Journal of Molecular and Pharmaceutical Sciences

Original Research Article

Multilayer perceptron networks for aid in the diagnosis of coronary heart disease

Mohamed Nadjib Boufenara [a*], Meriem Bahi [a]

[a] Pharmaceutical Sciences Research Center (CRSP), Constantine, 25000, Algeria.

Abstract

Heart disease, the predominant cause of morbidity and mortality worldwide, poses considerable diagnostic challenges. The imperative to identify and assess risks at an early stage requires the development of a robust and effective forecasting system. Machine learning enables accurate predictions, proving invaluable in various fields such as finance and healthcare. However, the varied performance of algorithms introduces complexities, with some achieving high prediction accuracy and others demonstrating comparatively lower precision.

In this work, we introduce a method using a multi-layer perceptron to provide decision support in the diagnosis of heart diseases in patients. This proposed model is systematically compared with four important machine learning counterparts: Random Forest, Logistic Regression, Naive Bayes, and SVM. Performance evaluations are conducted using a comprehensive dataset derived from a cardiovascular study including Massachusetts residents with coronary heart disease.

The obtained results suggest that the accuracy of the model based on a multilayer perceptron is superior compared to the four other machine learning models. The use of MLPs appears to be a promising and effective approach for enhancing diagnostic accuracy in the field of cardiovascular health when employing this type of data.

1 Introduction

The deployment of machine learning techniques in the medical and biomedical fields plays a central role in disease diagnostics. Notably, the survey by Lahmiri et al. highlights this role by using machine learning for the diagnosis of diseases such as Parkinson's disease, relying on measures of dysphonia [1].

Cardiovascular diseases (CVD) are among the deadliest diseases worldwide [2]. However, a considerable number of people suffering from this disease could have been saved if they had been diagnosed in time. Therefore, accurate diagnosis at an early stage of cardiovascular disease is imperative and should be performed with increased prediction accuracy [3]. Machine learning methodologies present a potential avenue to mitigate mortality rates, improve prediction accuracy, and reduce diagnostic time.

Deep learning enables computational models, characterized by multiple processing layers, to acquire data representations with varying levels of abstraction [4]. This approach excels in addressing challenges associated with non-linear relationships and intricate interactions between models and outcomes, surpassing conventional methods [5]. Among Deep Neural Network (DNN) models, the MultiLayer Perceptron (MLP)
is one of the most fundamental architectures. The latter is distinguished by its layered structure, comprising at least three main layers: an input layer, one or more hidden layers and an output layer. This layered architecture enables progressive, hierarchical modeling of data features, where each layer of neurons learns to represent increasingly abstract aspects of the input data. [6].

Motivated by recent advances in deep learning and with the aim of improving the predictive accuracy of coronary heart disease (CHD) diagnosis, we compare the performance obtained by the MLP model against alternative machine learning approaches. The rest of this paper is organized as follows: in section 2, we expose some research works which are interested in disease diagnosis prediction. Section 3 presents the proposed classification method. In section 4, we present the experimental results obtained by the models. Finally, section 5 concludes the document and gives some perspectives.

2 Literature Review

Disease prediction and diagnosis has aroused the interest of many researchers and has been the focus of several studies. In the work of Banu and Gomathy [8], the authors applied preprocessing to the data before storing it in the database. Then, a grouping of the processed data was applied using grouping algorithms which included the K-means grouping technique. Using the C 4.5 algorithm, the observed patterns were classified into different classes. After that, the measure of performances of the decision tree model was based on the entropy of the information. The experimental results demonstrated the ability of this method to predict a heart attack with satisfactory accuracy.

In the work of Tan et al. [9], a hybrid method was introduced using two machine learning algorithms, namely: SVM and genetic algorithms. Then a data mining was applied using tools like WEKA and LIBSVM to analyze them. In order to assess the performance of the proposed method, five different datasets were collected from the UCI repository. The hybrid method achieved an accuracy of 84.07% for heart disease and 86.12% for hepatitis. Another work carried out by MIAO et al. [10] presented a deep neural network to help increase the reliability of heart disease diagnosis and prognosis. The suggested model included a classification model based on training data and another model that could predict the diagnosis of patients. A data set of 303 patients diagnosed with coronary heart disease was used. This model achieved an accuracy of 83.67%.

In the work of Uhmn et al. [11], several machine learning techniques were used to predict chronic hepatitis: SVM, decision tree and decision rule. The prediction was made from Single Nucleotide Polymorphism (SNP) data.
The results obtained showed that the decision rule was able to distinguish chronic hepatitis and obtained the highest precision of the three techniques used; 73.20%.

In order to predict heart disease, Kumar et al. [12] compared the RIPPER classifier, decision tree, artificial neural networks and SVM. The experimental results showed that SVM proved to be the best technique among the four classification methods used for predicting heart disease with the highest accuracy.

Most of the related works uses the same dataset, accessible on the UCI repository, containing the medical records of 303 patients with heart failure. The difference between our work and related works lies in the use of a larger and a more complex dataset (over 4,000 clinical cases) [7] and also in the proposed neural network architecture to perform the classification.

3   Proposed Method

In this section, we present a method of diagnosing heart disease using deep learning to assess its ability to predict coronary heart disease compared to other popular machine learning methods. This method has already been used in previous work in other diseases [13] [14].The experimental results were compared with the results obtained using the different models: Random Forest, Logistic Regression, Naive Bayes and SVM. To evaluate each model, we used the same dataset provided by a cardiovascular study conducted by Massachusetts residents.

3.1   Description of the method

The proposed method consists mainly of three steps. The first includes the preprocessing of the raw data to remove or replace the missing values and transform the categorical data present in the dataset into numeric data in order to be usable by the model. In the second step, we proceed to the classification of the preprocessed data in the first step based on a DNN algorithm in order to build the model. The last step is dedicated to the test to evaluate the model built in the previous step.

3.2   Different steps of the method

As highlighted previously, the methodology is based on a structured sequence encompassing three discernible steps each playing a crucial role in developing the overall effectiveness of the predictive model:

Step 1: data preprocessing

In this step, we preprocessed the data by exploring the raw data to transform it into an understandable format because the raw data is often incomplete and can cause errors during classification. The data preprocessing applied in this method included the following two points:

• Missing values: they can lead to misclassification. There are several methods commonly used to impute these kinds of values [15]. One method is to calculate the mode, median or average of the entity and impute them with the missing values. It is an approximation that can add variance to the dataset. This method only works for numeric data. The coronary heart disease dataset used in this paper has several missing values.

• Categorical data: they are generally transformed into numeric data. This transformation is applied because numeric data is often easier to process than categorical data in order to avoid additional complexities due to the semantics relating to each category value. In the dataset used, some data were categorical and they were transformed into numerical data as explained in the following section.

Step 2: classification

In this step, the preprocessed data was used for MLP training in order to build the pre-diction model. Since there is no precise method for choosing the number of neurons and the number of layers of a deep neural network, we carried out several tests in order to fix the number of layers and the number of neurons in each layer.

The proposed neural network consists of five layers of neurons; three of them are hidden layers. The number of neurons in the input layer of this MLP is equal to the number of characteristics in the heart disease dataset. Therefore, the first layer is made up of 14 neurons and the output layer contains a single neuron. These should indicate whether heart disease is present 1 or absent 0. At the three hidden layers and the output layer, two activation functions are used as shown in **Fig. 1.**
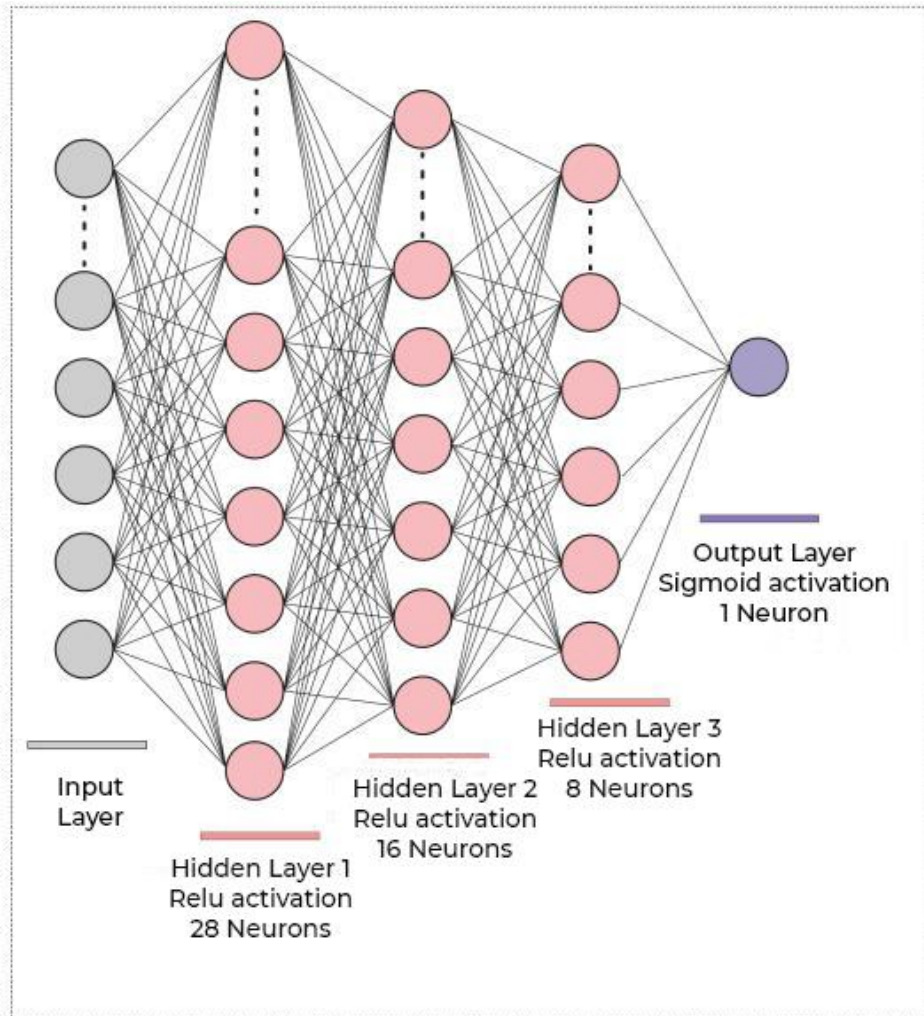

Fig. 1. Design of proposed multilayer perceptron

This figure illustrates the design of the proposed MLP for diagnosing CHD. This MLP uses the ReLu (Rectified Linear Unit) function [16] for hidden layers. The advantage of this function is that its gradient becomes 0 for all negative values and 1 for all positive values. During the learning process, this property is important because it allows to correct or avoid the problem of the leakage gradient.

The second function used in the MLP is the Sigmoid function [17]. This function is generally used in the case of a binary classification. It takes a real input value and transforms it into a real output value between 0 and 1. The equations of the function ReLU and Sigmoid are defined as follows:

- ReLU function:

$$f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \quad \dots (1)$$

- Sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \quad \dots (2)$$

At the end of this step, the classification of the preprocessed data was carried out. The model performance was evaluated in the next step.

Step 3: test

In this step, we evaluate the model's effectiveness, which results from the previous step, to diagnose heart disease using the test data section. We apply cross-validation to keep our data random across the training and testing set.

To evaluate the performance of the presented MLP, we used five performance metrics: accuracy, precision, recall, F-measure, and area under the ROC curve (AUC). These performances were calculated as follows:

- Accuracy $= \dfrac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$

- Precision $= \dfrac{True\ Positive}{True\ Positive + False\ Positive}$

- Recall $= \dfrac{True\ Positive}{True\ Positive + False\ Negative}$

- F-measure $= 2 \times \dfrac{Precision \times Recall}{Precision + Recall}$

- AUC $= \int_0^1 ROC(t)dt \ where\ t = 1 - specifity$

The experimental results obtained are presented and compared to the results obtained by other machine learning methods in the following section.

4 Experimentation and Results

As mentioned in the previous section, we used a dataset that comes from an in-court heart disease study in which patients are residents of the town of Framingham, Massachusetts. This dataset [7] includes 4,241 clinical cases in CSV format. Each of these patients exhibits 15 attributes. It provides statistical data on obesity, physical activity, smoking, sex, age, BMI, systolic and diastolic blood pressure, heart rate, blood glucose level, alcohol and diseases requested by the World Health Organization (WHO). The target attribute represents a Boolean. The 0 signifies the absence of CHD in the patient and 1 its presence. The dataset shows an imbalance between the number of patients with CHD and the number of healthy patients.

4.1 Experimentation

In order to prepare and clean the data for classification, we made sure that there were no duplicate entries or categorical variables to transform into a numeric value. How-ever, the dataset had some missing values. Columns that have missing values are shown in **Tab. 1.** and **Fig. 2.**

Tab. 1. Number of missing values per attribute

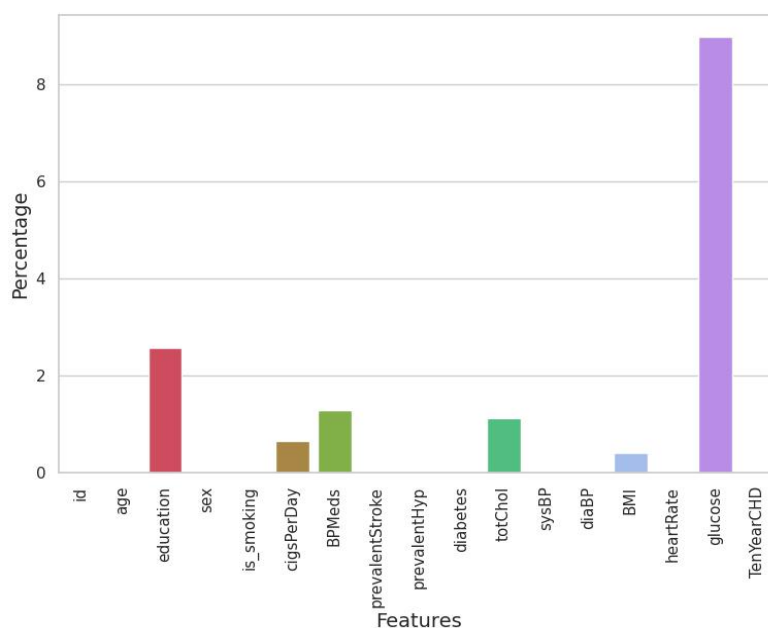| Columns	cigsPerDay	BPMeds	TotChol	BMI	heartRate	Glucose
Missing values	29	53	50		1	3800

Fig. 2. Percentage of Missing Data by feature

The column that shows the glucose level has the highest percentage of missing values. Despite the number of these values, they represented only about 12% of the total data and were therefore suppressed.
Missing entries only made up 12% of the total data and could therefore be deleted without losing much data. The number of registrations decreased from 4241 to 3751. However, the dataset was unbalanced with a number of negative cases diagnosed much greater than the number of positive cases, which may affect the quality of the classification. In order to balance the data, we used the synthetic minority oversampling technique (SMOTE) proposed in 2002 [18]. This allows creating a synthetic example rather than oversampling by replacements. SMOTE first selects an instance of minority class x1 at random and finds its k nearest minority class neighbors.

Once the method was applied, the dataset was much more balanced. In order to implement the MLP described in the previous section, we used Python (version 3), Pandas, TenserFlow and Keras. MLP uses the ReLu activation function for hidden layers and the Sigmoid function for the output layer. The three hidden layers of neural network are composed of 28, 16 and 8 neurons respectively.

We opted for using the Adam algorithm. It is an adaptive learning rate optimization algorithm that is designed specifically for the formation of MLP, with cross entropy function and 10-fold cross validation.

The models' performance was evaluated based on five performance measures: accuracy, recall, precision, F1-measure and AUC [19].

The same performance measures and dataset used to assess the effectiveness of the MLP were used to assess the performance of the models constructed using other ma-chine learning methods.

4.2 Results and discussion

The model constructed from the MLP was tested using 10-fold cross-validation. **Fig. 3** shows the accuracy curve of the model during training and testing.
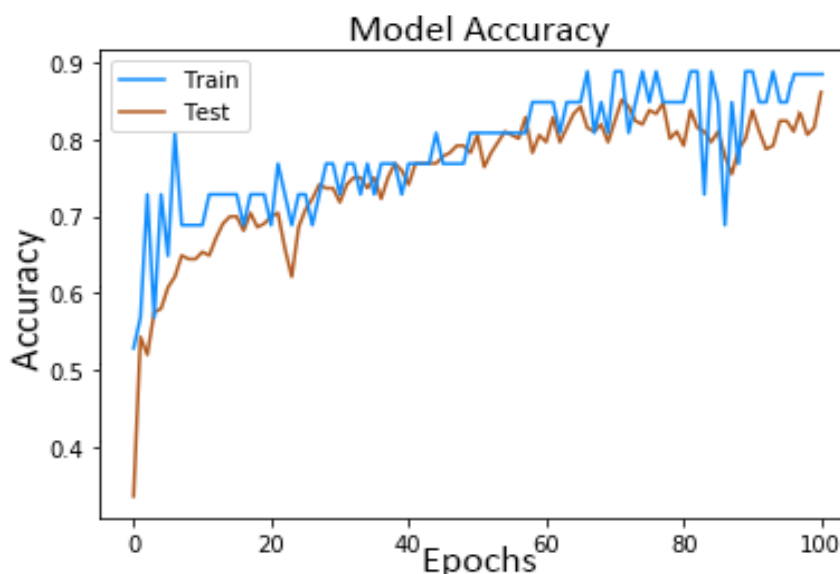

Fig.3 Accuracy Curve of the model

The other classifications were performed on the same heart disease dataset using other machine learning methods. The results obtained are presented in **Tab. 2**. The methods: Random Forest, Logistic Regression, Naive Bayes, SVM and MLP are named M1, M2, M3, M4 and M5 respectively in the table below.

Tab. 2. Performance of each machine learning method used as well as that of the proposed MLP.

Measure	M1	M2	M3	M4	M5
Accuracy	0.69	0.66	0.70	0.84	0.85
Recall	0.70	0.65	0.72	0.85	0.87
Precision	0.67	0.63	0.69	0.84	0.85
F1-score	0.68	0.64	0.70	0.84	0.86
AUC	0.76	0.73	0.82	0.91	0.92

Based on the five reported performance measures, the developed method obtained the best results compared to the four other techniques we used in this paper. The accuracy obtained by the proposed method was the highest (0.85), which suggests that the model constructed from the neuron network was the best classification algorithm. The SVM method was the most concurrent method with the MLP, but its precision remained lower (0.84). However,

the proposed method obtained only a slight improvement in recall compared to SVM. This is likely due to the MLP's ability to learn complex non-linear relationships between various features.

While the results seem promising, it's crucial to consider limitations:

- The study might not be generalizable to other populations with different characteristics.
- More research is needed to validate the findings and compare against other state-of-the-art deep learning models.

5 Conclusion

In this paper, we have presented a method based on deep learning to predict the diagnosis of coronary heart disease. The dataset of patients with CHD was used to evaluate the models. The data was prepared and cleaned by balancing it and removing any records with missing values so that the quality of the prediction would not be affected.

This dataset tested the ability of MLP to diagnose CHD compared to four other machine learning methods: Random Forest, Logistic Regression, Naive Bayes and SVM. The developed model obtained the best results compared to the other methods used with an accuracy of 0.85 and recall of 0.87. These results suggest that the use of MLP can increase the accuracy of the model in order to predict and inform the user of the system as quickly as possible. However, the accuracy and recall remain to be improved in order to build a system with good efficiency for health workers, and the MLP should be tested and generalized on other datasets. More research is needed to validate the findings and compare against other state-of-the-art deep learning models.

References

1. Lahmiri, S., Dawson, D. A., & Shmuel, A.: Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures. Biomedical engineering let-ters, 8(1), 29-39 (2018).

2. World Health Organisation, http://www.who.int/mediacentre/factsheets/fs317/en, Last ac-cessed 4 Oct 2020.

3. Ramaraj, M., & Thanamani, A. S. A.: Comparative study of CN2 rule and SVM algorithm and prediction of heart disease datasets using clustering algorithms. Network and Complex Systems, 3(10), 1-6 (2013).

4. Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E.: Deep learning for com-puter vision: A brief review. Computational intelligence and neuroscience, (2018).

5. Goldstein, B. A., Navar, A. M., & Carter, R. E.: Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European heart journal, 38(23), 1805-1814 (2017).

6. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, (2016).

7. D'Agostino, Ralph B., et al. "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation." Jama 286(2) 180-187 (2001)

8. Banu, M. N., & Gomathy, B.: Disease forecasting system using data mining methods. In 2014 International conference on intelligent computing applications, IEEE (pp. 130-133).

9. Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C.: A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 36(4), 8616-8630 (2009).

10. Miao, K. H., & Miao, J. H.: Coronary heart disease diagnosis using deep neural networks. Int. J. Adv. Comput. Sci. Appl., 9(10), 1-8 (2018).

11. Uhmn, S., Kim, D. H., Cho, S. W., Cheong, J. Y., & Kim, J.: Chronic hepatitis classification using SNP data and data mining techniques. In 2007 Frontiers in the Convergence of Bio-science and Information Technologies, IEEE (pp. 81-86) (2007).

12. Kumari, M., & Godara, S.: Comparative study of data mining classification methods in car-diovascular disease prediction 1. (2011)

13. Somasundaram, R. S., & Nedunchezhian, R.: Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications, 21(10), 14-19 (2011).

14. BOUFENARA, Mohamed Nadjib, BOUFAIDA, Mahmoud, et BERKANE, Mohamed Lamine. A Prediction Approach Based on Self-Training and Deep Learning for Biological Data. International Journal of Organizational and Collective Intelligence (IJOCI), 2020, vol. 10, no 4, p. 50-64.

15. BOUFENARA, M. N., BOUFAIDA, M., et BERKANE, M. L. A Machine learning technique dedicated for biological data. In : 2019 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS). IEEE, 2019. p. 1-4.

16. Agarap, A. F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018).

17. Agostinelli, F., Hoffman, M., Sadowski, P., & Baldi, P.: Learning activation functions to improve deep neural networks. arXiv preprint arXiv:1412.6830 (2014).

18. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: SMOTE: synthetic mi-nority over-sampling technique. Journal of artificial intelligence research, 16, 321-357 (2002).

19. Bekkar, M., Djemaa, H. K., & Alitouche, T. A.: Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl, 3(10) (2013).