Received: October 29, 2021. Revised: November 29, 2021. 453

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Classification and Prediction of Low-Density Lipoprotein Cholesterol LDL-C in The Palestinian Patients Using Machine Learning Techniques

Sanad Malaysha[1] Mohammed Awad[2]* Rami Hadrob[1]

[1]*Department of Computer Science, Arab American University, Palestine*
[2]*Department of Computer Systems Engineering, Arab American University, Palestine*
* Corresponding author's Email: mohammed.awad@aaup.edu

Abstract: Cholesterol is one of the major causes of health problems in Palestine and globally in the world, with percentages of 31.5% and 31.4%, respectively. Cholesterol has four main values which are the Total Cholesterol TCH, Triglycerides TG, Low-Density Lipoprotein Cholesterol LDL-C, and High-Density Lipoprotein Cholesterol HDL-C. The main level that is a major factor for Cardiovascular Disease CVD is LDL-C that is called bad cholesterol, it builds upon the arteries walls narrowing them and slowing the blood flow feeding the heart and brain, causing heart attacks and strokes. Machine learning (ML) techniques support recognizing and diagnosing the LDL-C, which is based on the past medical history and heuristic data. This research utilized ML techniques for classifying and predicting the LDL-C. Additionally, the techniques applied to the HDL-C classification and prediction. The utilized techniques are the Artificial Neural Networks ANNs, Recurrent Neural Networks RNN, Radial Basis Function Neural Networks RBFNN, Fuzzy Logic, Support Vector Machines SVM, Decision Tress DT, Logistic Regression LR, and a hybrid model of combining the ANNs with Fuzzy Logic for optimizing the results accuracy and reducing the classifying error. The dataset was collected from Palestine with cooperation with the Palestinian Ministry of Health MoH. Another additional supportive international dataset is utilized, which is collected by the Korean National Health and Nutritional Examination Survey KNHANES, it's used to generalize the results and compare with the other efforts. The obtained result outperformed the other efforts done on the same idea with a significant difference, it reached the accuracy of 97.10% in the international dataset and 95.55% in the national dataset.

Keywords: Cholesterol, Prediction, Classification, LDL-C, HDL-C, Triglycerides, Neuro-fuzzy, ANNs, SVM, Regression, RNNs, RBFNNs, Fuzzy logic.

1. Introduction

Cholesterol is an essential substance in the human body cells. The body and its cells need Cholesterol to make vitamins such as D, hormones such as steroids, acids such as bile -which helps digest the food-, and also is a substantial component in the cell membranes [1]. The lipid profile is a laboratory test used to measure the cholesterol levels in the blood, this test mainly covers the amounts of Total Cholesterol TCH, High-Density Lipoprotein Cholesterol HDL-C, Low-Density Lipoprotein Cholesterol LDL-C, and Triglycerides TG [2]. The clinical diagnosis decisions used to rely on the Cholesterol as total concentration, but the researchers have discovered that LDL-C is the actual value that has the medical meaning to the human health and diagnostic of the disease, so the LDL-C, replaced the TCH in the clinical decisions and became the high-risk indicator of the Cardiovascular Disease CVD [3]. This study will focus majorly on the prediction and classification of LDL-C as explained above, it's the main indicator, predictor, and reason for the CVDs, which is the kind of harmful disease that results in a significant percentage of deaths across the globe, including Palestine. So, if its concentration in the blood exceeds the acceptable range of around 160 milligrams (mg) per decilitre (dL) mg/dL, then it

needs accurate prediction and classification for such critical effect on human health and risking the life, and though to prescribe the right action and treatment for taking care of. Also, will consider the HDL-C as a secondary goal for the prediction and classification, because it represents the good Cholesterol where it supports to decrease the risk of the overflowing high concentration of the LDL-C [3]. Such important risk factor LDL-C of the CVDs, it's very critical to classify and predict its concentration in the blood accurately either the way is a laboratory or calculational, because if it exceeded the threshold there would be urgent actions and treatments to lower it since if it continues for a long time with high concentration would lead to the unwanted side effects of narrowing the arteries and vessels that feeding the heart and the brain. Measuring the LDL-C clinically is usually costive and time-consuming, so science started a time back to search for methods to predict its value in calculational ways such as the Friedewald equation, this and similar equations depend on calculations using the other lipid profile values of TCH, HDL, TG [4]. The studies have proved that the equations are not accurate because it relies on a fixed value of the ratio and changeable variables in the lipid profile that is not following regular patterns, and though the experiments and research started to find more efficient ways for detecting the LDL-C values even if the other parameters are chaotic, new methods where can classify and predict accurately based on the medical record of the patient including the most affecting risk factors in the LDL-C value [5]. To overcome the gap in the traditional equations of calculating the LDL-C and to avoid the costive clinical tests for measuring it, then researchers worked on the era of utilizing the Machine Learning ML for such scenarios especially in the healthcare sector because the ML science interest is to find the best techniques and methods for predicting the most accurate expected results based on datasets inputs representing the patient's profiles. Sometimes the available patient medical record data could be very complex on the expert abilities to find patterns or standardizations for diagnosing the case, but the machine can. So, the dataset itself would be the base of learning knowledge generation in the machine since the machine has the powerful abilities to match and analysis of complex data for the strong computational strength in its hardware and software [6]. As the ML has very efficient techniques for prediction and classification of outputs LDL-C value based on patterned or non-patterned medical data inputs, more specifically the medical record inputs including the lipid profile levels and other factors such as age, gender. So similar to the way the physician study the patient case, check symptoms, request labs, apply medical imaging, review the past medical history of the patient and family, relies on the knowledge gained from the medical field for evaluating the disease case. The same would be applied in expert systems that utilize ML techniques [7]. For the broad area of techniques that ML has for the prediction and classification purposes, so would traverse the most effective and commonly used methods similar to Logistic Regression LR, Support Vector Machine SVM, Decision Trees DT, Fuzzy Logic, and Artificial Neural Networks ANNs for the classification experiments of LDL-C and HDL-C. e prediction would use ANNs, Radial Basis Function Neural Networks RBFNNs, Recurrent Neural Networks RNNs, then study a combination of Fuzzy Logic and ANNs to optimize the prediction results. As explained, these techniques would need data for learning through training and testing phases of each algorithm, to be able to evaluate their results. The evaluation of classification results would use the accuracy percentage, and the prediction would use the Mean Square Error MSE which measures how far the predicted value from the actual value [8]. The used datasets belong to national and international sources. As the study target is the Palestinian Patients community so there was a cooperation with the Palestinian Ministry of Health MoH, that is to get the data related to the medical records of patients diagnosed with Cholesterol (High LDL-C), as there are specific fields and risk factors more affecting the LDL-C were required to be collected based on the studies [2, 4, 5, 7]. The international dataset source is from South Korea, it is a dataset collected by the Korean National Health and Nutritional Examination Survey KNHANES, and this dataset included the required factors but with limitation in the number of available records which didn't exceed a thousand records [9].

The paper is organized as follows. Section 2 introduces a set of related works. The datasets description presents in Section 3. Section 4 will illustrate the methodology. The implementation will be presented in Section 5. Section 6 shows the results and discussions; Conclusions and future works will be discussed in Section 7.

2. Related works

Cholesterol and especially the LDL-C has become an interest of researchers to carefully accurately predict and classify, that is for its harmful effect on human health and even on the human life,

it's a major cause for killing diseases such as CVD and relatives. Though, that made the interest of this research for the Palestinian Patients Cholesterol. Here would have a look at the closest and related efforts to the research goal, even it's very few efforts being done on the same idea. Milan et al. [2] have worked on predicting the LDL and other lipid profile levels using ANNs. Their goal is to predict, mainly based on age, gender, blood pressure, and obesity indicators such as Body Mass Index BMI. So, all the parameters are non-invasive which need no blood or any clinical laboratory tests.

In [10], this research interest is the LDL-C and additionally the HDL-C, so will focus on these attributes from the reviewed studies, by all means, their results showed a percent of accuracy by 79.29% for predicting LDL, and 21.23% for predicting HDL, which for sure need more works and optimization to increase the accuracy. Their results showed a percent of accuracy 79.29% for predicting LDL, which for sure need more works and optimization to increase the accuracy, so recommended to include the lipid profile test results additional to obesity.

Additional efforts were done in [4] by doing a comparison among Friedewald equation, regression, and the ANNs, where ANNs outperformed the others in the accuracy and achieved the average of Root Mean Square Error RMSE to be 24.00. The study had three inputs which are HDL-C, TCH, and TG; the output is the LDL-C value. They used ANNs to analyze the correlation between the fields of the lipid profile with the LDL-C which showed a high correlation of average equals 95%. It can include more risk factors such as obesity and blood pressure, which would increase the accuracy because the lipid test results can have additional considerations affecting the diagnosis.

In [5] toward predicting the LDL-C value without clinical laboratories, Taesic et al. utilized the concept of deep learning using the Deep Neural Network DNN to optimize the accuracy. The inputs to the model are the lipid profile levels which are the HDL-C, TCH, and TG, while the output is the major study goal which is the LDL-C. Based on the calculated MSE lowest MSE are around 6% and the highest accuracy among their studied options. Their study used the lipid test only, the more included risk factors would have a positive effect toward realistic prediction and be more accurate. We tried to include fields related to hypertension and body fat measurements.

The Back-Propagation ANNs (BP-ANN) technique is used in [7] for accurate prediction of LDL-C, more specifically in overweight people. The results showed a correlation mapping between the inputs and output around 94% with an ending condition of 1000 epochs.

The Genetic Algorithm GA used as an optimization technique for the BP-ANN inputs and structure construction in [11] by efforts from Tone et al., the inputs are Very Low-Density Lipoprotein VLDL, Intermediate-Density Lipoprotein IDL, LDL, and HDL, these values with all corresponding versions of the cholesterol, triglyceride, apolipoprotein A1 ApoA1, and apolipoprotein B ApoB. Their results showed LDL-C with a correlation of 60% and standard deviation SD of 1.5, also HDL-C with a correlation of 80% and SD of 0.3. Although they used all inputs of the blood plasma, the results showed no high correlation.

In this research, we focus on the Palestinian patients for predicting the bad Cholesterol which is the LDL-C, and the good Cholesterol which is HDL-C, utilized the maximum available possible data of the patient medical records, with data collected from local Palestinian hospitals. Additionally, worked in the study on the international dataset KNHANES to guarantee and compare the results of the local efforts with the other researches and works. The data included invasive and noninvasive input risk factors such as age, gender, hypertension, BMI, and the lipid profile laboratory measurements. And applied many ML classification and prediction techniques on both datasets so used the Fuzzy Logic to build Fuzzy Inference System FIS for the classification based on experts' feedback collected from Ibn Sina Hospital in Jenin City, Palestine. Then used the SVM, LR, and DT and compared it with the ANNS backpropagation version. The prediction used a different version of ANNs such as the BP-ANN, RBFNN, RNN, also combined the ANNs with the Fuzzy logic to build a hybrid model called the Neuro-Fuzzy that proved high-performance results in the prediction.

3. Dataset

The local data collected from across many hospitals in support from the Palestinian MoH, the source hospitals are distributed alongside the country from the north, south, and middle areas. The extracted patient profiles are 5484 records that include the features illustrated in Table 1, this number of records was extracted after a pre-processing, cleaning, transformation, and filtration process for the total received files from the hospitals under the MoH supervision. The files were received per the need of achieving acceptable results, each

Table 1. Features list with their values ranges, Palestine data

Feature Name	Value Range	Average
LDL	2.1 - 727.1	116
HDL	7 - 403	42
Total Cholesterol TCH	33.7 - 928.2	179
Triglyceride TG	4.3 - 5990	186
Age	0.5 – 104.33	53
Gender	Male - Female	--

Table 2. The Korean dataset selected features and their ranges.

Feature Name	Value Range	Average
LDL	20 - 252	115
HDL	8 - 84	42
Total Cholesterol TCH	92 - 372	208
Triglyceride TG	200 - 2115	310
Age	10 - 80	51
Gender	Male - Female	--
BMI	16.72 - 42.88	26
WHtR	0.37 - 0.75	0.5
SBP	84 - 204	124
DBP	49 - 135	80

hospital data is shared at different time, the experiments initial results of the first data grouped showed the need for extra data, so collected more hospitals data until reached the last data group and achieved the satisfying results.

The KNHANES dataset features are illustrated in Table 2. This dataset uses extra features such as Waist to Height Ratio (WHtR), Systolic Blood Pressure (SBP), and Diastolic Blood Pressure (DBP).

The picked international data is selected per what used in the related studies of the Cholesterol disease, so used KNHANES to extract the required data, the process filtered thousands of records until reached the total of 910 records that has all the required fields with available values, those records only that suited the experiments because all the records miss one or more values that are required to fulfill the experiments. The data is distributed into 580 males and 330 females, and these records are selected based on the data availability, as the records with missing values are excluded from the selection process, because a filtration stage done on the thousands of records to collect the profiles consisting of the minimal required features, the total features were 750 in the original file shared by the dataset source as mentioned in the previous section, but on the extraction process remained kept 10 fields.

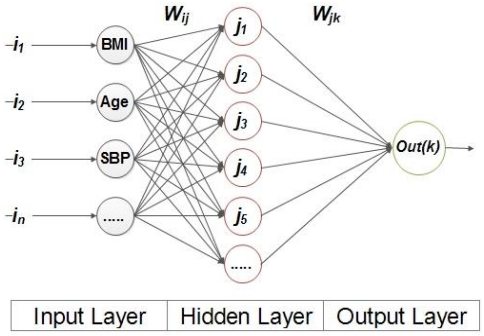

Figure. 1 (BP-ANNs) – structure

4. Algorithmic foundation

4.1 Back-propagation ANNs

The BP-ANNs algorithm is of the most common techniques used in the machine learning field, it is a powerful tool in the mapping between inputs-outputs of complex and non-linear data relations, similar to many needs that exist in human life such as objects classification and values predictions [12].

In the light of Fig. 1, the BP-ANNs perform their task depending on the following sequence with the mathematical correspondence [13, 14]:

- The input layer is represented by the input value $Xi[1 \ldots n]$.
- Hidden layer: each *Xi* is connected to the hidden layer neuron *j* with weight *Wij*
- Output layer: represented by the neurons that calculate the network output *Yk*
- the feed-forward stage to find the output is calculated per Eq. (1)

$$Output = f(\sum_{i=1}^{n} X_i . W_{ij}) \quad (1)$$

where: X_i: the *ith* input, W_{ij}: the *ith* corresponding weight between the *ith* input and the *jth* neuron, in this research, we use 50 neurons in the hidden layer, f: the activation function that could be linear or non-linear per the case complexity.

The BP stage is used to update the weights based on the error difference between the actual and objective outputs, it usually uses the gradient descent algorithm and relies on the Sigmoid activation function per Eq. (2)

$$w_{jk\,new} = w_{jk} + \Delta w_{jk} \quad (2)$$

where: $w_{jk\,new}$: the new weight that will be used in the next epoch, $w_{jk\,old}$: the old weight that is already used in the previous epoch.

4.2 Logistic regression LR

It's one of the machine learning techniques that is used in classification problems, especially the binary ones, like when classifying into two classes such as in this research scenario of LDL-C, it considers between two classes the high and normal. This technique took its name from the mathematical function used in, it is called the logistic function or sigmoid function which is shown in Eq. (3) [15].

$$f(x) = \frac{1}{1+e^{-x}} \quad (3)$$

where: x: the input to be classified and e: Euler's number which is a mathematical constant equal 2.71828. The output of the function would fail into the interval [0, 1], where the input value would be transformed into a value through the mentioned interval.

4.3 Support vector machine SVM

SVM is one of the most commonly used supervised ML techniques in pattern recognition problems, which is the mapping between inputs and corresponding class output in the complex patterns of input-output pair mapping. this research will utilize it in classifying the Cholesterol levels LDL-C and HDL-C between two major classes as if diagnosed with Cholesterol disease of high LDL-C or normal LDL-C concentration. It can generalize very well based on the training data and make prediction models for the new scenarios [16].

The general idea of the SVM is that there are data inputs of the form (x_1, y_1), (x_2, y_2), …, (x_n, y_n), where they have two patterns that need to find a separation area between them to classify as much accurately, the goal to find the marginal line equation and the support vectors equations with the maximum possible margins, that can guarantee the highest possibility of separation between the two patterns, which classify them into two areas with some percent of error comes from the pairs fail in the margin area or on the edge [17].

4.4 Decision trees DT

DT is the greedy algorithm where it has a top-down approach, starting from the root node reaching the leaf nodes by recursively creating the structure of the tree. So, it is more statistical methods that rely on the values but not on the scales or predefined assumptions about the data, and regardless of the data distribution or frequency, as it extracts the patterns based on the trends of the values, by this it's very flexible to handle linear and non-linear relations between the inputs and their target classes [18].

The Decision Tree algorithm is summarized in the following points:

- Start the tree from the root node R, which consists of all dataset instances.
- Find the best attribute that split the dataset into two main sub-trees left and right
- Repeat the previous step for the sub-trees
- The decision nodes for the splitting would contain the best attribute
- Keep splitting until reach the final node, where can't classify or more, that is the leaf node

4.5 Recurrent neural networks RNN

On the opposite side of the Feed-Forward ANNs that transfer the activation outputs as inputs to the next layer, there are the RNN where they have cycles as the neuron prior activation output would become again the input for the current output, considering the time-wise or ordinal wise, as per shown in Fig. 2, that figure illustrates one of the possibilities of the structure as there could be at least one cycle or many cycles, that maps between one-one, one-many, many-one or many-many pairs of inputs-outputs. Though the input value changes over time and sequence, so there could be varied expected outputs based on the activations cycling [19].

To represent the transition of the previous hidden state as input to the next or current hidden state, will translate it to Eq. (4), which use time sequencing to reflect how the prior output affects the new output as an addition to the current input [20], per show in the following:

$$h_t = \alpha\,(Wx_t + Uh_{t-1} + b) \quad (4)$$

where, x_t: the input at time t, h_t: the hidden state activation at time t, W: the weights for the current

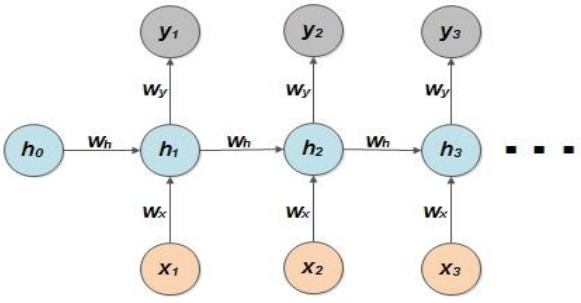

Figure. 2 RNNs structure

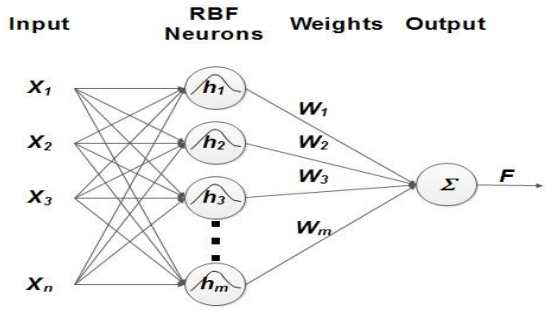

Figure. 3 RBFNN, structure

input, U: the weights for the recurrent input, b: the bias value of the neurons, in this research, we use 50 neurons in the hidden layer, and α: the activation function of the neurons.

4.6 Radial basis function neural network RBFNN

As in the below Fig. 3 shows the standard three layers of RBFNN, which is the input layer that supplies the high dimensional data inputs to the network, the hidden layer that re-distributes the inputs per the concept of clustering [21] to a set of centers, those centers that mainly utilize the Gaussian function as activation as it represents the radian functionality, by this the problem transformed into linearly separable, and the output layer define the separation and estimation of the outputs as single or many possible values [22].

The commonly used hidden layer activation function is the Gaussian function as the initial data mapping is non-linear, where it has mainly two parameters the center and the distribution spread, this spread that is controlled by the radial distance that defines the cluster circle diameter that would cover the data input per neuron [23]. In this research, we use 50 neurons in the hidden layer.

4.7 Neuro-fuzzy

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is a prediction model produced by Jang, around twenty years back, where it benefits from both Fuzzy logic and ANNs, Fuzzy Logic is important in the decisions to happen for uncertain cases that are very similar to the disease diagnosis because the diagnosis is always linked with the level of acuteness, also the ANNs have very powerful ability in the input-output mapping and recognition through the learning techniques with high accuracy. So, such a model will have a strong representation for the rules that can be recognized between the inputs and outputs, also for the rules that can be collected from the field experts [24]. This research utilized the input-output pairs collected from the

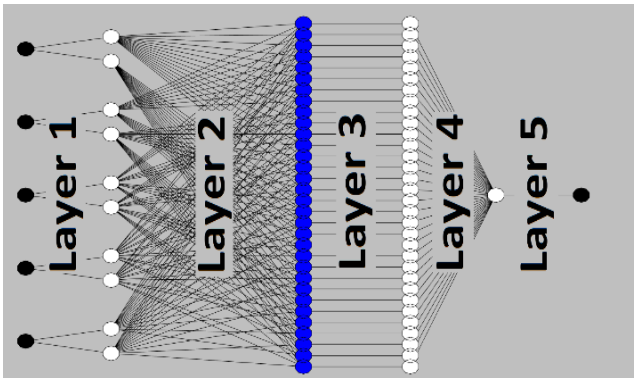

Figure. 4 Neuro-fuzzy structure

national and international data of Cholesterol patients, who are related to the Cholesterol disease caused by LDL-C exceeding the normal ranges.

Generally, the ANFIS structure consists of 5 layers in the form of the ANNs, where each layer would represent a phase of the Fuzzy Logic process, that is for reaching from the inputs to the expected output mapping. Each step represents one layer of the ANNs,

- Layer 1 represents the input variables to the ANNs, where each neuron at this layer transfer the input crisp value to the next layer
- Layer 2 (Fuzzification) represents the fuzzy sets, where the neurons in this layer represent the membership of each of the received inputs to which degree belongs to the represented fuzzy sets.
- Layer 3 (Fuzzy rules): each neuron in this layer represents a fuzzy rule that is generated based on the fuzzy sets' membership of the input crisp values.
- Layer 4 (aggregate layer): it combines the inputs from the previous layer using the operation of union per the fuzzy definition, though each neuron of this layer would make a probabilistic OR operation for all received inputs belonging to the same fuzzy set.
- Layer 5 (Defuzzification): it combines all the output of the previous layer representing its input, the union of the membership function degrees would be transformed into a crisp value, which is the average defuzzification value that represents the final membership value of the output fuzzy single set, that is corresponding the initial inputs from the different fuzzy sets.

5. Implementation

This work covered nine machine learning techniques, the ones used for the classification are Fine Trees and called Decision Trees DT, Logistic Regressions LR, Support Vector Machines SVM,

Back-Propagation ANN BP-ANN as liner output, Fuzzy Inference System FIS. On the other side, the techniques used for the prediction purposes are BP-ANN non-linear output (Sigmoid), RNN, RBFNN, and finally, the hybrid system of ANNs and Fuzzy Logic that is represented in a model called Neuro-Fuzzy. The computer that was used to run the experiments is HP ZBook Firefly 15 G7 Mobile Workstation, with the processor Inter (R) Core (TM) i7-10510U CPU of 2.3 GHz, the RAM is 16 GB and the HD is SSD 512 GB, the OS is Windows 10 Pro 64-bit.

Each technique was tried multiple times to get the average results, especially the cases when there are changeable attributes that affect the results, similar to the random initiation of weights in ANNs, those repetitions would give realistic outputs with no bias or chance-based accuracy. They are re-tried using a different number of neurons starting by 5 neurons, increasing fives, until reached 50 neurons, then to compare all rounds and select the best results for achieving the optimal classification per the conditions of the experiment, and used the five cross-validations to overcome the overfitting possibility. The experiments started with the local data that had fewer features than the needed and expected, while the results showed less accuracy than the international dataset because it has more features, but still both results are highly acceptable for the high accuracy approved in the experiments and high performance in the prediction. It is the first time such research gets applied on the Cholesterol classification and prediction using ML techniques on the Palestinian data.

Both local and international data files had, invalid records, non-formatted well to meet the requirements, so the first steps were to format them well. After the pre-processing and cleaning stage, the Palestinian data covered only 6 risk factors which are the lipid profile levels, in addition to age and gender. From around 60,000 Palestinian patient profiles, the considered records are 5484. And the correlation coefficient for the total inputs calculated for the output with an average of 0.9, that value means a perfect correlation between the input-output pairs, and that would lead to acceptable accuracy and mean errors in the prediction and classification of the Cholesterol levels targeted in this study, which are LDL-C and HDL-C. The same applied to the Korean dataset KNHANES, which consisted of 14,000 patient profiles, and each has 785 clinical fields, so after the filtration and formatting, the remained valid records are 910 profiles.

5.1 Performance metrics

This work relied mainly on the standard metrics available in the MATLAB tools, that is the Confusion Matrix, which is, by all means, the most commonly used and supported method, where it mainly represents four classes, those classes are the "True Positive TP", "False Positive FP", "True Negative TN", "False Negative FN". Hence, from the confusion matrix, then can calculate more metrics to evaluate the accuracy of the ML model abilities of classification and recognition. In the following, the list of used metrics definitions and equations:

- Accuracy: it is the general used metric to measure the overall results of the model, regardless of the positive and negative recognition

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

- Sensitivity: the percent of the TP instances

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

- Specificity: the percent of the TN instances

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

- Precision: the percent of TP instances to all instances of positive class

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

6. Results and discussions

This section will conclude the final results that achieved the best classification accuracy and prediction performance. There were many experiments on two different Cholesterol datasets, national Palestinian one Named PDS (Palestinian Dataset) and International Korean one named KNHANES, the test done for two different levels of the Cholesterol, which are the LDL-C as a major goal, in addition to the HDL-C.

There mainly are two different goals of the utilized Machine Learning techniques which are the classification and prediction, so divided the final discussion results into eight tables listed below and numbered 4 to 7, in each table will illustrate the best result done using which technique for which dataset and to which Cholesterol level if LDL or HDL.

Table 4. All techniques results comparison, PDS dataset for LDL-C classification

Technique	BP-ANN	LR	SVM	DT
Accuracy %	95.55	95.3	95.5	94.2
Sensitivity %	97.5	96.21	96.28	96.24
Specificity %	87.87	89.38	89.93	82.04
Precision %	96.17	98.39	98.48	97.02
G-Mean %	92.56	92.73	93.05	88.86
F-Measuring %	96.83	97.29	97.37	96.63

Table 5. All techniques results comparison, PDS dataset for HDL-C classification

Technique	BP-ANN	LR	SVM	DT
Accuracy %	91.8	91.7	90	90.4
Sensitivity %	98.43	0.93	0.91	0.93
Specificity %	29.23	0.72	0.61	0.56
Precision %	92.87	0.99	0.99	0.97
G-Mean %	53.64	0.82	0.74	0.72
F-Measuring %	95.57	0.95	0.95	0.95

Table 6. All techniques results comparison, KNHANES dataset for LDL-C classification

Technique	BP-ANN	LR	SVM	DT
Accuracy %	97.1	95.1	95.2	93
Sensitivity %	98.4	96.63	95.75	95.99
Specificity %	81.2	78.48	87.1	64.77
Precision %	98.43	97.93	99.02	96.22
G-Mean %	89.39	87.08	91.32	78.85
F-Measuring %	98.42	97.27	97.36	96.1

Table 7. All techniques results comparison, KNHANES dataset for HDL-C classification

Technique	BP-ANN	LR	SVM	DT
Accuracy %	97.8	95.8	95.7	93.4
Sensitivity %	99.77	96.33	95.71	96.03
Specificity %	0	54.55	50	13.79
Precision %	98.03	99.43	99.89	97.13
G-Mean %	0	72.49	69.18	36.39
F-Measuring %	98.89	97.85	97.76	96.58

The Two-Layered Feed-Forward Back-Propagation Artificial Neural Networks BP-ANN technique has always outperformed the other techniques in the classification, that is per shown in the tables 4.27 to 4.30, in the PDS dataset it achieved an accuracy of LDL-C classification as 95.55% and HDL-C classification as 91.8%, also for the KNHANES dataset achieved 97.10% for LDL-C classification and 97.80% for HDL-C. The other

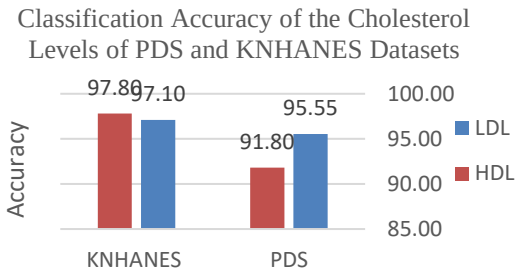

Figure. 5 The accuracy results of using BP-ANN for classifying LDL-C and HDL-C on the PDS and KNHANES datasets

techniques which are the Logistic Regression LR, Support Vector Machine SVM, and Decision Trees DT have also achieved very close accuracy to what is done by the BP-ANNs, but the BP-ANNs proved the highest ability in the classification. Figure 5 illustrates a comparison between the accuracy in the two different local and foreign datasets using the BP-ANN technique, that reached the best results.

On the side of the LDL-C and HDL-C prediction, the Neuro-Fuzzy technique has usually outperformed the other techniques, except in one case, the Recurrent Neural Network RNNs outperformed the Neuro-Fuzzy in the HDL-C prediction of the PDS dataset per shown in table 9, even that case the difference was not much in the prediction error between the actual and predicted values. Also, the Neuron-Fuzzy can outperform the RNN if the Neuro-Fuzzy model utilizes additional membership functions for the inputs, but that would increase the complexity.

Table 8. All techniques results comparison, PDS dataset for LDL-C prediction

Technique	MSE	Condition
BP-ANNs	0.00036	35 Neurons
RNNs	0.00039	20 Neurons
RBFNNs	0.00034	Spread Constant = 1.0 150 Neurons
Neuro-Fuzzy	0.00031	Gaussian Function Hybrid Optimization Linear Output

Table 9. All techniques results comparison, PDS dataset for HDL-C prediction

Technique	MSE	Condition
BP-ANN	0.00048	50 Neurons
RNN	0.00047	35 Neurons
RBFNN	0.00055	Spread Constant = 0.5 150 Neurons
Neuro-Fuzzy	0.00062	Gaussian Function Hybrid Optimization Linear Output

Table 10. All techniques results comparison, KNHANES dataset for LDL-C prediction

Technique	MSE	Condition
BP-ANNs	0.0009	40 Neurons
RNNs	0.0014	20 Neurons
RBFNNs	0.0012	Spread Constant = 1.0 150 Neurons
Neuro-Fuzzy	0.0001	Sigmoidal Function Hybrid Optimization Linear Output

Table 11. All techniques results comparison, KNHANES dataset for HDL-C prediction

Technique	MSE	Condition
BP-ANNs	0.0049	50 Neurons
RNNs	0.0058	35 Neurons
RBFNNs	0.0063	Spread Constant = 0.5 150 Neurons
Neuro-Fuzzy	0.0004	Trapezoidal Function Hybrid Optimization Linear Output

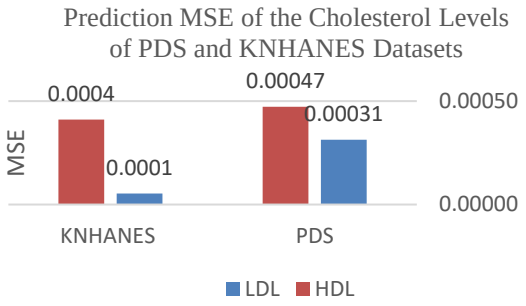

Figure. 6 The MSE of predicting LDL-C and HDL-C on the PDS and KNHANES datasets

The used measurement for the comparison between the results of the models is the Mean Square Error MSE. The whole used prediction results are illustrated in tables 8 to 11, where the results related to both PDS and KNHANES for predicting the LDL-C and HDL-C. As shown in the tables the used techniques in the prediction are the Neuro-Fuzzy, BP-ANNs, RNNs, and Radial Basis Function Neural Networks RBFNNs.

Actually, comparing the results between the PDS and KNHANES datasets would not reflect the real effect of the data, because the number of input fields and records is different between them, but Fig. 6 gives an initial indication for the performance of utilizing more fields or fewer records, especially in the LDL-C, as the MSE difference is very high between the two datasets to the right of the KNHANES, as it has more input fields representing very affecting risk factors in predicting the LDL-C, from those major risk factor is the hypertriton measurements of systolic and diastolic blood pressure SBP and DBP, respectively.

7. Conclusion

Cholesterol disease is a silent killer, where its diagnosis is mainly measured by the laboratory test of the lipid profile, especially the value of the Low-Density Lipoprotein Cholesterol LDL-C from the test results. The LDL-C is considered the major risk factor causing cardiovascular disease CVDs, where this disease is yearly costing the world billions of dollars and millions of deaths. Additionally, the LDL-C test itself is cost and time consuming, in addition to the invasive tests that require blood samples from the patient, sometimes after all that the lab results are not accurate because of human or device unintended error, hence the idea of utilizing the Machine Learning ML techniques to classify and predict the LDL-C has come to the light, where the accurate classification and prediction for sure would lead to the proper treatment, and through reducing the other side effects such as cost, time, samples, and the late diagnosis. For achieving the goal of the research which is to use the ML techniques for classifying and predicting the LDL-C in the Palestinian patients specifically, in addition to improving on the results done in the previous efforts in the field especially the international efforts. No local efforts were done to utilize ML in LDL-C diagnosis from before, though, ten ML techniques were applied on national and international datasets for referencing and comparison targets. The thesis used the Back-Propagation Artificial Neural Networks BP-ANNs, Recurrent Neural Networks RNN, Radial Basis Function Neural Network RBFNN, Fuzzy Logic, Support Vector Machines SVM, Logistic Regression LR, Decision Trees DT, also a hybrid model of ANNs and Fuzzy Logic that is called the Neuro-Fuzzy. The results outperformed other older efforts in the international works while proving high accuracy and low error in the national data, which is accepted to utilize locally in the medical sector for recognizing and diagnosing the Cholesterol disease and more specifically is the LDL-C concentration. On the margin, the experiments covered another lipid profile value which represents the good Cholesterol that is called the High-Density Lipoprotein Cholesterol HDL-C. In the classification area, the accuracy of the Palestinian dataset achieved a high percentage with 95.55% for the LDL-C and 91.80% for the HDL-C. in the Korean dataset, the highest accuracy achieved in the whole utilized models is 97.10% in the LDL-C and 97.80% for the HDL-C. on the other side, the

prediction measured per the error between the actual and predicted values using the Mean Square Error MSE, which also proved accepted results in the Palestinian dataset, which in the LDL-C prediction showed the lowest MSE to be 0.0003 and for the HDL-C is 0.0005, additionally, in the Korean dataset the prediction of LDL-C has lowest MSE of 0.001 and in the HDL-C the lowest value is 0.0004 for the MSE.

The designed Neuro-Fuzzy system in cooperation with the experts would require to continue the efforts in collecting more inputs and rules, in addition to supporting the Fuzzy rules with the classification phase using the Support Vector Machines SVM. It's a very priority to try Deep Learning by trying multiple hidden layers with a different number of neurons to achieve better classification accuracy, specifically if included more related risk factors especially the ones that were collected for this research preparation through the studying phase, such as body fat, Very Low-Density Lipoprotein VLDL, liver status, kidney status, asthma status, heart rate, diabetes test, and smoking.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Author 1: Data collection and curation, methodology, techniques analysis, solutions design, experimentation, and implementation.

Author 2: Conceptualization, Data collection, methodology, techniques validation and enhancement, orientation, review and editing, feedback and comments, criticize and revisit the work, supervision.

Author 3: Orientation, editing, review feedback, and comments.

References

[1] B. A. Ference, H. N. Ginsberg, I. Graham, K. K. Ray, C. J. Packard, E. Bruckert, and A. L. Catapano, "Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel", *European Heart Journal*, Vol. 38, No. 32, pp. 2459-2472, 2017.

[2] M. Vrbaški, R. Doroslovački, A. Kupusinac, E. Stokić, and D. Ivetić, "Lipid profile prediction based on artificial neural networks", *J. Ambient Intell. Humaniz. Comput.*, pp. 1-11, 2019.

[3] V. J. Carey, L. Bishop, N. Laranjo, B. J. Harshfield, C. Kwiat, and F. M. Sacks, "Contribution of high plasma triglycerides and low high-density lipoprotein cholesterol to residual risk of coronary heart disease after establishment of low-density lipoprotein cholesterol control", *Am. J. Cardiol.*, Vol. 106, No. 6, pp. 757-763, 2010.

[4] P. Thaisiam, J. Sothornwit, S. Charoensri, S. Pattanapairoj, P. Kotruchin, and C. Pongchaiyakul, "A New Low-Density Lipoprotein Cholesterol Estimation Model from a Linear Regression Model and an Artificial Neural Network", *J. Med. Assoc. Thail.*, Vol. 103, No. 4, pp. 346-352, 2020.

[5] T. Lee, J. Kim, Y. Uh, and H. Lee, "Deep neural network for estimating low density lipoprotein cholesterol", *Clin. Chim. Acta*, No. 489, pp. 35-40, 2019.

[6] J. Wiens and E. S. Shenoy, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology", *Clin. Infect. Dis.*, Vol. 66, No. 1, pp. 149-153, 2018.

[7] J. Ma, J. Yu, G. Hao, D. Wang, Y. Sun, J. Lu, and F. Lin, "Assessment of triglyceride and cholesterol in overweight people based on multiple linear regression and artificial intelligence model", *Lipids Health Dis.*, Vol. 16, No. 1, p. 42, 2017.

[8] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, Vol. 7, pp. 81542-81554, 2019.

[9] T. Lee, J. Kim, Y. Uh, and H. Lee, "Korean public and hospital data for estimating LDL-cholesterol", *Data Br.*, Vol. 22, p. 204, 2019.

[10] R. K. Wadhera, D. L. Steen, I. Khan, R. P. Giugliano, and J. M. Foody, "A review of low-density lipoprotein cholesterol, treatment strategies, and its impact on cardiovascular disease morbidity and mortality", *Journal of Clinical Lipidology*, Vol. 10, No. 3, pp. 472-489. 2016

[11] T. F. Bathen, J. Krane, T. Engan, K. S. Bjerve, and D. Axelson, "Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis", *NMR Biomed. An Int. J. Devoted to Dev. Appl. Magn. Reson. Vivo*, Vol. 13, No. 5, pp. 271-288, 2000.

[12] J. Liu, Y. Chen, L. Lan, B. Lin, W. Chen, M. Wang, and Y. Duan, "Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network", *European Radiology*, Vol. 28, No. 8, pp. 3268-3275. 2018.

[13] I. Hamdan, M. Awad, and W. Sabbah, "Short-Term Forecasting of Weather Conditions in Palestine Using Artificial Neural Networks", *J. Theor. Appl. Inf. Technol*., Vol. 96, No. 9, pp. 2494-2504. 2018

[14] H. Qteat and M. Awad, "Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabete", *International Journal of Intelligent Engineering and Systems*, Vol. 14. No. 3, pp. 11-22. 2021.

[15] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: an evaluation", *Mach. Learn.*, Vol. 68, No. 3, pp. 235-265, 2007.

[16] A. Alabdallah and M. Awad, "Using weighted support vector machine to address the imbalanced classes problem of intrusion detection system", *KSII Trans. Internet Inf. Syst.*, Vol. 12, No. 10, pp. 5143-5158, 2018.

[17] S. Suthaharan, "Machine learning models and algorithms for big data classification", *Integr. Ser. Inf. Syst*, Vol.36, pp. 1-12, 2016.

[18] Y. Y. Song and L. U. Ying, "Decision tree methods: applications for classification and prediction", *Shanghai Arch. Psychiatry*, Vol. 27, No. 2, p. 130. 2015.

[19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acousticmodeling", In: *Proc. of Fifteenth Annual Conference of the International Speech Communication Association*, pp. 338-342, 2014.

[20] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn", In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5457-5466, 2018

[21] M. Awad and A. Abuhasan, "A smart clustering based approach to dynamic bandwidth allocation in wireless networks", *Int. J. Comput. Networks Commun*, Vol. 8, No. 1, pp. 73-86, 2016.

[22] M. Awad and I. Qasrawi, "Enhanced RBF neural network model for time series prediction of solar cells panel depending on climate conditions (temperature and irradiance)", *Neural Comput. Appl.*, Vol. 30, No. 6, pp. 1757-1768, 2018.

[23] M. Awad, "Enhanced hybrid method of divide-and-conquer and RBF neural networks for function approximation of complex problems", *Turkish J. Electr. Eng. Comput. Sci.*, Vol. 25, No. 2, pp. 1095-1105, 2017.

[24] J. S. Jang, "ANFIS: adaptive-network-based fuzzy inference system", *IEEE Trans. Syst. Man. Cybern.*, Vol. 23, No. 3, pp. 665-685, 1993.