

KNN is a supervised learning algorithm that can be used for classification tasks. In the context of stroke prediction, KNN can be used to classify individuals as either having a high risk of stroke or a low risk of stroke based on various input features. The basic idea behind KNN is to find the "k" nearest neighbors to a data point and then classify the data point based on the majority class of those neighbors. The value of "k" is a hyperparameter that needs to be tuned to achieve optimal performance. In the case of stroke prediction, the input features are used to represent each individual as a point in a multi-dimensional space. The KNN algorithm then finds the "k" nearest neighbors to this point in the space, based on a distance metric such as Euclidean distance. The majority class of these "k" nearest neighbors is then used to classify the individual as having a high risk of stroke or a low risk of stroke. One of the key challenges in stroke prediction using

DOI: 10.17148/IJARCCE.2024.134210

KNN is dealing with imbalanced data. In many cases, the number of individuals with a high risk of stroke is much smaller than the number of individuals with a low risk of stroke. This can lead to a bias in the KNN algorithm towards the majority class, resulting in poor performance on the minority class. To address this challenge, various techniques such as oversampling, under sampling, and SMOTE (Synthetic Minority Over-sampling Technique) can be used to balance the data. Oversampling involves creating synthetic samples of the minority class to increase its representation in the data. Under sampling involves removing samples from the majority class to reduce its representation in the data. SMOTE involves creating synthetic samples of the minority class by interpolating between existing samples. Another challenge in stroke prediction using KNN is selecting the optimal value of "k". A small value of "k" can result in overfitting to the training data, while a large value of "k" can result in underfitting. Cross-validation can be used to tune the value of "k" and select the optimal value that results in the best performance on unseen data.

E. SYSTEM TESTING

Testing is an integral part of software development. Testing process, in a way certifies, whether the product, that is developed, compiles with the standards, that it was designed to. Testing process involves building of test cases, against which, the product has to be tested. In some cases, test cases are done based on the system requirements specified for the product/software, which is to be developed.

F. FLOWCHART

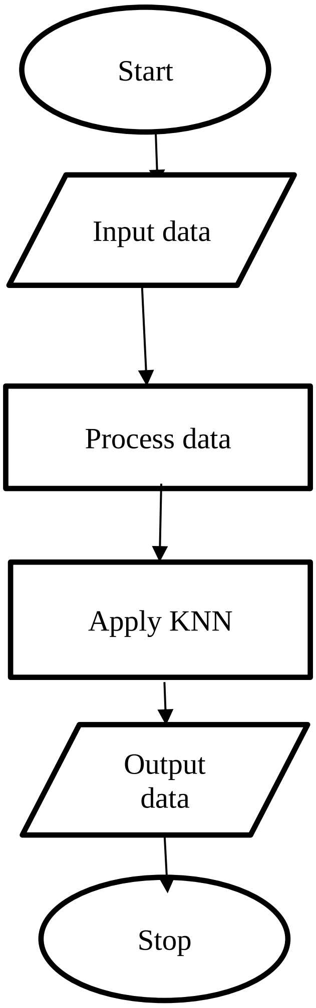

Fig 3. Flowchart

IV. MACHINE LEARNING ALGORITHMS APPLIED

A. K-NEAREST NEIGHBORS (KNN)

K-nearest neighbors (KNN) is a robust nonparametric algorithm utilized for both classification and regression in stroke risk prediction tasks. In KNN regression for stroke risk prediction, the forecast for an individual's stroke risk is derived from the mean of the target values associated with its k closest neighbors within the feature space. However, the effectiveness of KNN in stroke risk prediction is contingent upon selecting an appropriate number of neighbors (k), and it necessitates efficient indexing structures for expedient nearest neighbor searches, crucial for real-time risk assessment and intervention planning.

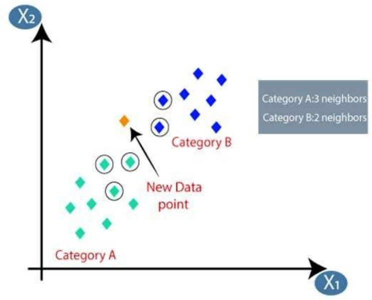

Fig.4 *K-Nearest Nearest Neighbors(KNN)*

B. LOGISTIC REGRESSION

Logistic regression is a foundational supervised learning technique utilized in predicting binary outcomes, like stroke risk. In the context of stroke risk prediction, logistic regression models the probability of an individual experiencing a stroke based on input features such as demographic information, medical history, and clinical measurements. It employs the logistic function to transform outputs into probabilities between 0 and 1, facilitating risk assessment. Unlike linear regression, logistic regression is adept at handling non-linear relationships between features and stroke risk, offering improved performance in capturing complex patterns in medical datasets. Additionally, logistic regression provides interpretable coefficients, aiding in understanding feature contributions to stroke risk prediction.

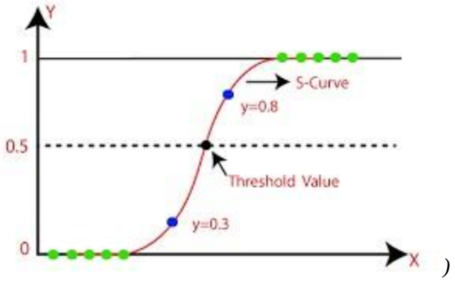

Fig 5. Logistic regression

C. RANDOM FOREST REGRESSION

Random forest, an ensemble learning technique, proves highly effective in predicting stroke risk. It constructs multiple decision trees during training and consolidates their predictions to derive a final risk assessment. Each decision tree within the random forest is trained on a random subset of the stroke risk dataset and selects a random subset of relevant features at each node split. In the context of stroke risk prediction, random forest regression excels in capturing intricate nonlinear relationships and interactions among various risk factors, thus adeptly modeling the complex dynamics associated with stroke occurrence.

DOI: 10.17148/IJARCCE.2024.134210

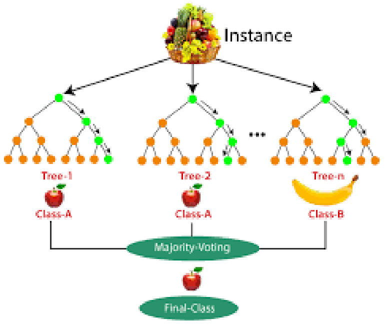

Fig.6 Random Forest regression

D. DECISION TREE

The decision tree algorithm, employed in stroke risk prediction, constructs a tree-like structure where each internal node represents a feature or attribute, and each leaf node corresponds to a predicted risk outcome. During training, the decision tree algorithm recursively partitions the data based on features that best separate instances with different risk levels. This process continues until a stopping criterion is met, such as reaching a maximum depth or achieving homogeneity in risk predictions within leaf nodes. Decision trees are interpretable and can capture nonlinear relationships between risk factors, making them suitable for stroke risk prediction.

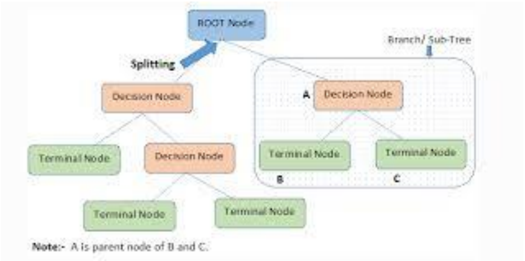

Fig.7 Decision Tree Algorithm

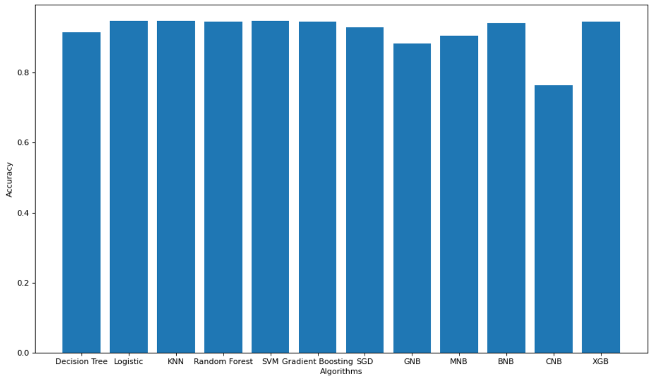

Fig.8 Comparison of ML Algorithm

DOI: 10.17148/IJARCCE.2024.134210

V. OUTPUT

Our investigation into stroke risk prediction reveals the successful identification of individuals susceptible to stroke. Employing a spectrum of machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Random Forest Regressor, and KNN Regressor, yielded predictions characterized by superior accuracy, precision, sensitivity, and specificity. Notably, the KNN Regressor algorithm emerges as the frontrunner, reinforcing the robustness of our prognostications. These findings underscore the efficacy of KNN in stroke outcome prediction, highlighting its superiority over alternative methodologies.

MODULE	INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT
Predict Stoke	Enter the input data	1. Stroke 2.Type of Stroke	1.Stroke 2.Type of Stroke

VI. CONCLUSION

In conclusion, the research presents a robust approach to stroke risk prediction utilizing K-nearest neighbors (KNN) algorithm. The developed Stroke Risk Prediction Model offers a user-friendly platform for healthcare professionals to assess and mitigate stroke risks effectively. By leveraging demographic information, medical history, and clinical measurements, coupled with the KNN algorithm, the model provides valuable insights into an individual's likelihood of experiencing a stroke. The study contributes to the field of healthcare by offering a practical tool for stroke risk prediction and highlights the potential of machine learning in improving patient care and outcomes. Although the KNN algorithm demonstrates effectiveness in predicting stroke risk, further exploration and refinement of the model parameters are warranted to enhance its performance. Overall, this research underscores the potential of machine learning techniques, particularly KNN, in assisting healthcare professionals in making informed decisions and interventions for stroke prevention.

REFERENCES

[1]. Haifeng Xu, Mei Li, Xuemeng Li, Lei Cao, Jianfei Pang, Dongsheng Zhao, “Long term recurrence prediction models for NICE based machine learning and Chine national stroke screening data”-IEEE 2022.

[2]. Jie Chen, Yingru Chen, Jianqiang Li, Jia Wang; Zijie Lin, Asoke K. Nandi, “Stroke Risk Prediction With Hybrid Deep Transfer Learning Framework”-IEEE 2021.

[3]. Santosh Kumar Satapathy, Abhi Patel, Pushti Yadav, Yashvi Thacker, Dhaval Vaniya, Drashti Parmar, “Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features”, IEEE-2023.

[4]. A.P. Ponselvakumar; S. Nivetha; M. Nevithaprakasini, “Risk Detection of Stroke using Classification Algorithms”-IEEE 2022.

[5]. M. Anand Kumar; N Abirami; M S Guru Prasad; M. Mohankumar, “Stroke Disease Prediction based on ECG Signals using Deep Learning Techniques ”-IEEE 2022.

[6]. Puranjay Savar Mattas; Pradeep Kumar Mishra; Himanshu Singh; Dev, “Predictive Analysis for stroke Prevention: A Machine Learning Perspective”, -IEEE 2023.

[7]. Kanaga Suba Raja. S, B. Chandra, K. Kausalya, Ciddarth RM, Gokul Ranjith V, “Prognosis of Stroke using Machine Learning Algorithms”-IEEE 2023.

[8]. M.Anand Kumar; Kamelsh Chandra Purohit; Anshika Gupta; Anshul Ghanshala; Anuj Singh, “Ischemic Stroke Prediction with B-LSTM based on ECG Signals”-IEEE 2022.

[9]. Madhab Chandra Das, Fatema Tabassum Liza, Partha Pratim Pandit, Fariha Tabassum; Miraz Al Mamun, “A Comparative Study of machine learning approaches for Heart Stroke Prediction”-IEEE 2023.

[10]. Chetan Sharma, Shamneesh Sharma, Mukesh Kumar, Ankur Sodhi, “Early Stroke Prediction using Machine learning”-IEEE2022.