CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Stroke Prediction by Deploying Predictive Models on Fog Devices

by

Nagina Razzaq

A thesis submitted in partial fulfillment for the degree of Master of Science

in the
Faculty of Computing
Department of Computer Science

2023

This thesis is dedicated to, The Almighty Allah, my strength and power, who led me in the right way for success. To my parents and family who have been my support system and to my supervisor who guided me and showed me the right path.

CERTIFICATE OF APPROVAL

Stroke Prediction by Deploying Predictive Models on Fog Devices

by

Nagina Razzaq

MCS191049

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Mussarat Yasmeen	COMSATS, Islamabad
(b)	Internal Examiner	Dr. Umair Rafique	CUST, Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood
Thesis Supervisor
February, 2023

Dr. Abdul Basit
Head
Dept. of Computer Science
February, 2023

Dr. M. Abdul Qadir
Dean
Faculty of Computing
February, 2023

Author's Declaration

I, **Nagina Razzaq** hereby state that my MS thesis titled "**Stroke Prediction by Deploying Predictive Models on Fog Devices**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

Nagina Razzaq
Registration No: MCS191049

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Stroke Prediction by Deploying Predictive Models on Fog Devices**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

Nagina Razzaq

Registration No: MCS191049

Acknowledgement

First and foremost praises and thanks to the Allah almighty, the Most Omnipotent, and the Most Merciful, for his shower of countless blessings throughout my life. I would like to express my deep and sincere gratitude to my research supervisor, Dr. Nayyer Masood for providing me the opportunity to do this research and guiding me in best possible ways. I am extremely grateful for his continued support and encouragement. It was a great privilege and honour to work and study under his guidance. I offer my sincere appreciation for the learning opportunities provided by SAFE-RH project. Special thanks and love to my parents for their love, prayers, cares and sacrifices for educating me. Also I express my thanks to my brothers, sisters, friends and fellows for their support and valuable prayers. Finally, thanks of gratitude to the Computer Science Department of the Capital University of Science and Technology,Islamabad for equipping me with knowledge.

Nagina Razzaq

Abstract

The rural regions of Pakistan lack healthcare services and many a times people have to travel a long way to get proper health-related treatment. In such situation, health monitoring becomes more critical for the elderly people as they have less resistance and more prone to illness. A common disease among elderly people is stroke which is a condition that affects the arteries leading to and within the brain, causing paralysis or death. Early detection or prediction of stroke can save the patient's life. For the provision of improved medical facilities to the ageing population, specially to those living in rural regions, remote health monitoring can be helpful in saving lives. The newly developed ML technology has been quickly used into several areas of medicine, including stroke. To increase diagnostic and predictive efficiency, thorough data gathering and integration are necessary for the construction of ideal ML systems. Different ML techniques has been used for feature extraction, reduction, and for classification. A concept for remote health monitoring is developed that utilizes data mining techniques to enhance secure IOT data processing for early identification and prediction of diseases. For this purpose, the medical condition of the suspected patient needs to be continuously monitored which is not possible manually. The purpose of this research is remote monitoring and early detection of stroke in elderly people under the SAFE-RH project. In this research we build two predictive models by using different statistical techniques (coefficient of correlation and selectKBest) to identify stroke risk factors, Machine learning (ML) models (logistic regression, support vector machine, decision tree, random forest and XGBClassifier) for classification and synthetic minority oversampling technique (SMOTE) for data balancing. The First model(M1) predicts on the original EHR data set with an accuracy of 94%. The second model (M2) predicts on modified data in which hypertension values(0,1) are converted to systolic and diastolic form(120/80), with an accuracy of 82%. The models are trained by using Electronic Health Record (EHR) dataset of 43401 patients. The MIS server then sends alerts to the caretakers or other stakeholders. This research presents development of the proposed ML model, discussion of obtained results and deployment of model on fog nodes of SAFE-RH.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
EHR	Electronic Health Record
BP	Base Paper
ECG	Electrocardiogram
ICT	Information and Communication Technology
ML	Machine Learning
RHM	Remote Health Monitoring
SMOTE	Synthetic Minority Oversampling Technique
WHMS	Wearable Health Monitoring System
IMU	Inertial Measurment Unit
WD	Wearable Devices
PCA	Principal Component Analysis
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
NN	Neural Network
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
KNN	K-Nearest Neighbour
LSTM	Long Short-term Memory Networks
RBFNN	Radial Basis Functions Neural Network
GUI	Graphical User Interface
BMI	Body Mass Index
MRI	Magneto Resonance Image

CT Scan	Computed Tomography Scan
IoT	Internet of Things
RQs	Research Question(s)

Chapter 1

Introduction

1.1 Background

The rural regions of Pakistan lack healthcare services and many a times people have to travel a long way to get proper health-related treatment. Many places do not have any support to cope with emergency situations. It takes lots of time moving sufferers to city regions for proper treatment which is highly-priced as well. Lack of fitness care centers has extreme results on sufferers. People need routine checking and monitoring, as correct and on time evaluation of any health associated problem is essential for prevention and remedy of the illness [1]. Health monitoring becomes more critical for the elderly people, that is, people aged above fifty. For the provision of improved medical facilities to the ageing population or to those living in places with poor medical facilities and providing the highest level of affordable medical facilities, new techniques and technology must be developed and put into practice. To improve the overall health of the aged people, remote healthcare monitoring (RHM) strives to provide simple, affordable methods [2]. Instead of paying for costly treatments (hospitals or nursing homes), RHM enables patients to be monitored and treated without going to medical centers. Thus, it offers a viable substitute for on-site clinical monitoring that is also less expensive. Remote health monitoring (RHM) comprises of medical devices, actuators and latest ICTs (information and communication technologies); it provides a speedy and low cost environment for the aged people [3]. It is a rapidly developing trend that extends beyond conventional healthcare and provides healthcare support close to the people living in far-flung areas. Numerous people now consider smartwatches and fitness trackers to be normal equipment for monitoring various aspects of their health, including heart rate, blood oxygen levels, irregular heartbeats and more. By constantly monitoring patients' illnesses and making preventative care decisions using information patterns and projections, healthcare practitioners are turning to technology to better serve their patients. For instance, connected inhalers are a smart health monitoring device which helps patients manage consumption and by alerting them when to consume a medication to keep a healthy routine, they save lives.

1.1.1 Smart Technologies

With the help of electronic media, remote care, and electronic health records technologies, patients and doctors are now more connected than ever before, revolutionising how people are being treated. An electrical equipment that can access, share, and communicate with its consumer and other connected devices is known as a "smart device," as the title suggests. Smart devices typically have the computer capability of only a few megabytes, despite being compact in size. Smart devices are interacting with electronic gadgets that help people in daily activities as well as understand the commands given to the people. Mobile phones, tablets, phablets, smartwatches, smart glasses, as well as other technology products are some of the smart devices that are most often used. Although many smart devices are not very large, easy to carry with, and linking and interacting ability to a network as well as transmit or interact remotely is what they are truly meant for.

To aid enhance general health, people can wear a variety of various portable goods or equipment, such as:

1. Biosensors

2. Smart thermometers

3. Connected inhalers

4. Smart watches

5. Fitness trackers (FitBits)

6. ECG monitors

7. Blood pressure monitors

Wearable biosensors, which are utilized by healthcare professionals all over the world to monitor patients and give preventative treatment, have been another significant advancement in medical technology [3]. In order to provide healthcare professionals with more knowledge and insights on the course of diseases, disorders, and general health, these sensors are worn on the body and monitor important health signals including temperature and heart rate. The importance of wearable sensors in remote health monitoring platforms has recently been the main focus of numerous researchers, business owners, and IT corporations.

A large selection of application-specific wearable physiological and activity monitoring devices. In addition, a variety of wearable commercial products, including the biometric shirt, are available right now [3].

Monitoring physiological signals regularly may make it easier to identify and treat a number of cardiac, neurological, and lungs illnesses as they first manifest [1, 4]. Additionally, the examination of gait patterns, posture, and sleep patterns may all benefit from real-time observation of a person's movement patterns. In most cases, wearable health monitoring systems (WHMS) consist of wide ranged electrical and MEMS (micro electro- mechanical systems) sensors, actuators, detached transmission devices, as well as different units for processing the network signals.

1.1.2 The SAFE-RH Project

SAFE-RH https://safe-rh.eu/ is a remote health monitoring system that offers health monitoring facilities in Pakistan's rural areas. By using Sensing, Artificial intelligence and Edge networking, it aims to provide health facilities through Rural Health monitoring. The major objective is to lower the risk of mortality for women and children, promptly handle maternity-related problems in such areas, and keep an eye on the elderly. By utilizing remote health monitoring, healthcare organizations can lessen the difficulties and burdens their patients face, such as the need to travel for specialty care-related transportation problems. Additionally, remote health monitoring can enhance systemic coordination, monitoring, and responsiveness. The overall architecture of SAFE-RH is given below in figure 1.1.

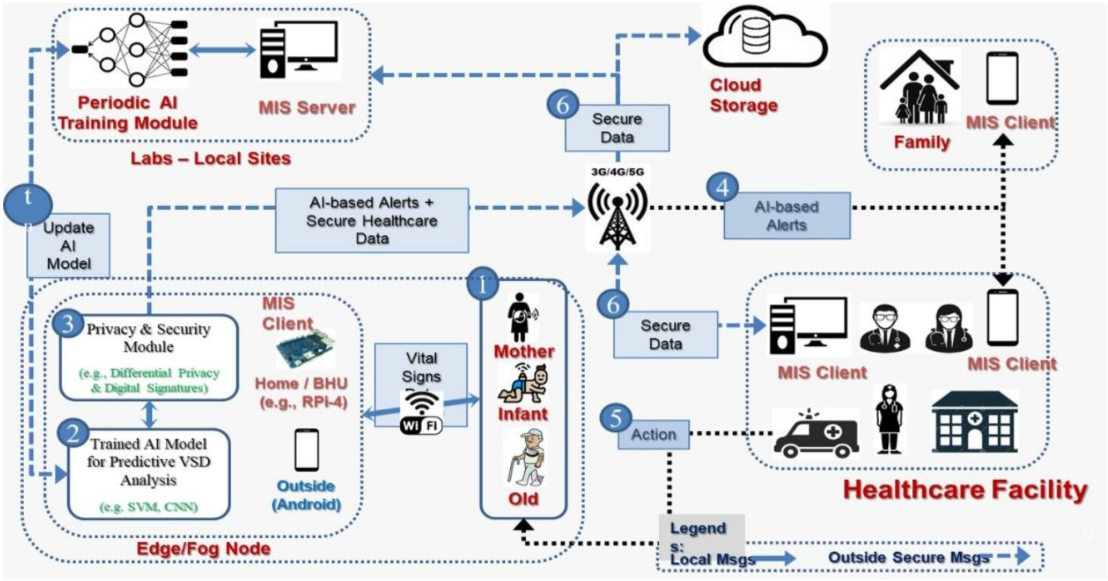


FIGURE 1.1: SAFE– RH Architecture

The aim of SAFE-RH in Pakistan is to provide health care facilities for patient of rural areas such as;

- Access to medical experts
- Minimizing frequent visits and re-admission to hospitals in case of serious issues
- Minimizing cost and time for consultation
- Ease for medical advice
- Reduce death rate due to lack or delay of medical help.

Table 1.1 shows some commonly used wearable sensors.

Table 1.1: Sensors and their Applications

Sr.	Device Name	Sr.	Device Name
1	Contactless Bed Sensors	7	Smart sleeves for Arms, Legs and Hands
2	Pulse Oximeter	8	Smart Watches for Health
3	Oxygen Cylinder	9	Epidermal Patch
4	Smart Eye Glass	10	Smart Rings
5	Smart Walker	11	Smart Thermometer
6	Nebulizer	12	Chest Straps and Smart Belts

1.1.3 SMART Devices Used in Stroke Prediction under SAFE-RH

1.1.3.1 BP Checker

A diagnosing instrument for determining the body's cardiovascular health is a blood pressure machine or digital BP monitor.

The scientifically reliable and precise measurements from the BLT Wireless Blood Pressure Monitor assist you in monitoring your normal heart health. Systolic and diastolic values, the pulse rate, the time, the user's icon, and a battery power indicator are all displayed.

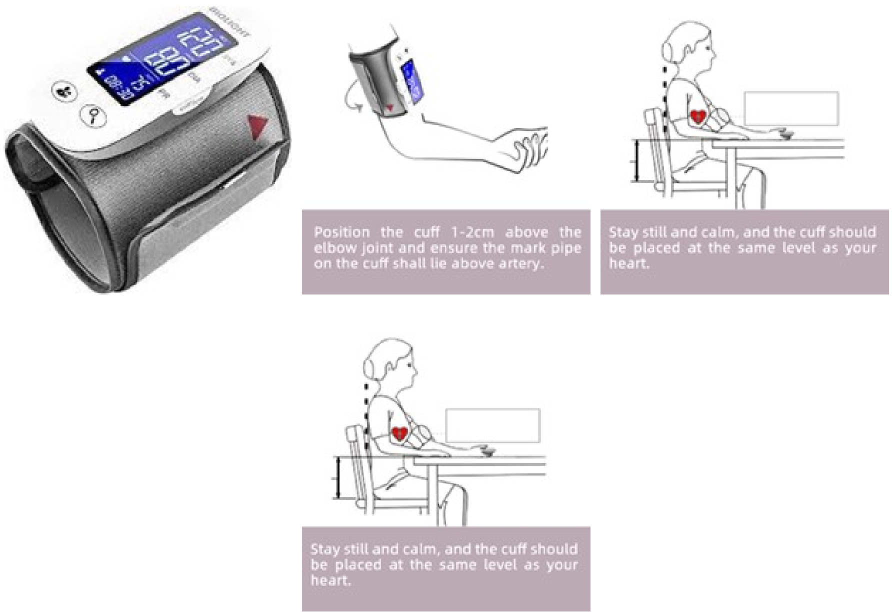

Position the cuff 1-2cm above the elbow joint and ensure the mark pipe on the cuff shall lie above artery.
Stay still and calm, and the cuff should be placed at the same level as your heart.
Stay still and calm, and the cuff should be placed at the same level as your heart.

Figure 1.2: BP Checker https://www.amazon.de/

1.1.4 Blood Glucometer

To determine how much glucose (sugar) is present in the blood, a blood glucose metre is a compact, portable device (also known as the blood glucose level).

In order to cope with their disease, people with diabetes frequently utilize blood glucose meters.

FIGURE 1.3: Blood Glucometer www.vivachek.com

1.1.5 Disease Prediction

The goal of our work is to make a clinical decisions, that is a highly specific and complicated process owing to a number of circumstances, particularly in the case of uncommon diseases or illnesses with overlapping features. Artificial intelligence (AI) in healthcare is a key subject. An artificial intelligence system would use the patient health information to suggest a set of suitable predictions. The system can forecast patients with illness and use medical characteristics like age, hypertension, blood glucose, etc. to predict the likelihood of patients having a disease. It can forecast a patient's chance of contracting a disorder. For the purpose of predicting illness, classification algorithms are utilized with a variety of features. The AI system was able to provide complicated answers, each with its own strengths in model interpretation, access to specific data, and accuracy. Critical or suspected patients may be identified. The treatment quality can then be improved and unexpected hospitalizations may not be needed as the doctors trying to control the risk. Latest developments in data analytic methodologies and tools have made it possible to use a wealth of semantic data for epidemic risk prediction, including demographic trends, diagnostic testing and assessment, health-related behaviors, diagnostic tests, medications, and service utilization.

1.1.5.1 Stroke Prediction

The stroke is a condition that affects the arteries leading to and within the mind. It is a primary cause of disability and death worldwide [5]. It is one of the most life-threatening diseases for elderly people especially in rural areas having inadequate medical facilities. It additionally has a poor effect on clinical offerings and the provision of beds. A stroke occurs when a blood artery that supplies the brain with oxygen and nutrients becomes blocked by a clot or breaks (or ruptures. It damages the brain similarly to a "heart attack," which damages the heart and is the third greatest cause of death in both developed and underdeveloped countries Whenever a stroke illness develops, it can potentially cause death as well as expensive medical care and permanent disability. Every 4 minutes, a stroke victim passes away, although up to 80% of strokes may be avoided if we could detect or anticipate a stroke in its early stages. Beyond the age of 45, the risk of stroke doubles, and after the age of 65, 70% of all strokes occur [6]. Through exercise and rising awareness of risk factors, stroke can be prevented before it occurs (obesity, diet, alcohol intake, and shortage of bodily activity) [06]. Underlying conditions, consisting of diabetes, hypertension, and cardiovascular illnesses, may cause stroke. Therefore, right self-control of those illnesses and the pursuit of a healthy life-style can also additionally save the prevalence of stroke. Patient records contain many beneficial prophetical factors, like patient demographic (e.g, age and gender), way of living for example diet and physical activity, and existing medical condition factors for example polygenic disorder and hypertension, which may cause stroke [7]. By systematically archiving the patients' scientific records, healthcare vendors create possibilities for higher affected person control with the growing utility of generation for scientific prognosis. With the aid of support vector machines , random forests , classification trees, and discriminant analysis, machine learning technology uses a variety of techniques for automatic data evaluation, including linear and logistic regression models. These techniques also allow for the non-linear combination of features and flexible selection boundaries [8]. The area of ML has massively developed with the improvement of numerous automated algorithms for sample popularity and statistical assimilation to enhance predictions, decisions, perceptions, and movements throughout numerous fields and serves as an extension to the conventional statistical approaches [9]. With the appearance of annotated dataset of scientific records, we are able to now use statistical mining strategies to perceive developments within the dataset. Such evaluation has helped the scientific practitioners to make a correct analysis of any scientific situations. It has caused an improved healthcare situation and decreased remedy costs. This assists the scientific practitioners to perceive the onset of illness at an advanced stage. Leveraging the capability of Machine Learning strategies to expect stroke and discover the hazard elements related to its incidence from numerous attributes of sufferers may be of exceptional assistance for doctors and healthcare professionals. The literature review reveals that existing models are being trained and tested on static data. These models have not been used in a real-time environment. None of the approaches have discussed the use of sensors' data.

1.2 Research Objective

This research aims to explore and leverage the potential of Machine Learning algorithms for the early diagnosis of stroke in the elderly using attributes of patients (gender, age, heart disease, hypertension, glucose level, marital status, smoking status, BMI, work type and residence type). The major contribution is getting some of these factors in real time from sensors and monitoring the health of the target person continuously. Number of output classes in stroke classification varies from 2 to 3 as discussed in the literature review section below.

1.3 Research Gaps

After detailed literature review following research gaps were identified:

- The base paper [10] used decision tree, random forest and neural network models for stroke prediction. Some of the models like XGBoost are missing
- Existing models are being trained and tested on static data.
- These models have not been used in real-time environment
- The baseline approach [10] uses sub-sampling approach for class-balancing that creates a small data set resulting in valuable data loss for the training of model
- None of the approaches have discussed use of sensors' data
- Unavailability of patients' data particularly in Pakistan context for model training

1.4 Problem Statement

Lack of early diagnosis of stroke is a leading cause of death and disability, which affects stroke patients and healthcare professionals. Machine learning (ML) has emerged as a pivotal field of study used in almost all walks of life including medical. The ML technique(s) can be used for early diagnosis of stroke that can help in saving the person from disability or death. This work aims to monitor a person continuously using sensors data to be used in an ML model for the prediction of stroke

1.5 Research Questions

This thesis have formulated the following research questions relying on the problem statement described above:

- RQ1: How can we enhance the performance of existing stroke prediction models
- RQ2: How can we use data obtained from sensor to continuously monitor the suspected stroke patient?

1.6 Scope

The scope of this research is stroke prediction in elderly people in the SAFE-RH Project by deploying predictive models on fog nodes. It involves:

- Building a machine learning model by using Electronic Health Record (EHR) data set available from Kaggle.
- Implementation of model on fog devices
- Use the model to continuously monitor the patients in the SAFE-RH project, where the patients will be monitored by using sensors.
- Improve accuracy for stroke prediction while keeping in view the optimal performance.

1.7 Significance of the Research

Early diagnosis of stroke in elderly using attributes of patients (gender , age, heart disease, hypertension, glucose level, marital status, smoking status, BMI, work type and residence type). The major contribution is getting some of these factors in real time from sensors and monitoring the health of target person continuously. Contribution in this research work is:-

- A comprehensive literature review in the area of stroke prediction

- An analytic approach for the prediction of stroke by using the techniques of machine learning in which analysis of stroke risk factors is performed by using correlation and SelectKBest feature selection techniques.
- SMOTE (Synthetic Minority Oversampling Technique) is applied for class imbalance. Machine learning model is developed for stroke prediction.
- The model is then deployed on fog devices where monitoring of patients is continuous. Data used is online and real time (live coming from sensors).

1.8 Thesis Organization

There are five distinct chapters that make up the entire booklet. The background information for the field, a brief introduction to the problem, the justification for the study, a problem statement, research questions, an objective, and motivation for the subject are all presented in chapter one. Chapter 2 contains the literature review. Following a review of the literature, we evaluated the methods previously employed to predict stroke. We also explain the identified research gap before outlining our proposed solution, a brief methodology, and a conclusion. In chapter three, we go into further detail about the deployment architecture and the suggested research strategy. Diagrams of the deployment architecture are also included. Experiments with different features and their results are presented in chapter four. The fifth chapter discusses the conclusion and future work.

Chapter 2

Literature Review

Minali et al. [9] The newly developed ML technology has been quickly used into several areas of medicine, including stroke.Deep learning has greatly improved the usability of machine learning techniques, and some more recent algorithms are reported to be as accurate as humans.A disease's diagnosis detection and treatment, including those of a stroke, are extremely complex and rely on a variety of clinical and individual criteria.To increase diagnostic and predictive efficiency, thorough data gathering and integration are necessary for the construction of ideal ML systems.Although ML algorithms have enhanced stroke treatment systems, uncritical reliance on such computerized technologies might result in incorrect prognostic predictions or misinterpretation.The optimum application of machine learning (ML) technologies is as "assists" for medical decision making, however they still need human monitoring to address pertinent clinical issues that the algorithm misses.

The model proposed by Dev et al. [10] is the baseline approach. For accurate stroke prediction, they analyzed electronic health records and concluded that the most crucial variables for identifying stroke in patients are "age, heart disease, average blood glucose level, and hypertension". The presented neural network classifier showed 78% accuracy. It provided a thorough examination of patient characteristics found in the electronic medical record for stroke prediction.They examined many aspects in a methodical manner. For the purpose of selecting an ideal collection of characteristics, they carried out feature correlation analysis and a step-by-step analysis. They concluded that there is little correlation between the many characteristics, and that a combination of just 4 features would be useful for stroke prediction. They also carried a principal component analysis. The research revealed that a larger variance requires practically all principle components to be present but the starting component with maximum variance can contribute to stroke prediction. The data was not balance, so they applied random-down sampling of data to handle class imbalance. Instead of recording and storing all the features, the data

M. S. Singh and P. Choudhary [6] have proposed a stroke prediction model by using Cardiovascular Health Study (CHS) dataset This has 5,888 samples, 3,228 of which are men and 2,660 of who are female. Additionally, it has over 600 characteristics. The dataset has 357 characteristics and 1824 examples with 212 stroke events after pre-processing.The feature selection process is carried out using the decision tree method, and the dimension reduction was done using the principal component analysis technique. In order to construct the classification algorithm, back propagation neural networks were utilized. The experimental results showed that the proposed method(*DT, PCA and NN*) has higher performance i.e., 97.7% accuracy than rest of models(*NN,PCA and NN*).Comparison showed that Neural Network provides more accurate categorization. management team can archive only those features that are essential for stroke prediction.

Emon et al. [11] used several machine learning for stroke prediction at early stage. The variables ("hypertension, BMI, heart disease, average glucose level, smoking status, previous stroke, and age") were used to suggest a weighted voting classifier for predicting stroke in this paper.Relying the predicted classes with the most votes, majority voting selects the most number of correctly predicted values. In each classification between different classes of results, voting weights should be different. The weight should be high for that specific class where the classification is performing well. In comparison to Logistics Regression (LR), Stochastic Gradient Descent (SGD), Decision Tree Classifiier (DTC), AdaBoost, Gaussian, Quadratic

Discriminant Analysis (QDC), Multi Layer Perceptron (MLP), KNeighbors, Gradient Boosting Classififier (GBC) and XGBoost (XGB) weighted voting provided the highest accuracy of around 97 percent, according to the performance evaluation. Deep learning-based imaging, such as brain CT scans and MRIs, may be offered in the future in conjunction with an existing model to improve performance metrics. The dataset is undefined, and implementation details are missing in the proposed system.

In a Chinese elderly population, Yafei Wu and Ya Fang [12] created a machine learning model for predicting stroke with unbalanced data. They used data balancing techniques such as random over-sampling (ROS), random under-sampling (RUS), and synthetic minority over-sampling method (SMOTE). To predict stroke using demographic, lifestyle, and clinical characteristics, researchers used regularised logistic regression (RLR), support vector machine (SVM), and random forest (RF). Based on the SMOTE-balanced data set, the top five factors for stroke prediction were chosen for each machine learning approach. With an area under curve of roughly 70%, the results were moderate in predicting stroke using patients' demographic, life style, and clinical characteristics.Self-reported stroke, limited data resulting to uncertainty in conclusions, discussion of only the three most commonly used methods (ROS, RUS, and SMOTE), and internal method validation are some of the limitations. Future studies will require external validation in big populations.

The model proposed by Azariadi et al. [13] uses Discrete Wavelet Transform (DWT) for feature extraction and reduction, and Support Vector Machine (SVM) for classification. The system has very low computational cost and it has shown accuracy of 98.9%. But it needs to be tested with large size of data.

Isina et al. [14] have suggested a system which uses Deep Convolutional Neural Network (named AlexNet) for feature extraction and Simple Back Propagation Neural Network for classification. The model was tested on three different datasets with correct recognition rate of 98.51%. Its drawback is that it classifies heartbeat into 3 classes only i.e. Normal, Paced and RBBB.

The system proposed by Xiong et al. [15] uses Novel 21-layer residual convolutional recurrent neural network named RythemNet Classifies into 4 classes i.e. Normal, Atrial Fibrillation, Noisy & Other. The main strength of this system is that it does not need any feature extraction, rather it can process raw ECG waveforms directly. Thus, it takes lesser time on training. But its F1 Accuracy is quite low i.e. 82% due to over-fitting on the cross-validation set.

Hammad et al. [16] suggested a model which used Modified Pan-Tompkins Algorithm in feature extraction. The system has utilized the combination of these classifiers: Feed Forward Neural Network, Multilayer Perceptron, SVM & kNN. The model classified the output into 2 types: Normal & Abnormal. Although the author has shown high accuracy results, but the model has been trained only on 48 records.

Yildrim et al. [17] have suggested a system based on 1D- Convolutional Neural Network (1D-CNN) which includes 7 layers of CNN among total 16 layers. The system is trained on long duration ECG segments i.e. 10 seconds each and it can detect 17 arrhythmia classes with an accuracy of 91.3%.

The approach suggested by Alfaras et al. [18] uses Echo State Networks (ESNs) and Reservoir Computing (RC). Its major is that strengths is that it exploits the Parallelism of GPU, resultantly reducing the cost of training. The system can work on single ECG lead.

Devi et al. [19] have suggested a model which uses statistical and dynamic features of ECG signals to detect Arrhythmia. Pan Tompkins QRS detection algorithm is used for dynamic features i.e. Heart Rate Variability (HRV). Support Vector Machine is used as a classifier. It has shown effectiveness of 92.2% when both features are considered. This system's effectiveness needs to be verified by training and testing on larger size of data.

The model proposed by Mousavi et al. [20] uses CNN to extract meaningful features from ECG signals. It utilizes Bidirectional RNN (BiRNN) which can process data in both forward and backward directions. Strength of Bidirectional RNN is that its current state has access to both previous and future input information simultaneously. It deals with Inter and Intra Patient Paradigms. The model is trained on imbalanced datasets which is its major advantage apart from BiRNN.

Tuncer et al. [21] have proposed a model in which feature extraction and reduction is done by using Discrete Wavelet Transform (DWT) and Neighbourhood Component Analysis (NCA). It classifies arrhythmia into 17 types with an accuracy of 95%.

The model proposed by Ghosh et al. [22] uses Multirate Cosine Filter Bank for ECG signal decomposition. It examines coefficients from ECG signals at various sub bands. Then Fractional Norm (FN) features are evaluated from extracted coefficients. Classification is performed by Hierarchical - Extended Machine Learning (H-ELM) in which hidden layer activation or feature vectors are evaluated with the unsupervised ELM-auto encoder (ELM-AE) learning. It has shown 99.4% accuracy and 98.77% sensitivity in detection of Atrial Fibrillation from single lead ECG signal.

Plawiak et al. [23] have proposed a novel three-layer (48+4+1) approach for detecting arrhythmia from the ECG signals. The approach is named as Deep Genetic Ensemble of Classifiers (DGEC). It combines advantages of Ensemble Learning (EL), Deep Learning (DL) and Evolutionary Computation (EC). For model training, 744 ECG Segments (each of 10 sec duration) of 29 different patients were used. The system can detect 17 Arrhythmia classes from Frequency components of the Power Spectral Density (PSD) of ECG signals. This system has shown accuracy of 94.62% for detection of 17 classes. The classifiers used are SVM, KNN, PNN and RBFNN.

Yao et al. [24] have proposed a model which divides ECG processing pipeline into two phases: Spatial information fusion based on CNN and Temporal information fusion based on LSTM and Attention mechanism and the technique is known as Attention-based Time-Incremental CNN (ATI-CNN). Authors have claimed that their model has reduced memory usage to half and decreased computational cost by around 90%. They have also shown 26.8% increase in accuracy for detecting

Paroxysmal Arrhythmia as compared to traditional other CNN based approaches. It has enhanced interpretability and enabled localization of irregular signal segments. It can detect 9 arrhythmia classes with an accuracy of 81.2%. Although, this model can detect 9 Arrhythmia classes, but the accuracy of model is quite low. The reason of lower accuracy is that the model was more focused on improving the performance (i.e. getting the results in lesser time), instead of accuracy. The proposed solution will be able to achieve higher accuracy for classification of Arrhythmia into 9 classes: (1) Normal (N); (2) Atrial fibrillation (AF); (3) First-degree atrioventricular block (I-AVB); (4) Left bundle branch block (LBBB); (5) Right bundle branch block (RBBB); (6) Premature atrial contraction (PAC); (7) Premature ventricular contraction (PVC); (8) ST-segment depression (STD); (9) ST-segment elevation (STE) [2]. This solution would significantly help the patients in self-monitoring their heart rate and predicting the Sudden Cardiac Arrest based on heart-rate abnormality.

Chiu et al. [25] have suggested a Inertial Measurement Unit (IMU) sensor-based fall detection system for elderly people using an accelerometer and a gyroscope, When an accidental fall is detected, the system transmits alarm messages to the data server via the wireless network. The experiments involve 120 falling and 450 non-falling actions of five participants. The results showed a fall detection accuracy of 95.44% and has the capacity to assist in the everyday healthcare of elderly people.

Senthil K et al. [26] have successfully constructed a forecast model using Python, Jupyter Notebook and Django to distinguish coronary illness dependent on the different calculations that are accessible for AI. The model was trained with the help of more than 40,000 data points from real word users, which resulted in successfully predicting the occurrence of Stroke for the user and also graphically representing the same for more detailed representation and understanding.

The model proposed by Li et al. [27] presents a method to integrate and package an accelerometer within a textile to create an electronic textile (e-textile). applied to measure limb movement angle. The angles measured by the e-textile movement sensor have been verified using a goniometer and a commercial 2-axis bending sensor. The approach shows significant promise in both the fabrication methodology, and the sensing results, providing a viable solution for movement monitoring which is unobtrusive and comfortable for the wearer, allowing a wide range of activity types to be identified.

In the study of Chun et al. [28] Cox models, machine learning, and ensemble models combining both methods were compared for the prediction of stroke in Chinese people. The ensemble technique demonstrated improved true negative rate (men: 76 percent, women: 81 percent), accuracy (men: 76 percent, women: 80 percent), sensitivity (men: 76 percent, women: 81 percent) (men: 26 percent , women: 24 percent).

Nigusse et al. [29] developed silver printed textile electrodes from knitted cotton and polyester fabric for ECG monitoring. Signals acquired at 15 mmHg pressure level with the textile electrodes provided a similar quality to those acquired using standard electrodes. The signal quality was low when the measurement was done in a moving body. When compared to the conventional Ag/AgCl gel electrodes, the ECG signals obtained with the dry cloth electrodes displayed unambiguous R-peaks without any spurious or missing peaks. The outcomes showed that when the retaining pressure rose, the quality of the ECG signal recorded using dry cloth electrodes improved. Signals obtained at a 15 mmHg pressure level are of comparable quality to those obtained with conventional Ag/AgCl gel electrodes. After ten washes, the silver-printed cotton electrodes produced usable waveforms despite a modest improvement in the surface resistivity. In addition to being environmentally stable, textile electrodes performed identically in terms of ECG detection when a measurement was taken after keeping the electrodes in an ambient air for six months. When the assessment was performed inside of a moving body, the signal quality was poor. By raising the holding power of the textile electrodes applied by employing elastic straps, the signal strength in motion may be increased. The created electrodes would work with portable ECG recorders to capture ECG and calculate heart rates in still, non-dynamic situations like persons lying down or sleeping.

Yuan et al. [30] proposed a fall detection algorithm and an classification algorithm for a wrist-worn wearable device(WD). Both algorithms are are based on an accelerometer which supports various interrupts and data buffering(FIFO). By processing accelerometer data completely locally, a WD does not have to stream massive sensor data out wirelessly thus saving both power and bandwidth. They can be successfully implemented in rechargeable batteries MCUs with constrained clock speeds and RAM since they are interrupt-driven. Because they are based on a contemporary digital MEMS accelerometer that supports a variety of interruptions and FIFO, both methods are hardware-reliant. They have the benefit of using less power than traditional algorithms, which must analyse and interpret each instance of accelerometer data.

Samira et al. [31] developed a concept for remote health monitoring that utilizes data mining techniques to enhance secure IoT data processing for early identification of combos of hypercholesterolemia, high blood pressure, and heart disease. Regarding the resource constraints in the IoT environment, an efficient light-weight block encryption scheme based on developing light-weight S-Boxes had also been introduced. Privacy and safety concerns are significantly important when transferring patients' crucial health information through IoT networks and recording them in cloud computing storages. According to experimental data, the K-star classification approach, which has 95% accuracy, 94.5% precision, 93.5% recall, and 93.99% f-score, produces the highest results for 10-fold cross-validation amongst RF MLP, SVM, and J48 classifiers.The results also demonstrated that, according to the evaluation variables involving bijection, rigorous avalanche condition, non - linearity, and algebraic degree, our suggested approach for creating dynamic S Boxes may be classified as a resilient crypt analysis. The proposed system complies with an effective development for remote medical monitoring to identify any potentially dangerous conditions in patients while protecting the privacy and safety of their private medical information, according to the experimentally obtained results.

Shaikh and Rao [32] in their study, described and contrasts the outcomes of several deep learning and machine learning techniques used to cancer prognosis.In particular, various patterns regarding the effectiveness of cancer prediction or outcome approaches have been observed, as well as trends regarding the same types of machine techniques to be applied and types of malignancies being examined. While ANNs are widely utilized, it is evident that at least three distinct cancer types are predicted using a wider variety of other learning methodologies.They enhanced numerous device classification systems' biological verification and field experiment, as well as the overall quality, replicability, and reproductive success of many systems. They come to the conclusion that if the standard of research keeps rising, deep learning classifiers and education gadgets will likely be used pretty often in several clinical and healthcare sectors.

Das et al. [33] created machine learning algorithms that, based on the description of the medicine and the drug review, may be used to predict the name of the disease.Users who lack the specialist guidance of medical professionals will find this to be of particular use. Additionally, users will be able to accurately guess the identity of a sickness by providing names of medications and specific reviews of them as inputs.The user will then be able to accurately buy medications online after the observed ailment and the projected disease have been successfully matched. After experimenting with different machine learning methods, we found that the model including the Random Forest technique produced the most accurate predictions. The standard machine learning techniques, in contrary, outperformed the deep learning models in terms of accuracy, but not otherwise since DL techniques frequently rely on enormous volumes of data to identify patterns.

Kumar and Pathal [34] proposed that with the right predictive modeling algorithm deployment, the disease might be predicted from the symptoms provided by the patients. In their study, they employed four machine learning techniques for prediction and attained an average accuracy of more than 95%, demonstrating a notable improvement and high accuracy over earlier work. This makes the system more dependable than the current one for this work and, as a result, offers the user greater satisfaction than the previous one. The system records the information supplied by the user and the name of the ailment the patient has in a database that can serve as a historical record and will be used for future treatments, making it simpler to monitor the patient's health. . They developed a user-friendly graphical user interface (GUI) to improve user engagement with the system. Their study demonstrates how an algorithm using machine learning can be utilized to predict illness with various models and features.

Keniya et al. [35] demonstrated a method for diagnosing a patient's condition based on their symptoms, age, and gender. Using the aforementioned criteria, the Weighted KNN model had the greatest accuracy of 93.5% in predicting diseases. Nearly every ML model made excellent accuracy results. After sickness prediction, the medical resources needed for the therapy could be managed simply.

Using EEG raw data, power values, and relative values, Choi et al. [36] suggested a method that facilitates the early identification and prediction of stroke illness via deep learning. . They concluded that raw EEG data alone, without the laborious procedure of extracting frequency domain features, may reliably predict the early diagnosis and onset of stroke illness. The approach suggested may offer valuable analytical data to medical professionals, stroke patients with a high risk of recurrence, or older persons with a high stroke incidence. The fact that stroke may be accurately predicted at a minimal cost during routine activities like walking conditions is a remarkable discovery. Before a person is admitted to the emergency hospital, the system can identify the danger of stroke early, giving access to therapy within the "golden time".

Kaur, M, et al. [37] developed a model to predict and diagnose stroke disease early. The prediction model can swiftly generate predictions just on online data for the early identification of strokes after gaining knowledge from offline data. Their methodology is based on a noninvasive technology that uses EEG signals as actual data to train a model and then predict whether or not a person will likely suffer a stroke soon. The proposed approach tests the flexibility of deep neural networks for stroke prediction using four different deep learning methods. It is discovered that all four deep learning methods can successfully detect strokes from biosignals, but also that GRU and biLSTM perform much better than LSTM and FFNN. In comparison to LSTM and FFNN, GRU and biLSTM provide more accuracy and have lower prediction error rates.

TABLE 2.1: Literature analysis summary

Ref	Title	Year	Dataset	Approaches, Representation Technique	Result	Analysis of Literature
[13]	ECG Signal Analysis and Arrhythmia Detection on IoT wearable medical devices	2020	N/A	Feature Extraction: Discrete Wavelet Transform (DWT) Classifier: SVM	98.9%	Trained & tested on smaller size of data
[14]	Cardiac arrhythmia detection using deep learning	2019	N/A	Feature Extraction: Deep CNN (AlexNet) Classifier: Simple Back Propagation Neural Network	92%	Only 3 classes: Normal, Paced & RBBB
[15]	ECG Signal Classification for the Detection of Cardiac Arrhythmias Using a Convolutional Recurrent Neural Network	2018	N/A	Novel 21-layer residual convolutional RNN named RythemNet 4 Classes: Normal, Atrial Fibrillation, Noisy & Other	82%	F1 Score: 82% Over-fitting
[16]	Detection of Abnormal Heart Conditions Based on Characteristics of ECG Signals	2018	N/A	Feature Extraction: Modified Pan-Tompkins Algorithm, Classifiers: FFNN, MLP, SVM, kNN, 2 Classes: Normal, Abnormal	N/A	Trained on 48 records only Sensitive to ECG signal quality

[26]	Stroke Predictions using Healthcare Dataset	2020	N/A	Python, Jupyter Notebook and Django	N/A	Successful prediction of stroke with more than 40,000 records from real world. Graphical representation of predictions
[19]	Machine learning and IoT based cardiac arrhythmia diagnosis using statistical and dynamic features of ECG	2019	N/A	Dynamic Feature Extraction: Pan Tompkins QRS Detection Algorithm Classifier: SVM	92.2%	Compared with only 3 other models
[21]	Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals	2019	N/A	Feature Extraction: DWT Feature Reduction: Neighborhood Component Analysis (NCA) Classifier: 1 Nearest Neighbour (1NN)	95%	Lower number of features involved
[6]	Stroke Prediction using Artificial Intelligence	2017	CHS	DT, PCA and NN	97.7%	Only 1824 records are used for experiments
[10]	A predictive analytics approach for stroke prediction using machine learning and neural networks	2022	EHR	Neural network (NN), Decision tree (DT) and Random forest (RF)	78%	Predictions with only four features. Sub-sampling technique for data balancing
[9]	Machine Learning in Action: Stroke Diagnosis and Outcome Prediction.	2021	N/A	Review of different ML studies	N/A	Stroke diagnosis and prediction using ML techniques

[38]	Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models	2021	National Health & Nutrition Examination Survey	Naive Bayes BayesNet Decision tree Random forest	96%	Easy to use stroke prediction model, build from labtests with high accuracy.
[28]	Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults	2021	N/A	Gradient boosted trees (GBT), Cox regression, Logistic regression, MLP	men: 76% women: 80%	Ensembling approach for identifying individuals at high risk of stroke
[11]	Performance Analysis of Machine Learning Approaches in Stroke Prediction	2020	N/A	LR, SGD, DTC, AdaBoost, Gaussian, Quadratic Discriminant Analysis QDC, MLP, KNeighbors, GBC, XGBoost, weighted voting	97%	Dataset and implementation details are missing

Summary: The literature study reveals that RHM is a hot research topic as it is need of current era. The monitoring activities span from simple messages, audio/video chats, prescribing medicine, viewing reports, monitoring through smart sensor devices and doing predictions of any critical situation through continuous monitoring. Stroke is one of the critical diseases, specially for the elderly people. Stroke prediction approaches have been mainly focusing developing ML models using static data. We have not found any approach for the real-time (or near real-time) monitoring of persons for the prediction of stroke. Such an approach will be highly valuable as an early detection/prediction of stroke can save person from death or disability. So it is highly required to work for such an approach.

Chapter 3

Proposed Research Methodology

We identified some research gaps in previous research during the critical analysis of the literature review, presented in the chapter 2. The work presented in this thesis addresses some of those gaps. The methodology of the proposed work has been described in this chapter in detail. For remote health monitoring patients need to be monitored continuously. In our solution, sensors are attached to the elderly people which are sending vitals of the patients continuously to the fog node. From fog node, this data goes to the MIS server of SAFE-RH where it is stored in order to maintain the historical record of person's health data. In our system, we have deployed different machine learning (ML) models on the fog node that are monitoring the elderly person continuously based on the vital data obtained from sensors and some personal attributes fetched from the MIS server. The personal data fetched online from MIS server and live data coming from sensors are merged to form a record that is passed through ML models to predict any chance of stroke. The vital signs captured from sensors include values of blood glucose level and hypertension for the prediction of any chance of stroke. The personal attributes include .Patient identity (id), Age , Heart Disease, whether the patient is married or not, type of work(occupation), Residence, Body mass index(BMI) and smoking information.

The purpose of this chapter is to describe in detail how the proposed solution achieves an effective stroke prediction model that is suitable for online prediction of stroke when deployed on fog nodes of SAFE-RH through a complete methodology which is explained in this chapter. Structure of chapter is as follows: Section 2 provides full discussion of the proposed system, whereas section 3 includes data description, section 4 describes data preprocessing, section 5 gives a detailed discussion of the methodological approach of research questions.

3.1 Proposed Solution

The proposed approach has been built surrounding two research questions (RQs). The methodology has been presented in the context of these RQs as described below:

RQ1: How can we enhance the performance of existing stroke prediction models?

RQ2: How can we use data obtained from sensor to continuously monitor the suspected stroke patient?

The objective of RQ1 is to evaluate the existing work on stroke prediction to identify a state-of-art approach (base approach) that can be adopted in the context of SAFE-RH and if possible to improve the performance of base paper before using it.

We proposed an approach to predict stroke using machine learning for remote health monitoring of elderly people under the project SAFE-RH in Pakistan. We used the same Electronic Health Record (EHR) dataset which was previously used in base paper [10] and is available at Kaggle. Following are major steps in the methodology for the RQ1:

1. Data collection
2. Statistical analysis is performed to identify stroke risk factors.
3. Preprocessing of data is done for data cleaning.

4. Our selected dataset is not balanced. The base paper used random-sub sampling for data balance which minimizes the majority class, resulting in less number of records for experimentation. We used SMOTE for class imbalance, as it creates synthetic replicas of data.

5. Feature selection is critical to any ML approach, the base paper used principal component analysis for selecting the features(or diseases) which are associated with the onset of stroke disease. We used Select KBest method for feature selection.

6. Data splitting is done by using f_classif, which split features into training and testing sets for training and testing our classification models

7. Machine learning models are used for classification.

8. Classification is done on original dataset of EHR datasets having values of hypertension in form of zeros and ones.

9. Classification with modified dataset having values of hypertension in systolic and diastolic form.

The methodology has been shown in the form of block diagram in figure 3.1.

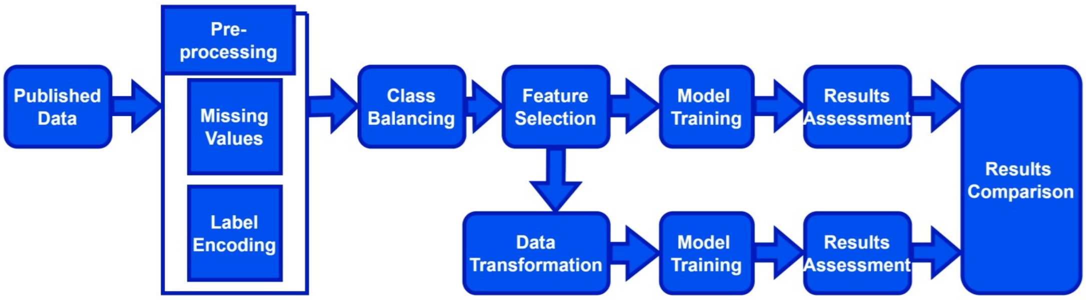


FIGURE 3.1: Block diagram for methodology of RQ1

As is shown in the above figure that we have adopted two approaches to address RQ1; in the first approach, we have used the data as such, that is, in the form in which it is published. In this format, the 'hypertension' attribute has been modeled as a binary one. This means that we are going to use/implement this model in the real-world scenario, then we will have to transform the blood-pressure reading into binary value. Which is obviously not difficult, but for better accuracy, we think, it would be better that if we use the actual values/reading of blood pressure in the model. For this purpose, we have transformed the binary 'hypertension' attribute of data into two attributes reflecting systolic and diastolic blood pressure. We have done this in consultation with the medical doctors. The base paper concluded that age, hypertension, heart disease and average glucose level are most important factors for stroke prediction and obtained highest accuracy with perceptron neural network. In this research we used the same dataset as used by the base paper. The base paper used random subsampling technique for for class imbalance, which has the disadvantage that it reduces the number of records for model training. Only 1096 out of 43401 records were left after random downsampling. We used SMOTE for data balancing that increased the minority class. After data balancing ,we used correlation of coefficient and selectKBest feature selection techniques and select six features(age, hypertension, heart disease, average glucose level,ever-married and smoking status) for stroke prediction and obtained highest accuracy with XGBoost.

3.2 Data Description

We have used the dataset collected by [10] and which is available on Kaggle https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

This dataset contains EHRs of 43401 patients comprising eleven features and one response variable. The brief description of eleven attributes is given below:

- Pid(): patient identifier
- Gender: male or female
- Age: infant, child, young, or old
- Hypertension: a binary attribute showing if the person has high blood pressure or not.

- Heat disease: a boolean attribute showing whether the patient has heart disease or not
- Marital status: to represent marital status
- Occupation type: Govt_job, Self-employed, Private,Children,
- Never_worked
- Residence: whether the person lives in urban or rural rural area)
- Average glucose level: over past five years period
- Bmi: body mass index value
- Patient's smoking_status: how often the person smokes?

The response variable is binary one indicating whether the patient suffered stroke attack or not.

Like any medical/disease based dataset, this dataset too is highly unbalanced pertaining to stroke, that is, most of the records in this EHR contain a negative or "No Stroke" label and relatively less record contain a positive or "Yes Stroke' label. However, the ethical requirements regarding the dataset have been fully complied as claimed by the publishers (Mckinsey and Company) [1] In this research we will use the remaining ten attributes of the patients as input and the target variable i.e., Stroke as output.

[1] https://datahack.analyticsvidhya.com/contest/mckinsey-analyticsonline-hackathon/

Table 3.1: Data Attributes

Sr.#	Attribute	Data Type
1	Patient ID	Numeric
2	Age	Numeric
3	Heart disease	Numeric
4	Avg glucose level	Numeric
5	Hyper tension	Numeric
6	Ever married	Categorical
7	Smoking status	Categorical
8	BMI	Categorical
9	Gender	Categorical
10	Work type	Categorical
11	Residence type	Categorical

3.3 Data Preprocessing

In order to improve accuracy after selecting a dataset for machine learning, we need to perform some prerequisite operations over it.

Initially, data cleaning is done by eliminating all irrelevant rows or rows with garbage data, and then normalize the data. Dataset normalization is the process of organizing data such that all records and fields have a consistent appearance. It converts the features of data to similar scale to enhance the efficiency of the model. At last, the best feature are selected that performs well in achieving the highest possible accuracy. So we see that all attributes have some role in achieving maximum accuracy. It seems that only the id column, which is used to uniquely identify the column, is useless, so we drop it, and it doesn't affect the accuracy, but all other features play a role, so dropping an individual feature leads to a decrease in accuracy, so we use all those features in our machine learning engine. The data mining approach called data preprocessing converts raw data into a form that can be used to create and train the machine learning classifiers.

Preprocessing involves preparing the raw data and making it useful and efficient to train a machine learning classifier.

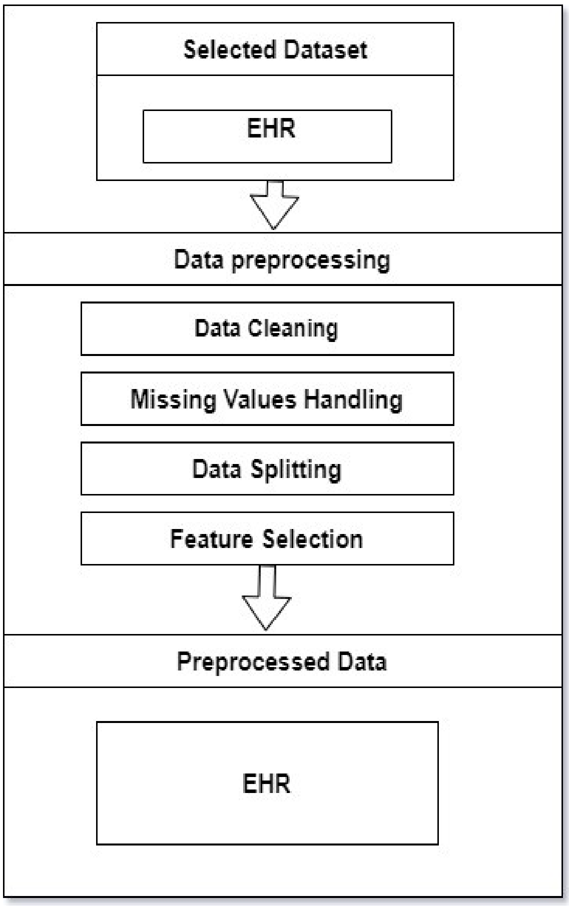


Figure 3.2: Data Preprocessing

Generally, real world datasets are incomplete, inaccurate and inconsistent. Because the data was extracted from an online source, hence it contains noise that is why, it cannot be used directly for analysis purposes, and therefore its preprocessing is required. For this purpose we perform missing values handling and label encoding technique in preprocessing step.

3.3.1 Cleaning the Data

Real word data may be incomplete or it may have irrelevant information. To make data complete, relevant and consistent, data cleaning is performed. Some of the data cleaning steps are as follows:

3.3.1.1 Handling the Missing Values

Some fields of data(attributes) may have missing values. Model's Accuracy may be negatively impacted because of missing values within dataset. There are different ways of dealing it. A few among them are:

Ignoring the Rows:

This technique only functions when a row contains to much missing values and a larger dataset.

Filling the Missing Values:

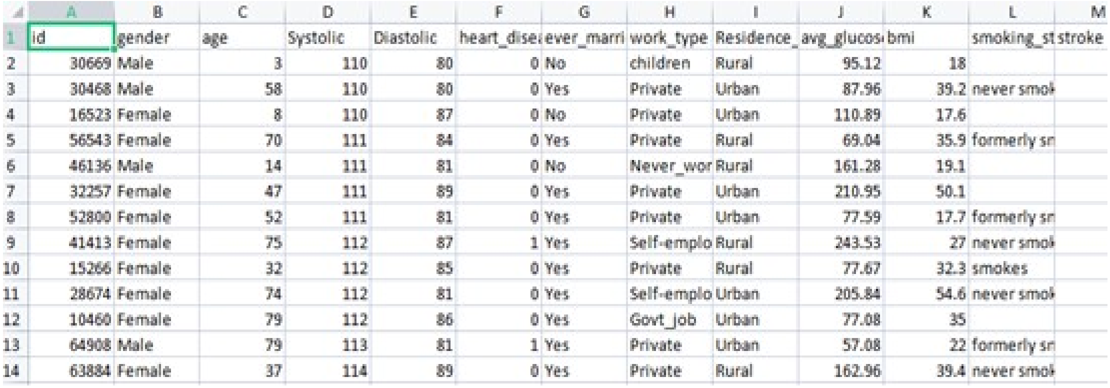
Missing values can be filled in different ways such as,The alternatives include manually inputting the lacking digits, using the attributes mean, or selecting the most probable value.Few rows in the Stroke dataset do not contain value of BMI attribute (missing values error). Among different techniques to handle missing values, a common one is replacing missing values with the mean value of the attribute. This does not cause any major change in the behavior of the data. So we adopt this technique for the missing values of BMI attribute.

3.3.1.2 Removing Noisy Data

Noisy lacks significance and is not understandable to computers. It may be result of a wrong data entry operator or due to faulty data.

3.3.2 Label Encoding

In machine learning, we often deal with datasets that include a number of tags in a single or multiple columns which can be expressed verbally or numerically. Our dataset, used in this work contains different features having values in categorical and numerical forms. As the categorical data is not understandable to machine learning classifiers, so it needs to be converted into numeric form. For this purpose we use label encoding technique. It involves converting labeling into the numeric format so that machines can read them. Machine learning tools are then able to identify how those labels are operating better. For such structured datasets, it is a key classification - based pre-processing step.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	gender	age	Systolic	Diastolic	heart_dise	ever_marri	work_type	Residence_	avg_glucos	bmi	smoking_st	stroke
2	30669	Male	3	110	80	0	No	children	Rural	95.12	18		
3	30468	Male	58	110	80	0	Yes	Private	Urban	87.96	39.2	never smok	
4	16523	Female	8	110	87	0	No	Private	Urban	110.89	17.6		
5	56543	Female	70	111	84	0	Yes	Private	Rural	69.04	35.9	formerly sn	
6	46136	Male	14	111	81	0	No	Never_wor	Rural	161.28	19.1		
7	32257	Female	47	111	89	0	Yes	Private	Urban	210.95	50.1		
8	52800	Female	52	111	81	0	Yes	Private	Urban	77.59	17.7	formerly sn	
9	41413	Female	75	112	87	1	Yes	Self-emplo	Rural	243.53	27	never smok	
10	15266	Female	32	112	85	0	Yes	Private	Rural	77.67	32.3	smokes	
11	28674	Female	74	112	81	0	Yes	Self-emplo	Urban	205.84	54.6	never smok	
12	10460	Female	79	112	86	0	Yes	Govt_job	Urban	77.08	35		
13	64908	Male	79	113	81	1	Yes	Private	Urban	57.08	22	formerly sn	
14	63884	Female	37	114	89	0	Yes	Private	Rural	162.96	39.4	never smok	

Figure 3.3: Label Encoding

Figure 3.3 shows that here are some attributes that are categorical in nature and contain terms as values, for example, 'smoking status' attribute contains 'unknown, never smoked, smokes, rarely smokes' as values. Classifiers generally do not handle such categories values properly and generate error. For this purpose, such categorical values are encoded. Table 3.2 gives the names of categorical attributes, their respective values and codes assigned to each of them. As can be seen from table, the numeric codes 0, 1, 2, 3 and 4 are used for five categorical attributes in the dataset. For example, the 'gender' attribute has 'male' and 'female' values that have been encoded as 0 and 1 respectively. Similarly, the other attributes

Table 3.2: Label Encoding

Attribute Name	**0**	**1**	**2**	**3**	**4**
Gender	female	male			
Ever_married	no	yes			
Residence Type	rural	urban			
Smoking_status	unknown	never	sometimes	smokes	
Work_type	children	never	Self	Govt	private

3.3.2.1 Label Encoding Limitation

Label encoding not only translates the information into machine-readable format but also assigns a unique number to each type of data (beginning at 0). Due to this, prioritization issues may surface during training with data sets. It is possible to give a label with a high value precedence over one with a low value. To avoid this problem we use nominal data type of the encoded feature' values.

3.4 Methodology for RQ1

3.4.1 Data Balancing

A total of 43401 medical records are included in our collection. Only 783 of these records are those of patients who have had a stroke; the remaining 42618 entries do not.

Table 3.3: Original Data

Data	Records	Stroke	No Stroke
Original	43401	783	42618

This dataset is incredibly imbalanced. This makes it difficult to use this data to train any machine-learning models directly. As a result, we employed a smooth strategy to lessen the negative effects of the dataset's imbalance. We classified the 783 entries of stroke patients as the minority class, and the rest 42618 instances without a stroke diagnosis as the majority class.

3.4.1.1 Subsampling

Subsampling is a technique for reducing the amount of data by choosing a smaller portion of the given data. The subset is defined by selecting the parameter n, which indicates each subsequent data point to be extracted.

The fundamental concept is to aggregate classifier predictions, typically via voting, after training them on various subsamples of the data. These strategies include bagging, where these examples get subsampled equally, and boosting, when examples that were incorrectly identified in earlier iterations are more likely to be chosen in subsequent iterations. Subsampling is required because of memory constraints. Subsampling strategies have been examined for the development of association rules in Knowledge Discovery for Datasets, while their validity gets evaluated on the entire dataset. Overlapping problem is employed to test the subsampling limitation.

Table 3.4: Random Downsampling of Data

Data	**Records**	**Stroke**	**No Stroke**
Original	43401	783	42618
After Random downsampling	1096	548	548

3.4.1.2 SMOTE

Synthetic Minority Oversampling Technique. Because there are insufficient instances of the minority class, unbalanced classification has the drawback that a model cannot efficiently learn the decision boundary. One solution to this problem is to oversample the occurrences of the minority class. Before model fitting, this may be accomplished by straightforward replication of samples from the minority class in the training dataset. Although it can balance the class distribution, this doesn't provide the classifier any new data. Synthesizing fresh minority-class instances is an advantage over using duplicate minority-class instances. SMOTE selects samples from the attribute subspace that are close to each other, draws a border between the samples, and creates a new sample along the border. This kind of data augmentation for tabular data can be extremely Successful as it synthesizes new examples from the minority class. SMOTE is arguably the most widely used technique for creating new samples. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space. After applying SMOTE data becomes balanced as both classes have 42617 instances each.Use of duplicating minority-class instances is preferable than synthesizing new minority-class samples. This type of data source enhancement has a great deal of promotion and development.

Table 3.5: Data After Smote

Data	Records	Stroke	No Stroke
Original	43401	783	42618
After SMOTE	85234	42617	42617

3.4.2 Feature Selection

Lowering the quantity of input parameters is a step in the feature selection procedure of a predictive model. In some cases, fewer input variables might improve the model's performance while simultaneously minimizing the cost of computing required for modelling. Random Forest, LASSO Regression and Analysis of variance techniques are mostly used for feature selection.

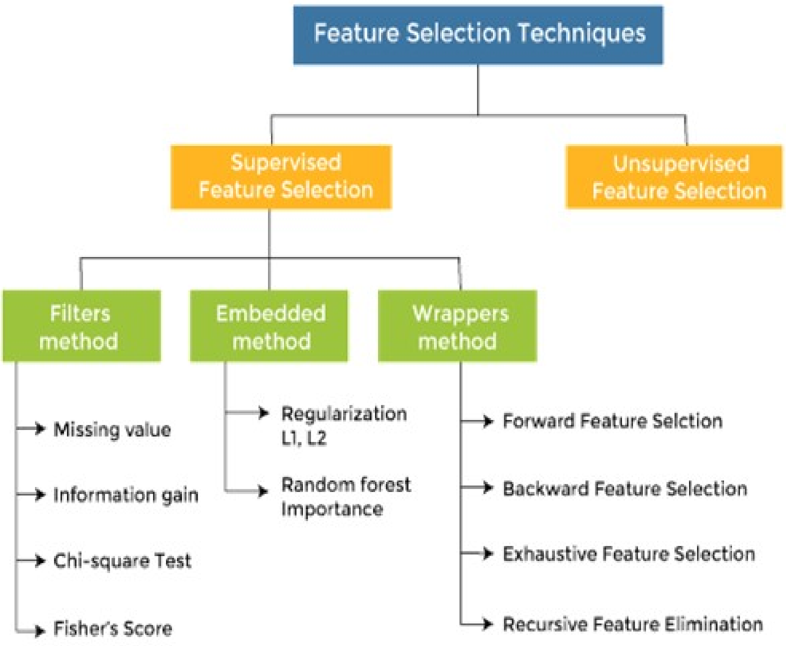


Figure 3.4: Feature selection (javatpoint.com)

Feature selection improves efficiency of the model. In this work, important features necessary for stroke prediction are selected by using correlation coefficient and SelectKBest technique. Fig shows some common feature selection techniques.

3.4.2.1 Correlation Coefficient

We analyse a dataset of electronic health records in this part. We do feature correlation analysis. The following are some ways that correlation analysis is helpful for selecting features:When two characteristics are highly correlated, one of them can be disregarded in the prediction of the likelihood that a stroke will occur since it adds no new information to a prediction model.

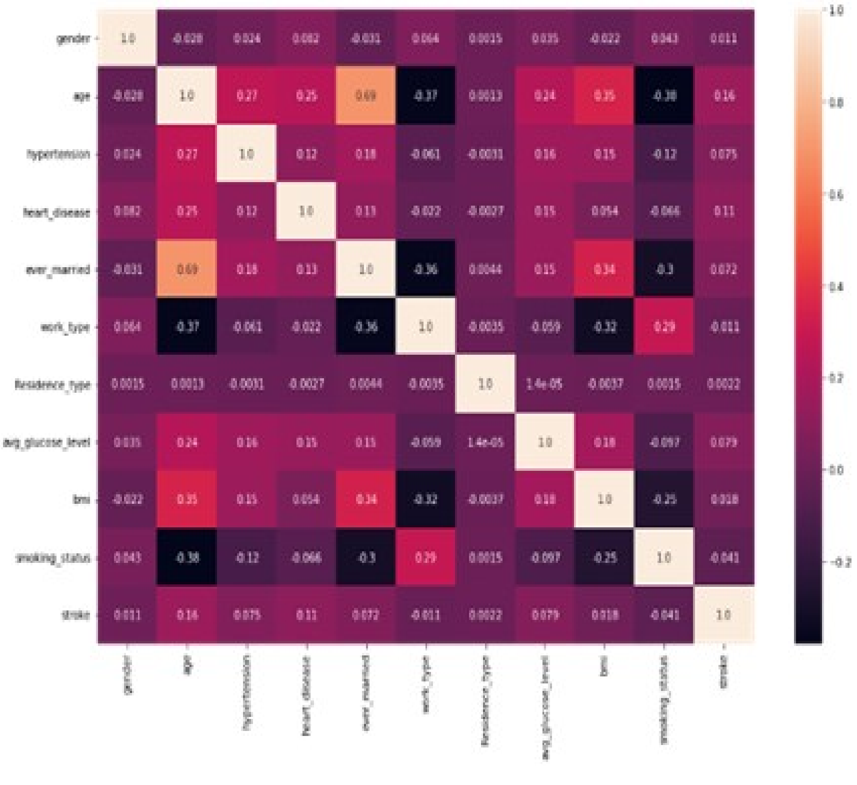


Figure 3.5: Correlation matrix for patient attributes

Fig. 3.5 illustrates the association between several patient characteristics using Pearson's coefficient of correlation.The values of coefficient will show how strongly the attributes are correlated with each other. The correlation between any feature and itself is unity, as would be expected. With a 0.5 correlation value, a patient's age and marital status are significantly correlated. Additionally, there is a significant association in between a patient's age, the nature of their profession, whether they have hypertension as well as heart disease, and his average blood sugar level (0.38 correlation index). Given that the majority of illnesses affect an ageing population, the association between the patient's age and other characteristics appears logical.There is no relationship between a patient's kind of housing and any other characteristic.With a 0.35 correlation score, the patient's line of job is positively correlated with their marriage status.

3.4.2.2 SelectKBest Technique

SelectKBest is a feature selection technique which selects the relevant features from a given dataset as compared to the target variable(that is Stroke in this research). SelectKBest will choose the K features from the initial set (EHR dataset) that are the most illuminating, therefore K must be a number larger than 0 and less than or equal to the entire feature set. The first k characteristics of X with the greatest scores are retained by SelectKBest.

We select the important features necessary for stroke prediction by using select KBest technique which calculates relationship between each input variable and target variable. Higher the score stronger the relationship. Table 3.6 gives score of all input variables with 'stroke'. We use above table values for feature selection in a step forward fashion. We get best performance of the model with the following six attributes: age, heart disease, avg glucose level, hypertension, ever married and smoking status.

Table 3.6: Scores Obtained by SelectKBest

Features	**Score**	**Attribute**	**Score**
Age	1083.17	Smoking_ status	73.88
Heart_disease	569.02	BMI	14.71
Avg_glucose	271.97	Gender	5.44
Hypertension	247.69	Work type	5.06
Ever_married	225.64	Residence type	0.22

3.4.3 Splitting Data

Data can be split based on data sampling methods such as:

Random sampling

During the data modelling process, this data sampling strategy prevents bias towards a number of potential data properties. The uneven distribution of the data, though, could cause issues with irregular splitting.

Stratified sampling at random

This method selects data samples at irregular intervals using preset parameters. It guarantees that the data is evenly distributed throughout the testing and training sets.

Nonrandom sampling

This method is frequently used by data modelers who want to use the most current data as the test set.

In machine learning, data splitting is done to avoid overfitting. Three or four sets of the original data are created. The three main typical sets are as follows:

1. The training dataset is that portion of data which is used to train the classifier. A trained model gives higher accuracy and is efficient as well.
2. Data data analysts frequently use this technique when they want the most recent data as the test set. It can aid in model selection and is intended to rank the performance of the models.
3. The unseen data set upon which evaluation of model is being done is called test data set.

In this research, we split the EHR data into two part split i.e., testing and training modules with a ratio of 70:30. In a two-part split, the model is often trained in one portion while another part is used to test or analyze the data. The table 4.2 gives values of evaluation metrics with two different splits i.e., 70:30 and 80:20.

3.4.4 Application of Algorithm On Original Dataset

The original dataset contains the values of the attribute hypertension in the form of zero and one. In this step we will perform classification on original dataset. Selection of the machine learning algorithm is an important thing because this plays the vital role in achieving accuracy. After data analysis through correlation and SekectKBest classifier, we employed the base paper algorithms (Decision Tree and Random Forest) along with SVM , Logistic regression, and XGBOOST with different number of features for our stroke prediction model. Selected algorithms are discussed below:

3.4.4.1 Decision Tree

One of the most widely used binary classification algorithms is the decision tree model. This approach entails creating a decision process that resembles a tree containing a number of condition tests, then using the tree to analyze the dataset of medical records.

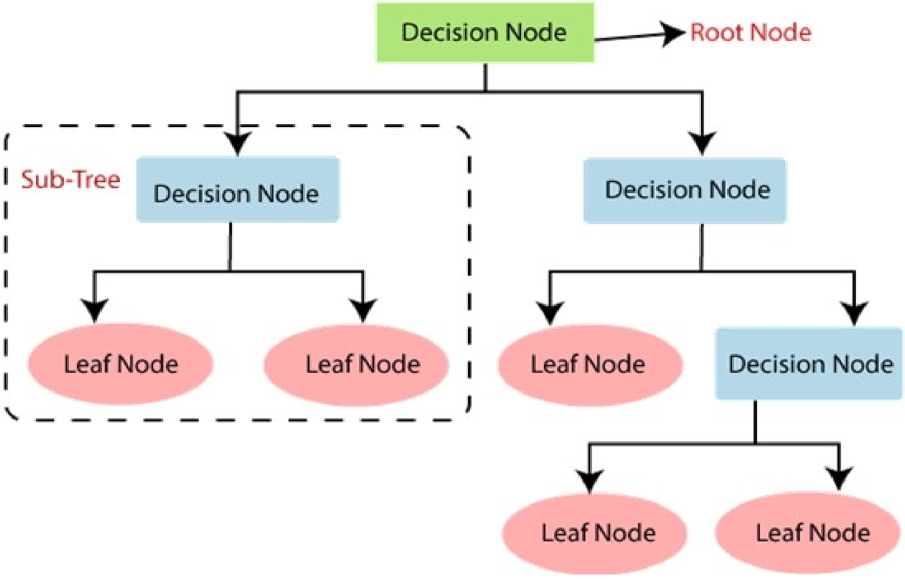


FIGURE 3.6: Architecture of Decision Tree (javatpoint.com)

The branches indicate the results of the tests, each node indicates a test and the leaf nodes indicate the class label. Such an algorithm is adaptable and accurate due to its pruning capability and is important in medical diagnostics.

3.4.4.2 Random Forest

We compare the dataset using a random forest technique as well. The random forest technique is useful in this application because it is adaptable and simple to use, and it consistently produces positive results even with only a small amount of hyper-parameter tuning. The quantity of trees present in the forest restricts the likelihood of over-fitting. Due of their tendency to overfifit to their training set, decision trees are improved by replacing them with random choice forests. The training phase of random forests involves the construction of several decision trees, each of which is then used to predict an output class. Additionally, random forest can offer sufficient indicators of how it rates the importance of each of these input variables.

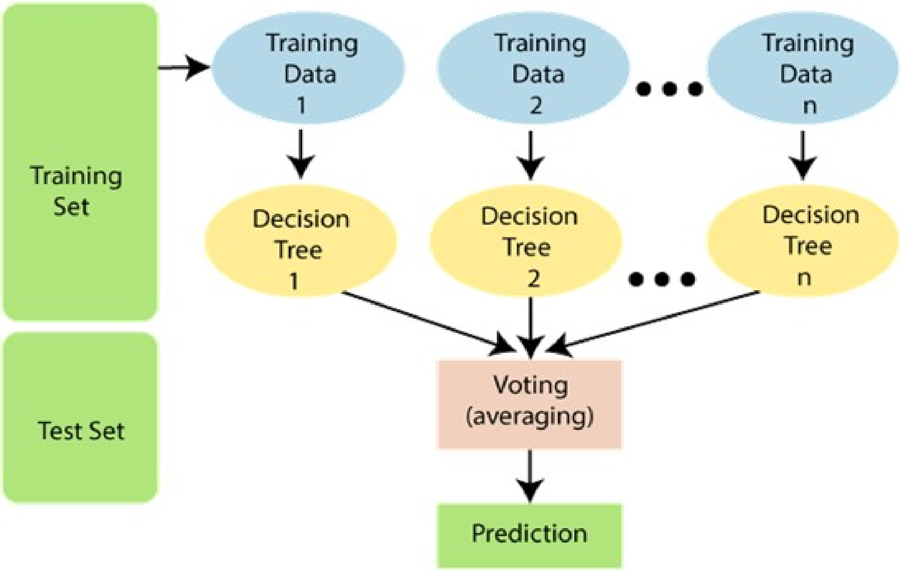


FIGURE 3.7: Architecture of Random Forest (javatpoint.com)

The time complexity and formula is being shown below Time Complexity = O(T.D)

$$RFfii = Pj\varepsilon alltreesnormfiijT$$

- RFfi sub(i) = the importance of feature i calculated from all trees in Random Forest model
- normfifi sub(ij) = the normalized feature importance for i in tree j
- T = total number of trees

3.4.4.3 Support Vector Machine (SVM)

SVM, is a linear model used in machine learning that performs analysis of data for classification and regression. It works well for many real-world issues and can solve both linear and non-linear problems.

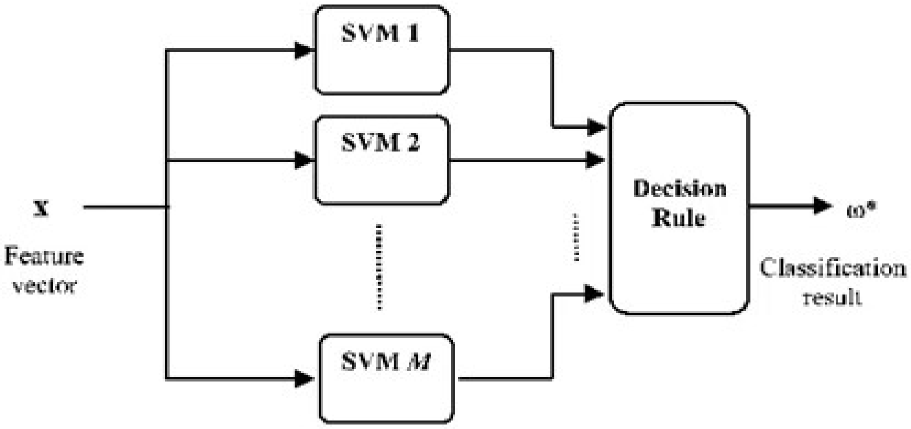


Figure 3.8: Architecture of SVM (javatpoint.com)

The Support Vector Machine (SVM) is a type of supervised learning that analyses data and places it in one of two categories. An SVM generates a map of the data after it has been sorted, with the margins between the two being as far apart as is practically possible.

3.4.4.4 Logistic Regression

A logistic function is applied to represent a binary dependent variable in the most basic version of the statistical model known as logistic regression. Logistic regression, in its simplest and basic form, is a statistical model that uses a logistic function to model a binary response variable.

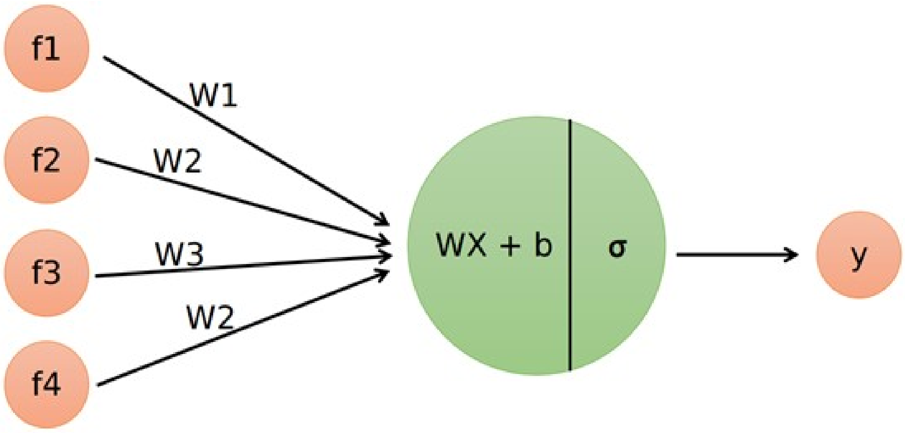


FIGURE 3.9: Architecture of Logistic Regression (javatpoint.com)

Estimating the features of a logistic model is the aim of the regression technology called as logistic regression (also known as logit regression).

3.4.4.5 XGBoost

Incorporating specific accurate approximation techniques, XGBoost is a tree-based ensemble machine-learning methodology that enhances the Gradient Boosting framework.

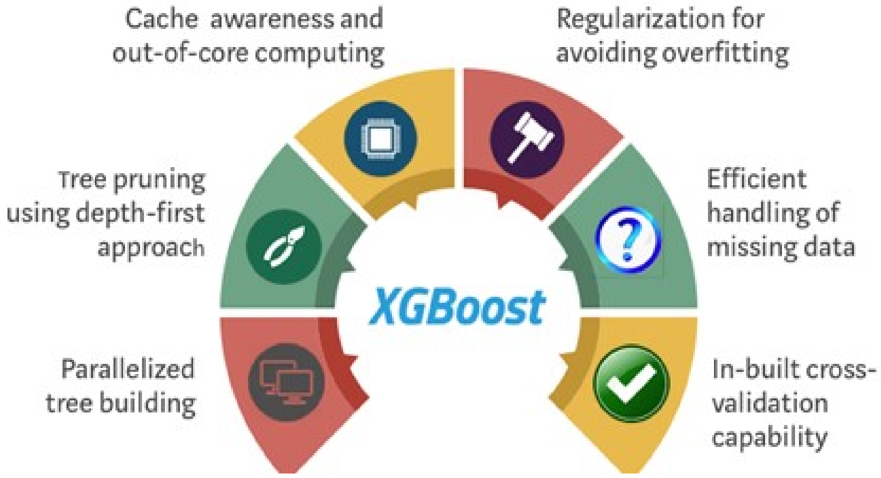

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

FIGURE 3.10: Architecture of XGBoost (javatpoint.com)

By avoiding overfitting and automatic tree pruning, it has increased predicting efficiency and accuracy. It selects the best tree model by making use of closer approximations.

3.4.5 Application of Algorithm On Modified Dataset

The original dataset contains the values of the attribute hypertension in the form of zero and one. After testing and training the model with EHR dataset from online source, the next step is to use the data obtained from the sensor to continuously monitor the suspected stroke patients and also to incorporate the runtime readings of hypertension attribute in the existing dataset. So in the first step we convert the values of the attribute hypertension which were in the form of 0 & 1 to corresponding systolic and diastolic values from the available EHR dataset.

3.4.5.1 Data Conversion to Systolic and Diastolic Values

To train the model for prediction of actual data of hypertension coming from sensors. We have to convert the available EHR dataset to systolic and diastolic values.

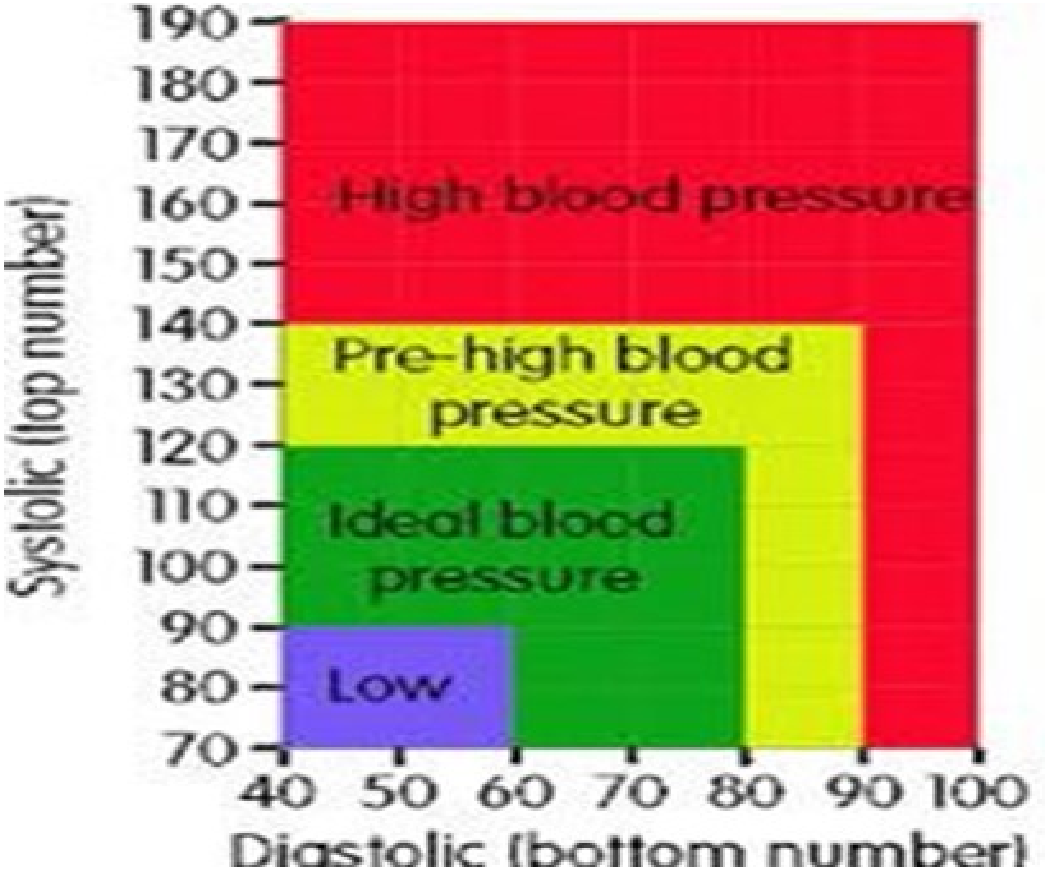
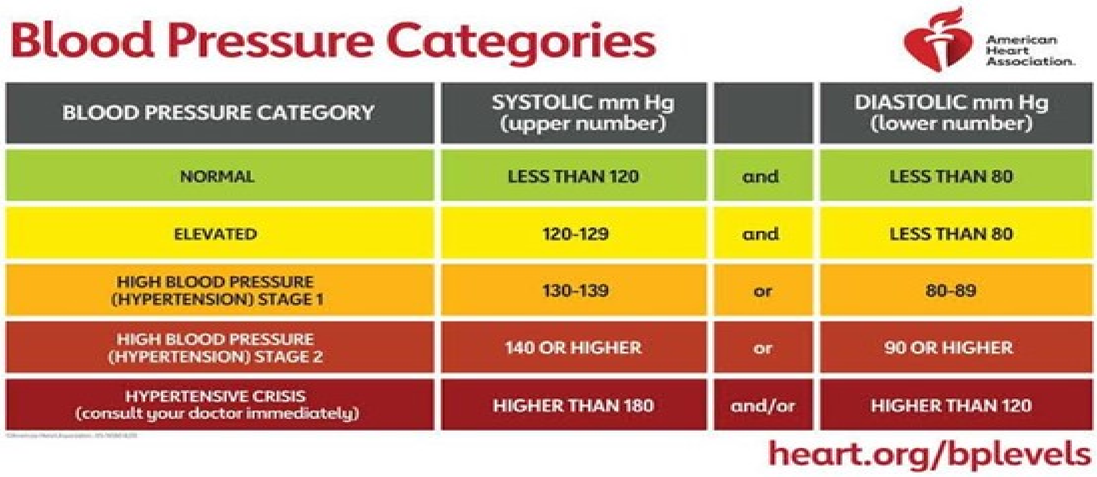


Figure 3.11: Systolic and diastolic Readings heart.org/bplevels



Blood Pressure Categories

American Heart Association.

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120-129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130-139	or	80-89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

heart.org/bplevels

Figure 3.12: Blood Pressure Ranges

3.4.5.2 Formula for Systolic Values (S)

Here, D = Age, F = Heart Disease, I = Smoking status

S=IF(AND(D2≥30,D2<40,F2=0,G2=”Yes”,I2=””), RANDBETWEEN(110,120),

IF(AND(D2≥30,D2<40,F2=1,G2=”Yes”,I2=””), RANDBETWEEN(120,125),

IF(AND(D2≥30,D2<40,F2=0,G2=”Yes”,I2=”Smokes”), RANDBETWEEN(120,130),

IF(AND(D2≥30,D2<40,F2=0,G2=”Yes”,I2=””), RANDBETWEEN(110,120),

IF(AND(D2≥40,D2<50,F2=0,,I2=””), RANDBETWEEN(110,120),

IF(AND(D2≥40,D2<50,F2=0,I2=”Smokes”), RANDBETWEEN(110,120),

IF(AND(D2≥40,D2<50,F2=1,I2=”Smokes”), RANDBETWEEN(120,130),

IF(AND(D2≥50,D2<60,F2=1,I2=”Smokes”), RANDBETWEEN(120,140),

IF(AND(D2≥50,D2<60,F2=1,I2=””), RANDBETWEEN(120,130),

IF(AND(D2≥50,D2<60,F2=0,I2=””), RANDBETWEEN(110,120),

IF(AND(D2≥50,D2<60,F2=0,I2=”Smokes”), RANDBETWEEN(120,130),

IF(AND(D2≥60,F2=0,I2=””), RANDBETWEEN(110,120),

IF(AND(D2≥60,F2=0,I2=”Smokes”), RANDBETWEEN(120,130),

IF(AND(D2≥60,F2=1,I2=”Smokes”), RANDBETWEEN(120,150),

IF(AND(D2≥F602=1,I2=”Smokes”), RANDBETWEEN(120,140))))))))))))))))

3.4.5.3 Formula for Diastolic Values (D)

Here, D = Age, F = Heart Disease, I = Smoking status

```
D = IF(AND(G4≥30,G4<40,I4=0), RANDBETWEEN(110,120),
IF(AND(G4≥30,G4<40,I4=1,), RANDBETWEEN(120,125),
IF(AND(G4≥40,G4<50,I4=0), RANDBETWEEN(110,120),
IF(AND(G4≥40,G4<50,I4=1), RANDBETWEEN(120,130),
IF(AND(G4≥50,G4<60,I4=0), RANDBETWEEN(120,130),
IF(AND(G4≥50,G4<60,I4=1), RANDBETWEEN(120,135),
IF(AND(G4≥60,I4=0), RANDBETWEEN(80,90),
IF(AND(G4≥60,I4=1), RANDBETWEEN(85,100))))))))
```

Data conversion is done under the supervision of doctors working under SAFE-RH project. Data is converted by considering vital signs of patients such as their age, bmi, smoking status, and heart disease.After data conversion,we use the same models that have been used for original data such as Decision Tree, Random Forest, SVM. However,we also use XGBoost classifier which gives good results for both categorical and numeric class labels. By forestalling over-fitting and permitting auto pruning of trees, it has expanded forecast power through parallelization and memory(Cache) streamlining execution. Implementing nearer approximations, it picks the best tree model. During the preparation cycle of ML model, different split proportions of preparing and test information are utilized.

3.5 Methodology for RQ2

RQ2: How can we use data obtained from sensor to continuously monitor the suspected stroke patient?

Most of the approaches build and test model on observed data with few examples.To incorporate the real time values of hypertension and blood glucose level obtained from the sendors(BP checker and blood glucometer) into the stroke prediction model, the model the model(repeated) needs to be deployed on fog nodes of SAFE-RH called Resberi Pi. We deployed both models on Respberi Pi, where they predict any chance of stroke on receiving input. data. The input data is combination of static and dynamic values. The static values are stored on MIS server and the dynamic values ara the readings of hypertension and blood glucose level coming live from sensors to Respberi Pi.

3.5.1 Setup for the Pilot Project

Figure 1.1 shows the overall architecture of the SAFR-RH project. The first module contains the sensors (BP checker ,Blood Glucometer etc) attached with the target group(Elderly, Maternal and infant). In the second module, trained AI models are deployed for stroke prediction. In the third module, the privacy and security of data is maintained. In the fourth module, AI-based alerts are send to the family and healthcare facility. In the fifth module, actions in response to the alerts are performed. In the module six, secured data is sent to both MIS client and cloud storage.

3.5.2 Flow of Work

As mentioned earlier, there are three target groups in SAFE-RH, that is, maternal, infants and elderly people. The focused target group in this work is elderly people which has been shown in block (1) of the figure 3.13. Edge/fog node is linked with all the sensors attached with elderly people. These sensors send data of vital signs of the target continuously. These sensors values go to fog node through Wi-Fi where our extracted machine learning models are deployed.

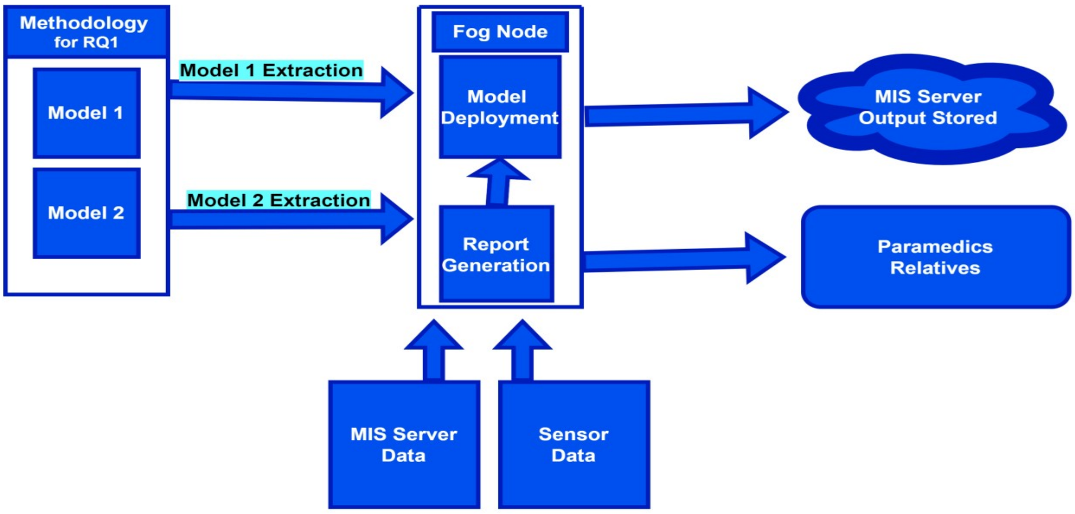


FIGURE 3.13: Block diagram for methodology of RQ2

The personal data of patient is bought from MIS (Management Information System) server and a record is created by combining sensor data and personal attributes of the patient. The final results predicted by the model are sent to the MIS server. The server then sends AI alerts to the patient attendee and to doctors. Actions in response of alerts are performed. Privacy and security of data is maintained through encryption. Secure data is then sent to cloud.

3.6 Tools and Technologies

To perform the complete experiments and suggestions of models we use the different tools and technologies mention below:

- Microsoft Visio – use for creation of diagrams
- Microsoft Excel – use for manage the dataset
- Python - For Programming and Machine Learning
- Sklearn – Machine Learning algorithms library
- Googlecolaboratory – For Python Programming

- Latex – For Thesis Write up
- Weka – For resampling and Initial experiments

After data analysis through correlation and SekectKBest classifier. We employed the base paper algorithms (Decision Tree and Random Forest) along with SVM, and XGBOOST for model creation.The performance of XGBClassifier model was the best among all so XGBClassifier is selected for stroke prediction.

Novality of this research work includes;

- Improved previous model's prediction capacity by using SMOTE instead of random downsampling technique for data balancing
- Use of different ML classifiers for stroke prediction
- Addition of extra features for stroke prediction
- Practicalability of records by transforming binary values of hypertension to corrosponding systolic and diastolic values
- Continuous monitoring of the patients(live)

Chapter 4

Experiments, Results and Evaluations

The proposed research methodologies based on research questions are explained in previous chapter. This chapter describes the results of experiments performed for our research work, followed by an evaluation of these results. An experimental setup is developed on the basis of a proposed system and selected methodology. Our next step is to gather the results of each experiment, which will be discussed further in order to determine the best one. Our research questions emphasize on improving the existing stroke prediction model proposed by the base paper and implementation of improved model on fog nodes of SAFE-RH for continuous monitoring of elderly people as well as early detection and prediction of any chance of stroke disease in the suspected patients. The focus of this research is early detection and prediction of stroke in elderly people by continuously monitoring them through remote health monitoring system. This study is beneficial for the patients, caretakers and medical personals by improving healthcare facilities, reducing cost and travel rate, etc. We use the python to perform different experiments by apply different machine learning algorithms and then we examine the results for each experiment. Details discussed below.

Experimentation and evaluation is done according to the research questions.

4.1 Evaluation

After defining the research methodologies, we conduct the experiments, evaluation of experiments are important to discuss because of bases of evaluation we suggest the right algorithm to use. For the evaluation of experiments, we find the accuracy of each algorithm than precision, recall and f1 score.

These all are base on the values of True Negative, False Positive, False Negative, True Positive. The description of these four values in the context of our work is as follows:

Actual Values: Positive = Stroke, Negative = No Stroke

True Positive (TP): A person having stroke is predicted stroke patient

True Negative (TN): A healthy person is predicted as healthy(No Stroke)

False Negative (FN): A stroke patient is predicted as a healthy person.

True Negative (TN): A healthy person is predicted as stroke patient.

Confusion Matrix

Table 4.1: Confusion matrix

		Actual True Values	
		Positive (Stroke)	**Negative (No Stroke)**
Predicted	**Positive**	TP (Stroke)	FP (Stroke)
Values	**Negative**	FN (No Stroke)	TN (No Stroke)

4.2 Evaluation Matrices Used for Stroke Prediction Model

We used accuracy, precision, recall, and F-score for model optimization.

4.2.1 Accuracy

The number of accurate predictions that your model was able to make for the entire test dataset is referred to as its accuracy. The following equation is used to determine its value.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.2.2 Precision

The degree to which a positive prediction is accurate is referred to as its precision. To put it another way, it means that if a result is predicted to be positive, how certain are you that the result will in fact be positive? It is determined by applying the formula that is as follows:

$$Precision = \frac{TP}{TP + FP}$$

4.2.3 Recall

The recall rate, also known as the true positive rate, is a metric that determines how many true positives are predicted out of the total number of positives in a Proposed Research Methodology's dataset. In certain instances, it is also referred to as the sensitivity. The following equation is used to calculate the value of the measure:

$$Recall = \frac{TP}{TP + FN}$$

4.2.4 F1-Score

The F1-score is the F-score that is utilized the vast majority of the time. It is a mixture of precision and recall, such as the harmonic mean of both of them. The formula below can be used to get an individual's F1 score:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

4.3 Experiments and Results Based on RQ1 (Original Data)

RQ1: How can we enhance the performance of existing stroke prediction model?

Now when the data is ready, we employ multiple classifiers for modeling which is a common practice in literature. We use most of the models that have been used in base paper [10] such as Decision Tree, Random Forest, SVM. However, we also use XGBoost classifier which gives good results for both categorical and numeric class labels. By including certain accurate approximation methods the tree-based ensemble machine learning technique known as XGBoost enhances the Gradient Boosting framework. By preventing over-fitting and allowing auto pruning of trees, it has increased prediction power through parallelization and cache optimization performance. Using closer approximations, it chooses the best tree model. During the training process of ML model, different split ratios of training and test data are used. We experiment with 70-30 and 80-20 ratios. For results evaluation, we use accuracy (Ac), precision (Pr), and F1-score (FS). Table 4.2 below gives values of these three measures with two differences splits.

Table 4.2: XGBOOST with different split ratios

Split	**70/30**			**80/20**		
Matrices	**Ac**	**Pr**	**F1**	**Ac**	**Pr**	**F1**
XGBoost	0.953	0.913	0.869	0.873	0.873	0.878

As shown in the table above, our approach performs better with 70/30 split which is the same ratio as used in [10]. We also compared our results with those in the base paper using the same four features.

The results of different classifiers along with proposed approach based on the XGBoost classifier are given in Table IV below. The metrics used are Precision (Pr), Recall (Re), F1-score (FS), Accuracy (Ac), Miss rate (MR), Fallout rate (FR).

Table 4.3: Base Paper With 4 Attributes

Matrices	**Pr**	**Re**	**FS**	**Ac**	**MR**	**FR**
Dec.Tree	0.78	0.71	0.74	0.75	0.20	0.21
RF	0.76	0.74	0.75	0.75	0.18	0.24
NN	0.78	0.71	0.74	0.75	0.19	0.20
XGBoost	0.83	0.89	0.86	0.86	0.06	0.17

The last row in table 4.3 presents the result of proposed approach using the same attributes as of [11]. It can be seen that the XGBoost classifier outperforms all results of the base paper.

Finally, we performed an experiment using the six features selected by our approach. We set the train/test split as 70/30 as it gives best results . We trained the model with 70% training after removing class imbalance and tested it on 30% test data. The confusion matrix for the test experiment is given below:

Table 4.4: Confusion Matrix For Model

		Predicted		
		Yes	**No**	
Actual	**Yes**	12348	289	P=12637
	No	1334	11600	N=12934
		P=13682	N=11889	T=25571

As can be seen from Table 4.4 above that Type 1 error (289) is quite less compared to Type 2 error (1334). However, this still needs to be further reduced and we set it for future work. The metrics calculated from this confusion matrix are given in table 4.5 below, in which we are comparing our results with those claimed best in base paper.

Table 4.5: Best Result Of [11] Vs Proposed Approach

	Pr	**Re**	**FS**	**Ac**	**MR**	**FR**
N.N	0.78	0.71	0.74	0.75	0.19	0.20
Proposed XGBoost	0.913	0.972	0.941	0.940	0.014	0.090

As shown in the table, the proposed approach uses XGBoost classifier which gives better results compared to the best results of the base paper.

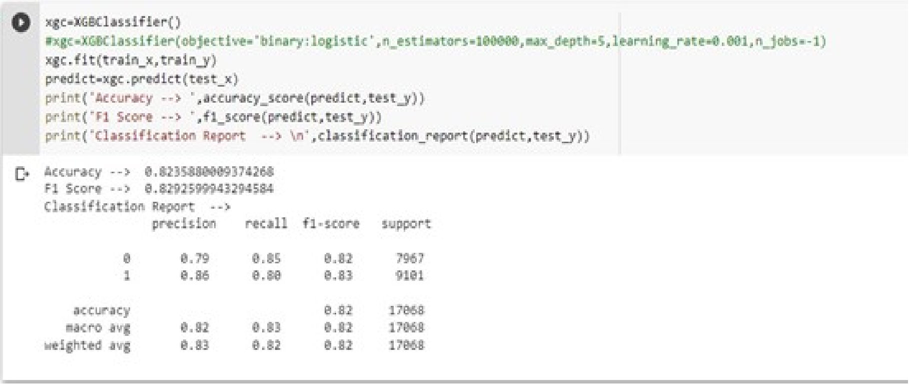
4.3.1 Experiments and Results Based on RQ1(Modified Data)

We experiment with 70-30 and 80-20 ratios. For results evaluation, we use accuracy (Ac), precision (Pr), and F1-score (FS). Table 4.6 below gives values of these three measures with two differences splits.

Table 4.6: XGBOOST with different split ratios

Split	**70/30**			**80/20**		
Matrices	**Ac**	**Pr**	**F1**	**Ac**	**Pr**	**F1**
XGBoost	0.81	0.82	0.81	0.82	0.86	0.83

As shown in the table above, our approach performs better with 80/20 split. We also compared our results with those from the original dataset. The results of different classifiers along with proposed approach based on the XGBoost classifier are given in Table 4.7 below. The metrics used are Precision (Pr), Recall (Re), F1-score (FS), Accuracy (Ac).

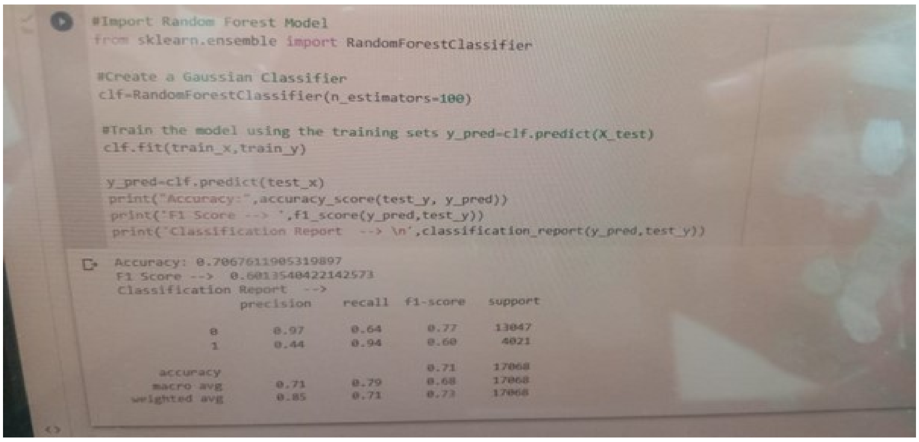


```
xgc=XGBClassifier()
#xgc=XGBClassifier(objective='binary:logistic',n_estimators=100000,max_depth=5,learning_rate=0.001,n_jobs=-1)
xgc.fit(train_x,train_y)
predict=xgc.predict(test_x)
print('Accuracy --> ',accuracy_score(predict,test_y))
print('F1 Score --> ',f1_score(predict,test_y))
print('Classification Report  --> \n',classification_report(predict,test_y))
```

```
Accuracy -->  0.8235880009374268
F1 Score -->  0.8292599943294584
Classification Report  -->
               precision    recall  f1-score   support

           0       0.79      0.85      0.82      7967
           1       0.86      0.80      0.83      9101

    accuracy                           0.82     17068
   macro avg       0.82      0.83      0.82     17068
weighted avg       0.83      0.82      0.82     17068
```

Figure 4.1: XGBoost

Figure 4.1shows the results of XGBoost classifier on the modified data with an accuracy of 82%.

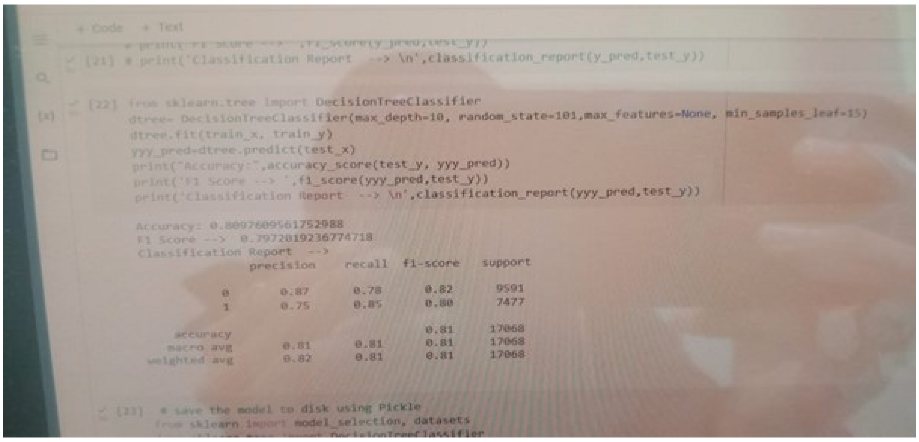


```
#Import Random Forest Model
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=100)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(train_x,train_y)

y_pred=clf.predict(test_x)
print("Accuracy:",accuracy_score(test_y, y_pred))
print('F1 Score --> ',f1_score(y_pred,test_y))
print('Classification Report  --> \n',classification_report(y_pred,test_y))
```

```
Accuracy: 0.7067611905319897
F1 Score -->  0.6013540422142573
Classification Report  -->
               precision    recall  f1-score   support

           0       0.97      0.64      0.77     13047
           1       0.44      0.94      0.60      4021

    accuracy                           0.71     17068
   macro avg       0.71      0.79      0.68     17068
weighted avg       0.85      0.71      0.73     17068
```

Figure 4.2: Random Forest

We also perform experiments with random forest classifier for stroke prediction with modified data. As shown in Fig.Random forest. The classifier gave an accuracy of 71%.



```
    # print('F1 Score --> ',f1_score(y_pred,test_y))
[21] # print('Classification Report  --> \n',classification_report(y_pred,test_y))

[22] from sklearn.tree import DecisionTreeClassifier
     dtree= DecisionTreeClassifier(max_depth=10, random_state=101,max_features=None, min_samples_leaf=15)
     dtree.fit(train_x, train_y)
     yyy_pred=dtree.predict(test_x)
     print("Accuracy:",accuracy_score(test_y, yyy_pred))
     print('F1 Score --> ',f1_score(yyy_pred,test_y))
     print('Classification Report  --> \n',classification_report(yyy_pred,test_y))
```

```
Accuracy: 0.8097609561752988
F1 Score -->  0.7972019236774718
Classification Report  -->
               precision    recall  f1-score   support

           0       0.87      0.78      0.82      9591
           1       0.75      0.85      0.80      7477

    accuracy                           0.81     17068
   macro avg       0.81      0.81      0.81     17068
weighted avg       0.82      0.81      0.81     17068
```

```
[23] # save the model to disk using Pickle
     from sklearn import model_selection, datasets
     from sklearn.tree import DecisionTreeClassifier
```

Figure 4.3: Decision Tree

Experimentation with decision tree for stroke prediction with modified data showed 81% accuracy.

Table 4.7: Base Paper With 4 Attributes

Matrices	**Pr**	**Re**	**FS**	**AC**
Dec.Tree	0.75	0.85	0.80	0.81
RF	0.44	0.94	0.60	0.71
XGBoost	0.86	0.80	0.83	0.82

The last row in table 4.7 presents the result of proposed approach using the modified data. It can be seen that the We set the train/test split as 80/20 as it gives best results (Table 4.6). We trained the model with 80% training after removing class imbalance and tested it on 20% test data. The confusion matrix for the test experiment is given below:

Table 4.8: Confusion Matrix For Model

		Predicted		
		Yes	**No**	
Actual	**Yes**	12348	289	P=12637
	No	1334	11600	N=12934
		P=13682	N=11889	T=25571

As can be seen from Table 4.8 above that Type 1 error (289) is quite less compared to Type 2 error (1334). However, this still needs to be further reduced and we set it for future work. The metrics calculated from this confusion matrix are given in table 4.9 below, in which we are comparing our results with those claimed best in base paper.

Table 4.9: Best Result of Modified vs Original Dataset

Classifier	**Pr**	**Re**	**FS**	**Ac**
XGBoost(120/80)	0.86	0.80	0.83	0.83
XGBoost(0,1)	0.91	0.97	0.94	0.94

As shown in the table, the modified dataset using XGBoost classifier gives better results with an accuracy of 83%.

4.4 Experiments and Results Based on RQ2

RQ2: How can we use data obtained from sensor to continuously monitor the suspected stroke patient?

The models build in RQ1 are deployed on the fog nodes of SAFE-RH, where they prediction is being done on receiving the input. The input is combination of patient's personal attributes (I.e, static values saved at one time) and dynamic values of the attribute hypertension and average glucose level coming from sensors.

4.4.1 Experiments with Original Data

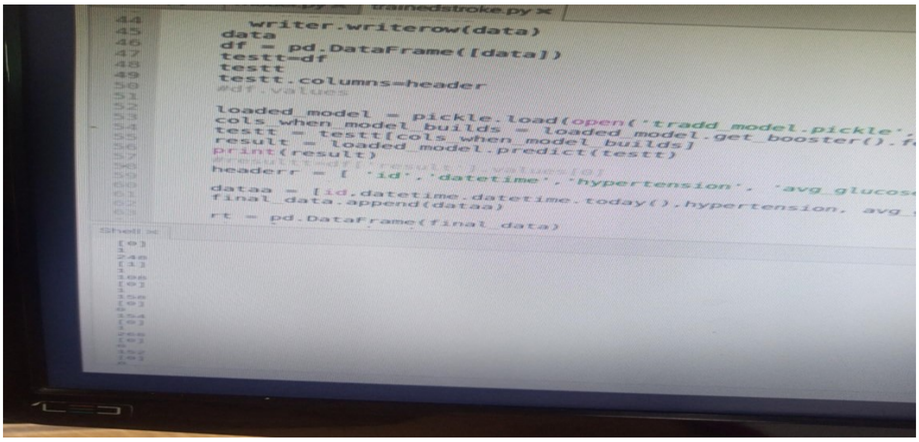
The sensors send data of vital signs of the target continuously. These sensors values go to fog node through Wi-Fi where our machine learning model for original data is deployed. In fog node,from the sensor data the value of hypertension is first converted to corresponding zero or one value and then used for prediction. The personal data of patient is bought from MIS server and a record is created by combining sensor(modified) data and personal attributes of the patient. The final results predicted by the model are sent to the MIS(Managing information system) server. The server then sends AI alerts to the patient attendee and to doctors. Actions in response of alerts are performed.

4.4.2 Experiments with Modified Data

The sensors send data of vital signs of the target continuously. These sensors values go to fog node through Wi-Fi where our extracted machine learning model is deployed. As the values of hypertension are in systolic and diastolic form, so directly used for prediction. The personal data of patient is bought from MIS server and a record is created by combining sensor data and personal attributes of the patient. The final results predicted by the model are sent to the MIS(Managing information system) server. The server then sends AI alerts to the patient attendee and to doctors. Actions in response of alerts are performed.

4.4.3 Experiments with Sensor Data

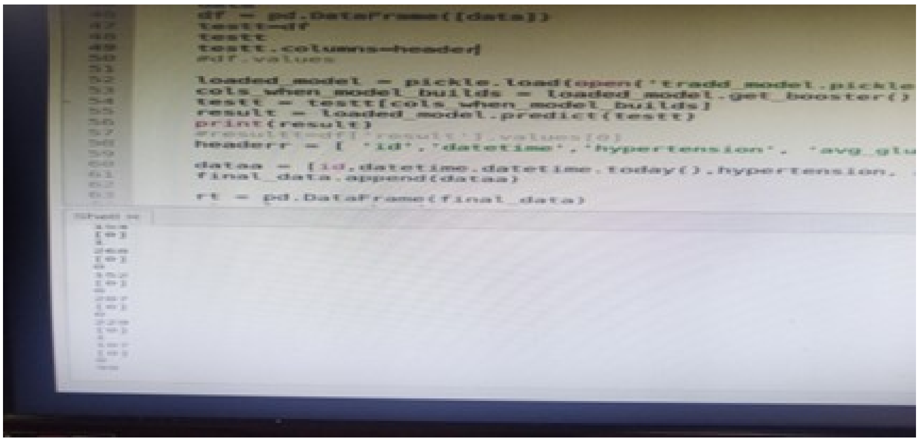
When sensor data (values of hypertension in systolic diastolic form and readings of average glucose level) is passed through the model on fog nodes for prediction, its working is fine. The model is picking data from MIS server, sensing sensor data, matching sensor data with MIS server data, doing predictions and storing the predicted values on MIS server. We are forwarding this stored data to doctors for evaluation. This process is going on.



```
    writer.writerow(data)
  data
  df = pd.DataFrame([data])
  testt=df
  testt
  testt.columns=header
  #df.values

  loaded_model = pickle.load(open('tradd_model.pickle',
  cols_when_model_builds = loaded_model.get_booster().f
  testt = testt[cols_when_model_builds]
  result = loaded_model.predict(testt)
  print(result)
  #resultt=df['result'].values[0]
  headerr = [ 'id','datetime','hypertension', 'avg_glucos

  dataa = [id,datetime.datetime.today(),hypertension, avg_
  final_data.append(dataa)

  rt = pd.DataFrame(final_data)
```

Figure 4.4: Fog Node Predictions

The fig 4.4 Fog Node Predictions shows predicted values after getting input from sensors. Here all outputs are zero, indicating no stroke disease in patients.



```
df = pd.DataFrame([data])
testt=df
testt
testt.columns=header
#df.values

loaded_model = pickle.load(open('tradd_model.pickle
cols_when_model_builds = loaded_model.get_booster()
testt = testt[cols_when_model_builds]
result = loaded_model.predict(testt)
print(result)
#resultt=df['result'].values[0]
headerr = [ 'id','datetime','hypertension', 'avg_glu
dataa = [id,datetime.datetime.today(),hypertension,
final_data.append(dataa)

rt = pd.DataFrame(final_data)
```

```
[illegible]
```

Figure 4.5: Fog Node Predictions

The fig 4.5 Fog Node Predictions shows predicted values after getting input from sensors. Here outputs are both zeros and ones, indicating some patients had stroke and some had no stroke.

The experimental analysis showed the performance of the proposed stroke predicting models with original EHR dataset, with modified dataset and also with sensor data.

Chapter 5

Conclusion and Future Work

The work we have done and thoroughly discussed in the previous four chapters is finally coming to an end in this chapter. Additionally, this chapter goes into detail on how we think this work should be expanded in the future. In this research, we propose two stroke prediction models based on the XGBoost classifier. The proposed approach is evaluated and tested on the EHR dataset. As the used dataset is unbalanced in terms of instances for each class for stroke prediction that is "Stroke" and "No Stroke", we apply to it a data balancing technique based on the SMOTE approach. The base paper used random down_sampling technique for data balancing which results in valuable data loss. Also the results produced are equally likely to be true are false.This technique is suitable when class imbalance is not very large like 60:40. SMOTE generates synthetic data points which are slightly different from original records. Then, the feature extraction is done by correlation analysis and SeleckKBest feature selection technique, allowing to select the attributes of the balanced dataset having the best scores for the prediction of stroke. The selected features(gender, age, hypertension, heart disease, avg glucose level, and ever married attributes) are used for prediction. We apply different classifiers which produce different results. Finally, an XGBoost classifier was used for classification purpose with original dataset. The obtained results are up to 94% in terms of accuracy, 91% in terms of precision and 94% for F1-score. The second XGBoost classifier was used for classification purpose with modified data set.

In the available EHR dataset, the values of the attribute hypertension are in the form of zero and one whereas the actual data of hypertension coming from sensors is in systolic and diastolic form(120/80). To train the model for prediction of actual data of hypertension we convert the available EHR dataset to systolic and diastolic values. The obtained results are up to 82% in terms of accuracy, 86% in terms of precision and 83% for F1-score. Both models are deployed on fog nodes of SAFE-RH for continuous monitoring of patients. The models predict any chance of stroke on receiving input. Actual data having systolic and diastolic values of the attribute hypertension is being collected by doctors. As perspectives, we will:

- Pass those actual values through existing models for stroke prediction.
- Build new model for actual data.
- Predict other diseases like fall detection and heart disease etc, under SAFE-RH.
- Evaluate the models further on different datasets.

Contribution

A conference paper has been published ”Nagina Razzaq, Nayyer Masood, Saba Nawaz, Nadeem Anjum, and Naeem Ramzan. ”Stroke Prediction in Elderly Persons using Remote Health Monitoring.” In 2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 1-4. IEEE, 2022”.

An extended abstract has been published ”Nagina Razzaq, Nayyer Masood. ”Stroke Prediction by Deploying Predictive Models on Fog Devices.” In International Conference on Computing Research (ICCoR-2022), December 17,2000

Funding

This research is funded by Project SAFE-Rural Health https://safe-rh.eu/ under the Erasmus+ programme of European Commission.

Bibliography

[1] V. L. Feigin, R. V. Krishnamurthi, P. Parmar, B. Norrving, G. A. Mensah, D. A. Bennett, S. Barker-Collo, A. E. Moran, R. L. Sacco, T. Truelsen, *et al.*, "Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the gbd 2013 study," *Neuroepidemiology*, vol. 45, no. 3, pp. 161–176, 2015.

[2] M. J. Deen, "Information and communications technologies for elderly ubiquitous healthcare in a smart home," *Personal and Ubiquitous Computing*, vol. 19, pp. 573–599, 2015.

[3] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 1, p. 130, 2017.

[4] Y.-J. Hong, I.-J. Kim, S. C. Ahn, and H.-G. Kim, "Mobile health monitoring system based on activity recognition using accelerometer," *Simulation Modelling Practice and Theory*, vol. 18, no. 4, pp. 446–455, 2010.

[5] N. Razzaq, N. Masood, S. Nawaz, N. Anjum, and N. Ramzan, "Stroke prediction in elderly persons using remote health monitoring," in *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–4, IEEE, 2022.

[6] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 158–161, IEEE, 2017.

[7] D. K. Arnett, R. S. Blumenthal, M. A. Albert, A. B. Buroker, Z. D. Goldberger, E. J. Hahn, C. D. Himmelfarb, A. Khera, D. Lloyd-Jones, J. W.

McEvoy, *et al.*, “2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines,” *Circulation*, vol. 140, no. 11, pp. e596–e646, 2019.

[8] D. K. Arnett, R. S. Blumenthal, M. A. Albert, A. B. Buroker, Z. D. Goldberger, E. J. Hahn, C. D. Himmelfarb, A. Khera, D. Lloyd-Jones, J. W. McEvoy, *et al.*, “2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines,” *Circulation*, vol. 140, no. 11, pp. e596–e646, 2019.

[9] S. Mainali, M. E. Darsie, and K. S. Smetana, “Machine learning in action: stroke diagnosis and outcome prediction,” *Frontiers in Neurology*, p. 2153, 2021.

[10] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, “A predictive analytics approach for stroke prediction using machine learning and neural networks,” *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[11] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1464–1469, IEEE, 2020.

[12] Y. Wu and Y. Fang, “Stroke prediction with machine learning methods among older chinese,” *International journal of environmental research and public health*, vol. 17, no. 6, p. 1828, 2020.

[13] D. Azariadi, V. Tsoutsouras, S. Xydis, and D. Soudris, “Ecg signal analysis and arrhythmia detection on iot wearable medical devices,” in *2016 5th International conference on modern circuits and systems technologies (MOCAST)*, pp. 1–4, IEEE, 2016.

[14] A. Isin and S. Ozdalili, “Cardiac arrhythmia detection using deep learning,” *Procedia computer science*, vol. 120, pp. 268–275, 2017.

[15] Z. Xiong, M. P. Nash, E. Cheng, V. V. Fedorov, M. K. Stiles, and J. Zhao, “Ecg signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network,” *Physiological measurement*, vol. 39, no. 9, p. 094006, 2018.

[16] M. Hammad, A. Maher, K. Wang, F. Jiang, and M. Amrani, “Detection of abnormal heart conditions based on characteristics of ecg signals,” *Measurement*, vol. 125, pp. 634–644, 2018.

[17] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, “Arrhythmia detection using deep convolutional neural network with long duration ecg signals,” *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.

[18] M. Alfaras, M. C. Soriano, and S. Ortín, “A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection,” *Frontiers in Physics*, p. 103, 2019.

[19] R. L. Devi and V. Kalaivani, “Machine learning and iot-based cardiac arrhythmia diagnosis using statistical and dynamic features of ecg,” *The journal of supercomputing*, vol. 76, no. 9, pp. 6533–6544, 2020.

[20] S. Mousavi and F. Afghah, “Inter-and intra-patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach,” in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1308–1312, IEEE, 2019.

[21] T. Tuncer, S. Dogan, P. Pławiak, and U. R. Acharya, “Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ecg signals,” *Knowledge-Based Systems*, vol. 186, p. 104923, 2019.

[22] S. K. Ghosh, R. K. Tripathy, M. R. Paternina, J. J. Arrieta, A. Zamora-Mendez, and G. R. Naik, “Detection of atrial fibrillation from single lead ecg signal using multirate cosine filter bank and deep neural network,” *Journal of medical systems*, vol. 44, pp. 1–15, 2020.

[23] P. Pławiak and U. R. Acharya, "Novel deep genetic ensemble of classifiers for arrhythmia detection using ecg signals," *Neural Computing and Applications*, vol. 32, no. 15, pp. 11137–11161, 2020.

[24] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network," *Information Fusion*, vol. 53, pp. 174–182, 2020.

[25] C.-L. Lin, W.-C. Chiu, F.-H. Chen, Y.-H. Ho, T.-C. Chu, and P.-H. Hsieh, "Fall monitoring for the elderly using wearable inertial measurement sensors on eyeglasses," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1–4, 2020.

[26] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 183–192, 2010.

[27] M. Li, R. Torah, H. Nunes-Matos, Y. Wei, S. Beeby, J. Tudor, and K. Yang, "Integration and testing of a three-axis accelerometer in a woven e-textile sleeve for wearable movement monitoring," *Sensors*, vol. 20, no. 18, p. 5033, 2020.

[28] T. Tuncer, S. Dogan, P. Pławiak, and U. R. Acharya, "Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ecg signals," *Knowledge-Based Systems*, vol. 186, p. 104923, 2019.

[29] A. B. Nigusse, B. Malengier, D. A. Mengistie, G. B. Tseghai, and L. Van Langenhove, "Development of washable silver printed textile electrodes for long-term ecg monitoring," *Sensors*, vol. 20, no. 21, p. 6233, 2020.

[30] J. Yuan, K. K. Tan, T. H. Lee, and G. C. H. Koh, "Power-efficient interrupt-driven algorithms for fall detection and classification of activities of daily living," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1377–1387, 2014.

[31] S. Akhbarifar, H. H. S. Javadi, A. M. Rahmani, and M. Hosseinzadeh, "A secure remote health monitoring model for early disease diagnosis in cloud-based iot environment," *Personal and Ubiquitous Computing*, pp. 1–17, 2020.

[32] F. Shaikh and D. Rao, "Prediction of cancer disease using machine learning approach," *Materials Today: Proceedings*, vol. 50, pp. 40–47, 2022.

[33] S. Das, S. Kumar Mahata, A. Das, and K. Deb, "Disease prediction from drug information using machine learning," *American Journal of Electronics & Communication*, vol. 1, no. 4, pp. 16–21, 2021.

[34] A. Kumar and M. A. Pathak, "A machine learning model for early prediction of multiple diseases to cure lives," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 6, pp. 4013–4023, 2021.

[35] R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang, and N. Mehendale, "Disease prediction from various symptoms using machine learning," *Available at SSRN 3661426*, 2020.

[36] Y.-A. Choi, S.-J. Park, J.-A. Jun, C.-S. Pyo, K.-H. Cho, H.-S. Lee, and J.-H. Yu, "Deep learning-based stroke disease prediction system using real-time bio signals," *Sensors*, vol. 21, no. 13, p. 4269, 2021.

[37] M. Kaur, S. R. Sakhare, K. Wanjale, and F. Akter, "Early stroke prediction methods for prevention of strokes," *Behavioural Neurology*, vol. 2022, 2022.

[38] E. M. Alanazi, A. Abdou, and J. Luo, "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models," *JMIR Formative Research*, vol. 5, no. 12, p. e23440, 2021.