AN EFFICIENT APPROACH TO PREDICT CORONARY HEART DISEASE USING DEEP NEURAL NETWORK

Maria Mohammad Yousef

Department of Computer Science, Al-al Bayt University, Mafraq, Jordan.
Email: Maria.yousef@yahoo.com

Abstract:

Coronary Heart disease (CHD) is one of the most common causes of death worldwide. A heart disease predicted at earlier stages not only helps the patients prevent it, but it can also help the medical practitioners learn the major causes of a heart attack and avoid it before its actual occurrence in a patient, which is the primary research objective of this study. Some researchers have created prediction systems that use a traditional neural network or another machine learning techniques, but the results were not effective and lead to misdiagnosis. Therefore, improving the accuracy of CHD prediction systems has been extensively considered in the literature. This research intends to develop a new prediction model that can be utilized to identify coronary heart disease using the medical history of the patient based on deep learning. In this study, the model is trained to predict using a deep neural network (DNN) that is based on a standard neural network with a different number of hidden layers. Binary classification has also been used in the DNN model to diagnose whether CHD is present (representative of '1') or missing (representative of '0'). To aid performance analysis, we used the Cleveland heart disease dataset from the UCI machine learning repository that consists of 13 features and 303 cases. Also, a comparative study of the performance of various machine learning algorithms like KNN, SVM, Naïve Bayes is done to demonstrate the effectiveness of the DNN algorithm in CHD prediction. The experimental results reveal that out of these four classification models, the DNN model was able to achieve the best accuracy of 96% when using four layers and 350 epochs.

Keywords: *Coronary Heart Disease, Deep Neural Network, Diagnostic models, UCI Machine Learning Repository*.

Introduction:

Coronary heart disease (CHD) Also known as Coronary Artery Disease or ischemic heart disease, is one of the world's most risky diseases. The National Heart, Lung, and Blood Institute (NHLBI) defines CHD as a kind of heart disease that occurs when the heart's arteries are unable to supply enough oxygen-rich blood to the heart and include symptoms such as angina, myocardial infarction, and hyperlipidemia. In 2019, the World Health Organization (WHO) estimated that 17.7 million people died from coronary heart disease (CHD) **(Nowbar et al., 2019)**.

Generally, medical experts make diagnoses based on electrocardiography, ultrasound, angiography, and blood test results. CHD is difficult to detect in the early stages of the disease **(Polonsky et al., 2010)**, yet it is critical for effective treatment. However, diagnoses are determined based on the doctor's personal experiences and understanding of the condition, which increases the likelihood of errors, delays appropriate treatment, extends treatment timeframes and drives up expenses. To handle these problems, many study have been performed on clinical decision-making systems utilizing various strategies such as machine learning and Data mining.

Traditional feature-based classification algorithm, as it is known, is less effective because their performance is typically dependent on the quality of handwritten features. To overcome this limitation, exploration of feature learning is utilized to improve the performance of traditional feature-based techniques. In this study, we propose using a deep neural network (DNN) **(Canziani et al., 2016)**. DNN, as everyone knows, can automatically learn a hierarchical feature representation from raw data, therefore it doesn't require any expert-crafted features because the features are automatically deduced and modified to achieve the desired result **(Takenaka et al., 2020)**.

The major purpose of this study is to reduce expert diagnosis time and improve diagnosis performance accuracy by proposing a novel CHD prediction model based on Deep learning algorithms. The UCI Machine Learning Repository Heart Disease Dataset is used in this study to classify normal and abnormal (CHD) using 14 attributes. The UCI heart disease dataset contains information on the sort of chest discomfort (angina) that is one of the most common CHD symptoms.

This paper is organized as follows. Section 2 discusses some of the related papers to our suggested approach and classification problem. The material and method of the DNN model are described in Section 3. The experimentation and findings of this effort are explained in Section 4. The final piece, section 5, concludes this study.

Related Works:

Nowadays many studies used data mining, machine learning, and deep learning in the disease prediction process but each research gives different strategy and this affect the performance accuracy.

(Sen, 2017) presented a new prediction system to perform automated prediction for heart disease. The system employed Naive Bayes, DT, and KNN as a classifier by using Weka Tool (www.cs.waikato.ac.nz/ml/weka/). Weka is a collection of a machine learning algorithm for data mining tasks developed by the University of Waikato Newzealand that implements data mining algorithms using JAVA language. The experimental results demonstrated that the Naive Bayes classifier is the best compared to DT, and KNN with 83.4% accuracy.

Two issues should be taken into consideration when implementing a heart disease decision support system. The first issue, determining the best subset of data with suitable and powerful features for the prediction process. The second issue, choose high-quality data mining techniques for indicating the presence of heart disease. This research **(Cherian & Bindu, 2017)** attempts to develop a new heart disease decision support system. This decision support system uses the Naïve Bayes algorithm for predicting whether a patient has heart disease or not, and uses a smoothing technique called Laplace smoothing as a feature selection method for increasing prediction accuracy by capturing important patterns in the data and avoiding noise. The Laplace smoothing method is a technique for the probability estimation and also known as Laplace correction or Laplacian estimator, it was able to decrease the number of features on the Cleveland Heart Disease database from 13 to 6 which are (age, gender, blood pressure, fasting blood sugar, cholesterol, and exercise-induced angina). Finally, the Naive Bayes classifier has been evaluated the result show that it has achieved 86% of accuracy in predicting heart disease.

In this study **(Gonsalves et al., 2019)**, the authors using one of the most common databases (South African Heart Disease dataset) to build a coronary heart disease prediction model based on three supervised learning techniques namely Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT). This database consists of 462 instances and 10 features that described historical medical data. experimental results using several performance evaluation metrics such as Accuracy, Sensitivity, Specificity, TPs, TNs, FPs, and FNs. NB algorithm reached 82% of accuracy which is the greatest accuracy amongst the three models.

(Gupta et al., 2019) used the random forest classification algorithm to predict heart disease based on the Cleveland Heart Disease dataset. The authors evaluate the accuracy of a random forest algorithm by changing the algorithm parameters like (number of the tree, number of splits, and a minimum number of the leaf node). Through experiments, it was observed that the accuracy could be increased by keeping the number of splits to be 20 and the number of the tree to be 75. The random forest algorithm achieved an accuracy of 85.8% by using previous parameters.

(Mohan et al., 2019) proposed a hybrid prediction model (HRFLM) that combining the characteristic of random forest algorithm (RF) as a classifier, and linear method (LM) as a new feature selection technique that aims to increase the performance of heart disease prediction by select a significant and best subset of features from among the 13 features. The hybrid model (HRFLM) was applied to the Cleveland Heart Disease database in two major phases. In the first phase, the LM feature selection algorithm decreased the number of features from 13 to 8. Moreover, in the second phase, the (RF) algorithm applied to the new subset of data with the most important 8 features and 303 instances to classify the patients of heart disease. They obtained an accuracy of 88.7%, which is better than used one algorithm as alone.

Materials And Method:

This section provides the major stages of the proposed prediction model for classifying coronary heart disease using a deep learning approach and four popular machine learning models SVM, kNN, LR, NB as a classifier. Also, the Heart disease dataset used in this research is explained in this section. Figure 1 presents the framework that has been proposed.

Figure - 1. Proposed Model Framework

Dataset

In this paper, we are using the Cleveland Heart Diseases datasets. This dataset is taken from the University of California, Irvine (UCI) Machine Learning Database. It consists of 303 cases (241 males, 62 females, mean age: 51 ± 9 years) and 14 features of healthy people and cardiac patients as described in Table 2. The Cleveland dataset contains some null values, as indicated in Table 1. Therefore we need some preprocessing approach to prepare this dataset before using it.

Table : 1 Heart Disease Database Properties.

Attributes	Description
Attribute Characteristics:	Categorical, Integer, Real
Associated Tasks:	Classification
Number of Instances	303
Number of Attributes	14
Missing or null Values	6 null values

Table : 2 Features Description.

Features	Descriptions
Age	Patients Age (in Year)
Sex	0 : female and 1 : male
Cp	Type of Chest pain Type 0: typical Angina Type 1: atypical angina Type 2: non-anginal pain Type 3: asymptomatic
Trestbps	Resting Blood sugar (in mm Hg on admission to the hospital).
Chol	serum cholesterols in mg/dl
Fbs	Fasting blood suger > 120 mg/dl. (1= true; 0= false)
Restecg	Resting ECG result
Thalach	Maximum heart rates Achieved.
Exang	Exercise induced angina.
Oldpeak	ST depression induced by exercise relative to rest.
Slope	Slope or peak exercise ST Segment. Value 1:upsloping Value 2: flat Value 3: downsloping
Ca	number of major vessels (0-3) colored by flourosop
Thal	3 = normal; 6 = fixed defect; 7 = reversable defect
class	The predicted attribute. 0: Yes; 1: No.

Data Cleaning

Data cleaning is the initial and important step of any data modeling and design process. It means modifying and filtering data to improve its quality so that it can be better examined and understood **(Alasadi et al., 2017)**. As indicated in Table 1, the training dataset for this experiment had 14 features, with 6 of them having a null values. The Filtering method assesses the mean of every attribute in which the data was null in our suggested system, and this mean value is replaced at the null value position.

Feature Scaling

Normalization is a Data Transformation technique that attempts to improve the classifier's functionality and performance by converting the original value into the proper format to make the dataset well-structured and acceptable for the next classification phase **(Patro et al., 2015)**. Feature scaling, which is a component of normalization, is a technique for standardizing the range of independent variables or data attributes. A min-max normalization is the simplest approach of feature scaling **(Borkin et al., 2019)**. In this strategy, the feature's minimum value is converted to a 0, the maximum value is converted to a 1, and all other values are converted to a decimal between 0 and 1. Eq (1) represents the generic formula.

$$a' = \frac{a - \min(a)}{\max(a) - \min(a)} \ldots\ldots\ldots\ldots (1)$$

Where:
a is original value, a' is normalized value, $\min(a)$ is minimum value of attribute, and $\max(a)$ is maximum value of attribute.

Learning Algorithms

1) SVM: A support vector machine is a type of model used to examine data and identify patterns in classification and regression problems **(Cotrtes et al., 1995)**. When your data contains exactly two classes, as shown in Figure 2, an SVM is used to classify it by determining the optimal hyperplane that divides all data points of one class from those of the other. The SVM technique is utilized in this study to predict this disease by displaying the training dataset with a hyperplane that classifies the data into two categories: presence and absence of heart disease.

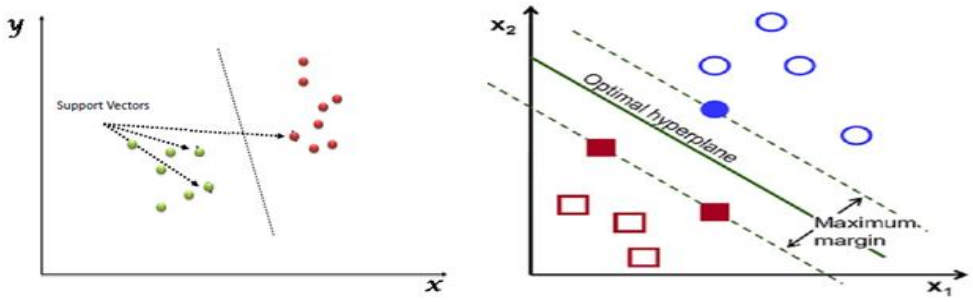

Figure - 2. SVM Algorithm

SVM is used to deal with class imbalance, as seen in Figure 2. When the total number of positive and negative classes is not equal, the class imbalance is a problem in machine learning, and the classifier will not perform effectively.

2) KNN: a nonparametric machine learning technique. It's a supervised learning algorithm. It refers to predicting the output from input data by calculating the distances between a query and all of the data's samples. The number of nearest neighbors to a new unknown variable that must be predicted or classified is defined by the symbol 'K' (Abu Alfeilat, 2019). The algorithm is depicted in Figure 3.

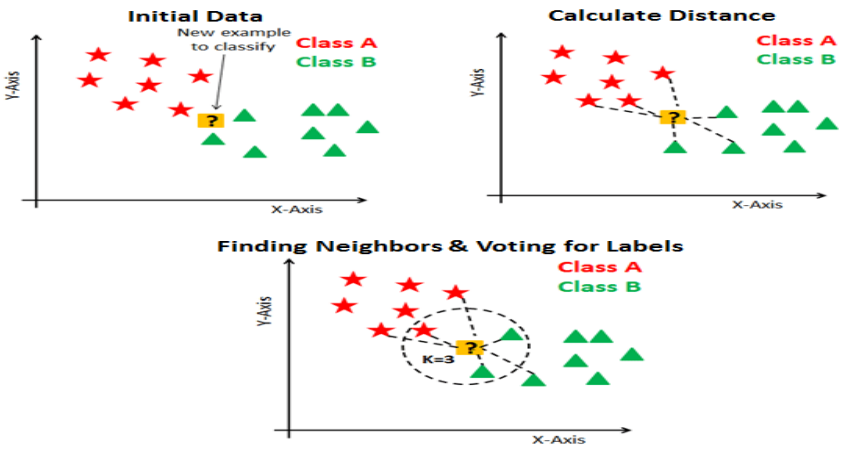

Figure - 3. KNN Algorithm

As shown in Fig. 3, we follow some steps to perform this algorithm:

- Decide on a K value.
- Calculate the distance between the unknown instance and all other cases.
- From the training data, choose the K- observations that are closest to the unknown data point.
- Using the most popular response value from the KNN, predict the response of an unknown data point.
- Stop.

3) Naive Bayes: The Bayes Theorem is used to create a probabilistic machine learning classification algorithm called Naive Bayes (Zhang, 2005). The Bayes Rule describes the probability of a feature based on prior knowledge of conditions that may be associated with that feature. It is a method of transitioning from P(X|Y) in the training dataset to P(Y|X) in the test dataset as the following Eq (2).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \ldots\ldots\ldots\ldots (2)$$

4) Deep Neural Network: The learning and generalization abilities of the neural network are inspired by human neural architecture **(Amato et al., 2015)**. A neural network is created by a group of neurons (or nodes) and its structure is designed by three layers: input layer, hidden layer, and output layer. Each neuron receives and analyzes information from other neurons before passing it on to the output layers **(Bouwmans et al., 2019)**. The network can be regarded as a black box or hidden system that accepts a vector of inputs with *Xia = Xi1, Xi2,..., Xim* and deliver a vectors of outputs with *Yib = Yi1, Yi2,..., Yin*. Figure 4 shows how the black box can be depicted in this work.

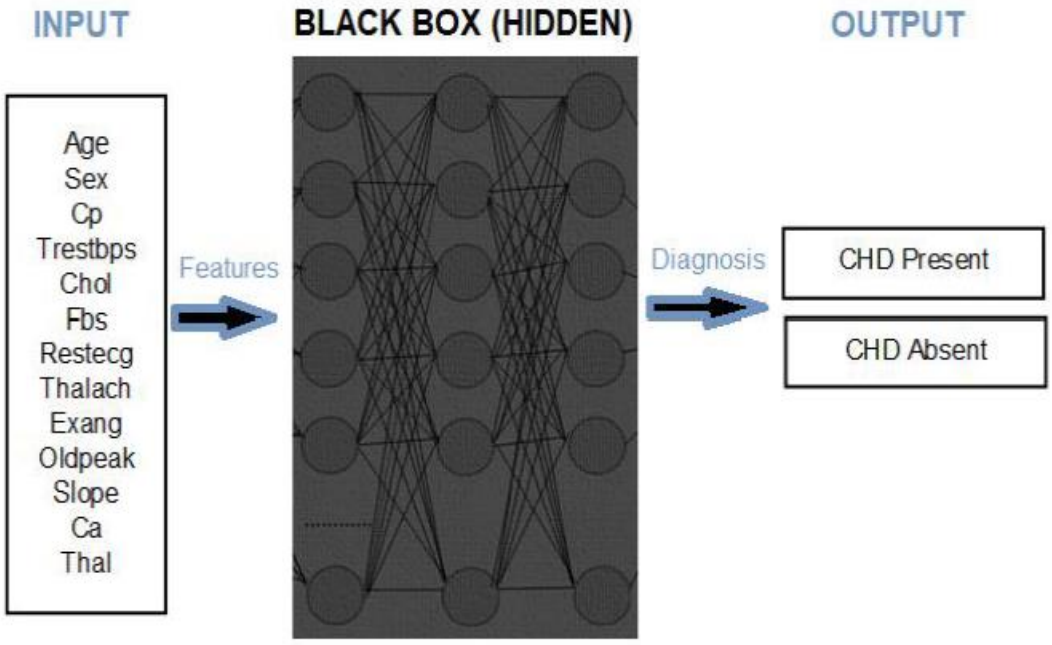

Figure - 4. The Neural Network's Black Box In CHD Prediction

From Fig. 4, The goal of the training manner is to approximate the function f within the vector Xia which represents the input (features of heart disease dataset) and Yib which is denoted to the output(CHD present, CHD absent).

$$Yib = f(Xia) \ldots\ldots\ldots\ldots (3)$$

The deep neural network (DNN) is a feed-forward neural network that is substantially larger and more complex than ordinary neural networks. The general deep framework is commonly used for classification via multiple hidden layers (more than two layers), and it enables the expression of complex hypotheses. Each layer only receives connections from its previous layer. The DNN model has a single input layer, many hidden layers, and a single output layer. One input layer used in the network model is the number of features of the heart disease dataset. Each hidden layer consists of nodes that have an activation function. The output layer has one node that gives the output regardless of whether the CHD is present or not. The sigmoid activation function in an output layer can be calculated as shown in eq.3 to predict the probability as an output. Since the probability of anything exists only between the range of 0 and 1.

Findings, And Discussions:

In this paper, experiments were conducted with several classifiers includes (K-NN, SVM, NB, and DNN) on Heart diseases data set to contrast them and obtain the best algorithm in the prediction process. The data set was divided into about 70% for training and 30% for testing. Table 3 illustrates the classification accuracy results of all algorithms where the accuracy metric is used to determine the performance of algorithms in the prediction process.

Table : 3 Comparison of Algorithm Accuracy.

Algorithm used	Accuracy
KNN	0.90
DNN	0.96
NB	0.85
SVM	0.82

Table 3 shows that the best accuracy of CHD predicting was achieved by the DNN model especially when using 4 hidden layers with 350 epoch numbers in the same number of nodes. To discuss the result, Table 4 depicts the outcomes of the DNN classifier of all layers, and clarifying the best performance of each layer in comparison to other layers with different epoch numbers, and reported that the best accuracy was achieved while using 4 hidden layers.

Table : 4 DNN Performance With One To Four Hidden Layers And Various Epoch Numbers

Hidden Layers	Epoch	Accuracy
1 Layer	200	0.77
	250	0.80
	300	084
	350	0.88
	400	0.91
2 Layers	200	0.84
	250	0.81
	300	0.88
	350	0.93
	400	0.81
3 Layers	200	0.78
	250	0.91
	300	0.88
	350	0.88
	400	0.87
4 Layers	200	0.89
	250	0.87
	300	0.93
	350	0.96
	400	0.89

One input layer utilized in the network model is the number of features of the heart disease dataset with a total of 13 nodes. To generate the input layer, the simulation uses 13 parameters as input. Furthermore, the number of hidden layers used in this study is determined by experimenting with which many hidden layers will best represent the DNN model. The number of hidden layers was varied from one to four, and the number of epochs was an experiment from 200 to 400. Each hidden layer consists of total of 100 nodes. The rectified linear unit (ReLU) activation function was utilized in each of the hidden layers because of its popularity. It has been employed in practically all neural networks and deep learning systems. ReLU refers to a unit in a neural network that uses the activation function $\max(0, x)$ (Dahl et al., 2013). The output layer has one node that gives the output regardless of whether the CHD is present or not. Table 4 shows that the DNN model performed best when four hidden layers with 350 epoch numbers were used in the same number of nodes.

Conclusion:

In a summary, in this paper, we proposed a new CHD prediction model that aims to assist healthcare providers in making correct and accurate decisions based on patients' symptoms and medical information. Moreover, the main objective of this study is to demonstrate the ability of the DNN algorithm in the prediction process and to explain how it works by performing multiple experiments with different numbers of layers and epochs at each time. In addition, we carried out an experiment to compare the predictive performance of different classifiers. We selected three popular classifiers considering their qualitative performance for the experiment which are NB, SVM, KNN and we compared its accuracy with the DNN algorithm based on the Cleveland heart disease dataset. DNN classifier achieved the best accuracy while using 4 hidden layers and 350 epochs.

References:

Abu Alfeilat, H. A. et al. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. Big Data, 7(4), pp. 221–248.

Alasadi, S. A. and Bhaya, W. S. (2017) 'Review of data preprocessing techniques in data mining', Journal of Engineering and Applied Sciences, 12(16), pp. 4102–4107.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis.

Borkin, D., Némethová, A., Michaľčonok, G., & Maiorov, K. (2019). Impact of data normalization on classification model accuracy. Research Papers Faculty of Materials Science and Technology Slovak University of Technology, 27(45), 79-84.

Bouwmans, T., Javed, S., Sultana, M., & Jung, S. K. (2019). Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Networks, 117, 8-66.

Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.

Cherian, V., & Bindu, M. S. (2017). Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique. International Journal of Computer Science Trends and Technology, 5(2), pp. 68–73.

Cotrtes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), pp. 273-297.

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013, May). Improving deep neural networks for LVCSR using rectified linear units and dropout. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8609-8613). IEEE.

Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: An experimental analysis. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (pp. 51-56).

Gupta, R., Kamal, R., & Suman, U. (2019). Heart Disease Prediction System Using Random Forest. IEEE Access, 6(2), 9206–9271.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, vol. 7, pp. 81542–81554.

Nowbar, A. N., Gitto, M., Howard, J. P., Francis, D. P., & Al-Lamee, R. (2019). Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. Circulation: Cardiovascular Quality and Outcomes, 12(6), e005375.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Polonsky, T. S., McClelland, R. L., Jorgensen, N. W., Bild, D. E., Burke, G. L., Guerci, A. D., & Greenland, P. (2010). Coronary artery calcium score and risk classification for coronary heart disease prediction. Jama, 303(16), 1610-1616.

Sen, S. K. (2017). Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. International Journal Of Engineering And Computer Science, 6(6), pp. 21623–21631.

Takenaka, K., Ohtsuka, K., Fujii, T., Negi, M., Suzuki, K., Shimizu, H., ... & Watanabe, M. (2020). Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology, 158(8), 2150-2157.

Zhang, H. (2005). Exploring conditions for the optimality of naïve bayes, International Journal of Pattern Recognition and Artificial Intelligence, 19(2), pp. 183–198.