Feature Selection Optimization of Risk Factors for Coronary Heart Disease

Ana Rita Antunes, Lino A. Costa(✉), Ana Maria A. C. Rocha, and Ana Cristina Braga

ALGORITMI Center, University of Minho, 4710-057 Braga, Portugal
id9069@alunos.uminho.pt,{lac,arocha,acb}@dps.uminho.pt

Abstract. Cardiovascular disease is a worldwide problem and is the main cause of mortality when coronary heart disease leads to a heart attack. Hence, it is important to evaluate how to prevent this disease considering the symptoms description and physical examinations.

This study points out the application and comparison of different performance measures for the classification of heart disease. Firstly, a feedforward neural network was applied to classify heart disease risk, using the well-known Framingham database. Feature selection optimization was performed to identify the most important variables to take into consideration, minimizing the Type II error and maximizing the accuracy. In addition, a multi-objective optimization algorithm was carried out to simultaneously optimize both performance measures. A set of non-dominated solutions representing the trade-offs between objectives were obtained, and gender, age, systolic blood pressure, and glucose level emerged as the principal factors to take into consideration to predict heart disease. The results obtained are promising and show the importance of considering more than one criterion to identify the most important variables.

Keywords: Feature selection · Optimization · Neural network · Heart disease

1 Introduction

Cardiovascular diseases are the main cause of mortality in the world and it is expected to be the most important cause of death by 2030 [28], despite recently the number of deaths caused by cardiovascular diseases has been decreasing over the decades [8]. Coronary heart disease (CHD) and coronary artery disease are cardiovascular diseases that involve heart and blood vessels, where CHD is a result of coronary artery disease [32]. CHD leads to a heart attack which occurs when the blood flow to the heart is cut off and there is a decrease in the supply of oxygen and nutrients [28].

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12953, pp. 413–428, 2021.
https://doi.org/10.1007/978-3-030-86976-2_28

In the European Union, Portugal has presented a low-risk of CHD for decades, but, in 2013, it was the second most common cause of death [25,30]. Therefore, it is essential to prevent heart attacks, taking into account raised blood pressure and glucose, physical inactivity, overweight, obesity, and tobacco use, since they are some risk factors [15,28].

Thus, to identify heart disease it is important to describe symptoms and make physical examinations [3]. In 2015, a study was carried out to identify which factors can be associated with the development of the disease, where high blood pressure, overweight, and hypercholesterolemia showed large increases in the incidence of heart disease [15]. Systolic, diastolic, and pulse pressure are risk factors that lead to heart failure, but systolic and pulse pressure have more impact [19].

Over the years, some authors have used the well-known Framingham database to study the factors that influence CHD. This database contains information about the residents of the city of Framingham, in Massachusetts, and comprises 15 variables on the demographic, behavioral, and medical history of more than 4000 patients. With this information, it is intended to verify whether the patient is at risk for future CHD [14]. According to Dawber and Kannel (1996), this was the first successfully detailed epidemiological study on heart disease and provided useful information [13]. However, a limitation of this study is that if other regions wish to classify the risk of CHD using Framingham data as a training dataset, they can not estimate the risk well, since the study uses only a restricted population, with daily habits that vary from region to region [11].

Since cardiovascular disease is a worldwide problem, it is important to understand what factors can be analyzed to prevent CHD. This work aims to identify which combinations of variables are capable of predicting whether the patient is at risk for future CHD, using the Framingham database. First, a feedforward neural network will be trained to learn with the available data. Then, a feature selection optimization will be carried out to identify the best subset of variables capable to predict the risk for future coronary heart disease. Finally, a multi-objective approach will be conducted to maximize accuracy and minimize Type II error, simultaneously. The computational environment MatLab® (version R2020b) will be used to obtain the results.

This paper is structured as follows. Section 2 presents a literature review, where some related works about cardiovascular diseases, feature selection, neural networks and performance measurement criteria are explained. Thereafter, the methods implemented and the parameters defined are in Sect. 3. The descriptive analysis and discussion of the results are presented in Sect. 4 and the main conclusions are reported in Sect. 5.

2 State of the Art

In this section, an analysis of some works related to cardiovascular diseases is presented to understand which methodologies were used and for what purpose.

Then, feature selection, neural networks, and performance measurement criteria are briefly described.

2.1 Cardiovascular Disease Studies

Cardiovascular diseases are the main cause of death worldwide and over the years this theme has been studied in several countries with different applications [28]. In most studies, the main purpose was to diagnose cardiovascular, heart, or artery diseases regarding the given datasets using different approaches. Some authors applied several machine learning techniques in order to evaluate the classification performance of different models taking into account performance measures such as accuracy, precision, recall, specificity, F-measure, and area under the ROC curve (AUC) [4,24,29]. In [5,6], a Genetic Algorithm (GA) was considered to optimize the weights of a Neural Network (NN) in order to improve performance. Feature selection using correlation matrix or Particle Swarm Optimization (PSO) was studied in [17,23]. Furthermore, cross-validation by splitting data into training, validation and test sets are also common in these works. The main goal of these models is learning from the available data.

A system that uses GA to optimize the NN weights to predict the risk of cardiovascular diseases is proposed in [6]. The dataset consists of heart disease information, the data was divided into training and test sets and the performance was measured in terms of accuracy. The accuracy reported for the test set was 94.17% [6].

A new hybrid model of NN and GA, using risk factors data of 50 patients, is used to diagnose heart diseases. The aim is to optimize the connection weights of the NN to improve performance. Data was divided into training, test, and validation and the accuracy obtained was 96.2%, 92%, and 89%, respectively [5].

PSO and NN feedforward backpropagation were used to rank factors of cardiovascular diseases. PSO was applied to minimize cost and maximize precision to select the most relevant features. The data is about Cleveland clinic and it was divided into training and test sets. Accuracy, recall, and precision were used to measure the performance of the model. The results achieved were an accuracy of 91.94%, a recall of 93% and a precision of 91.9% [17].

A system to predict the risk of cardiovascular disease, using data from 689 patients with cardiovascular disease symptoms, was developed in [29]. The data set was divided into training, validation and testing, and a logistic regression, Bayesian classification and quantum NN were applied to the Framingham dataset for validation purposes. In this work, the authors concluded that quantum NN obtained the best accuracy result (98.57%).

In [23], a NN was applied to predict CHD risk through feature selection, considering the correlation analysis. Korea's national health and nutrition examination survey was used to conduct the analysis and the performance was compared with the Framingham risk scores. Data were divided into training and validation sets, where the accuracy was 89% and 82.51%, respectively. Several methods (for example, Naive Bayes, Random Forest, NN) were studied to diagnose coronary artery disease, using three publicly available data, and different measures, like

sensitivity, specificity, F-measure, AUC, and running time. It is shown that accuracy is not the unique important measure to use to determine the performance of a classifier [24].

To predict the CHD risk in the Korean population, a deep NN was used in [4]. Tenfold cross-validation was used to split data. The model was compared with different algorithms, such as Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Random Forest. Moreover, different measures were used to assess the performance of the models (accuracy, precision, recall, specificity, F-measure, and AUC). The proposed deep NN achieved the best performance measure values, with the exception of specificity.

2.2 Feature Selection

Feature Selection is the process of selecting the best subsets of features to improve predictor performance. Nowadays, it is used in different research areas due to the exponential increase of data, where there are many variables to study. According to Iguyon and Elisseeff [18], feature selection can also help to understand the data to be analyzed. There are different methods capable of extracting variables, in which some include variable ranking and others the similarity between the variables. The most common methods for feature selection can be divided into filter, wrapper, or embedded [18,21,34].

In the filter method, the focus on the selection of features is based on a performance measure, where the first step is to find the best subset of features. Some well-known performance measures are correlation between variables, Chi-square and Fisher score [21]. In this method, the variables are ranked considering the measure chosen and the variables selected have useful information [12]. After the feature selection, the variables are used in the model. This method is also called the preprocessing step [18].

In the wrapper method, different combinations of features are used to find the model with the best performance, for example, with the highest accuracy [34]. This method uses the predicting performance to find the best subset of features [18]. In general, this method allows to obtain better results than the filter method since the subsets of features are evaluated using a modeling algorithm [21].

In the embedded method, the selection of features is made in the training process, without split data into training and testing, and aims to reduce the computation time [12]. The selection is made during the modeling algorithm's execution. Some methods consider objective functions to minimize fitting errors and a penalty is assigned to the features that do not contribute to the model [21]. Since the data is not split, a better use of the available data is observed and it is possible to obtain a faster solution when compared to the filter and wrapper methods [18].

2.3 Neural Networks

In order to make better decisions, many researchers investigate how to diagnose heart disease problems using intelligent systems such as Neural Networks [3]. NN is widely used in this area since it can extract more information about the system in the study due to the learning process [6].

In a NN, a neuron, also known as a unit or node, is the basic computational unit that can receive signals from other neurons and multiplies each signal by the corresponding weight (the connection strength). The weighted signals are then summed and passed through an activation function [27]. In Fig. 1 it is shown an individual neuron architecture.

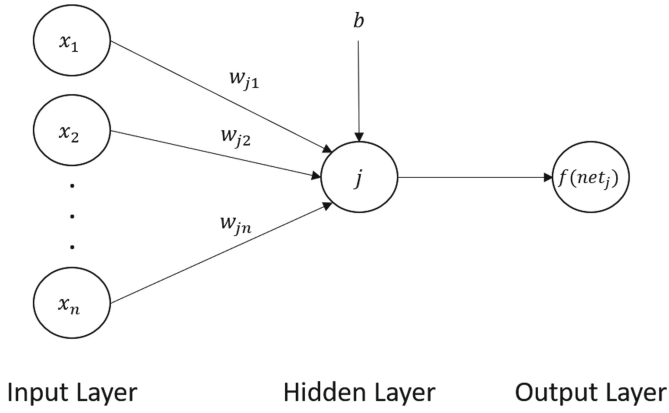

Fig. 1. Individual neuron architecture (based on [22]).

NN can have several different layers of neurons. In theoretical terms, the input layer is the first layer, the intermediate layer(s) is known as the hidden layer(s) and the last one is the output layer. Therefore, the output layer takes into account the number of values to be predicted. Thus, in Fig. 1 there is an input layer with n neurons, one hidden layer with one neuron, and one output layer. The input signal is denoted by a vector $\mathbf{x}$ (x_1, x_2, ..., x_n) and corresponds to the independent variables present in the data. Moreover, the weights of the neuron j are denoted by w_j (w_{j1}, w_{j2}, ..., w_{jn}) and f is the activation function. The *net* input to the neuron j is described in (1), where b is called bias [7,22].

$$net_j = \sum_n w_{jn} x_n + b \tag{1}$$

Most of the activation functions are nonlinear and the most widely used are hyperbolic tangent, sigmoidal and gaussian [7,22].

A multilayer NN has more than one hidden layer. In a feedforward NN the information propagates along the forward direction [7,22]. It is difficult to choose

the appropriate network size, i.e., the number of layers in the NN and the number of neurons per layer. Hence, the quality of the solution found, using NN, depends on the network size that can affect the complexity, learning time, and the ability to produce accurate results [10].

A NN can be used for supervised and unsupervised learning. In supervised learning, for each input, the target output is known. The NN weights are adjusted to produce the smallest error possible, considering the actual output and the predicted output. Furthermore, the generalized delta rule is used to minimize the error. In contrast, in unsupervised learning the NN adjusts the weights without knowing the associated output and the NN learns how to classify input patterns [3,10]. Backpropagation learning is the most common type of supervised learning used to optimize the weights. According to Ding, Su, and Yu [16] this optimization process can be stuck in a local minimum. The combination of backpropagation with a GA is one solution to this problem, since GA is a global optimization method. GA is an optimization algorithm that considers the principles of natural genetics and can escape from local minimums. The combination of a GA with the learning NN can provide a better predictive accuracy [6].

2.4 Performance Measures

The performance of classifier models, for a given dataset, can be assessed by different evaluation measures to describe how well the classification is done and to compare different models [1].

Thereby, inferential statistics are used to detect the effects of the independent variables regarding the variability that is inherent in the variable being measured (the dependent variable). Thus, hypothesis tests are, in general, used in inferential statistics in order to extract more information about the data under study. Two types of errors can be committed in the decisions, known as Type I and Type II errors. Type I error is committed when the null hypothesis is rejected, when in fact it is true. In contrast, Type II error is deciding not to reject the null hypothesis when it is actually false [9]. These types of errors have to be minimized but it is not always easy to do it. Type I error is the level of significance and can be controlled since it is the amount of risk that the authors are willing to take. On the other hand, the Type II error is related to the sample size, since it is sensitive to the number of observations in the sample. Thus, if the number of observations increases, the Type II error decreases [31].

A confusion matrix is a well-known tool for evaluating the classifier and takes into account the number of positive and negative instances correctly classified, also known as True Positive (TP) and True Negative (TN), respectively [1,9,33]. Hence, the confusion matrix also considers the number of instances that are predicted to be negative, but are actually positive, known as False Negatives (FN), and the number of instances that are predicted to be positive when they are negative, known as False Positives (FP) [1,20]. Table 1 presents an example of a confusion matrix that considers the hypotheses in the study and the decision made.

Table 1. Example of a confusion matrix

	Hypothesis	
Decision	H_0	H_1
Retain H_0	TP	FN (Type II Error)
Reject H_0	FP (Type I Error)	TN

FP and FN are the Type I and Type II Errors, respectively [33]. Therefore, the ratio for each error can be express as (2) and (3).

$$\text{Type I Error } (\%) = \frac{FP}{FP + TN} \tag{2}$$

$$\text{Type II Error } (\%) = \frac{FN}{TP + FN} \tag{3}$$

Another measure to take into account is the accuracy [33], in (4), which considers the ratio of observations that the model correctly classifies.

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

Accuracy does not consider instances that are misclassified. It can have serious implications, for example, in the health area. This can be an important limitation [20]. However, there are measures capable of filling this gap such as precision, recall rate, also known as sensitivity, specificity, F-Measure, and AUC. These measurements can be computed from a confusion matrix [20,33].

3 Methods

This section addresses the feature selection procedure, explaining the steps needed to achieve the main goals of the work. Besides that, the implementation details to identify the parameters considered are described.

3.1 Feature Selection Procedure

Feature selection optimization requires several steps in order to identify the subset of independent variables with better performance for classification. Figure 2 shows an overall summary of the different steps involved and what is defined in each step.

The first step refers to the definition of the independent and dependent variables. After that, the data is split into training (80%) and test (20%) sets, using the holdout method, in order to construct the feedforward NN and optimize the objective functions. This split intends to prevent overfitting to the Framingham dataset. Thereafter, the genetic algorithm is carried out to obtain the best subset of variables capable to predict the risk of CHD.

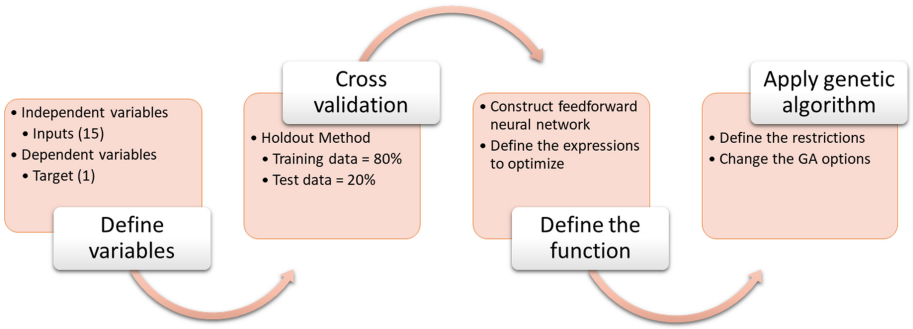

Fig. 2. Steps to build the solution for the feature selection optimization problem.

According to the literature review, accuracy is not the only measure to take into account to find the performance of a classifier [24]. Therefore, in this study, two criteria were selected to be optimized: maximization of accuracy and minimization of Type II error. Moreover, using the wrapper method it is expected to find the best subset of variables that can explain risk of CHD. The hypotheses defined were:

H_0: The patient does not have risk of CHD;
H_1: The patient has risk of CHD.

Thereby, Type II error is committed when the model predicted that the patient does not have risk of CHD, but he has. This type of situation should not happen and must be minimized. It provides incorrect information and it can be detrimental to patient health. Conversely, Type I error happens when the model predicts that the patient has risk of CHD, but he does not have. This situation is a false alert. These two situations can affect the patient's health, but also the time of doctors and the resources available. Considering these details, Type II error leads to a worse situation than Type I error. For this reason, Type II error and accuracy were the measures to take into account.

3.2 Implementation Details

In order to implement the proposed methodology, the software MatLab® [26] was used.

Firstly, a single-objective optimization was conducted, using the `ga` function from the Global Optimization Toolbox. In this approach, the objective functions were maximizing accuracy and minimizing type II error, separately. Three different feedforward NN were performed to predict the risk of CHD, using `feedforwardnet` function. The number of layers and the number of neurons per layer are the following: one hidden layer with eight neurons (NN_1), two hidden layers with eight and four neurons (NN_2), respectively, and three hidden layers with eight, four and two neurons (NN_3). With these three feedforward NN

it is intended to evaluate which one has the best performance. Besides that, the number of epochs was set to 750 and the training ratio parameter was defined as 1 to consider the same training set along the optimization. As a consequence, the validation and test ratio were defined as 0. Relatively to the optimization parameters concerning the genetic algorithm, the default values were considered, except the population size option set to 100.

Thereafter, a multi-objective optimization was addressed to maximize accuracy and minimize Type II error, simultaneously. The multi-objective optimization was performed using the `gamultiobj` function, where the standard options were used, except the use of the adaptive feasible mutation and the population size set to 100.

4 Results

This section begins by making a descriptive analysis of the Framingham dataset to be used in this work. Thereafter, a feature selection optimization using the feedforward neural network was performed. The results for single and multi-objective optimization are presented and discussed.

4.1 Dataset Description

Framingham dataset [2] contains information about 4240 patients. There are cases where variables are missing, so these cases were not considered in this study. Thus, only information about 3658 patients was analyzed. Table 2 presents the description of Framingham variables, where the Risk of CHD is the dependent variable and all the others are independent variables. The independent variables contain different types of information, namely demographic, behavioral and medical information. A codification for each variable is given in Table 2 to facilitate the identification of the selected variables.

In this data, 55.63% are women and 15.23% of patients were diagnosed with the risk of future CHD. The youngest patient is thirty-two years old and the oldest is seventy years old. Figure 3 shows the risk of CHD by gender. It can be observed that more men were diagnosed with the risk of future CHD, approximately, 18.92%. In contrast, 12.29% of women were diagnosed with CHD. With this representation, there are, by far, more patients that were diagnosed with no risk of CHD.

Table 2. Description of Framingham variables.

Type of information	Variable	Codification
Demographic	Gender	G
	Age	A
	Education level	EL
Behavioral	Current Smoker (Yes or No)	CS
	Cigarettes per day	CPD
Medical	Blood pressure medication (Yes or No)	BPM
	Had a stroke (Yes or No)	HS
	Hypertensive (Yes or No)	HYP
	Diabetes (Yes or No)	DIAB
	Total cholesterol level	TCL
	Systolic blood pressure	SBP
	Diastolic blood pressure	DBP
	Body mass index	BMI
	Heart rate	HR
	Glucose level	GL
	Risk of CHD (Yes or No)	CHD_Risk

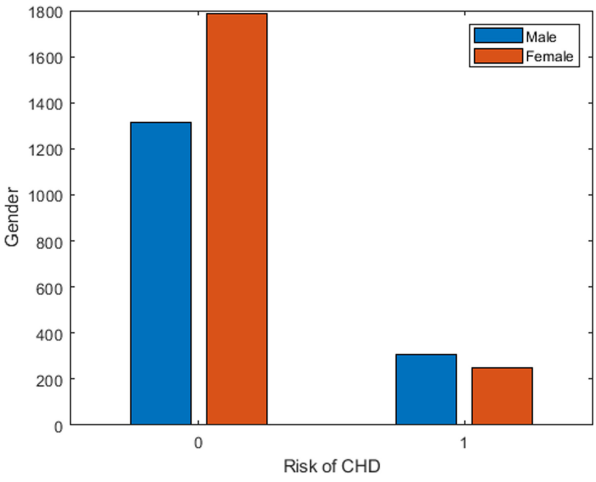

Fig. 3. Risk of CHD by gender.

4.2 Single-Objective Optimization

The different models constructed and the respective results for the single-objective optimization are presented in Table 3. First column reports the NN characterization showing the number of layers and the number of neuron per

layer. The following columns present the variables retained and the percentage values for Type II error and accuracy.

Table 3. Type II error and accuracy results for NN_1, NN_2 and NN_3.

NN	Variables	Type II Error	Accuracy
8	G, A, HYP, DBP, GL	–	84.82%
	G, A, EL, CS, CPD, HYP, DIAB, TCL, SBP, DBP BMI, GL	80.73%	–
8,4	BPM, HS, HYP, DIAB	–	85.64%
	G, A, EL, HYP, DIAB, SBP, DBP, BMI, HR, GL	84.11%	–
8,4,2	BPM, HS, HYP, DIAB	–	85.64%
	G, A, EL, CS, CPD, HYP, DIAB, SBP, DBP, BMI, GL	78.90%	–

The lowest value for Type II error was obtained with NN_3 and the highest value for accuracy was achieved with NN_2 and NN_3. Thereby, when the NN size increases the accuracy value is established. However, for the variation of Type II error was different. The Type II error value in NN_2 was the worst and it improves with NN_3. The results showed that Type II error criterion requires more variables than the accuracy criterion to predict the risk of CHD.

If it is important to just maximize accuracy, BPM, HS, HYP, and DIAB are the subset of variables to be taken into account. These variables are all qualitative. When the answer for these variables is no, there are 2206 in 3658 patients that do not have the risk of CHD. On the other hand, when minimizing the Type II error, G, A, EL, CS, CPD, HYP, DIAB, SBP, DBP, BMI, and GL are the variables to be taken into consideration. These results are in accordance with the literature that refers to the association of high blood pressure, overweight, and high cholesterol with large increases in the incidence of heart disease [15,19]. Type II error considers blood pressure (SBP and DBP) and overweight (BMI) as the most important variables.

4.3 Multi-objective Optimization

In the multi-objective optimization, the NN_3 with the best performance for single-objective optimization (see Table 3) was considered.

The Pareto front, in Fig. 4, provides information about the seven different non-dominated solutions achieved. This curve allows to analyze the trade-offs between solutions. It can be seen that the solution corresponding to a Type II error of 70.64% and an accuracy of 85.50% is a good compromise between the two criteria.

Table 4 presents the variables retained, Type II error and accuracy values for each non-dominated solution.

According to the results, it can be concluded that when the Type II error increases, the accuracy value also increases. Thus, G, A, SBP, and GL variables are present in the seven solutions. While the variables CS, CPD, BPM, and BMI

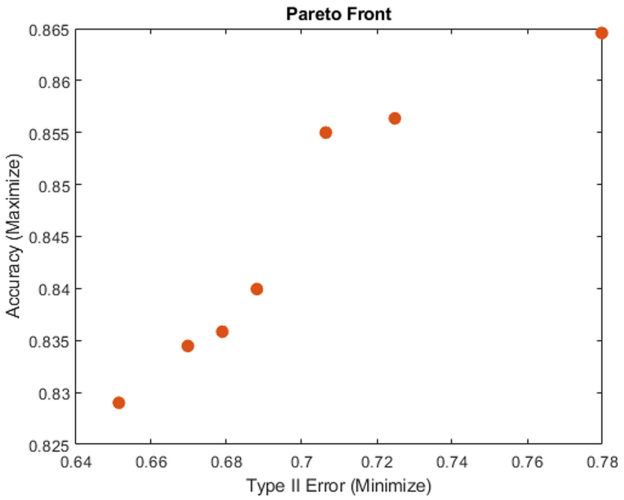

Fig. 4. Pareto front: Type II Error against Accuracy.

Table 4. Multi-objective optimization results.

Variables	Type II Error	Accuracy
G, A, EL, HS, HYP, DIAB, TCL, SBP, DBP, HR, GL	65.14%	82.90%
G, A, EL, SBP, DBP, HR, GL	66.97%	83.45%
G, A, EL, HS, SBP, DBP, HR, GL	67.89%	83.58%
G, A, EL, HS, HYP, DIAB, TCL, SBP, DBP, HR, GL	68.81%	83.99%
G, A, CPD, HYP, DIAB, SBP, DBP, HR, GL	70.64%	85.50%
G, A, CPD, HYP, SBP, DBP, GL	72.48%	85.64%
G, A, HS, TCL, SBP, HR, GL	77.98%	86.46%

do not enter into any model. Moreover, the variable TCL is only used in three solutions.

In addition, there are two solutions with the same variables, but the criteria values are different. The NN weights were optimized in each solution and this can be the reason for these results. The best values for Type II error and accuracy were 65.14% (accuracy is equal to 82.90%) and 86.46% (Type II error is equal to 77.98%), respectively. The "best" solution depends on whether the decision-maker gives more importance to Type II error or accuracy.

4.4 Discussion of Results

CHD leads to heart attacks, hence the importance of analyzing what factors can be measured to prevent this from happening. Over the years, different expensive applications have been developed to assist the decision-making.

Different machine learning techniques were carried out to assess which one presents the best results. Moreover, NN is commonly applied in the healthcare sector. Some researchers used optimization methods to optimize the NN weights and perform different feature selection methods. In this particular situation, it is more important to identify and understand the risk factors that lead to CHD. Kim and Kang identified that triglyceride and chronic renal failure were related to CHD [23]. In another study, sex, age, cholesterol, fasting blood sugar, resting electrocardiograph results, angina experience, ST depression, and slope of the peak exercise were the best subset of variables to predict heart disease [17]. Besides that, high blood pressure, overweight, hypercholesterolemia, SBP, DBP, and pulse rate were also identified as risk CHD factors [15,19].

This study combines a machine learning technique (NN) and feature selection optimization, whose objectives were to maximize accuracy and minimize Type II error. When a NN is performed in MatLab, the weights are optimized to minimize the mean square error. The wrapper method was performed to identify the best subset of variables capable to predict CHD.

In a first approach, a single-objective optimization was conducted and, an interesting remark is that Type II error requires more variables to predict the risk of CHD than the accuracy criterion. Moreover, the NN_3, with three layers, presented the best results in both cases (85.64% for accuracy and 78.90% for Type II error). Thereby, if it is intended to maximize accuracy, the variables to take into consideration are BPM, HS, HYP, and DIAB, whereas if only Type II error is taken into account, it considers G, A, EL, CS, CPD, HYP, DIAB, SBP, DBP, BMI and GL.

Furthermore, using multi-objective optimization, seven different solutions were obtained. In terms of risk CHD factors, G, A, SBP and GL variables are included in all of these solutions. This means these variables must be controlled to prevent CHD. Some other variables can be added to this list, but it depends on whether the Type II error is more important than accuracy or vice versa. The Type II error values are too high (65.14%–77.98%), which may be related to the sample size [31]. In Fig. 3 it is possible to see that there are more patients without risk of CHD, where only 15.23% of the patients were diagnosed at risk for CHD.

5 Conclusions

The main goal of this paper is to identify the best subset of risk factors that can be used to prevent CHD. In this study, the Framingham dataset was used for validation purposes. Firstly, a single-objective optimization was carried out in order to minimize the Type II error and maximize accuracy. Secondly, a multi-objective optimization was conducted to simultaneously optimize both objectives. Thereby, the wrapper method was used to obtain the more important risk factors.

Three NN characterizations were used in single-objective optimization, considering different numbers of layers and neurons per layer, in order to assess

which one gets the best results. In this approach, one interesting remark is that Type II error considers more variables than accuracy. The lowest and highest value for Type II error and accuracy was 78.90% and 85.64%, respectively. Risk factors G, A, EL, CS, CPD, HYP, DIAB, SBP, DBP, BMI, and GL must be taken into account when the criterion is to minimize Type II error. On the other hand, the subset of the variables BPM, HS, HYP, and DIAB must be considered when the criterion is to maximize accuracy. All of these variables are qualitative.

In the multi-objective optimization, seven non-dominated solutions were obtained, where G, A, SBP and GL variables belong to all of them. Therefore, these variables can be considered the risk factors to pay more attention to prevent CHD.

As future work, the implementation of different machine learning techniques can be performed, like logistic regression, support vector machine, and random forest, to identify which method has better results. Moreover, a distinct feature selection method, such as the embedded method, and the selection of other performance measurement criteria can also be implemented.

References

1. Aggarwal, C.C.: Data Classification: Algorithms and Applications, 1st edn. Chapman and Hall/CRC, Boca Raton (2014)
2. Ajmera, A.: Framingham heart study dataset (2018). https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset. Accessed 03 Dec 2020
3. Al-Milli, N.: Backpropogation neural network for prediction of heart disease. J. Theor. Appl. Inf. Technol. **56**(1), 131–135 (2013)
4. Amarbayasgalan, T., Van Huy, P., Ryu, K.H.: Comparison of the framingham risk score and deep neural network-based coronary heart disease risk prediction. In: Smart Innovation, Systems and Technologies (2020)
5. Amin, S.U., Agarwal, K., Beg, R.: Genetic neural network based data mining in prediction of heart disease using risk factors. In: 2013 IEEE Conference on Information and Communication Technologies, ICT 2013 (2013)
6. Amma, N.G.: Cardiovascular disease prediction system using genetic algorithm and neural network. In: 2012 International Conference on Computing, Communication and Applications, ICCCA 2012 (2012)
7. Annema, A.J.: Feed-Forward Neural Networks. Springer, Boston (1995). https://doi.org/10.1007/978-1-4615-2337-6
8. Araújo, F., Gouvinhas, C., Fontes, F., La Vecchia, C., Azevedo, A., Lunet, N.: Trends in cardiovascular diseases and cancer mortality in 45 countries from five continents (1980–2010). Eur. J. Prev. Cardiol. **21**(8), 1004–1017 (2014)
9. Bakeman, R., Robinson, B.F.: Understanding Statistics in the Behavioral Sciences. Psychology Press, Hove (2005)
10. Bebis, G., Georgiopoulos, M.: Feed-forward neural networks. IEEE Potentials **13**(4), 27–31 (2002)
11. Brindle, P., et al.: Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. BMJ **327**(7426), 1267 (2003)
12. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)

13. Dawber, T.R., Kannel, W.B.: The Framingham study. An epidemiological approach to coronary heart disease. Circulation **34**(4), 553–555 (1966)
14. Dawber, T.R., Meadors, G.F., Moore, F.E.: Epidemiological approaches to heart disease: the Framingham Study. Am. J. Public Health **41**(3), 279–286 (1951)
15. Dawber, T.R., Moore, F.E., Mann, G.V.: II. Coronary heart disease in the Framingham study. Int. J. Epidemiol. **44**(6), 1767–1780 (2015)
16. Ding, S., Su, C., Yu, J.: An optimizing BP neural network algorithm based on genetic algorithm. Artif. Intell. Rev. **36**(2), 153–162 (2011)
17. Feshki, M.G., Shijani, O.S.: Improving the heart disease diagnosis by evolutionary algorithm of PSO and feed forward neural network. In: 2016 Artificial Intelligence and Robotics, IRANOPEN 2016 (2016)
18. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
19. Haider, A.W., Larson, M.G., Franklin, S.S., Levy, D.: Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the Framingham Heart Study. Ann. Intern. Med. **138**(1), 10–16 (2003)
20. Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, Cambridge (2011)
21. Jović, A., Brkić, K., Bogunović, N.: A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings (2015)
22. Kamruzzaman, J., Begg, R., Sarker, R.: Overview of artificial neural networks and their applications in healthcare. In: Neural Networks in Healthcare: Potential and Challenges. Idea Group Publishing (2006)
23. Kim, J.K., Kang, S.: Neural network-based coronary heart disease risk prediction using feature correlation analysis. J. Healthcare Eng. (2017)
24. Kolukisa, B., et al.: Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease. In: Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018 (2019)
25. Levi, F., Lucchini, F., Negri, E., La Vecchia, C.: Trends in mortality from cardiovascular and cerebrovascular diseases in Europe and other areas of the world. Heart **88**(2), 119–124 (2002)
26. MATLAB: version 9.9.0.1495850 (R2020b). The MathWorks Inc., Natick, Massachusetts (2020)
27. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**(4), 115–133 (1943)
28. Mendis, S., Puska, P., Norrving, B.: Global atlas on cardiovascular disease prevention and control. World Health Organization (2011)
29. Narain, R., Saxena, S., Goyal, A.K.: Cardiovascular risk prediction: a comparative study of framingham and quantum neural network based approach. Patient Prefer. Adherence **10**, 1259 (2016)
30. Pereira, M., et al.: Explaining the decline in coronary heart disease mortality in Portugal between 1995 and 2008. Circ. Cardiovasc. Qual. Outcomes **6**(6), 634–642 (2013)
31. Salkind, N.J.: Statistics for People Who (Think They) Hate Statistics: Excel 2007 Edition. SAGE Publications, Inc., Thousand Oaks (2010)

32. Sanchis-Gomar, F., Perez-Quilis, C., Leischik, R., Lucia, A.: Epidemiology of coronary heart disease and acute coronary syndrome. Ann. Transl. Med. **4**(13), 256 (2016)
33. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009)
34. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)