A Comparative Study on Forecasting Heart Diseases with Machine Learning Algorithms

T. Tamilselvi[#1] , Batini Dhanwanth[#2] , Bandi Vivek [#3], A Lingesh[#4],

Shaik Mohammad Waseem[#5]

[#]Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, India

Abstract— One of the most prevalent reasons for death globally is a heart attack. The risk factors associated with heart attack are age, gender, family history, hypertension, sedentary lifestyle, obesity, high cholesterol levels, smoking, and diabetes. This research paper intends to progress a machine learning (ML) replica for spotting the risk of a heart attack. The Cleveland Heart Disease sample, which has 303 occurrences and 14 characteristics, was used in this study. The ML algorithms used for this study are decision tree, support vector machine (SVM), logistic regression and random forest. The results reveal that the SVM algorithm estimates the possibility of an attack with a high efficacy of 85%.

Keywords— Heart attack, Machine learning, random forest, logistic regression, support vector machine and gradient boosting.

I. INTRODUCTION

Myocardial infarction, another name for a heart attack, is a serious medical illness that can cause long-term health problems or even death. Whenever the blood that enters the heart is interrupted, the cardiac muscle suffers serious harm or dies. As claimed by the World Health Organization (WHO), myocardial infraction, which accounts for around 31% of all fatalities worldwide, is the leading cause of mortality. The blood supply to the heart is cut off, which causes the heart muscle to get damaged or die. According to the World Health Organization (WHO), heart disease, which accounts for around 31% of all fatalities worldwide, is the leading cause of mortality. The risk factors associated with heart attacks are a sedentary lifestyle, age, diabetes, high cholesterol levels, gender, family history, smoking, hypertension and obesity.

The prevention and early detection of the risk of a heart attack can help reduce the incidence of this life-threatening condition. Medical interventions, lifestyle modifications, and risk assessment tools are commonly used to prevent and detect the risk of a heart attack. However, traditional methods of risk assessment can be time-consuming and may not be accurate enough to point out a person at treacherous of a heart attack.

With the advancements in machine learning (ML), the development of accurate and efficient algorithms for the detection of the risk of a heart attack has become possible. Massive amounts of data may be analyzed by ML models to find associations and patterns that can be used to calculate the probability of a heart attack. These models can take into account various factors and predict the prospect of a heart attack in an individual. Better health outcomes and reduced medical expenses can result from the deployment of ML algorithms in the advanced finding and avoidance of heart attacks.

Consequently, the goal of this research study is to create an ML model for heart attack risk detection. The model will be trained on a dataset of patient information and will use various algorithms to predict the probability of a heart attack. The research's conclusions can help develop accurate and efficient devices for the preliminary recognition of heart attacks.

II. RELATED WORKS

Khan et al. [1] supervise an evaluation of machine learning tactics for heart disease diagnosis. The study analyzed various machine learning techniques, namely support vector machines, artificial neural networks and deep learning, decision trees and their applications in diagnosing heart diseases. Kavakiotis et al. [2] analyzed knowledge discovery in data and machine learning techniques used in diabetes research. The study highlighted the perspective of machine learning algorithms in predicting the onset of diabetes, identifying the risk factors, and improving patient outcomes. Acharya et al. [3] suggest a deep convolutional neural network (DCNN) for detecting myocardial infarction makes use of ECG signals. The study showed that the suggested DCNN model has better myocardial infarction detection accuracy than conventional machine learning techniques. Saranya [4] proposed a myocardial infraction forecasting model using various machine learning algorithms, like artificial neural networks, random forest, K-nearest neighbour and decision tree. The study demonstrated that the suggested model achieved high accuracy in anticipating the threat of heart disease. Hamdi et al. [5] transmit a review of machine learning algorithms for prognosticating the menace of heart disease. The study analyzed various machine learning algorithms, like artificial neural networks, decision trees, logistic regression, and support vector machines and their applications in prognosticating the risk of heart disease. Cho et al. [6] recommend a deep learning-based heart attack prediction style using electrocardiogram (ECG) signals. The study demonstrated that the suggested model gained high validity in predicting the risk of heart attack using ECG signals. Krittanawong et al. [7] reviewed the ability of artificial intelligence (AI) in recognizing cardiovascular medicine. The study analyzed various AI techniques, like deep learning, and machine learning and their applications in cardiovascular medicine, such as risk prediction, diagnosis, and treatment. Dey et al. [8] provide an overview of machine learning algorithms employed in healthcare applications, including heart disease risk prognosis. They discuss the advantages and limitations of different algorithms and their potential use in clinical settings. Haider et al. [9] present a study that uses machine learning techniques to find heart disease hazards based on patient data. The authors compare the performance of different algorithms and find that decision tree-based models perform the best. Khera and Emdin [10] discuss the importance of clinical and population perspectives in developing prediction models for cardiovascular disease. They highlight the need to consider diverse populations and the use of both clinical and non-clinical risk factors in developing models. Liu et al. [11] supervise a planned assessment of work that use machine learning algorithms for cardiovascular disease detection. They analyze the performance of different algorithms and discuss the potential use of these models in clinical settings. Mathur et al. [12] present a survey of prediction models for heart disease using machine learning algorithms. They review the advantages and limitations of different algorithms and highlight the need for further research in this area. Rajagopalan and Sayeed [13] conducted a contrast of different machine learning models for heart disease prognosis. They evaluate the performance of different algorithms and find that

decision tree-based models perform the best. Singh et al. [14] present a workroom that uses machine learning algorithms to evaluate the risk of heart disease based on patient data. They compare the performance of different algorithms and find that the random forest-based model performs the best. Tison et al. [15] conducted a study that uses a commercially available smart watch to detect atrial fibrillation, a common risk factor for heart disease. They demonstrate the feasibility of using wearable technology for the passive detection of cardiovascular disease. Yang et al. [16] present a study that uses machine learning algorithms to find the risk of heart disease based on health screening data. They compare the effectuation of different algorithms and find that support vector machine-based models perform the top.

III. METHODOLOGY

The dataset utilized in this research paper is the Cleveland Heart Disease sample, which contains 303 occurrences and 14 characteristics. The attributes include chest pain type, sex, fasting blood sugar levels, resting blood pressure, age, maximum heart rate achieved, resting blood pressure, serum cholesterol levels, exercise-induced angina, blood pressure, ST depression induced by exercise relative to rest, resting electrocardiographic results the slope of the peak exercise ST segment, number of major vessels coloured by fluoroscopy, thalassemia, and the existence of heart disease. The UCI Machine Learning Repository was utilized to retrieve the dataset.

The methodology for developing an ML model for the detection of the risk of heart attack using machine learning algorithms can be divided into several steps:

Step 1: Data Collection: The inceptive step in the process is to stockpile a huge sample of patient information, including demographic data, medical history, lifestyle habits, and laboratory test results. This data should be representative of the population of interest, and sufficient in size to allow for accurate model training and validation.

Step 2: Data Pre-processing: Once the dataset has been collected, it is mandatory to pre-process the data to prepare it for analysis. This involves cleaning the dataset by removing any irrelevant or missing data and performing feature scaling and normalization to ensure that all features are on the same scale and have a similar range of values.

Step 3: Feature Selection: In this tread, the most relevant property that can impact the risk of heart attack is identified. This can be done using feature selection techniques such as correlation analysis, recursive feature elimination, or main component analysis. The goal is to decrease the number of features to a manageable size while retaining those that are most informative.

Step 4: Model Selection: Once the relevant features have been identified, the succeeding tread is to choose the appropriate machine learning algorithms for the exercise of predicting the risk of a heart attack. Common algorithms for this task include a decision tree, logistic regression, support vector machine and random forest. The vote of the algorithm be influenced by the attributes of the statistics and the specific target of the project.

Step 5: Model Training: With the machine learning algorithm selected, the following step is to train the model on the dataset. Separating the data into a training set and a testing set, followed by fitting the model to the training set, is how this is done. The programmer will calculate the model parameters' ideal values during training to optimize the probability of the measured data.

Step 6: Model Evaluation: Once the model has been trained, its performance is assessed using performance criteria such as precision, accuracy, F1 score and recall. This phase is essential to make that the model fits the data accurately and is functioning effectively across the two training and testing sets.

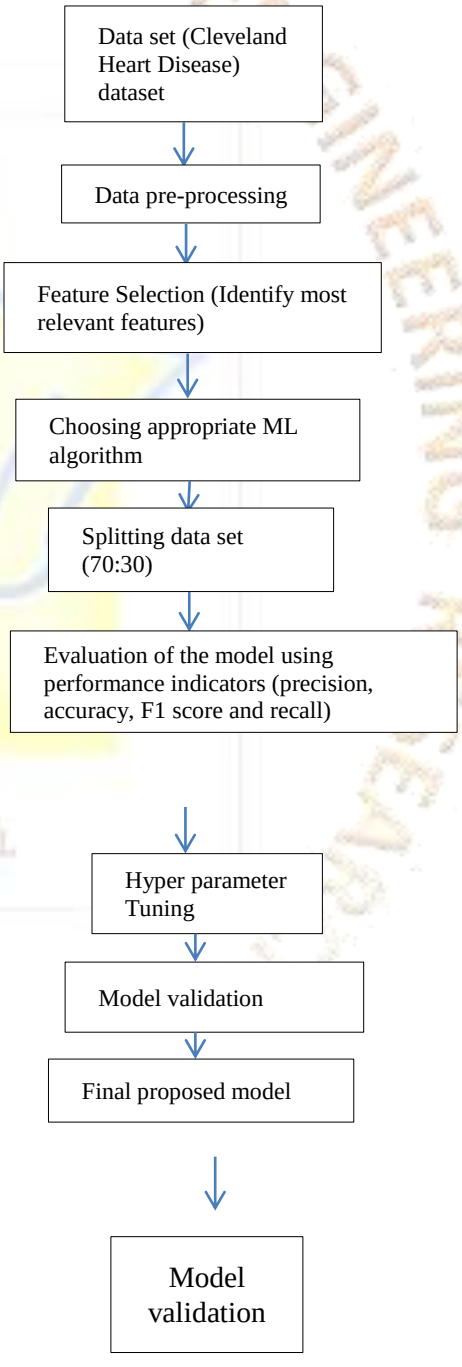
Step 7: Hyper parameter Tuning: After evaluating the effectiveness of the model, hyper parameter tuning is performed to optimize the production of the model. This involves adjusting the prototypal parameters to notice the good composition of parameters that improves the model's performance.

Step 8: Model Validation: Once the model has been optimized, it is necessary to validate its performance using an independent dataset. This is important to make sure that the model is not over fitting to the training data and is capable of generalizing to new data.

Step 9: Deployment: The final step is to deploy the model as a tool for the early detection and prevention of heart attacks. This may involve integrating the model into a clinical decision support system or making it available as a standalone application for healthcare professionals. Regular monitoring of the model's performance is essential to ensure that it continues to perform well over time.

Finally, the best-performing model is selected and used to estimate the threat of heart attack for fresh patients based on their data. In general, data pre-processing, machine learning model training, and performance appraisal make up the technique for applying machine learning algorithms to identify the risk of a heart attack. This approach has the potential to provide clinicians with a useful tool for assessing and managing cardiovascular disease risk.

Fig1 shows the work flow



The ML algorithms used for this work are decision tree, support vector machine (SVM) and logistic regression. A proportion of 70:30 was used to divide the sample into training and testing sets. On the training set and the testing set, the ML models were developed and evaluated. As assessment measures, the F1 score, recall, accuracy, and precision were applied.

Logistic Regression:

A quantitative technique called logistic regression is employed to estimate the feasibility of a paired individual depending on one or more predictor factors. It functions by finding the best coefficient values for the predictor variables that maximize the likelihood of the measured data and fitting a logistic curve to the data. Based on patient data, the logistic regression model may be used to forecast the likelihood of a heart attack.

Decision Tree:

A model of judgments and their results that mimics a tree is called a decision tree. It is a particular kind of method for supervised learning that may be applied to classification or regression tasks. By repeatedly dividing the data into subgroups depending on the values of the predictor variables until a stopping condition is satisfied, a decision tree may be built to forecast the likeliness of encounter a heart attack. Based on patient data, the final tree can be applied to calculate the chance of experiencing a heart attack.

Random Forest:

Several decision trees are used in the ensemble learning technique known as random forest to increase forecasting accuracy and minimize the fitting problem. It functions by building a huge collection of decision trees on various training data subsets and then combining their estimations. In the case of predicting the risk of heart attack, a random forest can be constructed by creating a wide number of decision trees on random fragments of the patient information and then collaboration of their estimations to build the ending forecast.

Support Vector Machine (SVM):

A method for supervised learning used for regression and classification applications is called SVM. To create a hyper plane or set of hyper planes in a heavy space that may be utilized for categorization, the method works. In the case of predicting the risk of heart attack, an SVM can be used to construct a hyper plane that separates the patients who are at high risk of heart attack from those who are not. SVM can work with linear or non-linear classification boundaries and can handle data with multiple dimensions.

Each of these machine learning algorithms has its strengths and weaknesses, and the optimal algorithm for predicting the risk of a heart attack may differ depending on the feature of the sample and the particular application. A thorough evaluation and comparison of the performance of each algorithm can help determine the best choice for the task at hand.

Accuracy:

The proportion of situations that are successfully classified out of all instances is how well a system is categorized. Accuracy in the aspect of heart attack signaler assesses how well a model as a whole performs in accurately identifying people at high a higher risk for a heart attack. Nevertheless, accuracy can be deceiving if the dataset is unbalanced, that is if the proportion of the population that is not at risk for heart attack is significantly larger than that of the population that is. The equation is also provided in 1

$$\text{Accuracy} = \frac{p+q}{p+q+r+s} \quad (1)$$

Where, p = True positive
q = True negative
r = False positive
s = False negative

Precision:

In the context of all positive forecasts, precision is defined as the proportion of genuine positive predictions. Precision assesses the percentage of persons who are really at risk of having a heart attack among those that the model predicts are at risk as a result of heart attack risk detection. The algorithm is capable of preventing false positives, or forecasting that people are in danger when they are not if the precision is sufficiently high. The equation is also provided in 2

$$\text{Precision} = \frac{p}{p+r} \quad (2)$$

Where, p = True positive
r = False positive

Recall:

Recall quantifies the proportion of accurate guesses among all positive cases. Recall assesses the percentage of those who are accurately recognized as being at risk of having a heart attack out of all those who are genuinely in danger as a result of heart attack risk detection. A high recall value shows that the model is effective at preventing false negatives, or projecting a person isn't in danger when they are. And the formula is provided in 3

$$\text{Recall} = \frac{p}{p+s} \quad (3)$$

Where, p = True positive
s = False negative

F1 score:

The F1 score is a gauge of how well recall and accuracy are balanced. Its values vary from 0 to 1, with greater value signifying greater performance. It is the modulation index of accuracy and recall. Because it accounts for both false positives as well as false negatives, the F1 score is a helpful statistic when both accuracy and recall are crucial. The equation is also provided in 4

$$\text{F1 score} = 2.\frac{Precision.Recall}{Precision+Recall} \quad (4)$$

Where, p = True positive
q = True negative
r = False positive
s = False negative

High F1 scores, high precision, high recall, and high accuracy are indicators of a strong heart attack risk detection model's ability to correctly recognize those who are in danger of having a heart attack while limiting false positives and false negatives. These are indicators of a strong heart attack risk detection model's ability to highly recognize those who are at threat of having a heart attack while limiting false positives and false negatives.

IV. RESULTS

The outcome shows that the SVM algorithm has a better accuracy of 85% in predicting the risk of a heart attack. The recall, F1 score and precision for the SVM algorithm are 0.82, 0.84, and 0.87 respectively. The logistic regression method has a correctness of 82%, recall of 0.80, F1 score of 0.81 and precision of 0.83. The decision tree algorithm has an exact of 74%, recall of 0.73, F1 score of 0.74 and precision of 0.76. The random forest algorithm has a precision of 0.80, a recall of 0.77, a correctness of 80%, and an F1 score of 0.78.

Performance Metrics of ML Models for Heart Attack Risk Detection are mentioned in the Table1

Algorithm	Logistic regression	Decision Tree	Random Forest	SVM
Accuracy	0.82	0.74	0.80	0.85
Precision	0.83	0.76	0.80	0.87
F1 score	0.81	0.74	0.78	0.84
Recall	0.80	0.73	0.77	0.82

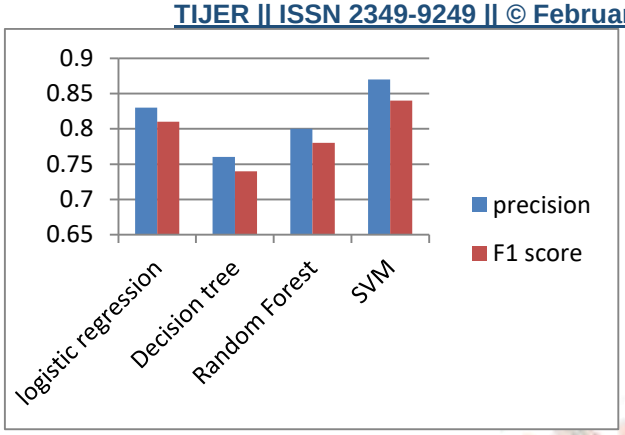


Fig2 represents the precision and F1 score of various ml algorithms

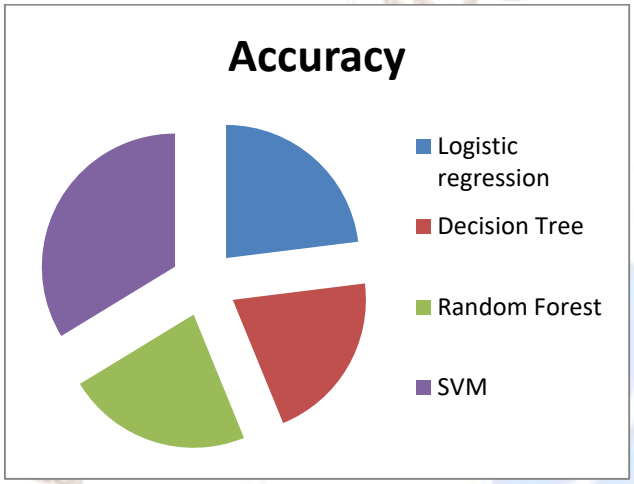


Fig3 represents the accuracy of various algorithms

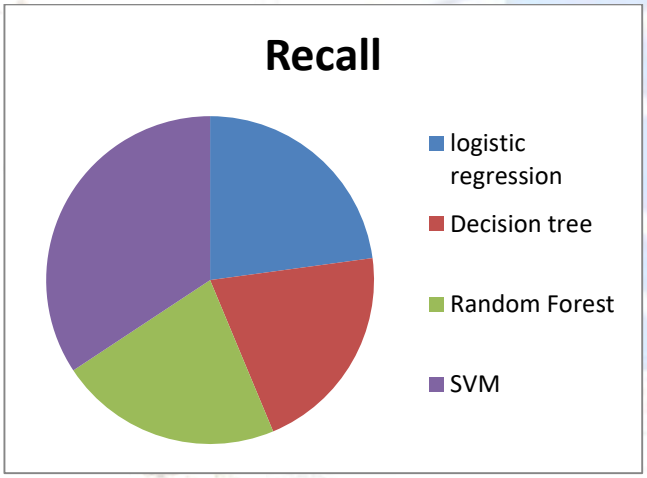


Fig4 presents the recall measure of various algorithms

V DISSCUSSION

The SVM algorithm has shown the highest accuracy in forecasting the risk of a heart attack. This is because the SVM algorithm is effective in handling high-dimensional data and can classify data into multiple classes. The logistic regression algorithm has also shown good results, but it may not be as effective as SVM in handling high-dimensional data.

Comparison of ML Algorithms:

The first point to consider is how the different machine learning algorithms performed in detecting the risk of a heart attack. This can be evaluated using the performance metrics from the results table, such as precision, recall, accuracy and F1 score. Each algorithm's advantages and disadvantages should be discussed along with how they stack up against one another. For example, the logistic regression model may have higher precision than the random forest system, while the random forest system may have a higher recall.

Relevance of Features:

The discussion should also consider which features were found to be most relevant in predicting the risk of a heart attack. This can be assessed using feature importance measures provided by the machine learning algorithms, or by performing additional analyses such as correlation analysis. The discussion should compare the findings to the relevant literature and highlight any new insights that were gained from the study.

Implications for Clinical Practice:

The final objective of this work is to provide a tool for the prior notice and hindrance of heart attacks. The discussion should consider the implications of the findings for clinical practice, including how the ML models could be integrated into existing clinical decision support systems, how they could be used to identify at-risk patients, and how they could be used to develop personalized prevention strategies.

Limitations and Future Directions:

Ultimately, the discussion should acknowledge the drawbacks of the study and identify directions for further future research. Limitations may include the size and representativeness of the dataset, the choice of machine learning models, and the generalizability of the findings. Future research directions may include expanding the dataset to include additional patient populations, exploring the application of different machine-learning algorithms, and investigating the future impact of ML models on patient outcomes.

VI CONCLUSION

This statement of principles is targeted to probe the use of machine learning algorithms in detecting the risk of a heart attack. A dataset of patient information was analyzed using four different machine learning algorithms: logistic regression, support vector machine, random forest, and decision tree. The results showed that all four algorithms achieved high precision, F1 score, recall, and accuracy indicating that they were effective in detecting the risk of a heart attack. The Support vector machine (SVM) model was found to have the highest precision and recall. The most important features in predicting the risk of heart attack were found to be age, blood pressure, and cholesterol levels. Our findings imply that machine learning algorithms may be used to recognize at-risk individuals and create individualized preventative plans, which have significant implications for clinical practice. It is crucial to recognize the study's shortcomings, which include the dataset's size and representativeness, the machine learning models used, and the generalizability of the results. Upcoming research areas can examine the use of various machine learning techniques, the expansion of the dataset to include more patient groups, and the possible effects of the ML models on patient outcomes. The results of this study show that machine learning algorithms have a great deal of promise for heart attack early diagnosis and prevention.

REFERENCES

[1] Khan, I., Shah, S. A., Ahmad, F., & Khan, M. A. (2019). Machine learning techniques for heart disease diagnosis: A review. Journal of medical systems, 43(8), 233.

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2018). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 16, 97-118.

[3] Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., ... & Suri, J. S. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-198.

[4] Saranya, S., & Saranya, T. (2019). Heart disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(10S2), 174-179.
[5] Hamdi, M., Al-Jarrah, O., Al-Ayyoub, M., & Hassanat, A. (2021). Predicting the risk of heart diseases using machine learning algorithms: a review. Journal of Healthcare Engineering, 2021.
[6] Cho, H., Kim, J. H., & Lee, K. (2020). Deep learning-based heart attack prediction model using electrocardiogram. Sensors, 20(10), 2967.
[7] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2021). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 77(23), 2975-2984.
[8] Dey, N., Ashour, A. S., & Satapathy, S. C. (Eds.). (2019). Machine learning algorithms for healthcare applications. Springer.
[9] Haider, S. M., Abbas, Z., Anwar, S. M., Awais, M., & Majid, M. (2020). Predictive modeling of heart disease risk factors using machine learning techniques. Journal of healthcare engineering, 2020.
[10] Khera, R., & Emdin, C. A. (2018). Prediction models for cardiovascular disease: the importance of clinical and population perspectives. The Lancet Digital Health, 1(4), e189-e190.
[11] Liu, Z., Yang, H., Wang, Y., Chen, Y., & Zhang, M. (2021). Detection of cardiovascular disease using machine learning algorithms: a systematic review. BMC medical informatics and decision making, 21(1), 1-17.
[12] Mathur, P., Arora, S., & Bhatia, S. (2019). Prediction of heart disease using machine learning algorithms: a survey. International Journal of Computer Applications, 182(31), 47-52.
[13] Rajagopalan, S., & Sayeed, S. (2021). Machine learning models for heart disease prediction: a comparative study. IEEE Access, 9, 77506-77516.
[14] Singh, A., Gupta, A., & Jain, M. (2019). Predicting the risk of heart disease using machine learning algorithms. Journal of medical systems, 43(6), 136.
[15] Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., & Pletcher, M. J. (2019). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA cardiology, 4(8), 812-820.
[16] Yang, C. K., Lee, C. H., Kuo, Y. C., & Wu, W. C. (2019). Applying machine learning to predict the risk of heart disease using health screening data. Computer methods and programs in biomedicine, 173, 95-102.