Toward a Model to Predict Cardiovascular Disease Risk Using a Machine Learning Approach

Khaoula Slime, Abderrahim Maizate, Larbi Hassouni, Najat Mouine

***Abstract*— Cardiovascular diseases (CVD) remain a major global health concern, contributing significantly to both death and morbidity. To avoid premature mortality, people with heart diseases who are more likely to have a crisis should be informed as soon as feasible. In the meanwhile, the main difficulties in determining the risks of heart disease diagnosis are related to inadequate information, poor data quality, imprecision, and uncertainty. In this research, we examined a wide range of variables, with particular attention to critical risk factors identified by medical professionals, such as age, body mass index (BMI), diastolic and systolic blood pressure (BP), alcohol usage, smoking, and physical activity.**

Furthermore, a range of feature selection techniques is used to evaluate their utility in the prediction of CVD. To help with the creation of a predictive model that can identify and evaluate cardiovascular risk and enable the early prediction of heart attacks, we also recommend conducting a comparison analysis. Using a combination of numerous dataset attributes, we empirically assess the performance of our system and attain 70% accuracy in both training and testing data, with the best score difference of 0.15%. When 99% accuracy in training data and 63% accuracy in test data are combined, the worst difference score comes out to 36%.

***Index Terms*—Cardiovascular, Deep learning, heart disease, Machine learning, Prediction**

I. Introduction

The World Health Organization (WHO) predicts that by 2030, the prevalence of heart disease will increase to 23.3% worldwide [1]. Furthermore, 90% of patients disregard their risk and vulnerability to a heart attack prior to experiencing a heart attack, according to the American Heart Association (AHA) [2]. Heart disease is one of the most avoidable, preventable, and controlled diseases [3], despite these consequences and despite being one of the illnesses that cause the greatest amount of mortality [4] and disability.

These claims encourage us to consider creating an intelligent system that can lower the risk of cardiovascular disease (CVD) and lessen the number of premature deaths brought on by a failure to recognize the early signs of a heart attack [5]. The medical industry stands to gain significantly from the potential of machine learning classification techniques, which can expedite and enhance the accuracy of disease diagnosis [6].

Manuscript received July 21, 2023; revised March 27, 2024.

Khaoula Slime is a PhD candidate at the National Institute of Electronic & Mechanic, Hassan II University, Morocco (phone: 212-63805-5009; e-mail: khaoulaslime@gmail.com)

Abderrahim Maizate is a professor of Mathematics & Informatics Department, Hassan II University, Morocco. (e-mail: maizate@outlook.com).

Larbi Hassouni is a professor of Mathematics & Informatics Department, Hassan II University, Morocco. (e-mail: lhassouni@hotmail.com).

Najat Mouine is a professor of Medicine at the military hospital, Mohamed V University, Morocco. (e-mail: mouine2@yahoo.fr).

Furthermore, researchers have proposed numerous contributions to heart disease prediction, including for instance: Sujata et al. applied decision tree, Naive Bayes, and K nearest neighbor algorithms for early prediction of heart disease [7]. To predict cardiac illness, Martins et al. [8] used a Bayesian optimization XG boost classifier and a one-hot encoding approach. Neural networks, support vector machines, multilayer perceptrons, C4.5, PART, and radial basis functions were used by Purushottam et al. [9] to identify the primary cause of cardiovascular disease and to explain the relationships between various patients.

Heart disease is analyzed and classified automatically using Convolutional Neural Networks (CNN) [10]. Using the Cleveland dataset, Haq et al. [11] propose a hybrid model for cardiovascular disease and use classification methods to find important variables. Shah et al. [12] used data acquired from Iranian patients and applied ten machine-learning algorithms to construct a prediction of a heart disease model. In their research, R. Thanga Selvi and I. Muthalakshmi proposed using an artificial neural network approach for the diagnosis of cardiac disease [13].

Furthermore, R. Katarya et al. [14] recommend comparing machine learning algorithms for cardiovascular illness prediction, while Chhillar [15] analyzes data mining predictive methods for heart disorders using WEKA (Waikato environment for knowledge analysis) tools. The authors of [16] created Intelligent Heart Disease Prediction Systems, a prototype for heart disease prediction, using data mining techniques. The importance of neural network-based cardiovascular disease prediction is discussed by Peng et al. [17]. In addition to Garg et al. [18], who conducted a comparison analysis of five machine learning algorithms on four datasets, Deshmukh et al. [19] offered a comparative study by applying kNN, SVM, DT, and ANN on two datasets. Using biometric data Mizuho created a device that prevents heart disease by identifying irregular heart sounds [20].

Even with the abundance of studies conducted in the field of cardiovascular medicine, algorithms are still open to development and improvement. The importance of data quality and the careful curation of particular traits, which have a significant impact on these algorithms' overall performance, highlight the need for this. Most previous research indicates that the quality of cardiovascular disease variables employed in machine learning algorithms considerably affects the predictive model's output [21, 22].

Thus, in order to guarantee better algorithm performance, we examined the features in our dataset and assessed how they affected our model in consultation with physicians. Machine learning and neural networks were employed in Javid's research effort [23] to improve the prediction accuracy of cardiac illness. My research team and I outlined the general architecture for applying deep learning techniques to patient monitoring with an emphasis on heart disease in our earlier work [24].

In this work, we present a machine learning-based cardiovascular disease (CVD) prediction model. The purpose

of this model is to aid medical practitioners in their first diagnosis-making process.

As a result, we suggest the following contributions in this study:

1. An awareness of the implications of each dataset parameter from a business standpoint.
2. A new and creative method for CVD early diagnosis.

The structure of our paper is as follows: Section 2 delineates our technique and comprehension of the data, while Section 3 delineates the steps involved in data preparation. Section 4 presents the experimental results along with a commentary, and Section 5 concludes our investigation.

II. Methodology and Data Understanding

Numerous domains, such as risk assessment, pattern recognition, and prediction, have found extensive uses for neural networks and machine learning techniques [25]. These methods have been applied in various studies [26, 27, 28, 29, 30] to construct an automated intelligent model for heart disease prediction. To construct our model for evaluating the risk of cardiovascular illness, we choose to express the research design as shown in Figure 1.

A. Dataset

We obtain the dataset from the Kaggle repository [31] for our investigation. It has 13 qualities and 70000 registers. Documents are gathered while the patient is being examined. Three categories include those attributes:

- Objective: The factual information.
- Examination: The results of the medical examination.
- Subjective: The patient provided the information.

In the dataset, the target class "cardio" equals to:

0 → Not present (if the patient is healthy)

1 → Present (when the patient has cardiovascular disease)

B. Data cleaning

The data is a big challenge for machine learning algorithms [32]. In this part, we have checked the data volume by checking if there are any missing values (Figure 2), we also check if there are any duplicated rows and finally, we verify the balanced data distribution. In our case, we have 0 duplicate values in the dataset and it didn't hold any missed values and doesn't have any duplicated lines.

One major obstacle for machine learning algorithms is the data [32]. We have examined the data volume in this section by looking for any missing values (Figure 2), duplicate rows, and, lastly, confirming a balanced data distribution. The dataset we have in our instance contains zero duplicate values, no missing values, and no duplicate lines

Table I describes the data columns.

TABLE I
Dataset Columns

Column	Description	Type
Id	Patient id	Objective
age	Age in number of days	Objective
gender	Gender (male or female) can have 0 or 1 as a value	Objective
height	Indicate the patient's height in cm	Objective
weight	Indicate the patient's weight in kg	Objective
ap_hi	Indicate systolic blood pressure	Examination
ap_lo	Indicate diastolic blood pressure	Examination
cholesterol	Indicate the patient's cholesterol	Examination
gluc	Indicate the patient's glucose	Examination
smoke	Indicate whether the patient smokes or not can have 0 or 1 as a value	Subjective
alco	Indicate whether the patient consumes alcohol or not can have 0 or 1 as a value	Subjective
active	Indicate if the patient is physically active or not. Can have 0 or 1 as a value.	Subjective
cardio	Indicate if the patient had a heart problem. Can have 0 or 1 as a value	Target

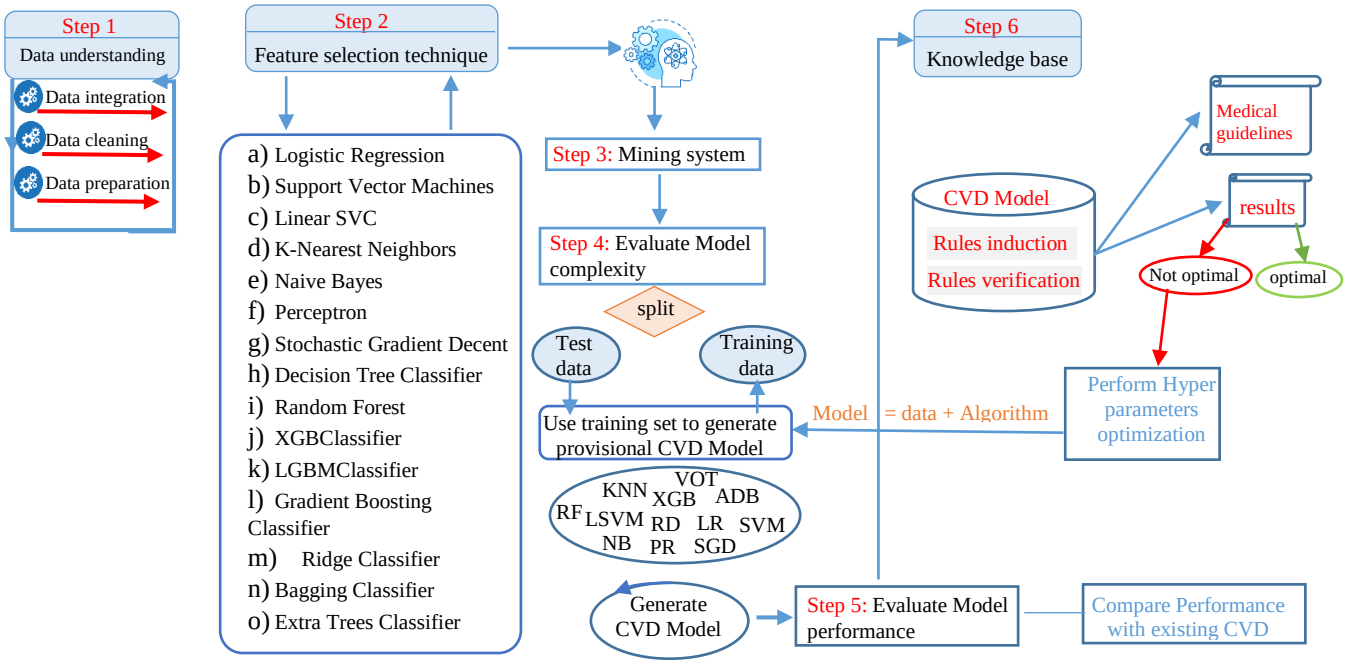

Fig. 1. Design for a cardiovascular risks prediction model

id	0	cholesterol	0	weight	0
age	0	gluc	0	ap_hi	0
gender	0	alco	0	ap_lo	0
height	0	active	0	smoke	0

Fig. 2. Check missing values

Figure 3 shows that our dataset is balanced and contains 50.03% of healthy patients (35021 records with Cardio = 0) and 49,97% of patients with CVD (34979 records with Cardio = 1)

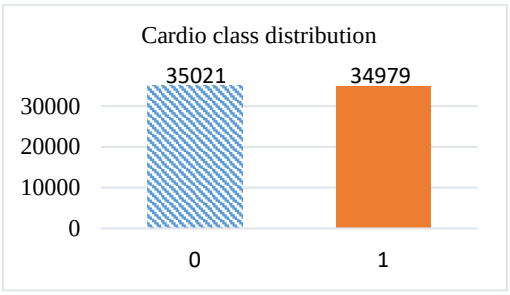

Fig. 3. Class distribution of the target variable '*cardio'*

III. *Data* Preparation

Following the data cleaning procedure, we can declare that the dataset is ready for analysis. We have three different kinds of input features, as was previously stated. We categorize the aspects in our analysis to help us better comprehend our data. First, we convert the age data from days to years, which are represented as categorical numbers. It is well known that aging has a major role in the development of cardiovascular disease (CVD). We chose to classify our data using the age groups supplied by the US National Library of Medicine National Institutes of Health in order to confirm this (Table II).

TABLE II
Age Groups

Age	Group
0 < age ≤ 2	Infants
2 < age ≤ 5	Pre School Child
5 < age ≤ 12	Child
12 < age ≤ 19	Adolescent
19 < age ≤ 24	Young
24 < age ≤ 44	Adult
44 < age ≤ 65	Middle Aged
65 < age	Aged

The data distribution is displayed in Figure 4. This graph leads us to the conclusion that the dataset primarily consists of data from middle-aged individuals, with only a small amount of adult data. We then decided to look at the relationship between age and the risk of CVD. (Figure 5 and Figure 6).

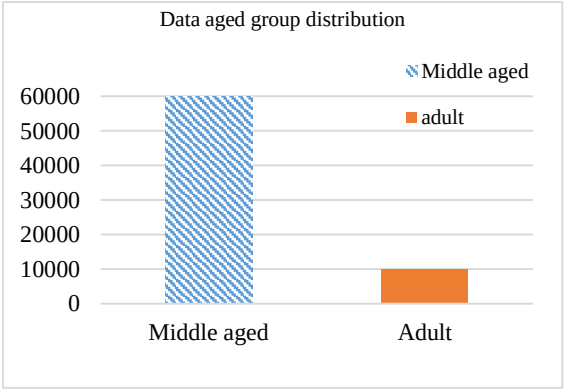

Fig. 4. Data age group

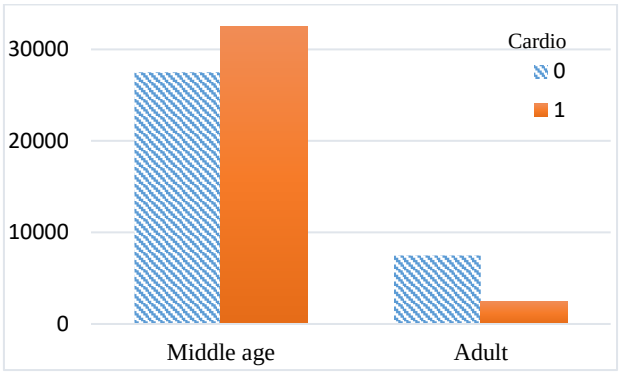

Fig. 5. CVD by age group

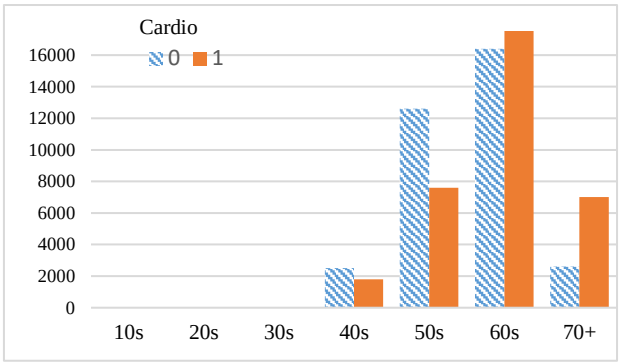

Fig. 6. Data age group by 10's

According to the graph, middle-aged people have a higher risk of cardiovascular disease (CVD) than do adults. Therefore, we can group people according to their age in 10-year intervals to better understand the association between age and CVD. This graph suggests that there aren't any CVD patients in the age groups of 10, 20, or 30. Furthermore, compared to people in their 40s and 50s, those in their 60s and 70s are more vulnerable to CVD. Our modeling was based on this finding.

Next, we'll look at how the genders are distributed. Table III indicates that the gender categorization in our dataset is represented by the binary digits 0 and 1.

TABLE III
Gender Ratio

Gender	Height	Alcohol consumption	Gender Ratio
1	161.355612	1161	45530
2	169.947895	2603	24470

Essentially, women and men have distinct characteristics that allow for distinction. We can determine the categories from the dataset by looking for the following characteristics.

- Height: Generally, men are taller than women
- Alcohol: Compared to women, men drink more of it.
- Gender ratio: Men outnumber women.

It is clear from the comparisons that Label 1 represents females and Label 2 represents males, with Label 2 showing greater values than Label 1 for females. Additionally, compared to females (Label 1), males (Label 2) consume more alcohol. The graph (figure 7) shows that while the gender population is not balanced, the disease label is spread equally. There are twice as many females as males (Label 1). We should take into account either the oversampling of men or the undersampling of women when creating an unbiased model.

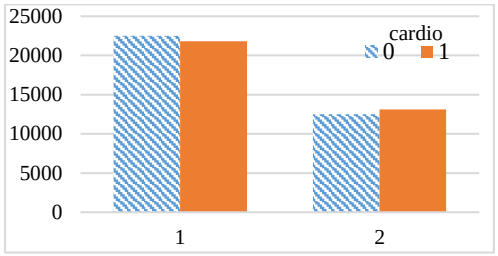

Fig. 7. Gender Ratio

The Body Mass Index (BMI) of the patient is a significant risk factor for heart disease. Based on height and weight, BMI is a measure of body fat that can be used to determine whether or not a person's health is within normal limits. To compute it, we utilize a particular formula: BMI is measured in kg/m2. kg: Kilograms of weight, m2: Squared meters of height The normal range for BMI is 18.5 to 25.

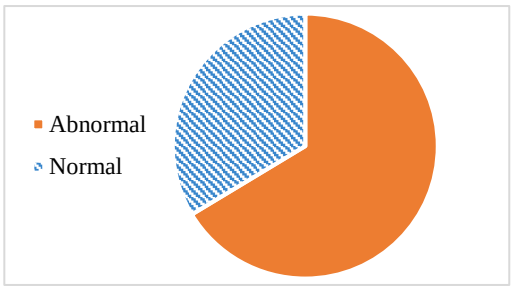

Fig. 8. BMI Graph

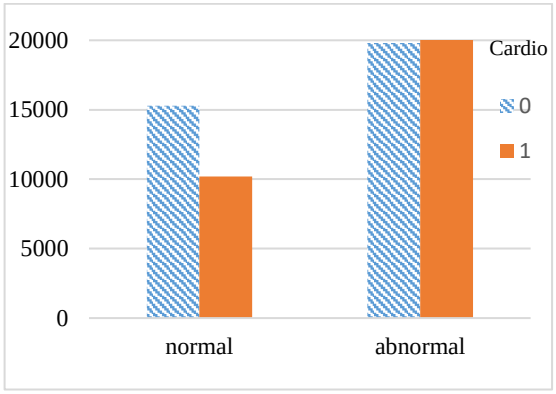

Fig. 9. Impact of BMI

A useful viewpoint is provided by this graph (Figure 9), which shows that people with aberrant BMI are more likely to have CVD, whereas people with normal BMI scores are less likely to develop CVD. The majority of the patients in our database have abnormal BMIs. Subsequently, to gain a more profound insight into the habits of patients and their potential contribution to increasing the risk of cardiovascular disease (CVD), we perform some analysis.

Based on the graphs presented in Figure 10, we may infer that for each of the three factors:

- **Alcohol**: Patients who do not drink are less than those who do, and their distribution among cardiac patients who are actively treated is reversed.
- **Smoking**: In our dataset, the number of smokers is lower than that of non-smokers. The majority of CVD patients are smokers.
- **Physical Activity**: Patients who engage in physical activity outnumber those who do not. Patients who engage in physical activity are less vulnerable to CVD.

We can conclude that alcohol consumption and smoking seem not to be contributing features to CVD classification.

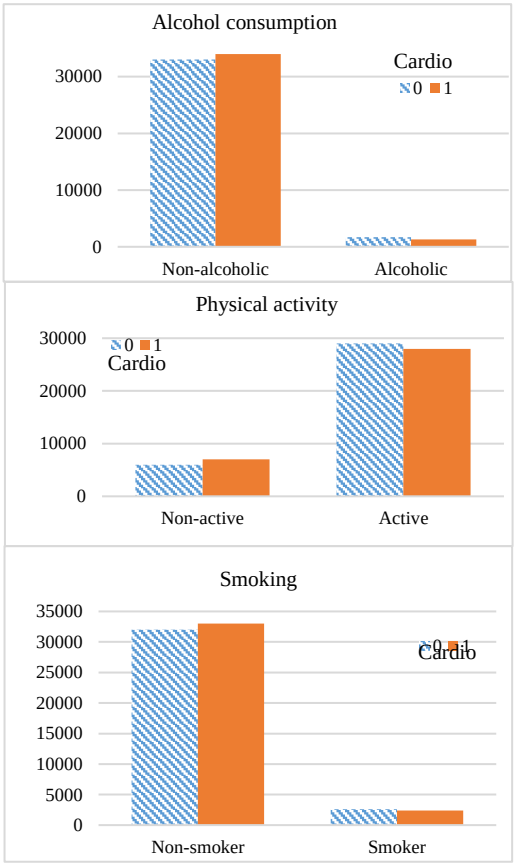

Fig. 10. Impact of patients' habits on CVD.

Afterward, we check the Blood Pressure (BP) impact. Blood pressure is indicated by two numbers.

- The first number is **systolic pressure:** which is the pressure in the arteries when the heart contracts or beats.
- The second number is **diastolic pressure:** Which corresponds to the pressure in the arteries when the heart relaxes.

Afterward, we check the Blood Pressure (BP) impact. Two numbers represent the blood pressure.

- **Systolic pressure,** or the pressure in the arteries during a heartbeat or contraction,
- **Diastolic pressure**, or the pressure in the arteries during a heartbeat

The normal interval for:

- **Systolic blood pressure**, the range should be **120 - 180 mm.**
- For **Diastolic blood pressure**, the range should be **80 - 120 mm.**

According to medical professionals, blood pressure might rise as high as 370/360 mm Hg. This indicates that if we remove ap_hi outlier values greater than 200, there is no chance of missing data. We examine which blood categories are most affected by CVD after eliminating outliers and classifying blood categories according to range.

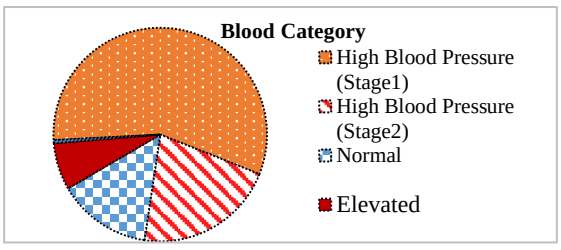

Fig. 11.Blood category

The plots in Figure 11 and Figure 12 show that high blood pressure stage 2 is more prone to CVD.

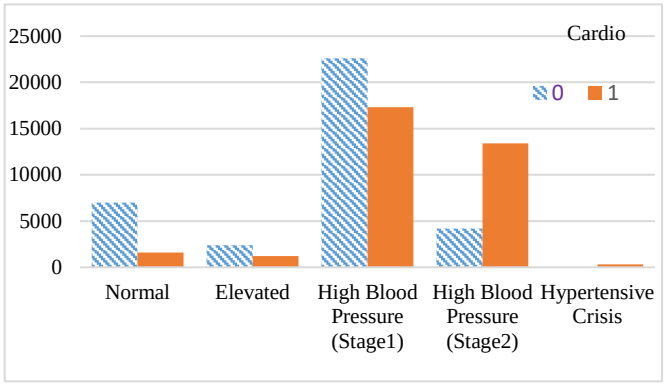

Fig. 12. Impact of blood category on CVD

Next, we examine the effect of cholesterol. In fact, an excessive amount of cholesterol in the blood can develop atherosclerosis, a type of heart disease, which is a build-up in the artery walls. Blood flow to the heart muscle is slowed down or blocked as a result of the arteries narrowing.

In contrast to patients with normal cholesterol levels, those who are well above normal and above normal have the largest risk for CVD, as demonstrated by the visualization (Figure 13).

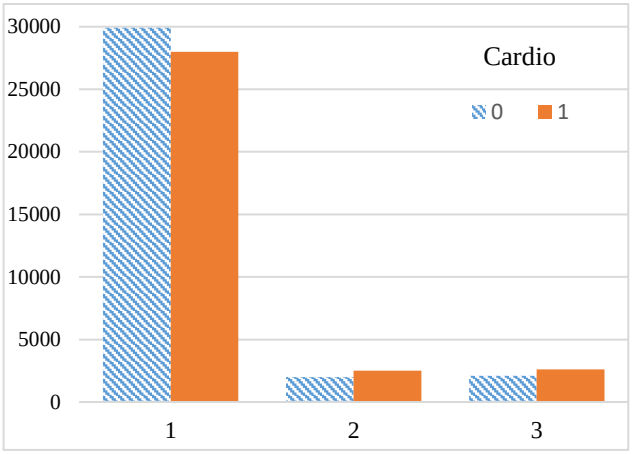

Fig. 13. Impact of Cholesterol

Glucose is the final factor in our dataset. The graph og Figure 14 demonstrates that individuals with normal glucose levels have a lower risk of cardiovascular disease (CVD) than do abnormal patients.

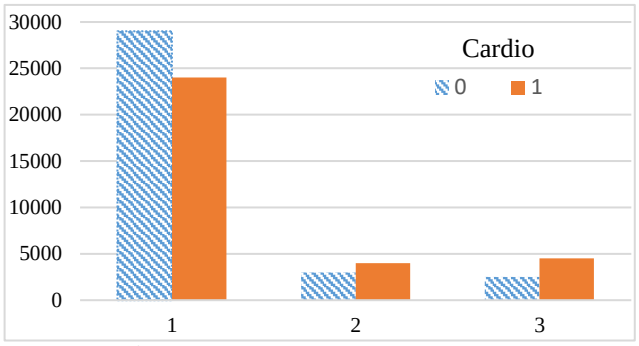

Fig. 14. Impact of glucose

Lastly, in order to prioritize the engineering features, we create a correlation graph. After that, we went on and eliminated attributes that were constant, redundant, and highly correlated.

The correlation between the most significant features is displayed in the graph below (Figure 15).

IV. Modeling and Experimental Results

A. Modeling

We will use a variety of machine learning and deep learning algorithms on our data at this point of the project, parameters, and evaluate the performance results of each technique.

The characteristics of gender_tees, BMI, cholesterol, glucose, physical activity, and blood category (blood pressure) are the most relevant for the early prediction of CVD, according to the analysis of each feature in our dataset that we conducted in the earlier section of this study (Figure 16).

- gender_tees
- gender
- BMI
- BMI_State
- Cholesterol
- glucose
- physical activity
- blood pressure

Fig. 16. Most important dataset features

	gender	age_tees	BMI	BMI State	cholesterol	gluc	active	blood category	cardio
gender	1.000	-0.012	-0.097	0.042	-0.036	-0.020	0.006	-0.062	0.008
age_tees	-0.012	1.000	0.083	-0.083	0.144	0.088	-0.009	-0.100	0.229
BMI	-0.097	0.083	1.000	-0.571	0.146	0.101	-0.014	-0.043	0.166
BMI State	0.042	-0.083	-0.571	1.000	-0.128	-0.081	0.002	0.033	-0.150
cholesterol	-0.036	0.144	0.146	-0.128	1.000	0.452	0.010	-0.008	0.221
Gluc	-0.020	0.088	0.101	-0.081	0.452	1.000	-0.007	-0.016	0.089
active	0.006	-0.009	-0.014	0.002	0.010	-0.007	1.000	0.015	-0.057
blood category	-0.062	-0.100	-0.043	0.033	-0.008	-0.016	-0.015	1.000	-0.057
cardio	0.008	0.229	0.166	-0.150	0.221	0.089	-0.036	-0.057	1.000

Fig. 15. Data correlation

To begin our medialization process we split data into 30 % in the test dataset and the remaining 70% in the traini ng dataset (Figure 17).

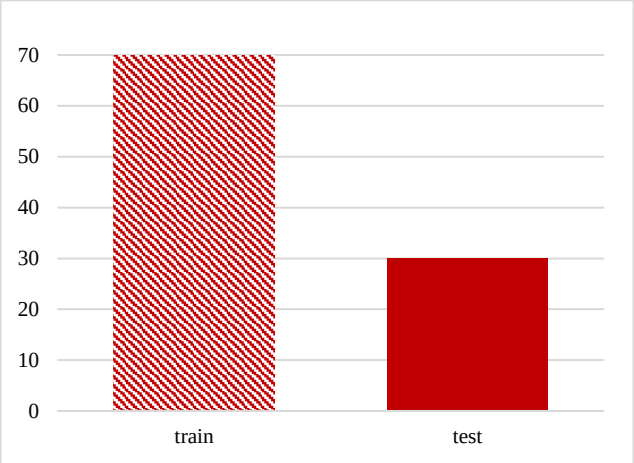

Fig. 17. Data splitting

We use a random forest classifier to calculate the Gini importance of each feature.

B. Model results

With more than 60 predictive modeling algorithms at our disposal, it's critical to comprehend the nature of the issue and the particular requirements of its solution. In this case, we are trying to find correlations between the input and output variables through both regression and classification.

a) The Logistic Regression algorithm's experimental results: all default values are maintained, and the max_iter attribute is fixed at 500 iterations. With the training data, we obtain an accuracy of 63.3%, and with the test data, 62.49%.
b) Support Vector Machines algorithm experimental results: 63.63% and 62.51% accuracy, respectively, are obtained from the SVM method using training and test data.
c) The k-Nearest Neighbors algorithm's experimental findings show that three is the ideal value for the K parameters. With training data, it yields an accuracy of 77.53%, whereas with test data, it yields 58.94%.
d) Gaussian Naive Bayes experiment results: The GNB algorithm achieved 64.2% accuracy in training data and 63.4% accuracy in test data.
e) Perceptron algorithm experimental results: a single-layer neural network is represented by a perceptron. In training data, its accuracy is 54.45%, and in test data, it is 54.11%.
f) Stochastic Gradient Descent experimental results: 61.96% and 61.89%, respectively, were the algorithm's accuracy in training and test data.
g) Decision Tree Classifier experimental results: The DT algorithm provides the highest accuracy of 91.7% for training data, but 51.54% for test data.
h) Random Forest algorithm's experimental outcomes: With the following values, the best accuracy was obtained: n_estimators = 300, cv = 5. In test data, accuracy was 52.83%, whereas in training data, it was 91.7%.
i) The XGB Classifier's experimental results: The XGB algorithm's optimal accuracy was obtained using the following parameters:
This algorithm's accuracy was 67.03% for test data and 70.8
j) LGB Classifier's experimental findings show that it performs 71.31% accurately on training data and 66.87% accurately on test data.
k) Gradient Boosting Classifier experiment results: the settings for GBC were {'max_depth': 9 'n_estimators': 867}, and the accuracy in test data was 57.68% and in training data was 82.22%.
l) Ridge Classifier experimental results: Using training data, this method yields an accuracy of 63.32%, and testing data, an accuracy of 62.55%.
m) Bagging Classifier experimental results: 89.64% accuracy in training data and 54.53% accuracy in test data.
n) Extra Tree Classifier experimental results: The ETC's optimal parameters were:
This algorithm's accuracy was 62.72% in test data and 62.8% in training data.
o) First Neural Network Experiment Results:
The various parameters of our neural network in this work are as follows:
- Inputs: 8 neurons (8 characteristics): blood pressure, glucose, cholesterol, gender, gender_tee, BMI, BMI_State, and physical activity
- Output: 1 neuron (1 output) CVD outcome.
- Layers: There are two dropout layers with 16 and 64 neurons each, and two hidden layers with 16 and 32 neurons each.
- The Hyper-parameters: Epochs = 500, Batch size = 64.
-The cost function and the optimizer: binary_crossentropy, Adam.
- Activation functions: RELu, sigmoid

Layer (type)	Output Shape	Param #
dense_24 (Dense)	(None, 16)	144
dropout_12 (Dropout)	(None, 16)	0
dense_25 (Dense)	(None, 64)	1088
dropout_13 (Dropout)	(None, 64)	0
dense_26 (Dense)	(None, 32)	2080
dense_27 (Dense)	(None, 1)	33

Total params: 3,345
Trainable params: 3,345
Non-trainable params: 0

Fig. 18. 1st Neural Network architecture

The NN gives us an accuracy of 66.27% in training data and 66.0% in test data.
p) Experimental results of 2nd Neural Network:
In this work, the different parameters of our neural network are as follows:
- Inputs: 8 neurons (8 characteristics).
- Output: 3 neuron (3output) CVD result.
- Layers: The network consists of 2 hidden layers containing 32 neurons and two dropout layers.
- The Hyper-parameters: Epochs = 120, Batch size = 32.
- The cost function and the optimizer: sparse_categorical_crossentropy, Adam.
- Activation functions: RELu, softmax.

This second NN algorithm gives us an accuracy of 68.45% in training data and 67.82% in test data.

Layer (type)	Output Shape	Param #
dense_28 (Dense)	(None, 8)	72
dense_29 (Dense)	(None, 32)	288
dropout_14 (Dropout)	(None, 32)	0
dense _30 (Dropout)	(None, 32)	1056
dropout _15 (Dense)	(None, 32)	0
dense_31 (Dense)	(None, 3)	99

Total params: 1,515
Trainable params: 1,515
Non-trainable params: 0

Fig. 19. 2nd Neural Network architecture

q) Experimental results of Adaboost Classifier: Parameters of this algorithm were: {'learning_rate': 0.0019, 'n_estimators': 482}. It gives us an accuracy of 60.75% in training data and 60.68% in test data.

r) Experimental results of Volting Classifier – Hard Volting: The hard volting classifier gives us an accuracy of 75.41% in training data and 61.02% in test data.

s) Experimental result of Volting Classifier – Soft Volting: The soft volting classifier gives us an accuracy of 84.68% in training data and 58.3% in test data.

C. Models evaluation

TABLE IV
OUTCOMES OF ALGORITHMS COMBINING VARIOUS HEART DISEASE RISK FACTOR COMBINATIONS

N°	Model	Score train	Score test	Score diff
14	ExtraTreesClassifier	62.8	62.76	0.04
20	AdaBoostClassifier	60.75	60.68	0.07
6	Stochastic Gradient Decent	61.96	61.89	0.07
5	Perceptron	54.45	54.11	0.34
17	Neural Network 2	68.45	67.82	0.63
12	RidgeClassifier	63.32	62.55	0.77
2	Linear SVC	63.31	62.53	0.78
4	Naive Bayes	64.2	63.4	0.8
0	Logistic Regression	63.3	62.49	0.81
1	Support Vector Machines	63.63	62.51	1.12
16	Neural Network 1	66.27	67.82	1.55
9	XGBClassifier	70.89	67.03	3.86
10	LGBMClassifier	71.31	66.87	4.44
18	VotingClassifier-hard voting	75.41	61.02	14.39
3	k-Nearest Neighbors	77.53	58.94	18.59
19	VotingClassifier-soft voting	84.68	58.3	26.38
11	GradientBoostingClassifier	89.22	57.68	31.54
13	BaggingClassifier	89.64	54.53	35.11
8	Random Forest	91.7	52.83	38.87
7	Decision Tree Classifier	91.7	51.54	40.16

The native dataset, which has more parameters, was subjected to the same techniques. The effectiveness of several feature combinations in early heart disease prediction is shown in Table V.

TABLE V
SECOND APPLICATION OF ALGORITHMS SCORES

N°	Model	Score train	Score test	Score diff
4	Naive Bayes	70.31	70.46	0.15
6	Stochastic Gradient Decent	64.18	64.42	0.24
10	LGBMClassifier	72.31	72.6	0.29
19	AdaBoostClassifier	71.1	71.45	0.35
0	Logistic Regression	71.6	71.99	0.39
5	Perceptron	65.16	64.72	0.44
1	Support Vector Machines	60.17	59.72	0.45
14	ExtraTreesClassifier	71.02	71.51	0.49
15	Neural Network 1	50.68	50.11	0.57
12	RidgeClassifier	72.19	72.77	0.58
16	Neural Network 2	49.32	49.9	0.58
2	Linear SVC	71.99	72.67	0.68
17	VotingClassifier-hard voting	76.63	72.41	4.22
9	XGBClassifier	79.72	73.43	6.29
18	VotingClassifier-soft voting	83.49	72.81	10.68
3	k-Nearest Neighbors	81.41	66.86	14.55
11	GradientBoostingClassifier	97.93	71.6	26.33
13	BaggingClassifier	98	69.48	28.52
8	Random Forest	99.99	71.22	28.77
7	Decision Tree Classifier	99.99	63	36.99

In training data, Random Forest and Decision Tree Classifier yielded the best accuracy of 99.99%. and the highest accuracy of 73.43% in test data that the XGBClassifier was able to get. In contrast to the first application, the score difference is altered, and the sequence of algorithms based on the least score difference is altered as well.

Figure 20 illustrates the graphical representation of the 20 algorithms utilized in both the training and testing datasets. The graphical representation offers a thorough summary of their performance based on several indicators.

The comparison makes it possible to comprehend each algorithm's performance in more detail, which helps to identify its advantages and disadvantages.

The findings indicate that combining various parameters (id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoking, alcohol, active, and cardio) yields optimal outcomes. As a result, it is clear from several earlier research that artificial intelligence is important in the healthcare industry [33].

We use the Python Jupyter Notebook online application to implement our model. Figure 21 displays the home screen of the mobile application, along with a user interface that allows patients to enter their data. The program then uses the entered parameters to determine and show the risk level for heart disease. When the estimated risk level is abnormally high, an automatic alarm is set off and is immediately forwarded to medical professionals.

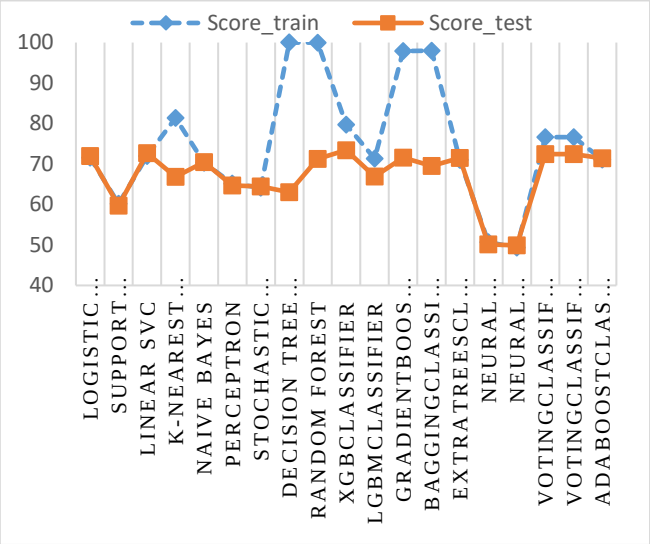

Fig. 20. Score of 20 popular models for training and test datasets

Additionally, the patient directly receives individualized advice via their mobile application, which improves prompt interventions and proactive health care. Healthcare professionals are empowered by the user interface's simplicity, which makes it simple for them to identify patients who are at a high risk of heart disease. This is accomplished by automatically having clever sensors record simple attributes, or by having patients submit them. The design is easy to use and makes identification quick and easy. This makes it easy for medical professionals to analyze and manage cardiovascular risks.

We have implemented our concept in both desktop and mobile applications. Our model functions flawlessly on a variety of platforms, guaranteeing consumers' usability and accessibility on desktop computers as well as mobile devices.

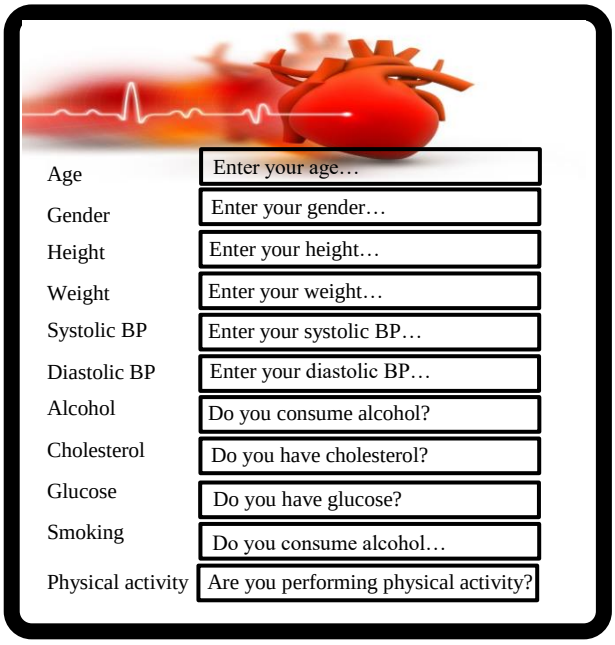

Fig. 21. Mobile application interface

V. Conclusion

In this work, we created a risk assessment model that, from a business understanding of important risk variables, allows the first prediction of cardiovascular illness. Based on the recommendations of cardiologists, we performed our analysis using a variety of feature selection techniques. Next, we determined the top 20 most popular algorithms for healthcare forecasting.

We evaluate various algorithms including Logistic Regression, Support Vector Machines, k-nearest Neighbors, Gaussian Naive Bayes, Perceptron, Stochastic Gradient Descent, Decision Tree, Random Forest, XGB, LGB, Gradient Boosting, Ridge, Bagging, Neural Network, Extra Tree, Adaboost, Volting Classifier (both Hard and Soft), by computing the score difference between training and test data, assessing each algorithm's performance, and confirming its accuracy. Next, we conduct a comparison analysis of the outcomes of 20 algorithms running on the treated dataset using just the eight features that were chosen, and the same 20 algorithms running on the original dataset using all of the features. Even though certain features might seem insignificant, we have found that the best outcomes come from combining several criteria.

In addition, we have created a user-friendly smartphone application that makes it easier for patients to enter important health information. The interface is user-friendly, enabling users to easily track and update their data. Furthermore, the program analyzes the supplied data using sophisticated algorithms, going beyond simple data entry. This gives customers important information about their cardiovascular health by enabling real-time monitoring and timely notifications in the event of possible cardiac problems.

Personalized dashboards, another element of the mobile app, provide patients with a thorough picture of their health patterns and help them make better decisions regarding their well-being. Furthermore, it encourages proactive healthcare management by providing personalized advice on lifestyle decisions and preventive actions based on each person's unique health profile. Our smartphone application combines data analytics, tailored insights, and an intuitive design to improve patient experience while promoting proactive heart health care.

References

[1] W. H. Organization, Global Status Report on No Communicable Diseases 2010, World Health Organization, Geneva, Switzerland, 2011.

[2] P. Kakria, N. K Tripathi, et P. Kitipawang. A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors. International journal of telemedicine and applications, vol. 2015, p. 8-8, 2015.

[3] K.M Sturgeon, L. Deng, SM. Bluethman, et al. A population-based study of cardiovascular disease mortality risk in US cancer patients. European heart journal, vol. 40, no 48, pp. 3889-3897, 2019.

[4] T. A. Assegie, "An optimized k-nearest neighbor based breast cancer detection," J. Robot. Control, vol. 2, no. 3, pp. 115–118, 2021.

[5] Singh, A., & Prakash, N. A Review of AI Models for Prediction and Detecting Heart Diseases for Improved Wellbeing. Vivekananda Journal of Research, Vol. 10, Special Issue, pp. 14-25, October 2021.

[6] M. A. A. R. Asif, M. M. Nishat, F. Faisal, et al., "Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease," Engineering Letters, vol. 29, no. 2, pp. 731–741, 2021.

[7] S. Joshi, & M. K Nair, Prediction of heart disease using classification based data mining techniques. In Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014, pp. 503-511. Springer India 2015.

[8] B. Martins, D. Ferreira, C. Neto, A. Abelha, & J. Machado, Data mining for cardiovascular disease prediction. Journal of Medical Systems, Vol 45, pp 1-8, 2021.

[9] K. S. Purushottam, K. Saxena, and R. Sharma, "Efficient heart disease prediction system," Procedia Computer Science, vol. 85, pp. 962–969, 2016.

[10] Al-sharu, Wafaa N., Alqudah, Ali Mohammad, QAZAN, Shoroq, Alqudah, A. Detection of Valvular Heart Diseases Using Fourier Transform and Simple CNN Model. IAENG International Journal of Computer Science, vol. 49, no 4, p. 985-993, 2022.
[11] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garcia-Magarino, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, vol. 21, pp. 1-21, 2018.
[12] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods," Computers & Electrical Engineering, vol. 84, p. 18, 2020.
[13] R. Thanga Selvi, et I. Muthulakshmi, an optimal artificial neural network based big data application for heart disease diagnosis and classification model. J Ambient Intell Human Comput. 2020.
[14] R.Katarya, et S.K Meena, Machine learning techniques for heart disease prediction: a comparative study and analysis. Health and Technology, vol. 11, p. 87-97, 2021.
[15] R.S. Chhillar, et al. Analyzing predictive algorithms in data mining for cardiovascular disease using Weka tool. International Journal of Advanced Computer Science and Applications, vol. 12, no 8, pp:144-150, 2021.
[16] S. Palaniappan, R. Awang. "Intelligent heart disease prediction system using data mining techniques," IEEE/AC International Conference on Computer Systems and Applications, Doha, pp: 108–115, 2008.
[17] C. C. Peng, C. W. Huang, and Y. C. Lai, "Heart Disease Prediction Using Artificial Neural Networks: A Survey," in 2nd IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2020, ECBIOS 2020, pp. 147–150, May 2020.
[18] S. B. Garg, P. Rani, and J. Garg, "Performance analysis of classification methods in the diagnosis of heart disease," in Lecture Notes in Networks and Systems, vol. 140, pp. 717–728, 2021.
[19] J. Deshmukh, M. Jangid, S. Gupte, and S. Ghosh, "Heart disorder prognosis employing knn, ann, id3, and SVM," in Advances in Intelligent Systems and Computing, vol. 1141, pp. 513–523, Feb. 2021.
[20] Tagashira, Mizuho et Nakagawa, Takafumi. Biometric authentication based on auscultated heart sounds in healthcare. IAENG International Journal of Computer Science, vol. 47, no 3, p. 343-349, 2020.
[21] T. A. Assegie, P. K Rangarajan, N.K Kumar, et al. An empirical study on machine learning algorithms for heart disease prediction. IAES International Journal of Artificial Intelligence, vol. 11, no 3, p. 1066, 2022.
[22] R.Katarya, et S.K Meena, Machine learning techniques for heart disease prediction: a comparative study and analysis. Health and Technology, vol. 11, p. 87-97, 2021.
[23] Javid, Irfan, Alsaedi, Ahmed Khalaf Zager, et Ghazali, Rozaida. Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method. International Journal of Advanced Computer Science and Applications, vol. 11, no 3, 2020.
[24] K. Slime, A. Maizate, L. Hassouni, et al. A System Architecture to Implement Deep Learning Techniques for Patients Monitoring with Heart Disease: Case of Telerehabilitation. Conf. BML'21, pp: 483-487, 2022.
[25] Y.Wang,, Y.Duan, Y.Li, H. Wu. "Smoke Recognition based on Dictionary and BP Neural Network". Eng. Lett. 31, 554–561, 2023.
[26] J. H. Joloudari et al., "Coronary artery disease diagnosis; ranking the significant features using a random trees model," Int. J. Environ. Res. Public Health, vol. 17, no. 3, p. 731, Jan. 2020.
[27] L. Rasmy, et al., "A Study of Generalizability of Recurrent Neural Network-Based Predictive Models for Heart Failure Onset Risk Using a Large and Heterogeneous EHR Data Set." J Biomed Inform, 84, pp. 11-16, May 2018.
[28] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," J. Big Data, vol. 7, no. 1, Dec. 2020.
[29] S. Elyassami, A.A Kaddour, Implementation of an incremental deep learning model for survival prediction of cardiovascular patients. IAES International Journal of Artificial Intelligence, vol. 10, no 1, p. 101, 2021.
[30] X. Liu, et al., "Coronary Artery Fibrous Plaque Detection Based on Multi-Scale Convolutional Neural Networks," J. Sign Process. Syst., 92, 3, pp. 325-333, 2020.
[31] S. Ulianova, "Cardiovascular Disease dataset", Kaggle repository, kaggle.com, Jan 2019.
[32] J.S Rumsfeld, K.E Joynt, et T.M Maddox, Big data analytics to improve cardiovascular care: promise and challenges. Nature Reviews Cardiology, vol. 13, no 6, p. 350-359, 2016.
[33] Isha Baokar, and Lili He, "Memristive Neural Networks Application In Predicting of Health Disorders," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2023, 5-7 July, 2023, Hong Kong, pp94-99.