DOI: http://dx.doi.org/10.26483/ijarcs.v15i4.7110

Volume 15, No. 4, July-August 2024

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A PREDICTION MODEL FOR STROKE BASED ON MACHINE LEARNING ALGORITHMS

Abdelhafid Ali I. Mohamed
Department of Computer Science
Libyan Academy for Postgraduate Studies,
Libya, Benghazi

Ahmed Alwirshiffani
Department of Computer Science
College of Computer Technology
Libya, Benghazi

Ramadan A.M. Elghalid
Department of Computer Science
College of Computer Technology
Libya, Benghazi

Zeiad A. Abdelnabi
Higher Institute of Science and Technology -
Libya, Suluq

Rafea. M. Almejarab
Department of Computer Science
College of Computer Technology
Libya, Benghazi

Abstract: In this modern era, people are working hard to meet their physical needs and non-effective their ability to spend time for themselves which leads to physical stress and mental disorder. Many reports state that stroke is caused when blood flow to a part of the brain is stopped abruptly. Without the blood supply, the brain cells gradually die, and disability occurs depending on the area of the brain affected. According to the World Health Organization (WHO), stroke is the greatest cause of death and disability globally. Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke. In this research work, with the aid of machine learning (ML), several algorithms has been used and evaluated for the long term risk prediction of stroke occurrence. We have collected datasets to analyze data and mining using 8 algorithms of machine learning to predict whether the patient suffers from stroke or not. This paper used a dataset retrieved from kaggle repository, which consists of 12 attributes (Features). This work is implemented using K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (TD), Neural Network (NN) and eXtreme Gradient Boosting (XGB) algorithms. Results showed eXtreme Gradient Boosting (XGB) gave the best result with an accuracy of up to 95.14%.

Keywords: Stroke disease, machine learning, Prediction, Classification Algorithms

I. INTRODUCTION

According to the World Stroke Organization[1] , 13 million people get a stroke each year, and approximately 5.5 million people will die as a result. It is the leading cause of death and disability worldwide, and that is why its imprint is serious in all aspects of life.

Stroke not only affects the patient but also affects the patient's social environment, family and workplace. In addition, contrary to popular belief, it can happen to anyone, at any age, regardless of gender or physical condition[2] . it is now possible to anticipate the onset of a stroke by utilizing ML techniques. the algorithms included in ML are beneficial as they allow for accurate prediction and proper analysis.

the majority of previous stroke-related research has focused on, among other things, the prediction of heart attacks. Brain stroke has been the subject of very few studies. The main motivation of this paper is to demonstrate how ML may be used to forecast the onset of a brain stroke.

the most important aspect of the methods employed and the findings achieved is that among the four distinct classification algorithms tested, eXtreme Gradient Boosting (XGB) fared the best, achieving a higher accuracy metric in comparison to the others. the implementation of four ML classification methods is shown in this paper.

Numerous academics have previously utilized machine learning to forecast strokes. [3] used data mining and a machine learning classifier to classify stroke disorders in 507 individuals. they tested a variety of machine learning methods for training purposes, including Artificial Neural Network (ANN), and they found that the SGD algorithm provided the greatest value, 95 percent. [4, 5] performed research to predict a stroke occurrence. They classified 50 risk variables for stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia, and alcohol consumption in 807 healthy and unhealthy individuals. they used two of the most accurate methods: the c4.5 decision tree algorithm (95 percent accuracy) and the K-nearest neighbor algorithm (94 percent accuracy). [6] conducted research to determine the predictability of a stroke patient death. they identified the stroke incidence using 15,099 individuals in their research. they detected strokes using a deep neural network method. the authors utilized PCA to extract information from the medical records and predict strokes. they have 83 percent area under the curve (AUC). [7] [conducted research using artificial intelligence to predict strokes. they employed a new technique for predicting stroke in their research using the

cardiovascular health study (CHS) dataset. Additionally, they used the decision tree method to do a feature extraction followed by a principal component analysis. In this case, the model was built using a neural network classification method, and it achieved 97 percent accuracy. [8]conducted research to determine the accuracy of an automated early ischemic stroke detection. The major objective of their research was to create a method for automating primary ischemic stroke using Convolutional Neural Network (CNN). they amassed 256 pictures for the purpose of training and testing the CNN model. they utilized the data lengthening technique to increase the gathered picture in their system's image preparation. their CNN technique achieved a 90 percent accuracy rate. [9] conducted research to establish a stroke severity index. they gathered data on 3577 patients who had an acute ischemic stroke. they utilized a variety of data mining methods, including linear regression, to create their predictive models. their ability to predict outperformed the k-nearest neighbor method (95% confidence interval). [10] used machine learning to predict the functional prognosis of an ischemic stroke. they tested this method on a patient who died three months after admission. they obtained an AUC value of greater than 90. [11] conducted research to determine the risk of stroke the authors of the research analyzed the data to predict strokes using Naive Bayes, decision trees, and neural networks. they assessed their pointer's accuracy and AUC in their research. they categorized all of these algorithms as decision trees, with naive Bayes providing the most accurate results. [12]
conducted research to determine the classification of an ischemic stroke. they categorized ischemic strokes using two models: the k-nearest neighbor method and the decision tree technique. In their study, the decision tree method was found to be more useful by medical experts when used to categorize strokes. The majority of studies had an accuracy rate of around 90%, which was considered to be quite good. However, the novelty of our research is that we used several well-known machine learning methods to get the best result. eXtreme Gradient Boosting (XGB), Random Forest (RF), Decision Tree (DT) and Naïve Bayes (NB) were the most successful algorithms, with 95.14, 90.72, 87.17, and 80.97 percent F1-scores, respectively. the accuracy percent of the models used in this research is much greater than the accuracy percent of the models used in previous investigations, suggesting that the models used in this investigation are more trustworthy. they have been shown to be resilient in many model comparisons, and the scheme may be generated from the results of the study's analysis. As mentioned earlier, the major contribution of this research is that we have used different machine learning models on a publicly available dataset. In the previous work, most of the researchers used a significant model to predict the stroke disease. However, we used four different models, and also, we compared the results with the previous work. All the results and comparisons are briefly discussed in the following section. The rest of this article is set out as follows: Materials and Methods are described in Section II; data set have been discussed in section III; the results and discussions are provided in Section IV and conclusions have been discussed in Section V.

II. Materials and Methods

Several algorithms were utilized to predict Stroke among which K nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (TD), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression (LR), , Random Forest (RF)and eXtreme Gradient Boosting(XGB). These algorithms are applied to a stroke dataset taken from the kaggle repository including 5110 samples (patients records). The dataset includes Stroke features. To enhance the performance of the algorithms, these features are analyzed, and the features' importance scores, Accuracy, Sensitivity, and Specificity are considered. the designed system's block diagram is shown in Figure 1.
All the components of the block diagram have been discussed in the following subsections.

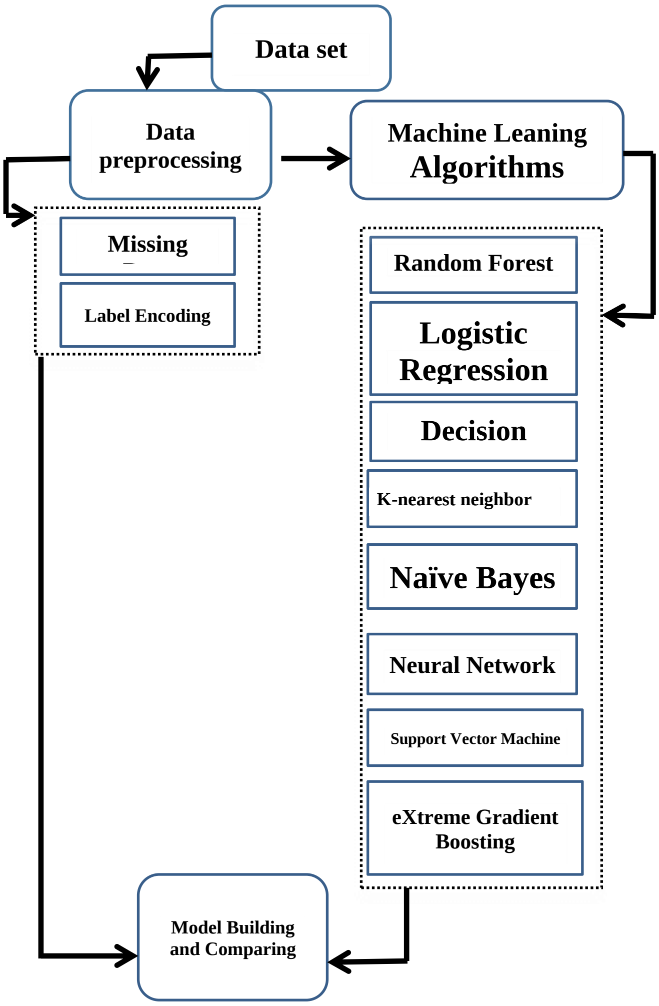

Figure 1: Proposed system's block diagra

A. K-nearest neighbour classifier (KNN)

The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows[13]. KNN calculations use the data and characterize new data points dependent on resemblance measures (e.g., distance function). The KNN calculation accepts that comparative things are close to one another. In KNN, Classification occurs by considering the majority vote to its neighbours. The data point goes to the class that has the most intimate neighbours. As we increment the number of nearest neighbours, the estimation of k and accuracy may increment.

B. Naïve Bayes classifier

The next algorithm is known as Naïve Bayes. it's also a supervised learning classification model, which classifies the info by computing the probability of independent variables. After calculating the probability of every class, the high probability class does assign for the entire transaction[14].and medical data mining[15]. It works by using the values for independent variables and predicting a predefined class for every record.

C. Decision Tree classifier

The decision tree algorithm has a tree structure. It divides the dataset into smaller subsets. a choice node has two or more branches. A resolution may be a target node. the basis node is the top node of the choice tree. This algorithm uses entropy and knowledge gain. Entropy is employed to calculate the homogeneity of a sample. Building a choice tree is about finding the attributes that give the biggest information gain. Finally, select the attribute with the very best information gain as the decision node and the zero universe branch as the end node[16].

D. Neural Network classifier

The neural network is an iterative process. It uses nonlinear data. The most goal is to reduce the difference between the actual production and the cost of the forecast production. A random weight is assigned to each of the entries. Then the corresponding performance is calculated and compared with the specified performance. The difference between them gives an error this algorithm minimizes the error with successive iterations by adjusting the input parameters. The advantages of neural network include adaptive learning, fault tolerance etc. Several Neural Network methodologies have been developed like the classification methodology called an artificial neural network, which may be a combination of a forward and backward propagation algorithm for predicting Stroke[17]. Many world problems can be solved using this methodology.

E. Support Vector Machine classifier (SVM)

Support Vector Machine (SVM) is one of the supervised learning methods, that can be widely used in statistical classification and regression analysis. SVM belongs to generalized linear classifiers, which are characterized by their ability to minimize empirical error and maximize geometric edge region at the same time. Therefore, another name for SVM is the maximum edge region classifier[18].

F. Logistic Regression classifier

Logistic regression is another kind of classification model, which learn and predict the parameters in the given dataset using regression analysis[19]. The learning and prediction processes are based on measuring the probability of binary classification. The logistic regression model requires class variables that should be binary classified. Likewise, in this dataset the target column has two types of binary numbers, "0" for the patient who has no chances of Stroke, and "1" for the patients who have been predicted as Stroke patients. On the other side, the independent variables can be binary classified, nominal, or polynomial types.

G. Random forest classifier

The random forest is the next model chosen and implemented in this study. Since this model belongs to the classification family, it is also known as the teacher training algorithm. In the training phase, this model first generates a few random trees called a forest[20]. for example, if the dataset contains an "x" number of attributes, first select some features, randomly named "y". Using all possibilities; (ie "and"), create nodes using the simpler rift method. Also, the algorithm will work to create the entire forest by repeating the steps above. Then, in the forecasting process, the algorithm tries to shuffle the trees using the estimated result and the voting procedure. the goal of combining random trees by voting in the forest is to eliminate the use of the highest predicted tree, which can improve the accuracy of forecasts for future data.

H. eXtreme Gradient Boosting(XGB)

eXtreme Gradient Boosting (XGBoost) is a scalable and improved version of the gradient boosting algorithm (terminology alert) designed for efficacy, computational speed and model performance. It is an open-source library and a part of the Distributed Machine Learning Community. XGBoost is a perfect blend of software and hardware capabilities designed to enhance existing boosting techniques with accuracy in the shortest amount of time.[21]

III. DATASET DESCRIPTION

The dataset used in this research paper was collected from the UCI platform. This dataset contains a set of features associated with stroke-like age, type of work, type of residence, history of high blood pressure and heart disease, body mass index (BMI) and of course each has its own data type.

In this dataset, there are more than 5,000 records of people, some of whom are infected and some of them are healthy.

The data set was analyzed using the Python language to obtain a greater understanding of the existing data, as it turned out that it needs a set of processing operations in order to become more balanced and give highly reliable results.

Table 1: Description of Features

Data #	Feature	Description
1	ID	unique identifier
2	gender	"Male", "Female" or "Other"
3	age	age of the patient
4	hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5	heart_disease	0 if the patient doesn't have any heart diseases,

		1 if the patient has a heart disease
6	ever_married	"No" or "Yes"
7	work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8	residence_type	"Rural" or "Urban"
9	avg_glucose_level	average glucose level in blood
10	BMI	body mass index
11	smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"*
12	stroke	1 if the patient had a stroke or 0 if not

IV. Results and Discussions

In this research, a set of techniques were used to obtain the best performance of the previously mentioned models to reach the best model. Where the data set was loaded and then a set of operations were performed to analyze and process the data, and as a result of these operations the ID feature was deleted as it has no effect as a serial number, and the empty data were also processed, and after the analysis and processing process it became clear that the data set Unbalanced and therefore will not give the desired results, and to address the imbalance of the data set, the SOMTE algorithm was used, which balances the data set, after that the data set was separated into the training set and the test set, and the SMOTE algorithm was applied to each set separately, and after applying the prediction algorithms, the results shown in the following table:

Table 2: Results of used algorithms

Model Name	Accuracy %	Sensitivity %	Specificity %
K-nearest neighbour classifier (KNN)	70.00	59.59	80.41
Naïve Bayes	**80.97**	**90.55**	**71.38**
Decision Tree	**87.17**	**78.41**	**95.93**
Neural Network	76.38	74.28	78.48
Support Vector Machine (SVM)	58.62	24.69	92.55
Logistic Regression	79.62	84.69	74.55
Random Forest	**90.72**	**82.07**	**99.38**
eXtreme Gradient Boosting(XGB)	**95.14**	**90.34**	**98.48**

Moreover, the following figure shows the results of the algorithms used in terms of accuracy, sensitivity, and specificity.

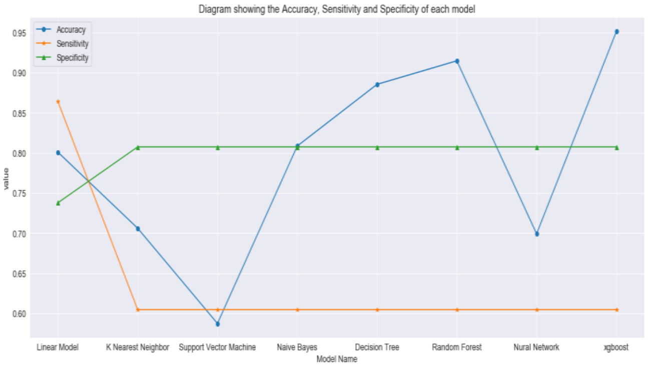

Figure 2: Showing the Accuracy, Sensitivity and Specificity of each model

It is clear from the previous table that the eXtreme Gradient Boosting (XGB) algorithm achieved the best accuracy with a rate of 95.14%, followed by the Random Forest algorithm with an accuracy rate of 91.48%, and then came the Decision Tree algorithm with an accuracy rate of 88.55%.

According to Tables 3, 4 and 5, age is important feature in stroke detection and prediction. In addition to gender and residence_type.

Table 3: Shows Important Features for xgbModel

xgbModel	
Feature	**importance**
age	0.07068
smoking_status	0.05324
gender	0.04013
avg_glucose_level	0.03275
Residence_type	0.03002
work_type	0.02559
bmi	0.02495
hypertension	0.01369
heart_disease	0.0102
ever_married	0.00583

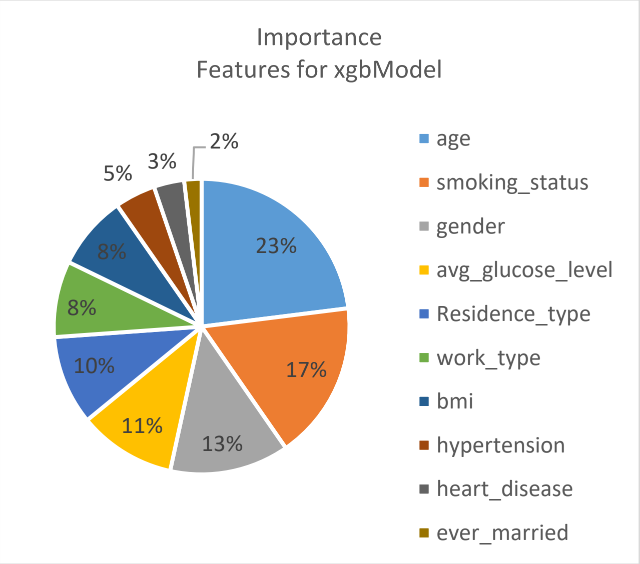

Figure 3:Imprtant Features for xgbModel

Table 4: Shows Important Features for Random Forest Model

Random Forest Model	
Feature	**importance**
age	0.08722
Residence_type	0.06271
gender	0.05957
hypertension	0.05119
work_type	0.03709
avg_glucose_level	0.03111
bmi	0.01991
smoking_status	0.01955
ever_married	0.01829
heart_disease	0.01478

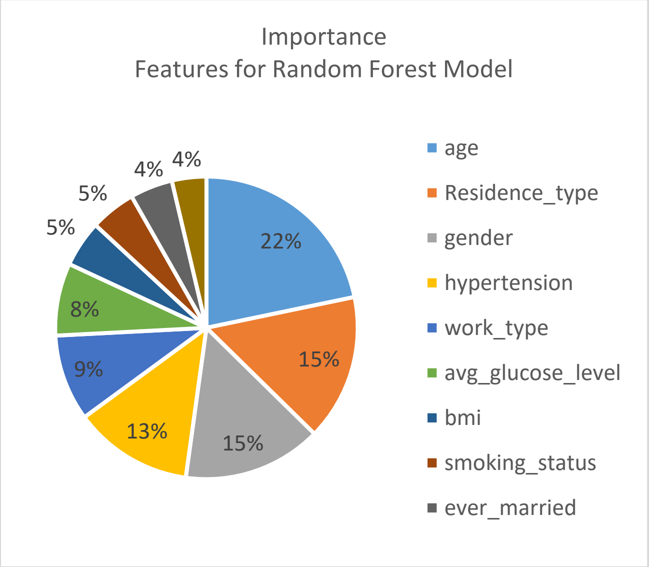

Figure 4:Imprtant Features for Random Forest Model

Table 5:Shows Imprtant Features for Decision Tree Model

Decision Tree Model	
Feature	**Importance**
Age	0.22298
Residence_type	0.12718
Gender	0.08493
work_type	0.07986
avg_glucose_level	0.06834
smoking_status	0.06775
Hypertension	0.04855
ever_married	0.0414
Bmi	0.03093
heart_disease	0.00607

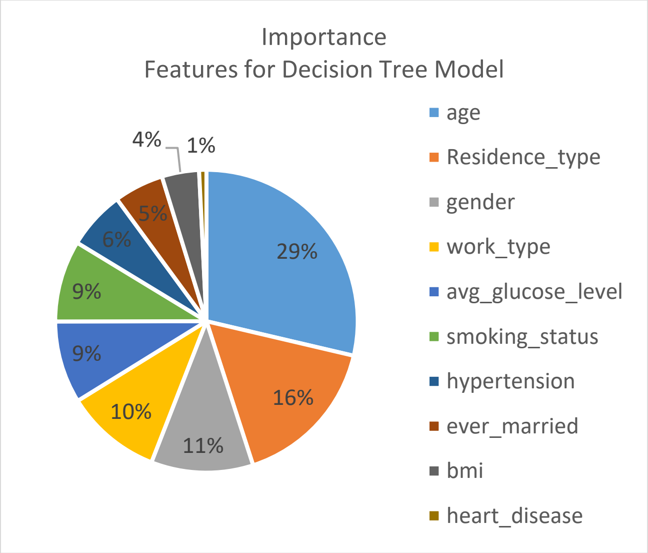

Figure 5: Important Features for Decision Tree Model

Table 6 shows the four main features based on feature importance and correlation value for the three best algorithms:

Table 6:Shows Feature ranking for Stroke

#	**Best Algorithms**		
	eXtreme Gradient Boosting(XGB)	**Random Forest**	**Decision Tree**
1ST	age	age	Age
2nd	smoking_status	Residence_type	Residence_type
3rd	gender	gender	Gender
4th	avg_glucose_level	hypertension	work_type

V. CONCLUSION

Machine learning techniques help to reduce the effort and time for medical officers to conduct early predictions for healthcare management purposes. Stroke is a life-threatening medical illness that should be treated as soon as possible to avoid further complications. As the number of deaths increases due to stroke, a machine learning technique system can help predict stroke accurately an effectively. This paper showed that applying machine learning techniques in making early predictions of stroke may have the potential to improve the healthcare management system.

After comparing the algorithms that were used in the process of predicting stroke, Neural Network, Support Vector Machine and Logistic Regression seem to achieve the best performance score compared to other techniques. As a scope of future work, a hybrid of machine learning techniques with optimization algorithms with more data will be examined. This will help in increasing the accuracy.

REFERENCES

1. Elloker, T. and A.J. Rhoda, The relationship between social support and participation in stroke: A systematic review. African Journal of Disability, 2018. **7**(1): p. 1-9.
2. Govindarajan, P., et al., Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 2020. **32**(3): p. 817-828.
3. Amini, L., et al., Prediction and control of stroke by data mining. International journal of preventive medicine, 2013. **4**(Suppl 2): p. S245.

4. Cheng, C.-A., Y.-C. Lin, and H.-W. Chiu. Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. in ICIMTH. 2014.
5. Reza, S.M., M.M. Rahman, and S. Al Mamun. A new approach for road networks-a vehicle xml device collaboration with big data. in 2014 International Conference on Electrical Engineering and Information & Communication Technology. 2014. IEEE.
6. Cheon, S., J. Kim, and J. Lim, The use of deep learning to predict stroke patient mortality. International journal of environmental research and public health, 2019. **16**(11): p. 1876.
7. Zheng, W., Y.-H. Chen, and M. Sawan. Longitudinal Data to Enhance Dynamic Stroke Risk Prediction. in Healthcare. 2022. MDPI.
8. Chin, C.-L., et al. An automated early ischemic stroke detection system using CNN deep learning algorithm. in 2017 IEEE 8th International conference on awareness science and technology (iCAST). 2017. IEEE.
9. Sung, S.-F., et al., Developing a stroke severity index based on administrative data was feasible using data mining techniques. Journal of clinical epidemiology, 2015. **68**(11): p. 1292-1300.
10. Monteiro, M., et al., Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018. **15**(6): p. 1953-1959.
11. Kansadub, T., et al. Stroke risk prediction model based on demographic data. in 2015 8th Biomedical Engineering International Conference (BMEiCON). 2015. IEEE.
12. Adam, S.Y., A. Yousif, and M.B. Bashir, Classification of ischemic stroke using machine learning algorithms. International Journal of Computer Applications, 2016. **149**(10): p. 26-31.
13. Bashir, S., et al. Improving heart disease prediction using feature selection approaches. in 2019 16th international bhurban conference on applied sciences and technology (IBCAST). 2019. IEEE.
14. Razaque, F., et al. Using naïve bayes algorithm to students' bachelor academic performances analysis. in 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS). 2017. IEEE.
15. Qasim, O. and K. Al-Saedi, Malware Detection using Data Mining Naïve Bayesian Classification Technique with Worm Dataset. Int. J. Adv. Res. Comput. Commun. Eng, 2017. **6**(11): p. 211-213.
16. Rani, K.U., Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626, 2011.
17. Ambati, N.S.R., et al. Performance Enhancement of Machine Learning Algorithms on Heart Stroke Prediction Application using Sampling and Feature Selection Techniques. in 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS). 2022. IEEE.
18. Wang, J. Heart Failure Prediction with Machine Learning: A Comparative Study. in Journal of Physics: Conference Series. 2021. IOP Publishing.
19. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, Applied logistic regression. Vol. 398. 2013: John Wiley & Sons.
20. Bashar, S.S., et al. A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm. in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). 2019. IEEE.
21. Malik, S., R. Harode, and A. Singh, XGBoost: A Deep Dive into Boosting (Introduction Documentation). 2020.