Research Article

Cardiovascular Disease Prediction Using Artificial Intelligence

Ayodele Martin Dossou,[1] Prof. Dr. Ali Okatan[2]

[1]Prof. Dr. at Department of Civil Engineering, Istanbul Arel University, Istanbul, mehmetfatihaltan@arel.edu.tr

[2]Department of Civil Engineering, Istanbul Aydin University, Istanbul, hakkicanalcan@stu.aydin.edu.tr
ORCID: 0000-0002-1329-61760000-0002-1329-6176

ABSTRACT

The likelihood of contracting a disease rises with the size of the human population. Globally, there are numerous ailments, and one of the main issues facing hospital systems today is the lack of technology to detect illness in patients. Cardiovascular disease, or CVD, is one such illness. Any cardiovascular, vascular, or blood vessel ailment is referred to. More people globally die from CVDs than from any other cause, according to the WHO. More so in low- and middle-income nations. When ill, it can be quite difficult for persons who live alone to contact the hospital. As a result, we created a model that can recognize when a patient is ill and send a report to the hospital. Currently, the system merely detects and informs the hospital about patients with heart disease. We chose to focus on heart disease detection because it's one of the worst diseases and there's a significant chance that people may pass away from it. Predicting whether a patient has cardiac disease or not is a categorization issue. We consider several variables, including age, blood sugar level, cholesterol level, and many more, and then we provide the result based on the input.

Keywords:*Cardiovascular Disease, Heart Disease Detection, Illness Recognition, Hospital Reporting, Categorization.*

[1]Department of AI & Data Science, Istanbul Aydin University, Istanbul, mammarkhan@stu.aydin.edu.tr

[2]Prof. Dr. at Department of AI & Data Science, Istanbul Aydin University, Istanbul, aliokatan@aydin.edu.tr,ORCID: 0000-0002-8893-9711-0

1. INTRODUCTION

The research primarily focuses on the numerous category segregation techniques used to forecast cardiac disorders. A poor lifestyle, drinking alcohol, eating a lot of fat, triggering hypertension, and not getting enough exercise all contribute to heart disease. The heart of a human controls blood flow throughout the body. The majority of the human body is made up of it. Heart irregularities are a severe cause for concern because they have an impact on many different body organs. Heart illness can be thought of as irregularities or abnormalities in how the heart usually beats. Heart congestion and disease are the primary causes of the majority of fatalities in today's fast and busy environment. [1] The WHO estimates that more than 10 million people worldwide pass away each year as a result of heart disease. The only strategies to stop heart-related illnesses are through early identification and a healthy lifestyle. Today's world makes it exceedingly challenging for hospitals to provide individualized care to every patient in need because of the growing population. In light of this, we have decided to try to notify the hospitals when a patient has an illness. We generally concentrate on heart illness [2] because there is a very high possibility that a patient would suffer a serious injury or pass away as a result of it. Additionally, it shortens the time it takes to detect cardiac disease. This initiative is crucial because, if a patient is living alone and suffering from heart illness, he may not be able to ask the hospital staff for assistance. Our effort goes a long way towards assisting such people. We faced difficulties and hurdles because we lacked sufficient data sets to produce pleasing results, which we overcame by producing synthetic data. To train the algorithm and determine if the patient is ill or not, we employ machine learning modules. Even if our experiments produced solid results, we still need to put the system into practice, which means creating an app that collects patient data and sends the output to hospitals Future implementation of this is possible.

Cardiovascular detection using deep learning has been a widely researched topic in recent years. Several research studies have been conducted to investigate the feasibility and accuracy of using deep learning models for cardiovascular disease detection. Some of the related works in this area are:

"Cardiovascular Disease Detection Using Deep Learning: A Review" by Liang et al. [4] (2020): This paper provides a comprehensive review of the recent advancements in cardiovascular disease detection using deep learning. The authors discuss the different types of cardiovascular diseases and the deep-learning techniques used to detect them. They also provide an overview of the datasets used in these studies and the performance metrics used to evaluate the models."Automated Detection of Cardiovascular Disease Using Deep Learning Techniques" by Attia et al. (2019): This study [5] used a convolutional neural

network (CNN) to classify patients with and without cardiovascular disease based on **t**heir echocardiogram images. The authors achieved an accuracy of 80.8% in detecting cardiovascular disease, which outperformed the accuracy of traditional machine learning models. "Cardiovascular Disease Prediction using Deep Learning Algorithm" by Gupta et al. (2020): In this study [6], the authors used a deep learning model to predict cardiovascular disease based on demographic, lifestyle, and medical history data. The authors achieved an accuracy of 88.23% in predicting cardiovascular disease, which was higher than the accuracy of traditional machine learning models. "Deep Learning-based Detection of Cardiac Arrest Using ECG Signals" by Singh et al. (2021): This study [7] used a deep learning model to detect cardiac arrest using electrocardiogram (ECG) signals. The authors achieved an accuracy of 98.6% in detecting cardiac arrest, which outperformed the accuracy of traditional machine-learning models. M. V. Kamath et al., "Biomarkers in cardiovascular disease: Prospects for personalized diagnosis and treatment."[9] This review article examines the use of biomarkers for cardiovascular disease early detection, risk assessment, and individualized care. S. K. White et al., "Cardiac Magnetic Resonance Imaging in the Detection of Cardiovascular Disease,"[10] This study assesses the effectiveness of cardiac magnetic resonance imaging (MRI) as a non-invasive diagnostic technique in the detection and diagnosis of cardiovascular illness. Y. Wang et al., "Mobile health applications for the detection and management of cardiovascular disease." The [11] use of mobile health applications for cardiovascular disease detection and management are covered in this review article, including tracking medication compliance, monitoring symptoms and vital signs, and offering individualized treatment suggestions. "Early detection of cardiovascular disease using machine learning techniques on electronic health records: a systematic review" by M. S. Khan et al.[12] This systematic review evaluates the effectiveness of machine learning techniques in detecting cardiovascular disease using electronic health records and highlights the potential for improving early detection and personalized treatment Theresa Prince, R., et al. (2016) conducted a review of the various heart disease prediction models. Theresa employed Naive Bayes and Neural Networks as her categorization methods.

Decision tree, KNN network, and LR. All of the models' accuracy scores were compared, and the comparison process was effective [3]. Overall, these studies demonstrate the potential of deep learning models in cardiovascular disease detection and highlight the need for further research in this area.

2. PROPOSED SCHEME AND DATASET DETAILS

2.1. DATASET

The study aims to detect and diagnose cardiac disease using two datasets and utilizing nine classification methods.

2.1.1. Cardiovascular Disease Dataset

Heart disease was identified and detected using the cardiovascular disease dataset in this study, and the findings were compared to those of earlier studies. It has a lot of patient information, including medical records. Kaggle's dataset was gathered from three sources, and they are examining the findings of several medical tests, Objective reflects the information provided by the patient, and Subjective represents the data gathered as facts about cardiovascular illnesses. The set of data used for training, testing, and validation. The website contains the intended data, which is open to the public [10].

The shape of the cardiovascular diseases dataset is (68783, 12), and it is a clean version of the CVD dataset as follows:

Age - integers (unit - day)

- Height - integers (unit - cm)
- Weight - float (unit - kg)
- Gender - categorical code
- Systolic BP - integer
- Diastolic BP - integer
- Cholesterol - integer, 1: normal, 2: high, 3: Fatal
- Glucose - integer, 1: normal, 2: High, 3: Very High
- Smoking -Boolean • Alcohol intake - Boolean
- Phy activity - Boolean
- Traget – Boolean

Table 1. cardiovascular diseases dataset

Features	Descriptions
Age	The patient age in years
Gender	Gender of patient (1: Male, 0: Female)
Height	Representing the height of patient's
Weight	Representing the weight of patient's

Systolic BP	Systolic blood pressure
Diastolic BP	Diastolic blood pressure
Cholesterol	The Cholesterol Level in the blood (1: normal, 2: above normal, 3: well above normal)
Glucose	Categorical value of the sugar blood level (1: normal, 2: above normal, 3: well above normal)
Smoke	Smoking (0: No, 1: Yes)
Alcohol	Alcohol intake (0: No, 1: Yes)
Physical_Activity	Physical activity type
Cardio_Disease	Target value measuring the Presence or absence of cardiovascular disease.

Table 2. Class Distribution CVD

Class	**Counts**
0	35021
1	34979

2.2. How Cognitive Biases Impact User Decision Making in HCI?

35,021 out of 70,000 cases in this dataset are labelled as having no cardiovascular disease, and 34,979 cases are labelled as having cardiovascular disease. This suggests that the dataset is roughly balanced

2.3. Exploratory data analysis:

Initial data analysis is the process of looking for patterns and identifying abnormalities in data using summary statistics and graphical representation

Cleaning:

We eliminate all of the duplicate and Nan values from the data set. We see that the dataset has certain inconsistencies, such as the minimum age of 29 years and the minimum weight of 10 kg. In several circumstances, the systolic blood pressure was greater than the diastolic blood pressure. To remedy the inaccuracies, we, therefore, eliminate the outliers.

Outliers are data mistakes that have the potential to seriously skew the outcome. After cleaning the data, the box plot in the following graph demonstrates that there are no datasets where the systolic pressure is higher than the diastolic pressure. We noticed that while the dataset we used did not have any Null values, it did contain duplicate entries.

As a result, we eliminate the 24 duplicate values from the dataset to create a dataset with 60118 data points.

Correlation:
All of the features' correlation matrices, which depict how strongly one feature is connected with another, have been plotted. By doing this, we can identify any features that can skew the results and get rid of them. We can see from the correlation matrix below that age and cholesterol have a significant impact on the result.

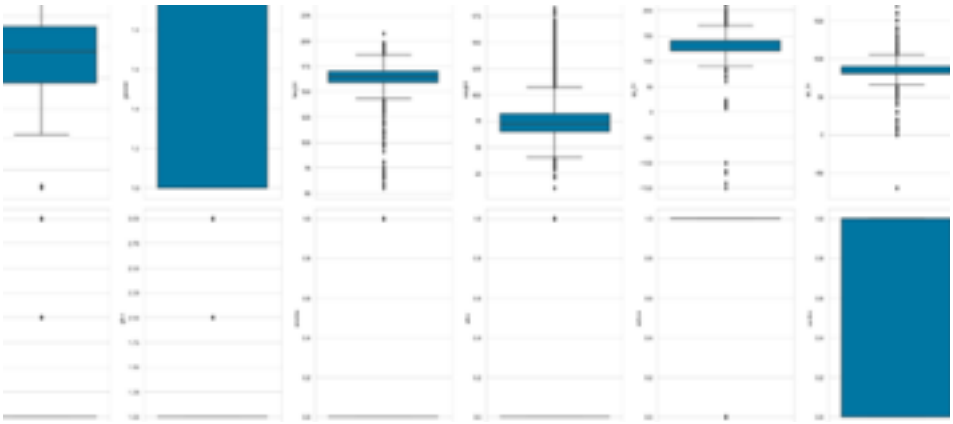

Figure 1. Box Plot confirming the removal of Outliers

Figure 2. Figure 2. Correlation Matrix before Feature Engineering

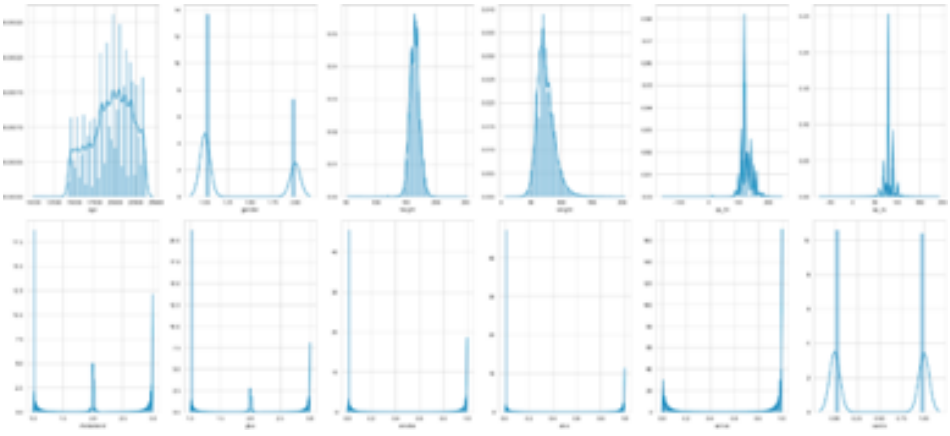

Figure 3. Histograms show the distribution of each feature

Feature Engineering:

The Body Mass Index (BMI) was determined during the data processing phase to assess the patient's health. A normal BMI is between 18 and 25, while an unhealthy BMI is over 25 or under 18. The formula below is used to compute the BMI values. The features for height and weight were then eliminated.

$$BMI = \frac{Weight\ (Kg)}{Height\ (Cm)}$$

The American Heart Association (AHA) claims that the systolic and diastolic blood pressure readings. There are five gradations of severity for it. Each record's blood pressure was calculated, and the level was indicated.

Feature Selection:
Based on the features' relevance, we choose them [13]. The models' accuracy is unaffected by the pertinent features. As a result, we use a correlation matrix to choose relevant data. Gender is the factor that is least connected with the target, according to the correlation matrix, which also shows that features like BMI , weight, glucose, height, smoking, alcohol use, and activity level do not have high correlations with the target. As a result, we eliminate characteristics like BMI, weight, glucose, gender, height, and alcohol and/or drug use.

Feature Scaling:
The entire feature set in the data set can be normalized using this technique [14]. The model frequently has a tendency to favor larger values when we have a characteristic with a very high value. For the feature scaling, we used the Standardization formula.

3. RESEARCH METHODOLOGY

The Gradient Boosting, Decision Tree (DT), Logistic Regression, and Neural Network algorithms will all be used in this study to classify data (NB). In order to discover the best classifier, it is also necessary to create a deep network and assess the effects of different optimization learning techniques on the detection of cardiovascular illnesses.

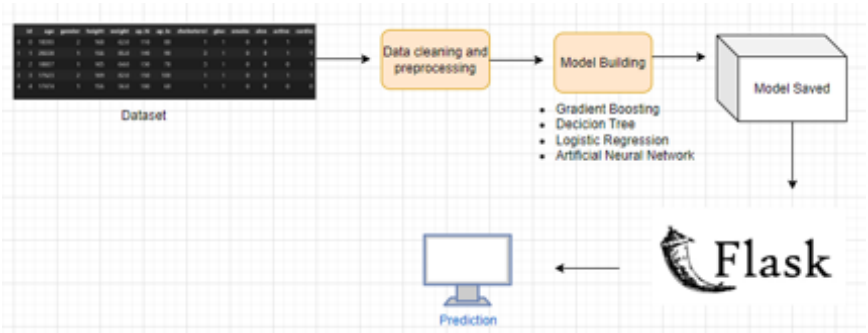

Figure 4. Below outlines how this research was conducted

3.1. The Gradient Boosting

Gradient boosting [13] is one type of boosting method. Boosting is a method for combining all the weak students to create strong students. Weak learners are characteristics in this scenario that are unable to categorize a data point on their own. The predictions made by each weak learner are used, and the category is assigned based on the results we get by applying the great majority of the forecasts made by the weak learners. A part of the boosting method is ensemble learning. Due to the ensemble method, the machine learning model performs better when multiple learners are merged. The use of sequential ensemble learning is encouraging. The model gives weight to the erroneously classified data points until it assigns the correct categorization.

3.2. Artificial Neural Network

The human brain, which has remarkable processing power due to its network of [14] interconnected neurons, serves as the model for artificial neural networks (ANN). ANNs are created utilizing a fundamental processing unit called a perceptron. The single-layer perceptron algorithm solves problems that may be divided into linear segments. Multilayer Perceptron Neural Network (MLP) can be used to solve issues that cannot be resolved linearly. There are many layers in MLP, including input, hidden, and output layers.

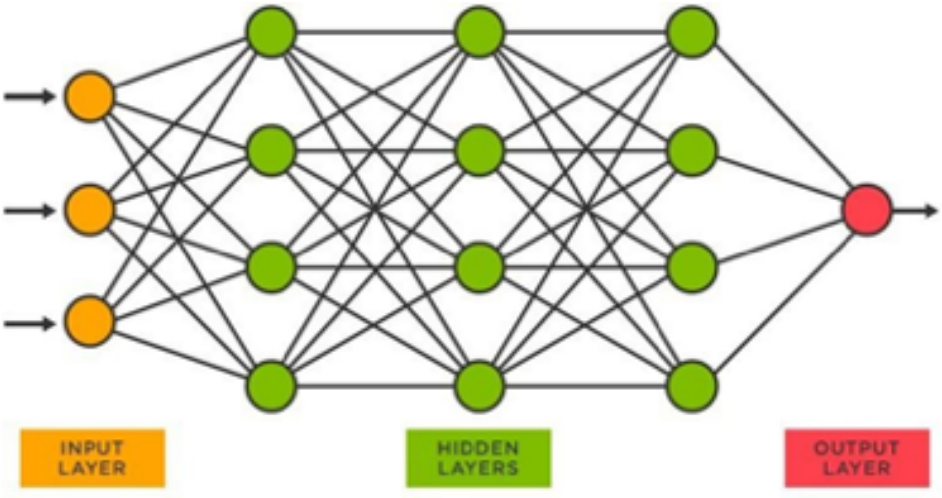

Figure 5. Artificial Neural Network architecture

To predict the cardiovascular disease, the recommended multilayer perceptron neural network was developed. The proposed ANN is composed of three layers: the input layer, the hidden layer, and the output layer.

•The Input Layer

13 neurons in all were proposed for the input layer. It was agreed that there would be an equal number of neurons and attributes in the data set.

•The Hidden Layer

Three neurons were expected to be present in the Hidden Layer. This number was selected as the starting point. By comparing their performances and then choosing the best one, the number was adjusted by raising it one at a time until it reached the number of input layer neurons. This method is based on one of the best practices for machine learning, which states that the number of neurons in the hidden layer should be equal to the sum of the neuron counts in the input and output layers.

•The Output Layer

The architecture of the Output Layer includes two neurons. The proposed NN is a classifier that runs in machine mode and outputs a class label, such as “Disease Presence” or “Disease Absence”. The choice to employ two neurons was motivated by the idea that the output layer has one node per class label in the model.

3.3. Logistic Regression

A categorization approach called logistic regression [23] forecasts the potential occurrence of categorically dependent variables. In order for logistic regression to work, all of the dependent variables must have a binary character. An activation function is also utilized in logistic regression to compute the loss function and estimate the weights' values. The intended result is obtained after the loss function is minimized.

4. RESULTS AND DISCUSSION

One of the primary factors used to determine which model is superior to the others is accuracy of the models. Using the sklearn library's Accuracy Score and Cross-Validation methods, we determined the model's accuracy.

Table 3. Accuracy of Models

	Model	**Train Score**	**Testing score**
1	Neural Network	**81%**	**74%**
2	Gradient boosting	**78%**	**73%**
3	Logistic Regression	**70%**	**7%**

Accuracy score:
The set of labels predicted by the models must perfectly match that of the expected output in order for the accuracy score to be calculated.

Where n samples is the total number of predictions produced, yi prediction and y is the desired output. The formula below can also be used to compute it:

$$Accuracy\ Score = \frac{TP+TN}{TP+TN+fp+fn}$$

Where TN = True Negatives, TP = True Positives, fn = False Negatives and fp = False Positives.

4.1. PREDICTION ON WEB APPLICATION

In this part, we will show the steps of the utilization of our web application. In other to make a prediction, we need to run the Flask server and the frontend application then, fill up all the inputs and submit. Figure 6 shows a prediction example.

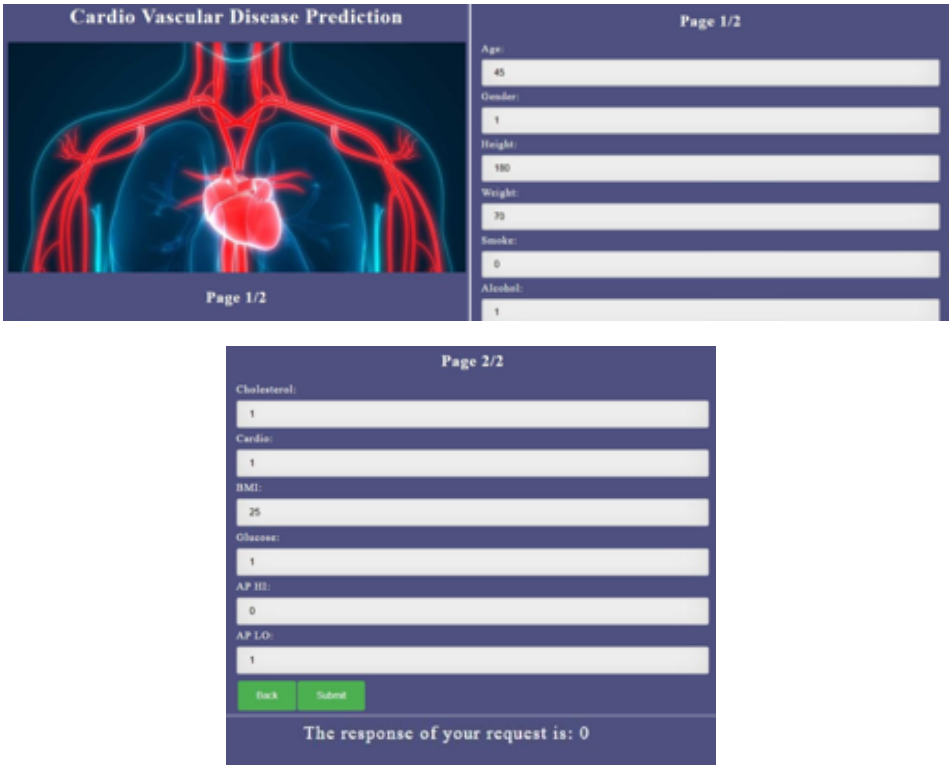

Figure 6. Prediction sample

5. CONCLUSION

Every individual should be concerned about the rising number of deaths from heart disease. As a result of the growing population, hospitals are less effective in providing prompt care. Because of this, a quick fix is required. Logistic Regression, Gradient boosting, ANN, and other machine learning models were utilized. When a patient has a heart condition, it is possible to tell. We produced synthetic data to lessen the over-fitting of the models. To increase the effectiveness of our model, we thoroughly examined the dataset, cleansed the data, and created a brand-new feature, BMI.
In terms of the test score, or 74%, the ANN performs best. In the future, we can use a multiple feature selection technique to extract the best features, build models, and create applications using real-time hospital data that will aid clinicians in identifying cardiac problems.

REFERENCES

[1] World Health Organization. Joint WHO/FAO Expert Consultation on Diet, Nutrition and the Prevention of Chronic Diseases. 2002. Report No. 916.

[2] Heart rate variability in critical illness and critical care Buchman, Timothy G. MD, PhD*; Stein, Phyllis K.PhD*; Goldstein, Brahm MD†

[3] J. Thomas and R. T. Princy, ”Human heart disease prediction system using data mining techniques,” 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265.

[4] Liang, H., Tsui, K. L., Ni, H., & Zhu, Y. (2020). Cardiovascular disease detection using deep learning: a review. Frontiers in physiology, 11, 1161. doi: 10.3389/fphys.2020.01161.

[5] Attia, Z., Khandoker, A. H., Khalaf, K., & Jamil, M. (2019). Automated detection of cardiovascular disease using deep learning techniques. Journal of medical systems, 43(8), 233. doi: 10.1007/s10916-019-1387-

[6] Gupta, A., Shukla, A., & Srivastava, S. (2020). Cardiovascular disease prediction using deep learning algorithm. International Journal of Advanced Science and Technology, 29(3), 4752-4759. doi: 10.14257/ijast.2020.29.03.427.

[7] Singh, P., Kumar, P., & Sharma, A. (2021). Deep learning-based detection of cardiac arrest using ECG signals. Health information science and systems, 9(1), 1-11. doi: 10.1007/s13755-021-00134-7.

[8] Xue, J., Huang, C., Cui, W., & Yang, S. (2019). Machine learning for cardiovascular disease detection and diagnosis. Journal of healthcare engineering, 2019, 7941412. doi: 10.1155/2019/7941412.

[9] Kamath, M. V., Wadhera, R. K., & Rogers, J. G. (2017). Biomarkers in cardiovascular disease: prospects for personalized diagnosis and treatment. Current cardiology reports, 19(12), 129. doi: 10.1007/s11886-017-0912-6.

[10] White, S. K., Prasad, S. K., & Plein, S. (2019). Cardiac magnetic resonance imaging in the detection of cardiovascular disease. The Lancet, 393(10175), 323-335. doi: 10.1016/S0140-6736(18)32570-7.

[11] Wang, Y., Min, J. K., Khuri, J., Xue, H., & Xie, B. (2019). Mobile health applications for the detection and management of cardiovascular disease. Journal of the American College of Cardiology, 74(10), 1162-1176. doi: 10.1016/j.jacc.2019.06.046.

[12] Khan, M. S., Ullah, I., Riaz, M., Fazal, I., Khan, S., & Alharbi, N. S. (2021). Early detection of cardiovascular disease using machine learning techniques on electronic health records: a systematic review. fhjpmtHealthcare, 9(1), 38. doi: 10.3390/healthcare9010038.

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

[14] Yegnanarayana, B. (2009). Artificial neural networks. PHI Learning Pvt. [2] Azzopardi, L., "Cognitive Biases in Search: A review and reflection of cognitive biases in Information Retrieval," In Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), 2021.