ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

Stroke Prediction using Machine Learning

Divya P.[1]; Ajmal Ahamed R.[2]; Duraiarasi S.[3]; Jaisuriya N.[4]; Jayashree A.[5]

[1,2,3,4,5]Department of Electronics and Communication Engineering,
Sri Krishna College of Technology, Coimbatore, India

Publication Date: 2025/03/01

Abstract: In this work, we aimed to predict the incidence of strokes using machine learning approaches. The dataset includes demographic and health-related variables such as age, gender, heart disease, hypertension, and smoking status. After pre-processing the data, which included encoding categorical variables and handling missing values, we trained several classification techniques, including Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier. To evaluate the models' performance, we employed metrics such as F1-score, recall, accuracy, and precision. The Random Forest Classifier achieved the highest accuracy of 94.72% on the test set. In order to solve the issue of class imbalance, we also employed techniques like Random Over Sampler and SMOTE (Synthetic Minority Over-Sampling Technique), which improved the models' capacity to predict the recurrence of strokes. All things considered, our findings suggest that machine learning algorithms, with the Random Forest Classifier showing promising accuracy results, may be able to predict the incidence of strokes based on demographic and health-related data.

Keywords: *Brain Stroke, Cerebrovascular Accident, Oxygen and Nutrients, Ischemic Stroke*.

How to Cite: Divya P.; Ajmal Ahamed R.; Duraiarasi S.; Jaisuriya N.; Jayashree A. (2025). Stroke Prediction using Machine Learning. *International Journal of Innovative Science and Research Technology*, 10(2), 1040-1045. https://doi.org/10.5281/zenodo.14944955.

I. INTRODUCTION

When the blood supply to the brain is suddenly cut off, there is a lack of oxygen and nutrients, which leads to brain stroke, also known as cerebrovascular accident (CVA), a dangerous medical condition. This disturbance may result from a hemorrhagic stroke, which is a rupture of a blood vessel inside the brain, or an ischemic stroke, which is a blockage in the blood arteries supplying the brain. Since the brain is particularly vulnerable to oxygen deprivation and every minute counts to reduce potential harm, early and accurate diagnosis of a brain stroke is crucial for initiating timely medical treatment. Advances in medical technology, especially imaging techniques like computed tomography (CT) scans and magnetic resonance imaging (MRI), have significantly improved the capacity to detect and assess the severity of brain strokes. Since timely medical attention can have a major impact on the prognosis and recovery of those suffering from this potentially fatal illness, it is equally important to understand the warning signs and symptoms of a stroke in order to implement early intervention.

A. Brain Stroke

A brain stroke, officially referred to as a cerebrovascular accident (CVA), is a severe and occasionally lethal event characterized by a sudden stoppage of blood flow to the brain. A obstruction of the blood vessels supplying the brain can result in either an ischemic or hemorrhagic stroke. The brain depends on a constant supply of oxygen and nutrients to function, and its sudden disruption can have detrimental effects. It is crucial to identify and treat brain strokes as soon as possible since the brain is particularly susceptible to oxygen deprivation and the duration of the disruption plays a significant role in determining the extent of damage. Our ability to recognize and understand the intricacies of brain strokes has significantly increased because to medical imaging technology like magnetic resonance imaging (MRI) and computed tomography (CT) scans. To reduce any possible repercussions and hasten their recovery, those affected by this neurological emergency must recognize the symptoms and indicators and seek prompt medical assistance.

B. Cerebrovascular Accident

A cerebrovascular accident, often known as a stroke, is a sudden and often fatal medical event that occurs when there is a disruption in the blood supply to the brain. This disturbance may result from a hemorrhagic stroke, which is a rupture of a blood vessel inside the brain, or an ischemic stroke, which is a blockage of the blood arteries supplying the brain. The brain, which regulates all body functions, requires a constant flow of oxygen and nutrients to sustain its intricate functions, therefore a cerebrovascular accident (CCVA) can have major consequences. A cerebrovascular accident (CCA) can cause a range of impairments, including motor, sensory, and cognitive limitations, depending on the area of the brain that is damaged. Prompt identification and action are crucial to minimize potential injury and increase recovery chances. Given that cerebrovascular accidents are a leading cause of death and permanent disability, it is imperative that the general public and medical professionals understand their

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

nature. This emphasizes how important preventative care and proactive healthcare are.

C. Oxygen and Nutrients

The intricate ballet of physiological processes in the human body depends on oxygen and nutrients, the two fundamental elements of life. Cellular respiration, which provides the energy needed for the body's many functions, is powered by one key gas, oxygen. However, nutrients are a diverse group of substances, including proteins, fats, carbohydrates, vitamins, and minerals, that support the body's growth, maintenance, and repair. Given that both are essential for optimal brain function, the brain is a perfect illustration of the symbiotic relationship between nutrients and oxygen. The brain needs an astounding 20% of the oxygen supply to support its energy-intensive processes, even though it only makes up 2% of the body weight. This delicate balance between oxygen and nutrients is essential for maintaining cellular function, and any disruption to it can have detrimental effects on overall health. This highlights the significance of a way of life and medical practices that priorities adequate oxygenation and nutritional support.

D. Ischemic Stroke

An ischemic stroke is a critical indicator of vascular damage in the intricate environment of the human brain. This medical condition is defined as a blockage of blood flow to a specific region of the brain, often brought on by a blood clot or plaque accumulation in the arteries supplying the organ. The damaged brain tissue begins to degenerate without the oxygen and nourishment it requires, which could have detrimental effects. Ischemic strokes account for the majority of stroke cases, highlighting their significance as a leading cause of mortality and disability globally. Understanding the underlying causes of ischemic strokes, such as hypertension, atherosclerosis, or cardiac problems, is crucial for both focused therapy and preventative measures. Current medical technology, particularly imaging techniques like magnetic resonance imaging (MRI) and computed tomography (CT) scans, has greatly enhanced our ability to detect and treat ischemic strokes promptly. This demonstrates how important early intervention is to minimizing long-term brain damage.

II. LITERATURE REVIEW

According to Yong Yuet al., strokes' devastating consequences on the central nervous system have made them a serious global public health problem in recent years. Ischemic and hemorrhagic strokes, two of the most prevalent types, pose significant risks and challenges. The World Health Organization (WHO) reports that 87% of stroke cases are ischemic strokes, 10% are intracerebral hemorrhages, and 3% are subarachnoid hemorrhages. The need for accurate stroke detection, classification, and prediction has led to an increase in the application of machine learning (ML) approaches in the medical field. Despite the increasing application of machine learning in healthcare, current research has limited capacity to fully predict risk factors associated with different forms of stroke. To close this gap, this work offers a novel Stroke Prediction (SPN) technique. The system uses an improved random forest technique to study and assess stroke risk levels. In particular, the work uses state-of-the-art machine learning techniques to increase the prediction accuracy of the Stroke Predictor (SPR) model. With an impressive prediction accuracy of 96.97%, the proposed SPN method shows significant gains over existing models. This significant development demonstrates how machine learning may revolutionize stroke prediction, enabling improved risk assessment and preventative measures. In addition to contributing to the field of medical informatics, the research has the potential to enhance patient outcomes and reduce the global burden of strokes through early and accurate prediction. Machine learning in healthcare is likely to be crucial as technology advances, improving patient care in general and expanding our knowledge of complex diseases like strokes.

Usman Wazir [2] et al. claim that this paper Early detection and prevention are crucial in addressing the potentially catastrophic consequences of strokes. In light of the gravity of this problem, the current study proposes a comprehensive strategy to stroke prediction by integrating many machine learning techniques. Important factors that significantly affect the incidence of strokes are considered in the study, such as age, body mass index, smoking status, heart disease, hypertension, average blood sugar levels, and previous stroke history. To maximize the predictive power of the chosen features, ten distinct machine learning classifiers were used: Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer Perceptron Classifier, K Neighbors Classifier, Gradient Boosting Classifier, and XGBoost Classifier. Diverse classifiers enable a more comprehensive analysis of the data, revealing a range of patterns and relationships that support stroke prediction. We demonstrated the effectiveness of this ensemble approach in raising prediction accuracy by combining the results of the individual classifiers using a weighted voting process. The study's overall accuracy of 97% was remarkable, and the weighted voting classifier performed better than the individual basic classifiers. The weighted voting classifier is a promising stroke prediction tool that provides patients and physicians with a reliable method of identifying and managing potential stroke risks due to its high accuracy and robust performance. Notably, the study evaluated the model's efficacy using additional metrics, such as the area under the curve (AUC) value.

Priti Narwal [3] et al. have suggested in this study With its intricate structure, including the brainstem, cerebrum, and cerebellum, and its protective skull, the human brain is a vital part of human physiology. Because of the intricacy of the brain, strokes the second-leading cause of death worldwide pose a major risk. In order to prevent or diminish the severity of strokes and, consequently, reduce death rates, early identification is essential. Minimizing brain damage requires prompt and effective therapy. In this sense, it seems feasible to use machine learning algorithms to identify risk variables associated with strokes. The current study presents a methodology designed to accurately predict brain strokes. For this model to be successful, proper data handling is required, including efficient data collection, pre-processing, and

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

transformation processes. The use of a particular "brain stroke dataset" serves as the foundation for both model construction and training. To ensure the veracity of the information the model delivers, we normalize the data using standardization techniques. Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) are the three classifiers that are employed during the training and testing phases. The model's performance is comprehensively evaluated using key metrics such as accuracy, sensitivity (SEN), error rate, false-positive rate (FPR), false-negative rate (FNR), root mean square error, and log loss. Interestingly, the test results show that the Random Forest (RF) classifier performs exceptionally well.

Nitin Kumar Sharma [4] et al. Because stroke is a fatal brain disease that can kill a person three to ten hours after it begins, it emphasizes the critical need for quick action to avert mortality. Smart health systems provide a workable solution to this urgency by enabling prompt identification and response to the kind of stroke. Creating a machine learning model that can forecast a patient's risk of having a stroke is the aim of this work. The Random Forest classifier outperforms cutting-edge models such as Logistic Regression, K-NN, and Decision Tree Classifier (DTC). There were 5110 observations and 12 attributes among the datasets employed in the experimental framework. exploratory data analysis (EDA) for pre-processing to improve the robustness of the model, and feature engineering techniques to balance the datasets. The emphasis on dataset balancing is crucial to avoiding biases and ensuring the model's generalizability in a range of scenarios. The study gradually integrates a cloud-based mobile application into the system. This smartphone software serves as a data collection tool that allows users to enter relevant data for analysis. The model's predictions offer the possibility of stroke detection based on user-provided data. At 96%, the system's precision, recall, and F1-score all show off its exceptional accuracy. Making accurate predictions, minimizing false alarms, and optimizing the system's utility in real-world scenarios all depend on this high level of precision. The suggested approach's ease of use is one of its primary advantages.

Ziad Ismail [5] et al. claim that this paper Predicting a person's risk of stroke has garnered a lot of attention globally as governments and researchers have recognized the potential benefits of early knowledge for prevention and treatment. Governments in particular have started collecting medical data in large quantities so that they might utilize artificial intelligence tools to make accurate projections. Even though black-box methods have shown a high level of prediction accuracy, they usually lack transparency and provide minimal justification for a given projection. For medical professionals, explanations are crucial because they allow them to understand the factors that affect the assessed risk level. This work addresses the need for precise and comprehensible stroke risk prediction. The proposed method outperforms existing methods and offers information on the most plausible causes of a high risk of stroke, according to the Dempster-Shafer theory of plausibility. Furthermore, the method demonstrates robustness in managing incomplete medical records, a common issue in real healthcare datasets. The testing dataset is from the regional hospital in Okayama, Japan, where legal requirements for annual physicals are in place. Real-world data enhances the findings' applicability to actual healthcare settings. This study's experiments compare the results of the Dempster-Shafer approach with those of other well-known machine learning techniques, including Multilayer Perceptron's, Support Vector Machines, and Naive Bayes. Remarkably, the Dempster-Shafer method outperformed the others, proving its ability to deal with missing data. This is a major benefit in hospital settings, where poor recordkeeping are common.

III. EXISTING SYSTEM

A stroke is a condition in which there is insufficient blood flow to the brain, causing the cells to die. It is today one of the main causes of death worldwide. Numerous risk factors believed to be related to the cause of stroke have been identified through examination of the afflicted persons. A number of studies have used these risk factors to predict stroke illnesses. Most of the models use data mining and machine learning technology. Based on medical report data and a person's physical condition, we have used five machine learning algorithms in this work to identify strokes that may occur or have already occurred. We have collected a significant number of entries from the hospitals to address our problem. The outcome is suitable for use in real-time medical reports based on the classification result. We believe that machine learning algorithms can help us better understand diseases and be a helpful healthcare companion.

IV. PROPOSED SYSTEM

The proposed solution makes full stroke predictions using machine learning techniques. The dataset first undergoes a thorough investigation and preparation process to handle missing values, eliminate unnecessary features, and encode categorical variables for machine learning techniques. Then, using feature selection techniques—which may include feature scaling and outlier treatment to further refine the dataset—the raw data is converted into useable features. Following training, the efficacy of a number of classification algorithms—such as Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier—in predicting strokes is evaluated. Hyperparameter tuning with techniques like Randomized Search CV optimizes model performance utilizing assessment metrics including accuracy, precision, recall, and F1-score used to measure efficacy. The dataset's imbalance prompts research into oversampling techniques like SMOTE and Random Over Sampler to reduce class imbalance and increase model accuracy.

A. Load Data

Importing the dataset into the system is the responsibility of this module. The dataset includes a number of health and demographic characteristics that are crucial for predicting strokes. After being loaded from structured sources like databases or CSV files, the data is saved in an appropriate format for additional processing. Investigating the dataset's structure initially entails examining its distribution of characteristics, data types, and missing values.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

B. Data Preprocessing

This module ensures quality and consistency by cleaning and transforming the raw dataset as needed. Improving model efficiency involves removing duplicate or irrelevant features and handling missing values with imputation approaches. Label encoding and one-hot encoding are two suitable methods for encoding categorical information. Additionally, to guarantee consistency in feature scales and avoid bias in model training, data normalization or standardization is used.

C. Feature Extraction

The goal of this module is to improve the predictive power of the model by choosing and altering the most pertinent features. Techniques for feature selection aid in identifying crucial characteristics while removing those that have little bearing on stroke prediction. It is possible to refine feature representation by using techniques like Principal Component Analysis (PCA), mutual information, and correlation analysis. Furthermore, methods like feature scaling and outlier detection improve the quality of the data even further.

D. Training and Testing

This module assesses the model's capacity for generalization by dividing the preprocessed dataset into training and testing sets. The training set is used to train a variety of classification algorithms, such as Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier. In order to maximize model performance, hyperparameter adjustment is required. To determine how well the trained models can predict the occurrence of strokes, the testing phase entails evaluating them on unknown data.

E. Model Evaluation

This module uses a variety of assessment metrics, including accuracy, precision, recall, and F1-score, to evaluate the performance of trained models. Oversampling methods such as SMOTE and Random Over Sampler are used to increase prediction reliability because of the dataset's imbalance. To find the best algorithm for stroke prediction, a comparative study of many models is carried out, guaranteeing a reliable and accurate method.

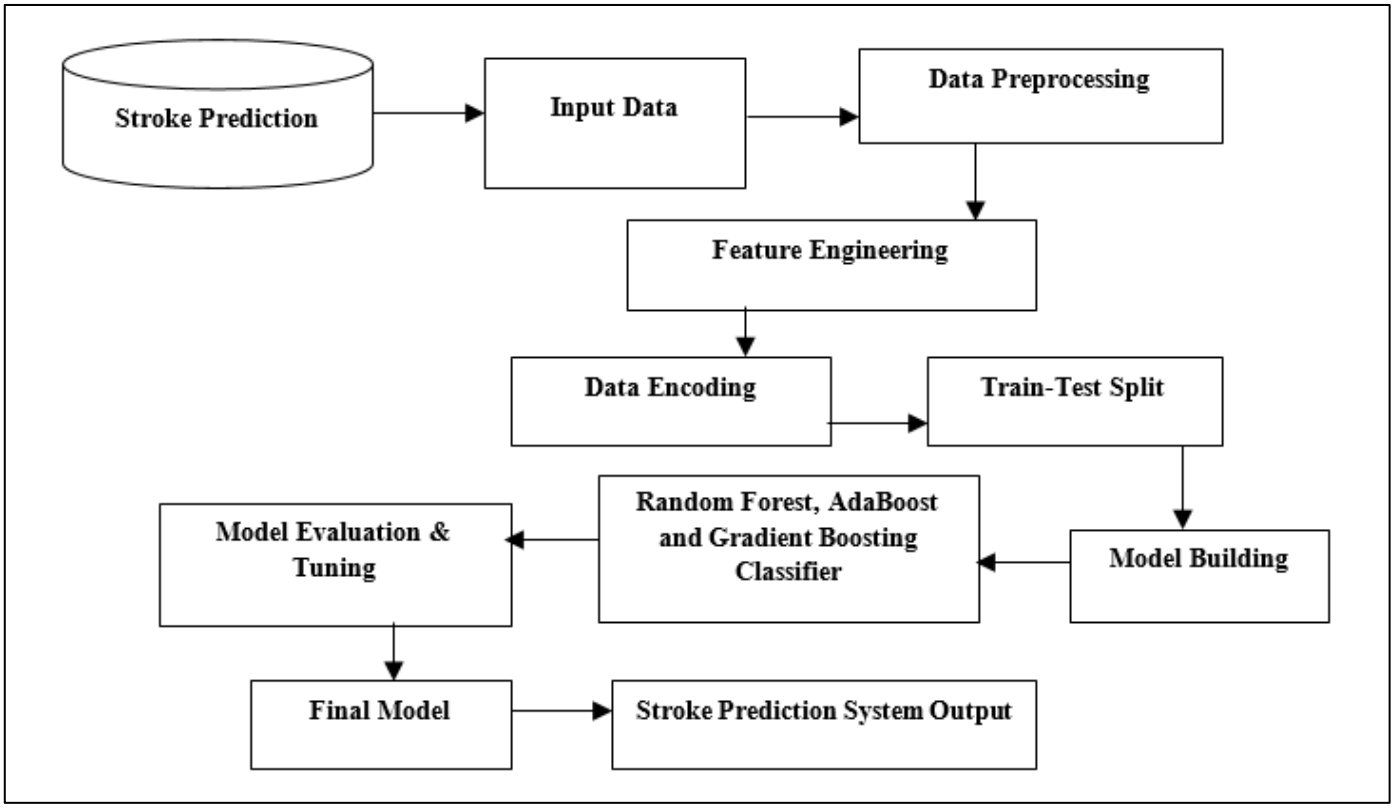

Fig 1: System Flow Diagram

V. RESULT ANALYSIS

The Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier were among the classification algorithms used to assess the efficacy of the suggested stroke prediction system. Accuracy, precision, recall, and F1-score were among the evaluation criteria used to gauge the model's efficacy. The Random Forest Classifier outperformed the other models in terms of predictive power, obtaining the highest classification accuracy for stroke incidents. By reducing the effects of data imbalance, oversampling methods such as SMOTE and Random Over Sampler greatly increased model reliability. By fine-tuning model parameters, hyperparameter tuning further improved performance. The findings show that machine learning models may predict the occurrence of strokes by analyzing demographic and health-related data, and the suggested method performs robustly and effectively in classification.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

➢ *Precision*

Precision measures the accuracy of the positive predictions made by the system. It is the ratio of correctly predicted positive observations to the total predicted positives.

Precision=TP+FP/TP

➢ *Recall*

Recall (also known as Sensitivity or True Positive Rate) measures the ability of the system to identify all the true positive observations. It is the ratio of correctly predicted positive observations to all actual positives.

Recall=TP+FN/TP

➢ *F1-Score*

F1-Score is the harmonic mean of precision and recall. It is used to measure the balance between precision and recall, especially when the class distribution is imbalanced.

F1-Score=2×Precision+Recall/Precision ×Recall

Table 1: Metrics

METRICS	VALUE
Precision	0.95
Recall	0.80
F1-Score	0.92
F-Measure	0.89

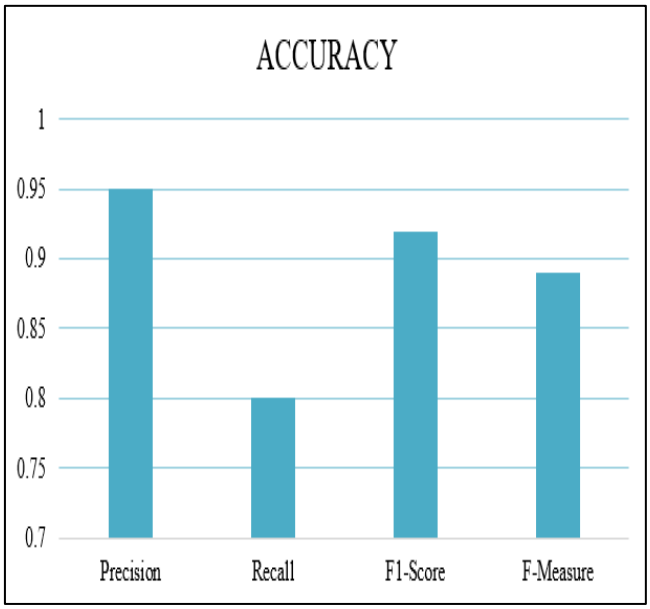

Fig 2: Comparison Table

Table 2: Algorithm

ALGORITHM	ACCURACY
EXISTING	85
PROPOSED	90

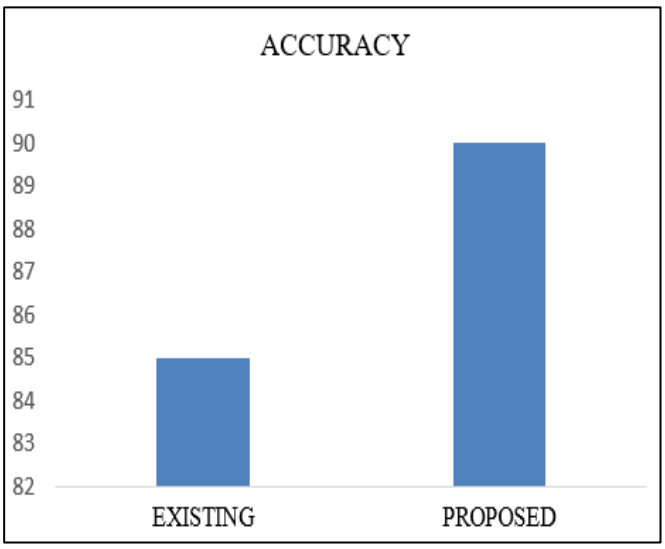

Fig 3: Comparison Graph

VI. CONCUSION

We end by demonstrating the accuracy with which machine learning models predict the incidence of strokes based on demographic and health-related factors. Through careful data analysis, preprocessing, feature selection, model building, and evaluation, we have shown how algorithms such as Random Forest Classifier, AdaBoost Classifier, and Gradient Boosting Classifier can efficiently identify individuals at risk of strokes. Additionally, by using oversampling strategies like SMOTE and Random Over Sampler to solve class imbalance, we have increased the resilience of our predictive models. We intend to enhance and broaden our predictive models for stroke occurrences in future studies by incorporating more features and looking into more complex machine learning techniques. Combining genetic and biomarker data, in particular, may contribute to the development of more personalized risk prediction models and provide more profound understanding of an individual's susceptibility to stroke. Furthermore, we plan to examine the potential impacts of environmental variables and socioeconomic factors on stroke risk in order to have a more thorough understanding of the underlying causes.

REFERENCES

[1]. N. Hatami, L. Mechtouff, D. Rousseau, T.-H. Cho, O. Eker, Y. Berthezene, and C. Frindel, "A Novel Autoencoders-LSTM Model for Stroke Outcome Prediction using Multimodal MRI Data," arXiv preprint arXiv:2303.09484, March 2023.

[2]. L. García-Terriza, J. L. Risco-Martín, G. Reig Roselló, and J. L. Ayala, "Predictive and diagnosis models of stroke from hemodynamic signal monitoring," arXiv preprint arXiv:2306.05289, May 2023.

[3]. M. Bahrami and M. Forouzanfar, ‘‘Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms,’’ IEEE Trans. Instrum. Meas., vol. 71, pp. 1–11, 2022

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14944955

[4]. L. Ismail and H. Materwala, "From Conception to Deployment: Intelligent Stroke Prediction Framework using Machine Learning and Performance Evaluation," arXiv preprint arXiv:2304.00249, April 2023.
[5]. S. H. Lee, C. S. Chan, S. J. Mayo, and P. Remagnino, "An Exploration on the Machine-Learning-Based Stroke Prediction Model," Frontiers in Neurology, vol. 15, p. 1372431, 2024
[6]. F. Ren, W. Liu, and G. Wu, "Using an Interpretable Classifier to Predict Stroke Risk," IEEE Access, vol. 7, pp. 122758–122768, 2019.
[7]. D. Lai, ‘‘Prognosis of sleep bruxism using power spectral density approach applied on EEG signal of both EMG1-EMG2 and ECG1- ECG2 channels,’’ IEEE Access, vol. 7, pp. 82553–82562, 2019,
[8]. T.-Y. Kim and S.-B. Cho, ‘‘Predicting residential energy consumption using CNN-LSTM neural networks,’’ Energy, vol. 182, no. 1, pp. 72–81, Sep. 2019
[9]. F. Rundo, S. Conoci, A. Ortis, and S. Battiato, ‘‘An advanced bio-inspired photoplethysmography (PPG) and ECG pattern recognition system for medical assessment,’’ Sensors, vol. 18, no. 2, pp. 1–22, Jan. 2018.