Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques

Elias Dritsas, Sotiris Alexiou and Konstantinos Moustakas
Department of Electrical and Computer Engineering, University of Patras, 26504 Rion, Greece

Keywords: CVDs, Machine Learning, Risk Prediction.

Abstract: Cardiovascular diseases (CVDs) are the leading cause of death worldwide and a major public health concern, with heart diseases being the most prevalent ones, thus the early prediction is being considered as one of the most effective measures for CVDs control. The risk evaluation for CVD occurrence on participants (men and women) especially aged older than 50 years with the aid of Machine Learning (ML) models is the main purpose of this research paper. The performance of supervised ML models is compared in terms of accuracy, sensitivity (or recall) in identifying those participants that actually suffer from a CVD and Area Under Curve (AUC) score. The experimental analysis demonstrated that the Logistic Regression classifier is the most appropriate against Naive Bayes, Support Vector Machine (SVM) and Random Forest with 72.1% accuracy, recall and 78.4% AUC.

1 INTRODUCTION

The term ”cardiovascular disease” contains a wide range of disorders, including all pathological changes involving the heart and/or blood vessels. These diseases include hypertension, coronary heart disease, heart failure, angina, myocardial infarction and stroke (Kumar and Ramana, 2021). Cardiovascular diseases have been the leading cause of death in developing countries for the past 15 years, and by 2030 deaths will exceed 20 million per year. A taxonomy of CVDs is presented in Table 1.

Table 1: Taxonomy of CVDs.

Heart Disease	Description
Coronary	disease of the blood vessels supplying the heart muscle
Cerebrovascular	disease of the blood vessels supplying the brain
Peripheral arterial	disease of blood vessels supplying the arms and legs
Rheumatic	damage to the heart muscle and heart valves from rheumatic fever caused by streptococcal bacteria
Congenital	malformations of heart structure existing at birth
Deep vein thrombosis Pulmonary embolism	blood clots in the leg veins, which can dislodge and move to the heart and lungs
Heart attacks, strokes	acute events mainly caused by a blockage that prevents blood from flowing to the heart or brain

Heart diseases and stroke constitute one of the biggest causes of morbidity and mortality among the population worldwide (Roth et al., 2017), with the most important behavioral risk factors being unhealthy diet, sedentary lifestyle, smoking and excessive use of alcohol. The effects of behavioral risk factors may show up as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. In detail, the main risk factors for the occurrence of cardiovascular diseases include (Wilkins et al., 2017), (Abdalrada et al., 2022):

- Obesity/High BMI: obesity alone is a high-risk factor for CVDs
- Physical activity: sedentary lifestyle increases the risk of CVDs
- Alcohol consumption: excessive alcohol use can raise blood pressure levels and increase levels of triglycerides, thus increasing the risk for cardiovascular diseases
- Smoking and secondhand smoke: nicotine raises blood pressure
- Hyperlipidemia: also known as High cholesterol or hypercholesterolemia
- Dyslipidemia: abnormal level of fat or cholesterol in human’s blood vessels
- Family history, Psychosocial stress, the coexistence of other chronic conditions: Type 2 diabetes, Arterial hypertension

The ”2013 ACC/AHA Guideline on the Assessment

Dritsas, E., Alexiou, S. and Moustakas, K.
Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques.
DOI: 10.5220/0011088300003188
In *Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2022)*, pages 315-321
ISBN: 978-989-758-566-1; ISSN: 2184-4984

of Cardiovascular Risk” (Yancy et al., 2013) provides detailed recommendations for estimating cardiovascular disease risk in the clinical practice, considering several factors including age, gender, race, cholesterol and blood pressure levels, diabetes and smoking status, and the use of blood pressure-lowering medications. In Europe, the 10-year risk factor of fatal CVD is estimated based on different charts established by the European Society of Cardiology for high-risk and low-risk populations across Europe, which may be further adapted to national or regional specific charts based on published mortality data.

In the literature, the CVDs risk prediction is addressed with either appropriate risk tools or the aid of machine learning. In (Gale et al., 2014), the Framingham cardiovascular disease risk score and incident frailty studied on English cohort data for ageing participants. Moreover, the systematic coronary risk evaluation (SCORE) has been suggested to predict the 10-year risk of cardiovascular death in Europe or the QRISK to predict the composite outcome of coronary heart disease and ischaemic stroke. Others employ machine learning techniques, also aiming at predicting potential risk of CVDs (Mohan et al., 2019), (Yang et al., 2020).

ML is a branch of artificial intelligence (AI) and a powerful tool in the medical field, as it can help predict various diseases. In (Dinesh et al., 2018), various data-driven approaches are presented to predict diabetes and cardiovascular disease with ML models. Here, we will solely focus on its application to cardiovascular medicine (Haq et al., 2018). Our purpose is to identify predictive data patterns and high-risk CVD groups among the elderly. Moreover, we aim to create personalized risk models that will be part of the predictive AI tools integrated into the SmartWork (Kocsis et al., 2019) and GATEKEEPER systems. The presented method for the risk prediction of CVDs occurrence was developed and validated independently with a publicly available dataset and, in parallel, as part of the projects with pilot data. The incorporation of the ML models into the Long-term Risk Prediction tools of the SmartWork system aims to design a smart age-friendly healthy living and working environment for office workers. The GATEKEEPER system pursues to sustain, as healthy as possible, the life of older people living at home, preventing the occurrence of CVD, type 2 diabetes mellitus (T2DM)(Fazakis et al., 2021), high cholesterol, hypertension (Dritsas et al., 2021), chronic obstructive pulmonary disease-COPD (Hussain et al., 2021) (chronic conditions related to Metabolic Syndrome-MetS).

Given that MetS combines risk factors that promote the development of cardiovascular disease (CVD) and type 2 diabetes (T2DM)(Hoyas and Leon-Sanz, 2019), as a first approach, our paper aims to present a methodology for correctly identifying those at risk of diagnosed with a CVD in long-term. For this purpose, the classification performance of various ML models is estimated on each test instance from a CVD dataset. The ML models that achieve the highest recall (namely, high sensitivity) and Area Under Curve (AUC) show that the CVD class can be predicted correctly. The main contribution of this work is a comparative evaluation of different ML models on a balanced dataset and the proposal of a Logistic Regression model for the long-term CVD risk prediction. In the upcoming sections, the main steps of the employed process are demonstrated.

The rest of this paper is organized as follows. Section 2 presents the main parts of the methods for the long-term risk prediction of CVD. Section 3 makes an analysis of the dataset features and Section 4 describes the pre-processing steps for the design of the training and testing dataset and feature ranking. Section 5 presents the experimental set up and the classification performance of ML techniques. Ultimately, Section 6 concludes the paper and notes future directions of the current outcomes.

2 MACHINE LEARNING METHODS

Data science and especially machine learning has been widely used in the field of medicine for the risk analysis of several chronic conditions. The most common application of these models aims to determine the most suitable factors for the long-term risk prediction to avoid serious health complications (due to certain symptoms) and support health care management.

In this study, the forecasting performance of four different machine learning models is presented. In particular, the Naive Bayes, SVM, Logistic Regression and Random Forest are utilized to estimate the long-term risk of an older person being diagnosed with cardiovascular disease.

The dataset is separated into a training set of size M, a test set of size N. A categorical variable c which captures the class label of an instance i in the dataset. In the context of this work, the investigating problem has two possible classes, e.g., c = ”CVD” or ”Yes” or c= ”Non-CVD” or ”No”. The features vector of an instance i is captured by $\mathbf{f}_i = \left[f_{i1}, f_{i2}, f_{i3}, \ldots, f_{in}\right]^T$ (with $M \gg n$).

Our aim is to achieve high recall or sensitivity and Area Under Curve (AUC) through supervised machine learning, meaning that the CVD class can be predicted correctly. Our methodology for CVD prediction includes the following models which are explained below.

2.1 Naive Bayes

Naive Bayes (Dinesh et al., 2018) is a simple classifier founded on the Bayes theorem that supposes highly independent attributes known as predictors to achieve probability maximization. The main interest is to find the posterior probabilities

$$P(c|f_{i1},...,f_{in}) = \frac{P(f_{i1},...,f_{in}|c)P(c)}{P(f_{i1},\cdots,f_{in})} \quad (1)$$

where $P(f_{i1},...,f_{in}|c) = \prod_{j=1}^{n} P(f_{ij}|c)$ is the probability of predictors given class, $P(f_{i1},...,f_{in})$ is the prior probability of predictors and $P(c)$ is the prior probability of class. The testing data is classified based on the probability of association:

$$\hat{c} = \arg\max P(c) \prod_{j=1}^{n} P(f_{ij}|c)$$

for $c \in \{CVD, Non-CVD\}$,

2.2 Support Vector Machine

The SVM is a machine learning algorithm that has been used in medicine due to its high classification performance. It can be used to solve a binary classification problem (as the one investigated here), either linear or non-linear using a Kernel function to map the nonlinear data to high dimensional feature space. In the linear case, the instances are separated with a hyperplane, called support vector, of the form $\mathbf{w}^T\mathbf{f}+b$, where $\mathbf{w}$ is n dimensional coefficient vectors normal to the hyperplane of the surface and b is offset value from the origin. The value of $\mathbf{w}$ and b are calculated, and the linear discriminant function can be written as $g(\mathbf{w}) = sign(\mathbf{w}^T\mathbf{f}+b)$.

2.3 Logistic Regression

Logistic Regression predicts the class label of the input features based on their values using a binary logistic regression model. Assuming $p = P(c = \text{`}CVD)$, $log_b(\frac{p}{1-p}) = \beta_0 + \sum_{j=1}^{n} f_{ij}\beta_j$, with $(\beta_1, \beta_2, \ldots, \beta_n)$ be the regression weights attached to features row vector $f_i, i = 1,2,\ldots,M$. Isolating p, if it is greater than the threshold, it is set to CVDs class (Yes); otherwise, it is set to Non-CVDs class (No). The β coefficients values of the logistic regression algorithm are estimated from training data using maximum-likelihood estimation.

2.4 Random Forest

The Random forest method (Yang et al., 2020) is a supervised learning algorithm that can be employed for the classification of instances as either CVD or Non-CVD. A Random forest algorithm is an ensemble of decision trees that are created on data samples. The more decision trees are considered, the more robust is the constructed forest. It combines the prediction from each tree separately to finally select the best outcome by means of majority voting.

3 DATASET DESCRIPTION

The training and test dataset for the cardiovascular diseases risk prediction model was constructed based on the CVDs dataset (an open-source dataset derived from Kaggle), which consists of 70000 participants. It is a balanced dataset of approximately equal healthy and diagnosed with CVD participants, which contains 11 features (4 demographic, 4 examination, and 3 social history) which include age (years), gender, weight (Kg) and height (m^2) from which $BMI = \frac{weight}{(height)^2}$ was derived, cholesterol and glucose levels characterized as normal, above normal or well above normal, physical activity, drinking and smoking habits with values yes or no, systolic and diastolic blood pressure (mmHg), physical activity. Notice that the blood pressure measurements have been recorded at the moment of medical examination. To identify linear correlations between the features and the target class we employed Pearson's correlation coefficient (CC) (Mukaka, 2012) defined as

$$C = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \quad (2)$$

where x_i and y_i are the values of features x and y for the i-th individual.

The results presented in Table 2 were derived setting an alpha value equal to 0.01 to compute a 99% confidence interval. From Table 2, we verify high linear dependence among BMI and weight features as the corresponding coefficient is equal to 0.7620. Also, a moderate correlation value of 0.4520 was observed between glucose and cholesterol. A low correlation value of 0.3400 was noted between smoke and alcohol. Focusing on the CVD class and the features that capture age and cholesterol, low correlation values are recorded (since $0.2 \leq C \leq 0.39$)from the current dataset.

Furthermore, in Tables 4-8, the distributions of selected participants considering different combinations

Table 2: Pearson Correlation Coefficients Matrix between the features and target CVD class.

	Age	Gender	Height	Weight	BMI	SBP	DBP	Chol	Glucose	Smoke	Alcohol	Physical Activity	CVD
Age	**1.0000**	-0.0230	-0.0810	0.0540	0.0850	0.0210	0.0180	0.1540	0.0990	-0.0480	-0.0300	-0.0100	0.2380
Gender	-0.0230	**1.0000**	0.4990	0.1550	-0.0970	0.0060	0.0150	-0.0360	-0.0200	0.3380	0.1710	0.0060	0.0080
Height	-0.0810	0.4990	**1.0000**	0.2910	-0.2910	0.0050	0.0060	-0.0500	-0.0190	0.1880	0.0940	-0.0070	-0.0110
Weight	0.0540	0.1550	0.2910	**1.0000**	0.7620	0.0310	0.0440	0.1420	0.1070	0.0680	0.0670	-0.0170	0.1820
BMI	0.0850	-0.0970	-0.2910	0.7620	**1.0000**	0.0250	0.0350	0.1460	0.1010	-0.0270	0.0140	-0.0140	0.1660
SBP	0.0210	0.0060	0.0050	0.0310	0.0250	**1.0000**	0.0160	0.0240	0.0120	-0.0010	0.0010	0	0.0540
DBP	0.0180	0.0150	0.0060	0.0440	0.0350	0.0160	**1.0000**	0.0240	0.0110	0.0050	0.0110	0.0050	0.0660
Chol	0.1540	-0.0360	-0.0500	0.1420	0.1460	0.0240	0.0240	**1.0000**	0.4520	0.0100	0.0360	0.0100	0.2210
Glucose	0.0990	-0.0200	-0.0190	0.1070	0.1010	0.0120	0.0110	0.4520	**1.0000**	-0.0050	0.0110	-0.0070	0.0890
Smoke	-0.0480	0.3380	0.1880	0.0680	-0.0270	-0.0010	0.0050	0.0100	-0.0050	**1.0000**	0.3400	0.0260	-0.0150
Alcohol	-0.0300	0.1710	0.0940	0.0670	0.0140	0.0010	0.0110	0.0360	0.0110	0.3400	**1.0000**	0.0250	-0.0070
Physical Activity	-0.0100	0.0060	-0.0070	-0.0170	-0.0140	0	0.0050	0.0100	-0.0070	0.0260	0.0250	**1.0000**	-0.0360
CVD	0.2380	0.0080	-0.0110	0.1820	0.1660	0.0540	0.0660	0.2210	0.0890	-0.0150	-0.0070	-0.0360	**1.0000**

Table 3: Distribution per gender group of healthy and diagnosed with CVD in the dataset.

	CVD		
Gender	**No**	**Yes**	**Total**
Female	32,73%	32,31%	65,04%
Male	17,30%	17,66%	34,96%
	50,03%	**49,97%**	**100,00%**

Table 4: Distribution per age group of healthy and diagnosed with heart disease in the dataset.

	CVD		
Age Group	**No**	**Yes**	**Total**
30-34	0,01%	0,00%	0,01%
35-39	0,45%	0,13%	0,58%
40-44	9,62%	4,12%	13,74%
45-49	7,04%	5,41%	12,46%
50-54	14,95%	12,64%	27,59%
55-59	10,23%	12,68%	22,91%
60-64	7,56%	14,65%	22,21%
65-69	0,16%	0,35%	0,50%
	50,03%	**49,97%**	**100,00%**

Table 5: Distribution per blood pressure category of healthy and diagnosed with heart disease in the dataset.

	CVD		
Blood Pressure	**No**	**Yes**	**Total**
Elevated	3,02%	1,45%	4,46%
Hypertension I	31,44%	25,60%	57,05%
Hypertension II	4,89%	19,87%	24,76%
Normal	10,68%	3,05%	13,73%
Total	**50,03%**	**49,97%**	**100,00%**

Table 6: Distribution of healthy and diagnosed with CVD in relation to smoke and alcohol features in the dataset.

		CVD		
Smoke		**No**	**Yes**	**Total**
No		45,40%	45,79%	91,19%
Alcohol	**No**	44,10%	44,36%	88,46%
	Yes	1,30%	1,43%	2,73%
Yes		4,63%	4,18%	8,81%
Alcohol	**No**	3,16%	3,01%	6,16%
	Yes	1,47%	1,18%	2,65%
		50,03%	**49,97%**	**100,00%**

Table 7: Distribution of healthy and diagnosed with CVD in relation to glucose level feature in the dataset.

	CVD		
Glucose level	**No**	**Yes**	**Total**
Normal	41,90%	32,94%	74,84%
Above Normal	5,43%	8,21%	13,64%
Well Above Normal	2,70%	8,82%	11,52%
	50,03%	**49,97%**	**100,00%**

of the features and the target classes (yes or no) are presented. In Table 3, the 32.31% of participants that suffer from CVD are women which are almost twice greater than men with CVD.

Moving on to Table 4, it is shown that approximately 70% of the total participants are elderly (age ≥ 50) which is the target group of people that concern the current study. From older participants, about 40% belongs to the CVD class.

Table 5 presents a classification of participants according to the values of systolic and diastolic blood pressure. The current data has shown that about 45% of the participants have been diagnosed with a CVD and are categorized as hypertensive (I or II).

Table 6 shows the distribution of participants in Non-CVD and CVD classes according to smoke and alcohol features. Here, it should be noted that a small percentage of participants (1.18%) diagnosed with CVD are simultaneous smokers and consume alcohol.

Tables 7, 8 show the distribution of participants in Non-CVD and CVD classes according to glucose

Table 8: Distribution of healthy and diagnosed with CVD in relation to cholesterol level feature in the dataset.

	CVD		
Cholesterol level	**No**	**Yes**	**Total**
Normal	41,90%	32,94%	74,84%
Above Normal	5,43%	8,21%	13,64%
Well Above Normal	2,70%	8,82%	11,52%
	50,03%	**49,97%**	**100,00%**

and cholesterol features. Here, it should be noted that a small percentage of participants (1.18%) diagnosed with CVD are simultaneous smokers and consume alcohol. Moreover, there is no knowledge of the amount of consumption to understand to what extent the participants' habits are harmful to their health.

Table 9 records the distribution of participants in Non-CVD and CVD classes according to physical activity and BMI classes. It is observed that 11,55% of participants although they belong to healthy BMI class and are physically active, they have been diagnosed with CVD. Also, in obese and overweight classes, the number of participants diagnosed with CVD is approximately similar. Both in this case, there is no knowledge of exercise-related characteristics (intensity, duration, frequency, type) to understand to what extent physical activity can be beneficial for CVD patients health.

Finally, it should be highlighted that the small percentage of participants with CVD who smokes, consumes alcohol and is physical activity in Table 3 is reflected as negatively low correlation with CVD class.

4 DATA PREPROCESSING

The data preprocessing was evaluated using Stata V.14 tool kit. Stata is a general-purpose statistical software package developed by StataCorp for data manipulation, visualization, statistics, and automated reporting. The constructed training and testing dataset, which is based on the CVDs dataset, was created by studying all attributes that are correlated and potentially relevant for the identification of Cardiovascular Diseases. In order to establish a unified set of attributes for each participant, we derived harmonized variables for each attribute, by transforming all numeric values to nominal values, according to predefined attribute rules. Moreover, the class was created based on CVD's dataset cardiovascular diseases feature. The class distribution is balanced containing in total 70,000 observations.

In order to examine all attributes' correlation with the class, we used a feature selection method that ranks all attributes with respect to the relevance to the specific class. More specific, we utilized a feature selection method based on a variation of Random Forests (Genuer et al., 2010) and according to this method, the attributes are ranked using the Gini importance score of the model's trees. The Gini index (Sundhari, 2011) is estimated as follows

Table 9: Distribution of healthy and diagnosed with CVD in relation to Physical Activity level and BMI class in the dataset.

		CVD		
Physical Activity		**No**	**Yes**	**Total**
No		**9,11%**	**10,52%**	**19,63%**
BMI Class	**Underweight**	0,10%	0,05%	0,15%
	Overweight	3,22%	3,71%	6,93%
	Obese III	0,16%	0,45%	0,61%
	Obese II	0,41%	0,90%	1,31%
	Obese I	1,22%	2,27%	3,49%
	Healthy	4,01%	3,14%	7,15%
Yes		**40,92%**	**39,45%**	**80,37%**
BMI Class	**Underweight**	0,56%	0,21%	0,77%
	Overweight	14,41%	14,63%	29,04%
	Obese III	0,68%	1,41%	2,09%
	Obese II	1,61%	3,36%	4,97%
	Obese I	5,71%	8,29%	14,00%
	Healthy	17,95%	11,55%	29,51%
		50,03%	**49,97%**	**100,00%**

$$Gini = 1 - \sum_{i=1}^{c} p_i^2, \tag{3}$$

where c is the number of classes and p_i is the relative frequency of class i in the dataset. In our case parameter $c = 2$ captures the CVDs and non-CVDs classes.

5 EXPERIMENTS AND EVALUATION

5.1 Experiments Setup

The experiments were evaluated using WEKA. WEKA is a JAVA-based data mining toolkit created and free software tool distributed under the GNU General Public License that provides a large library of methods and models for classification, clustering, prediction, feature selection. For the purposes of this experiment, we split the dataset into two parts, 30% testing and 70% training. To do that, we first randomized the dataset, in order to create a random permutation. Then, we applied the RemovePercentage method with 30% and saved the resulting dataset as training. We also applied the same filter choosing invertSelection in order to pick the rest of the data (30%) which is the testing dataset.

Table 10: Performance of ML models for CVDs risk prediction.

Algorithm	Accuracy	Recall	AUC
Naive Bayes	59.59%	59.60%	69.4%
SVM	70.61%	70.60%	70.6%
Random Forest	70.86%	70.90%	76.6%
Logistic Regression	72.06%	72.10%	78.4%

5.2 Experiments Results

For the specific experiment, accuracy, recall and AUC were considered as performance metrics. Also, the 10-fold cross-validation procedure was used to evaluate four different classifiers: Naïve Bayes, SVMs, Logistic Regression and Random Forests. The models' evaluation is executed based on the confusion matrix, namely TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). More specifically, the measures used for the calculation of the accuracy and recall are defined as

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

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

The performance results of the ML algorithms, in terms of the employed metrics, are shown in Table 10. The predictive ability of the Logistic Regression model in cardiovascular diseases is promising and superior in terms of accuracy, recall and AUC against the rest of the ML models. The higher the AUC, the better the performance of the model at distinguishing between CVD and Non-CVD classes. In particular, AUC shows that there is a 78.4% chance that the Logistic Regression model will be able to distinguish between CVD and Non-CVD classes.

6 CONCLUSIONS

In the context of this research study, a supervised machine learning methodology was employed and properly designed to assess the long-term risk of occurring CVD. The outcomes of the study may provide useful information in the clinic viewpoint and assist clinicians in how to interpreting data and implementing optimal algorithms for the dataset (Al'Aref et al., 2019).

It is ongoing research in which, as a first step, several traditional models were developed to investigate data quality and determine that model with the best predictive performance. The current outcomes will be used to design the next steps of our research and identify the optimal models to improve the performance metrics. The evaluation results presented similar accuracy and recall, a fact that is justified by the balanced distribution of participants into two classes. A promising direction for enhancing the achieved outcomes (accuracy, recall, AUC) is the utilization of deep learning models and techniques (Swathy and Saruladha, 2021), as they can provide complex decision boundaries and thus better fit the training data. Finally, our aim is to focus our analysis on i) anomaly detection techniques to identify instances with incorrect values and ii) dimensionality reduction techniques to optimize the performance of ML models.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Union's H2020 research and innovation programme SmartWork under grant agreement No 826343, SC1-DTH-03-2018 - Adaptive smart working and living environments supporting active and healthy ageing and GATEKEEPER under grant agreement No 857223, SC1-FA-DTS-2018-2020 - Smart living homes-whole interventions demonstrator for people at health and social risks.

REFERENCES

Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., and Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. *Journal of Diabetes & Metabolic Disorders*, pages 1–11.

Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., Maliakal, G., Van Rosendael, A. R., Beecy, A. N., et al. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, 40(24):1975–1986.

Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., and Mareeswari, V. (2018). Prediction of cardiovascular disease using machine learning algorithms. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–7. IEEE.

Dritsas, E., Fazakis, N., Kocsis, O., Fakotakis, N., and Moustakas, K. (2021). Long-term hypertension risk prediction with ml techniques in elsa database. In *International Conference on Learning and Intelligent Optimization*, pages 113–120. Springer.

Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., and Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access*, 9:103737–103757.

Gale, C. R., Cooper, C., and Sayer, A. A. (2014). Framingham cardiovascular disease risk scores and incident frailty: the english longitudinal study of ageing. *Age*, 36(4):1–9.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236.

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., and Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018.

Hoyas, I. and Leon-Sanz, M. (2019). Nutritional challenges in metabolic syndrome. *Journal of clinical medicine*, 8(9):1301.

Hussain, A., Ugli, I. K. K., Kim, B. S., Kim, M., Ryu, H., Aich, S., and Kim, H.-C. (2021). Detection of different stages of copd patients using machine learning techniques. In *2021 23rd International Conference on Advanced Communication Technology (ICACT)*, pages 368–372. IEEE.

Kocsis, O., Moustakas, K., Fakotakis, N., Vassiliou, C., Toska, A., Vanderheiden, G. C., Stergiou, A., Amaxilatis, D., Pardal, A., Quintas, J., et al. (2019). Smartwork: designing a smart age-friendly living and working environment for office workers. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 435–441.

Kumar, M. D. and Ramana, K. (2021). Cardiovascular disease prognosis and severity analysis using hybrid heuristic methods. *Multimedia Tools and Applications*, 80(5):7939–7965.

Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., et al. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1):1–25.

Sundhari, S. S. (2011). A knowledge discovery using decision tree by gini coefficient. In *2011 International Conference on Business, Engineering and Industrial Applications*, pages 232–235. IEEE.

Swathy, M. and Saruladha, K. (2021). A comparative study of classification and prediction of cardio-vascular diseases (cvd) using machine learning and deep learning techniques. *ICT Express*.

Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Leal, J., Luengo-Fernandez, R., Burns, R., Rayner, M., and Townsend, N. (2017). European cardiovascular disease statistics 2017.

Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Drazner, M. H., Fonarow, G. C., Geraci, S. A., Horwich, T., Januzzi, J. L., et al. (2013). 2013 accf/aha guideline for the management of heart failure: a report of the american college of cardiology foundation/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 62(16):e147–e239.

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., and Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern china. *Scientific reports*, 10(1):1–8.