Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

E-ISSN: 2687-6167 ***Number 60, March 2025***

RESEARCH ARTICLE

Receive Date: 04.11.2024 *Accepted Date: 25.11.2024*

Comparison of the effects of features and classifiers on performance in the cardiovascular disease detection system

İzzet Emir[a,*], Yıldız AYDIN[b]

[a]*Department of Cardiovascular Surgery, Erzincan Binali Yıldırım University, Erzincan, Turkey, ORCID: 0000-0002-1098-4889*
[b]*Department of Computer Engineering, Erzincan Binali Yıldırım University, Erzincan, Turkey, ORCID: 0000-0002-3877-6782*

Abstract

This study aims to analyze the effects of features and classifiers in detecting cardiovascular diseases (CVD), which remain the leading cause of morbidity and mortality worldwide. Early and accurate detection of CVD significantly affects treatment outcomes. Therefore, the proposed method aims to automatically detect cardiovascular diseases via artificial intelligence. In this research, the performances of artificial intelligence methods for the cardiovascular disease detection problem are analyzed. The dataset used in this study was sourced from the publicly available Kaggle platform. It used for performance analysis in the developed application includes the features of 70000 patients such as age, gender, height, weight, blood pressure, cholesterol, glucose, smoking and alcohol use. These features were classified with Gradient Boosting, XGBoost, SVM, Random Forest, Logistic Regression, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) methods and performance comparison was performed. In the experimental results, the highest accuracy rate of 72.55% was obtained using the Gradient Boosting method, demonstrating its superior performance in cardiovascular disease detection. In addition, it was observed that the classification performance decreased when the high blood pressure attribute was removed from the dataset, while the removal of other features did not significantly affect the performance.

Keywords: Cardiovascular disease (CD); artificial intelligence; detection; CD risk factors.

1. Introduction

Artificial Intelligence (AI) is a set of algorithms that demonstrate human skills such as noticing, learning, classifying and generating new ideas with the help of a computer. AI algorithms attempt to imitate humans through deep learning and machine learning. With the rapid development of technology, AI applications have found widespread application in many areas of science, technology and medicine. The use of advanced computing power of AI in medicine has been going on for about 60 years [1]. Artificial intelligence-based systems in cardiovascular medicine; cardiovascular imaging has found new applications in cardiovascular risk prediction and treatment process. As a result of these applications, faster and objective diagnosis can be provided for different types of cardiovascular diseases [2].

Despite all the advances in medicine, cardiovascular diseases (CVD) remain the leading cause of morbidity and mortality worldwide. The main reasons for this are; Mistakes in life and nutrition styles, inadequacy of preventive treatments and delays in diagnosis due to the increasing number of people, poor quality of life due to socio-economic problems, physicians' inability to apply treatments in accordance with the guidelines, inadequate patient follow-up, and low patient-physician compliance. Cardiovascular diseases are common conditions that affect the heart and blood vessels. Acute coronary syndrome includes stroke, heart failure, coronary heart disease, cardiomyopathies, and peripheral vascular diseases [3]. It causes approximately 1/3 of deaths worldwide [4]. Despite advances in medical science, delays in diagnosis can cause negative effects on the treatment process. Accurate and early detection of CVD is critical to improving treatment outcomes and reducing mortality rates. Therefore, AI-supported diagnostic softwares are one of the important systems needed.

Diabetes, dyslipidemia, obesity, hypertension, gender, low or lack of physical activity, alcohol and cigarette use, and body mass index are risk factors associated with cardiovascular diseases [5]. Dataset used in this study includes 70,000 patient records with features such as age, gender, height, weight, blood pressure (high and low), cholesterol, glucose levels, smoking status, alcohol consumption, and physical activity which is available on the open access Kaggle website. These variables were chosen as they represent well-established cardiovascular risk factors frequently cited in medical literature. Their inclusion ensures a comprehensive analysis of both modifiable and non-modifiable risk factors, enabling the development of a robust and accurate cardiovascular disease detection system

This is how the remaining part of the paper is organized. A overview of the recent, pertinent literature is provided in Section 2. Section 3 provides a detailed explanation of the recommended strategy for detecting cardiovascular disease. Section 4 offers numerous appropriate examples to illustrate the application of the recently proposed methodology. Ultimately, Section 5 concludes the paper.

2. Related work

In the first studies for the use of AI in cardiovascular medicine, a mathematical model was defined for the diagnosis of congenital heart disease. Thus, although there were several limitations, congenital heart disease could be diagnosed with an accuracy comparable to that diagnosed by a physician [6], [7]. AI has also been used for personalized drug therapy, reducing the risk of side effects and enhancing treatment effectiveness. For instance, Li et al. utilized a backpropagation neural network to predict the warfarin maintenance dose for heart valve replacement patients [8]. Moreover, deep learning-based artificial intelligence systems have potential breakthrough applications in drug discovery, personalized drug therapy, and precision medicine [8], [9].

It can be extracted using AI and offer new ideas in the diagnosis and treatment of cardiovascular diseases. In another study; It was used to diagnose different heart diseases and characterize the New York Heart Association (NYHA) class, showing high accuracy rates in the analysis, which examined the results of more than 10,000 patients over a nearly 20-year period [10]. Using a database of a specific population, Kakadiaris et al. showed that a machine learning-based method outperformed the American College of Cardiology (ACC) and American Heart Association (AHA) risk calculators in disease risk analysis, and that the AI application was better at predicting cardiovascular disease [11].

Modifiable risk factors like dyslipidemia and diabetes significantly impact cardiovascular disease, with diabetes posing a two-to-fourfold increased risk of CVD due to elevated blood glucose levels [12], [13]. Recent studies also highlight the importance of features like BMI, cholesterol, and systolic blood pressure in predicting CVD risk using AI models [14].

Artificial intelligence is an operating system whose effectiveness is increasing in every computational situation to realize, learn and solve problems in every aspect of human life. In this paper, we have tried to create an advancement of AI for the detection of cardiovascular diseases. In this way, we think that it can help find diseases faster and easier and lead to advances in their treatment.

The point to note here is that artificial intelligence helps to create some correlations with complex computer engineering methods and generates results according to the sensitivity of the method taught to the device. These are knowledge generators to make things easier and use energy in different areas. Therefore, it is important that studies using AI are thoroughly linked to the clinic. Nevertheless, AI is a technology with innovative potential in healthcare. With the application of diagnostic tools such as cardiovascular disease detection, biochemical analysis and imaging, AI offers exciting innovations in the early diagnosis and treatment of many cardiovascular diseases [15]

3. The proposed model

In the proposed method, research has been carried out in two sub-sections: the effect of features on the disease detection system for cardiovascular disease detection and the effect of the classifiers used on the disease detection system. The flow diagram of the experiments carried out within the scope of this study is given in Figure 1.

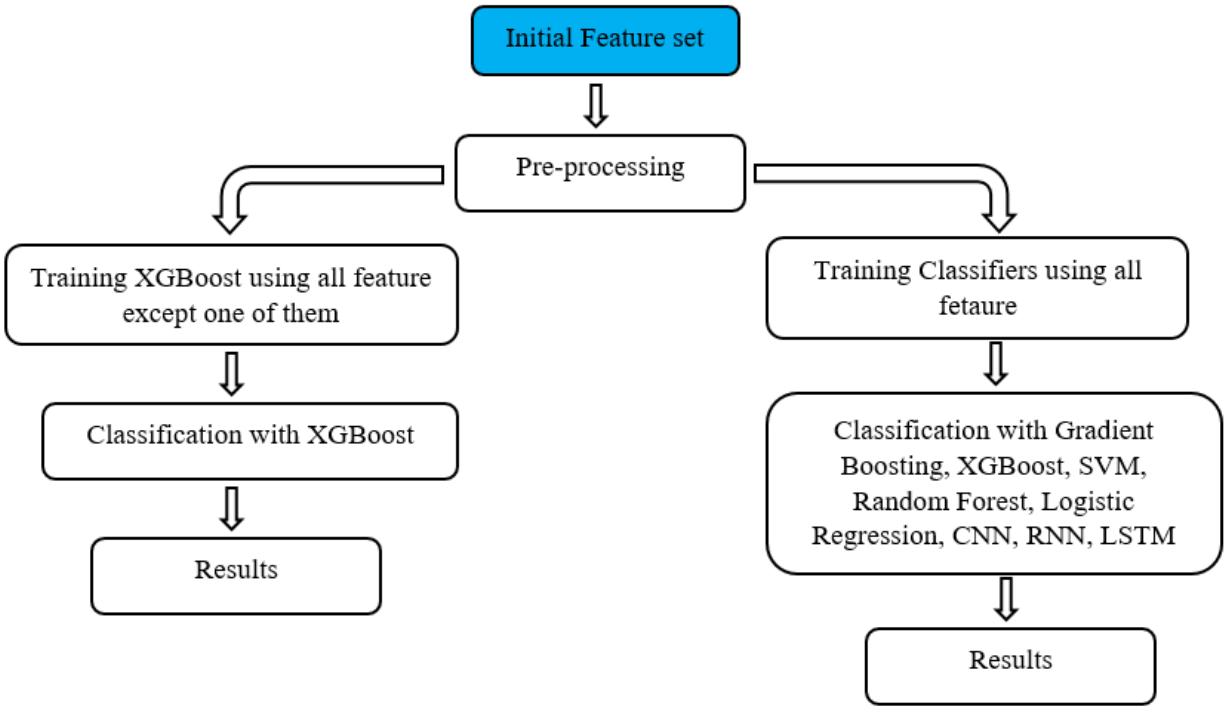

Fig. 1. Experimental flow diagram illustrating the steps for analyzing feature significance and classifier performance in cardiovascular disease detection.

In the first subsection, the effect of features on the cardiovascular disease detection system is analyzed. In each experiment, one feature is removed and the classification is done with the remaining features and the relative effects of these features on the model performance are evaluated. The aim is to determine which features are more critical in disease detection. In order to observe the effect of the features on the disease, in the first step, classification was made with XGBoost using all the remaining features except one feature. The XGBoost method, which has become a popular algorithm due to its high success rate and consistent results in classification problems, was preferred in the feature removal process. Since the experiments carried out in this process are iterative, the computational efficiency and consistent performance of this method provide advantages and the effect of the removed features on the classification performances was analyzed. The second subsection focuses on comparing the performances of classifiers used for cardiovascular disease detection system. In the experiments where all features are used, the performance of different classifiers such as Gradient Boosting, XGBoost, SVM, Random Forest, Logistic Regression, CNN, RNN and LSTM is analyzed. The aim is to determine the classifier that achieves the highest accuracy rate and to highlight the advantages of different methods.

4. Experimental results

Comparative results of the experiments carried out for the cardiovascular disease detection problem within the scope of this study are given in this section. Python was used as the scripting language for the experiments conducted in this study. Scikit-learn library was used for classical machine classification methods such as Gradient Boosting, Support Vector Machines (SVM), Random Forest and Logistic Regression. TensorFlow and Keras libraries were used for Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) networks. Matplotlib and Seaborn libraries were used for visualization of model results, while Pandas and NumPy libraries were used for data preprocessing and numerical calculations.

In the experiments conducted for cardiovascular disease detection, the data set available on the open-access website Kaggle [16] was used. In this data set, a total of eleven features are given for 70000 patients, including age, gender, height, weight, high blood pressure, low blood pressure, cholesterol, glucose, smoking status, alcohol use status and active. The dataset underwent a series of preprocessing steps to ensure data integrity and model robustness. In order to improve the classification performance, some preprocessing was performed on the dataset. Missing and repetitive records were checked and removed from the dataset if detected. The age attribute expressed in days in the dataset was converted into years. Since height and weight attributes are not meaningful separately, body mass index (BMI) was calculated and BMI attribute was used by removing height and weight columns. Variables such as gender, cholesterol and glucose were categorized. Figure 2 shows the value ranges of each attribute used in this study. Since the dataset does not consist of two parts, train and test, in order to evaluate the performance, it was divided into training and test sets using train_test_split function with 80/20 ratio and 25% of the data was reserved for testing.

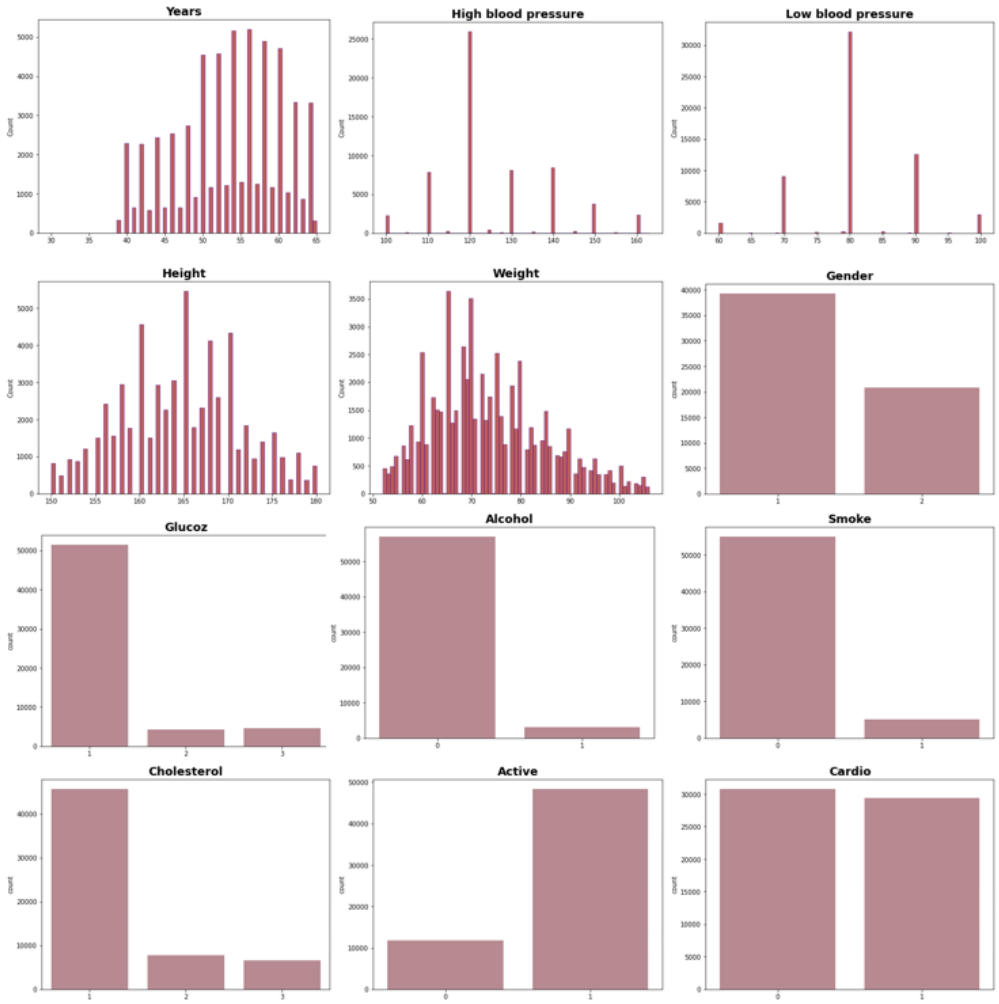

Fig. 2. Descriptive statistics of variables.

Precision, recall, F1 and accuracy metrics were used to measure the performance of the methods used in the proposed method. Precision is the value that shows how many of the people labeled as patients are actually patient, and recall is the value that shows how many of the people who are actually patients are labeled as patients.

Here, it is necessary to detect those who are not patient as well as those who are patient. Labeling someone who is not patient as patient can have bad consequences, especially in the case of cardiovascular disease, which is a serious disease. Therefore, there must be a proportion between precision and recall values, and in order for the method to be considered successful, one should not be too high and the other should not be low. F1 value is a value obtained using precision and recall values. Lastly, the accuracy metric used expresses the correct prediction rate in the results. The formulas of precision, recall, F1 and accuracy metrics are given in Equation 1, Equation 2, Equation 3 and Equation 4, respectively.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F1\ Score = 2 * \frac{Recall*Precision}{Recall+Precision} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

The experiments carried out within the scope of this study can be discussed in two parts. In the first part, in order to observe the effect of the features on the disease, experiments were carried out by removing only one feature and using the remaining ten features and the XGBoost classifier. The performances of these experiments are given in Table 1.

Table 1. Identification results of the experiment which was carried out by removing one feature and using the remaining features.

Classifier	Unused Feature	Accuracy
XGBoost	-	71.94
	Age	71.46
	Height	72.06
	Weight	71.78
	High Blood Pressure	68.42
	Low Blood Pressure	71.89
	Gender	71.68
	Cholesterol	71.44
	Glucose	71.87
	Smoke	71.71
	Alcohol	71.74
	Active	71.67

In order to determine the importance of the features in disease detection, experiments were conducted in which they were systematically extracted one by one and the results of these experiments are presented in Table 1. For example, the high blood pressure (ap_hi) feature was removed to examine its importance as it is a known critical factor in cardiovascular risk assessment. The experiment excluding this feature resulted in a decrease in the accuracy metric from 71.94% to 68.42%, indicating that this feature makes a significant contribution to classification performance.

Other features, such as smoking and alcohol consumption, were also removed to assess their roles. In the experiments where these features were removed, the accuracy did not change much (71.71% and 71.74%) and

therefore had a relatively small impact. This suggests that these attributes contribute to cardiovascular risk, but have less impact on the prediction performance of the model than features such as blood pressure.

This process supported the decision to keep all features in the model by revealing the differences in the importance levels of the features. Thus, a more comprehensive approach to cardiovascular disease prediction was provided while preserving the best performance of the model.

In the second part of the experiments, the performances of different classification methods were compared using all the features. The performance results of the experiments performed using Gradient Boosting (GB), XGBoost, Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) methods's results are given in Table 2 and Figure 3.

Table 2. Identification results of the experiment which was carried different classifier.

Method	Precision	Recall	F1-Score	Accuracy
Gradient Boosting	72.75	72.44	72.41	**72.55**
XGBoost	72.15	71.83	71.80	71.94
SVM	72.72	72.12	72.04	72.27
Random Forest	72.63	72.00	71.90	72.15
Logistic Regression	72.08	71.57	71.49	71.71
CNN	72.93	72.34	72.26	72.49
RNN	72.41	72.28	72.28	72.35
LSTM	72.64	72.43	72.43	72.52

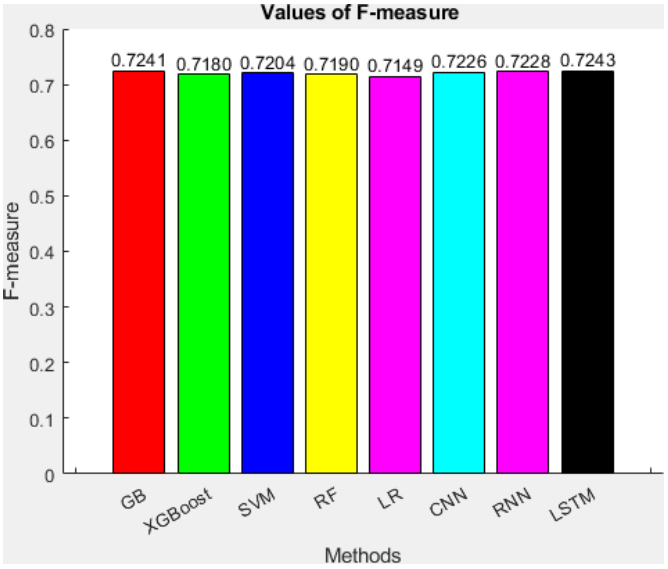

Fig. 3. Results of methods.

When Table 2 and Figure 2 are examined, it is seen that the performances of the methods are quite close to each other. The Gradient Boosting method achieved the highest accuracy value.

5. Conclusions

Experiments were carried out to observe the effect of features and classifiers on the disease detection system in the cardiovascular disease detection system. In order to observe the effect of the features used in the cardiovascular disease detection system, classification was made without using one feature in each experiment. In this study, systematic experiments were conducted in which each attribute was extracted individually to examine the effect of the attributes used on the classification performance. It was observed that clinically important attributes such as high blood pressure contributed significantly to the classification success rate. In particular, the removal of the high blood pressure feature resulted in a significant decrease in accuracy, indicating that this feature plays a critical role in classification. Other features such as cholesterol and glucose levels had a moderate effect, while variables such as smoking and alcohol consumption had a more limited effect on accuracy. This analysis clearly shows the relative importance of the features, indicating that not all variables contribute equally to the model performance.

In the second part of the experiments, the performance of the classifiers was compared with the use of all features. The experimental results show that the performance of the classifiers is similar, but Gradient Boosting gives the best results. The success of Gradient Boosting is based on its ability to combine multiple weak learners (decision trees) into a powerful predictive model. This method optimizes performance by iteratively reducing the errors of previous models, resulting in higher accuracy at each step.[17].

In future studies, it is aimed to develop a more successful cardiovascular disease detection system by using more features to increase classification performance.

Author contribution

Yıldız AYDIN and İzzet EMİR actively participated in conducting the experimental studies and writing the manuscript.

Acknowledge

The authors declare that they have no conflict of interest.

References

[1] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
[2] P. Liu, L. Lin, J. ZHANG, T. HUO, S. LIU, and Z. YE, “Application of Artificial Intelligence in Medicine: An Overview,” *Curr. Med. Sci.*, vol. 41, no. 6, pp. 1105–1115, 2021, doi: https://doi.org/10.1007/s11596-021-2474-3.
[3] vA. Shaito *et al.*, “Herbal Medicine for Cardiovascular Diseases: Efficacy, Mechanisms, and Safety,” *Front. Pharmacol.*, vol. 11, no. April, pp. 1–32, 2020, doi: 10.3389/fphar.2020.00422.
[4] E. J. Benjamin *et al.*, *Heart Disease and Stroke Statistics’2017 Update: A Report from the American Heart Association*, vol. 135, no. 10. 2017.
[5] Q. Cheng *et al.*, “Sex-specific risk factors of carotid atherosclerosis progression in a high-risk population of cardiovascular disease,” *Clin. Investig.*, vol. 46, no. 1, pp. 22–31, 2023, doi: 10.1002/clc.23931.
[6] A. F. Toronto, L. G. Veasy, and H. R. Warner, “Evaluation of a computer program for diagnosis of congenital heart disease,” *Prog. Cardiovasc. Dis.*, vol. 5, no. 4, p. 1963, 1963.
[7] O. I. Al-Sanjary and G. Sulong, “Detection of video forgery: A review of literature,” *J. Theor. Appl. Inf. Technol.*, vol. 74, no. 2, pp. 207–220, 2015.
[8] Q. Li *et al.*, “The Prediction Model of Warfarin Individual Maintenance Dose for Patients Undergoing Heart Valve Replacement, Based on the Back Propagation Neural Network,” *Clin. Drug Investig.*, vol. 40, no. 1, pp. 41–53, 2020, doi: 10.1007/s40261-019-00850-0.
[9] A. A. Kalinin *et al.*, “Deep learning in pharmacogenomics: From gene regulation to patient stratification,” *Pharmacogenomics*, vol. 19, no. 7, pp. 629–650, 2018, doi: 10.2217/pgs-2018-0008.
[10] G. P. Diller *et al.*, “Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients,” *Eur. Heart J.*, vol. 40, no. 13, pp. 1069–1077, 2019, doi: 10.1093/eurheartj/ehy915.
[11] I. A. Kakadiaris, M. Vrigkas, A. A. Yen, T. Kuznetsova, M. Budoff, and M. Naghavi, “Machine learning outperforms ACC/AHA CVD risk calculator in MESA,” *J. Am. Heart Assoc.*, vol. 7, no. 22, 2018, doi: 10.1161/JAHA.118.009476.
[12] C. P. Cannon, “Mixed Dyslipidemia, Metabolic Syndrome, Diabetes Mellitus, and Cardiovascular Disease: Clinical Implications,” *Am. J.*

Cardiol., vol. 102, no. 12 SUPPL., pp. 5L-9L, 2008, doi: 10.1016/j.amjcard.2008.09.067.
[13] R. Huxley, F. Barzi, and M. Woodward, "Excess risk of fatal coronary heart disease associated with diabetes in men and women: Meta-analysis of 37 prospective cohort studies," *Br. Med. J.*, vol. 332, no. 7533, pp. 73–76, 2006, doi: 10.1136/bmj.38678.389583.7C.
[14] H. Chu *et al.*, "Roles of Anxiety and Depression in Predicting Cardiovascular Disease Among Patients With Type 2 Diabetes Mellitus: A Machine Learning Approach," *Front. Psychol.*, vol. 12, no. April, pp. 1–8, 2021, doi: 10.3389/fpsyg.2021.645418.
[15] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.
[16] "Skin Cancer: Malignant vs. Benign." https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign (accessed on).
[17] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1002/9781118445112.stat08190.