Arab American University-Palestine

Faculty of Graduate Studies

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

By

Sanad Ahmad Anees Malaysha

Supervisor **Prof. Dr. Mohammed Awad**

Co-Supervisor **Dr. Rami Hadrob**

This Thesis was submitted in partial fulfillment of the requirements for the Master's degree in Computer Sciences.

June 2021

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

By

Sanad Ahmad Anees Malaysha

This thesis was defended successfully on **07th/07/2021** and approved by:

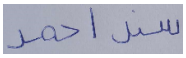
Committee Members	Signature
1. Supervisor: **Prof. Dr. Mohammed Awad**	.....
2. Co-Supervisor: **Dr. Rami Hadrob**	.....
3. Internal Examiner: **Dr. Ahmed Ewais**	.....
4. External Examiner: **Dr. Emad Natsheh**	.....

Declaration

I declare that this thesis with the title “Classification and Prediction of Low-Density Lipoprotein Cholesterol LDL-C in the Palestinian Patients using Machine Learning Techniques” is done by my efforts and works solely by myself, and does not contain any work from other researchers except the referenced ideas, and has not been submitted for any other degree or scientific community.

Name: Sanad Ahmad Anees Malaysha

Date: 11/08/2021

Signature: 

Dedication

I dedicate this work to my mother, father, sisters, brothers, and all supporters who made me enthusiastic and ambitious for finishing the thesis to move on to the next level for my dreams.

Acknowledgments

I would like to emphasize my sincere thanks and appreciation for the efforts that have been done by my supervisors Prof. Dr. Mohammed Awad and Dr. Rami Hadrob to have this thesis reach the capstone. It was a long journey under their kind support with care and advice to overcome any challenges that faced me across the path until ended with submitting it for the reviews and discussions.

Abstract

Cholesterol is one of the major causes of death locally 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 the 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 brain strokes.

The science of Artificial Intelligence (AI) and Machine learning (ML) has become a main player that supports in recognizing and diagnosing the LDL-C. The ML depends on the past medical history and heuristic data of the patients diagnosed with Hyperlipidemia. The Hyperlipidemia is the case when the LDL-C exceeding the acceptable and healthy normal ranges threshold of 160 Milligrams per deciliter mg/dL. Classifying and predicting the LDL-C using the ML techniques would guarantee accurate approximation and diagnosis for the disease, avoiding human error, costive laboratory materials, time-consuming waiting for the results. So, the ML has a significant positive effect on the diagnosis and hence the treatment. For this purpose, this thesis 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 error. These methods require a dataset for training and testing the used techniques. Since the

study targets the Palestinian community of Cholesterol patients, so cooperation with the Palestinian Ministry of Health MoH had taken a place, they provided the needed medical history data and risk factors related to the Palestinian patients of Cholesterol. 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 used and produced methods outperformed the other efforts done on the same idea with a significant difference, the older efforts done in the research titled "Lipid profile prediction based on artificial neural networks", reached 80% accuracy as they relied on a fewer number of fields limited to the lipid profile values only, where in this thesis it reached to the accuracy of 97.10% in the international dataset and 95.55% in the national dataset. Though this thesis recommends including more affecting risk factors and fields, especially the ones related to the noninvasive works, as it will lower the cost, efforts, time, and increase the accuracy.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Network-Based Fuzzy Inference System
ANNs	Artificial Neural Networks
ApoA1	Apolipoprotein A1
ApoB	Apolipoprotein B
ATHENA	Analysis Tool for Heritable and Environmental Network Associations
AUC	Area Under the Curve
BMI	Body Mass Index
BP-ANN	Back Propagation - Artificial Neural Networks
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
COA	Centroid of Area
CVD	Cardiovascular Disease
CVDs	Cardiovascular Diseases
DBP	Diastolic Blood Pressure
DNN	Deep Neural Network
DT	Decisions Trees
EVs	Expression Variables
FIS	Fuzzy Inference System
GA	Genetic Algorithm
GENN	Grammatical Evolution Neural Networks
GESR	Grammatical Evolution Symbolic Regression
GUIs	Graphical User Interfaces
HDL	High-Density Lipoprotein
HDL-C	HDL Cholesterol
HHD	Hypertensive Heart Disease
HR	Heart Rate
IDL	Intermediate-Density Lipoprotein
IHD	Ischemic Heart Disease

KNHANES	Korean National Health and Nutritional Examination Survey
LDL	Low-Density Lipoprotein
LDL-C	LDL Cholesterol
Max	Maximum
MF	Membership Function
mg/dL	Milligrams Per Deciliter
Min	Minimum
ML	Machine Learning
MLP	Multilayer Perceptron
mmHg	Millimeters of Mercury
MoH	Ministry of Health
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NCEP	National Cholesterol Education Panel
NLP	Natural Language Processing
PAD	Peripheral Artery Disease
PDS	Palestinian Dataset
RBF	Radial Basis Function
RBFNN	Radial Basis Function Neural Networks
ROC	Receiver Operating Characteristic
RMSE	Root Mean Square Error
SADHtR	Sagittal Abdominal Diameter to Height Ratio
SBP	Systolic Blood Pressure
SNPs	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
TCH	Total Cholesterol
TG	Triglycerides
VLDL	Very Low-Density Lipoprotein
VLDL	Very Low-Density Lipoprotein
WHtR	Waist to Height Ratio
X-Ray CT	X-Ray Computed Tomography

Chapter 1

Introduction

Chapter 1

1.1 Introduction

Cholesterol is an essential substance in the animal cell that is including the human body cells. Cholesterol is similar to the other contents that flow through the blood flood, especially in the plasma fluid. The body and its cells need Cholesterol to make vitamins such as D, hormones such as steroids, acids such as bile -which helps digesting the food-, and also is a substantial component in the cell membranes. The liver generates the Cholesterol, attaches it to carrier particles that transport it through the arteries to the needy cells [1].

The lipoprotein particles are the carriers of the Cholesterol substance, which from their names can know that each particle consists of lipid -fat- and protein. The particles are classified into mainly five types: Chylomicrons, Intermediate-Density Lipoprotein IDL, Very Low-Density Lipoprotein VLDL, Low-Density Lipoprotein LDL, and High-Density Lipoprotein HDL. Though, the Cholesterol levels description dependent on the attached carrier particle name. The major interest of studies focuses on LDL Cholesterol LDL-C as if it exceeds the normal ranges of its concentration in the blood will become a dangerous risk factor of cardiovascular diseases CVDs. Such diseases annually cost the globe around $863 billion, and 31.4% of deaths cause [2]. To have a wider view of the possible diseases related to the CVDs would mention some of them that include Ischemic Heart Disease IHD, Coronary Artery Disease CAD, Coronary Heart Disease CHD, Hypertensive Heart Disease HHD, Peripheral Artery Disease PAD, and many others to tell, where we can see how harmful and dangerous such factor the LDL-C [3].

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 [4]. 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 CVD [5].

Generally, the LDL-C is called the bad Cholesterol and HDL-C is called the good Cholesterol, actually not the Cholesterol itself that is the bad or good but the attached particles of the lipoprotein types which are the LDL and HDL. LDL-C is called bad because it passes the Cholesterol through the arteries to the needy cells and parts, on the way it usually builds up with other substances what called plaque on the artery walls. The plaque narrows the arteries and causes slowness of the blood flow especially in the coronary arteries causing the CVDs in the heart, and in neck arteries that down the blood flow to the brain leading to strokes. On the other side, HDL-C called good because it removes and returns the overflowing Cholesterol from the membranes and the cumulated on the arteries walls to the liver, that gets rid of the excess Cholesterol amounts from the body, that excess amounts mostly caused by the LDL-C cumulation on the arteries walls [1] [2] [6].

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 deciliter (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 [5].

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 the science started a time back to search for methods to predict its value in calculational ways such as Friedewald equation 1.1.1 [7], this and similar equations depend on calculations using the other lipid profile values of TCH, HDL, TG [8].

$$LDL = TCH - \frac{TG}{5} - HDL \qquad \text{[9] (1.1.1)}$$

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 [10]. And there are a lot of factors that can use in evaluating the LDL-C in addition to the lipid profile levels, such as age, gender, Body Mass Index BMI, Waist to Height Ratio WHtR, Sagittal Abdominal Diameter to Height Ratio SADHtR, Systolic Blood Pressure SBP, Diastolic Blood Pressure DBP, Heart Rate HR, family history, smoking history, diabetes, and many more to mention. Generally, the risk factors are important indicators to the physicians and practitioners for evaluating any disease, as it's not an independent view to the symptom of a disease, but the integrated and complementary view to include especially the most affecting factors in diagnosing the disease such as the Cholesterol, and more specifically the elevation in the LDL-C which in other words called the Hyperlipidemia [11].

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 [12].

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, get laboratory tests, apply medical imaging, review past medical history of 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 the ML techniques [13]. This combination of the medical profiles and ML techniques introduced the concept of medical expert systems, where it relies on the medical history of the patient and the physician's experience for evaluating and diagnosing the diseases by predicting the clinical measures and classifying the results levels. The system will work as supportive for the medical staff to shorten the process of checking symptoms and laboratory testing to decide the patient's case. The goal from the techniques to exploit the heuristic data and experts studies to solve such complex problems of approximating the LDL-C by machines, and usually the data will be in the

form of input-output pairs for representing the risk factors and their corresponding diagnosis, the same data will be divided and used for training and testing ML techniques [14].

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, and Artificial Neural Networks ANNs for the classification experiments of LDL-C and HDL-C, in addition, would build a Fuzzy Inference System FIS using the Fuzzy Logic for the classification purposes, to be considered as expert system. For the prediction would use ANNs, Radial Basis Function Neural Networks RBFNN, Recurrent Neural Networks RNN, 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 [15].

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 [4] [8] [10] [13]. An official letter was submitted to the MoH and they fulfilled the request with high cooperation, while there was only one limitation which is the shortage in the number of available collected values in the patient records, the requested factors are gender, age, height, weight, abdominal diameter, waist, SBP, DBP, TCH, TG, HDL-C, and LDL-C. While the received are gender, age, TCH, TG, HDL-C, and LDL-C, which by all means would prove accurate results by including all of them in the inputs to predict the LDL-C. 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 [16].

1.2 Objective

The major goal is to utilize the ML and AI prediction and classification techniques on the Palestinians diagnosed with Cholesterol disease, the dataset collected from different hospitals across the country. The main used techniques are the ANNs and Fuzzy logic, and a combination of them to increase the accuracy and decrease the error between the actual and predicted measurements, and also the diagnosis classification to be accurate using these techniques, as the wrong diagnosis would cause major side effects of the high LDL-C concentration on the human life threat, because such increase is a major risk factor in the CVDs. Where the precise evaluation would lead to the right treatment of the disease which reflects on the level of public health in Palestine. And generally, would summarize the mentioned objective in the following points:

- Collect Palestinian dataset for the Cholesterol study purposes.
- Apply ML techniques on the Palestinian dataset of Cholesterol.
- Increase the ML techniques accuracy and decrease the error in the Cholesterol classification and prediction, specifically the LDL-C level.
- Support the Palestinian medical staff with a medical expert system for recognizing and diagnosing the Cholesterol disease.
- Use international dataset to verify the results.
- Compare the thesis results with the other studies in the field.

1.3 Contribution

This thesis works done on two sources of data, the national dataset of Palestine that is collected from the local hospitals by support from the Palestinian MoH, and the international standard dataset of KNHANES that is based in South Korea. The applied modules of SVM, DT, LR, Fuzzy Logic, ANNs, RNN, RBFNN, and the hybrid module of Neuro-Fuzzy showed better accuracy and less error in the results. Especially in comparing the international data with other works, because this thesis utilized more input fields and techniques combination. On the side of the local data, it is a novel application of using those techniques and hybrids for approximation of and categorizing the LDL-C in the Palestinian patients, while the available risk factors and collected data are limited, but the results reached a level of precision which is very acceptable. That would lead to a more realistic evaluation and treatment for the patients. Additionally, it will lower the efforts and expenditure in the area of Cholesterol care, as the high accuracy of the results mean to be proactive in the diagnosis at early stages so the disease would not become chronic, also reduce the laboratory needs of tests and materials to get the clinical results of LDL-C. Moreover, will lower the cost on the patient and save the physician time in the diagnosis, because the result of the hybrid system is ready within seconds. The clinical and laboratory option has many drawbacks such as slowness, much preparations, samples, and the human error is a major reason for failures in diagnosis, and the laboratory device maintenance could lead to the same.

1.4 Organization of Thesis

The rest of this thesis is ordered starting from Chapter 2 to Chapter 5, with supportive appendices. Chapter 2 introduces a background about the study works, makes a literature review for the related works of the field, previews the datasets including both national and international sources. The related works review would cover the closest and most related studies to the field of classifying and predicting LDL-C using the ML techniques. Chapter 3 will explain the data extraction and preprocessing phase including the selection and normalization, in addition to the data fields explanation and defining. Then will move to explaining and reviewing the proposed methods and utilized techniques such as ANNs, RBFNN, RNN, Fuzzy Logic, SVM, LR, DT, and the Neuro-Fuzzy model, for each model the methodology and implementation, then the performance measurements metrics. Chapter 4, illustrated the model's experiment results, studying and analyzing their effect, discussing and comparing the advantage among them. The results were introduced in a way to show the performance for each model on each dataset of the Palestinian and Korean one, that is to show how accurate and precise each of them in the view of the other. Chapter 5 summarized the conclusion, the future works, and recommendation, which is based on what previewed in the results discussions. And Finally, appendices A, B, and C covered all the detailed results of the experiments tests for both datasets, in addition to any supportive documents, such as the official letters that were submitted to the Palestinian MoH and the local hospitals through the times of preparations.

Chapter 2

Background

Chapter 2

2.1 Background

In Palestine, Cholesterol disease has become widespread across [17], so need to have an efficient, proactive, precise, quick, and cheap way for evaluating the disease. Though introducing and using suitable ML techniques in the disease diagnostic would achieve the target. On that basis a data collected from across the country hospitals related to patients with Cholesterol diagnosis, and used them as a base to evaluate the utilized ML techniques and hybrid modules of ANNs different structure, Fuzzy Logic, SVM, and others to guarantee that the results of the local data are properly correct, so also applied the used techniques on the international data KNHANES, then compared the results with other works applied on the same dataset, which proved promising optimization on the accuracy and performance per the comparison of the results, it reached to a level where this thesis results outperformed the other works, and the details will be illustrated in the discussion of the results.

2.2 Datasets

As the country of study is Palestine there was an efforts to collect the required medical records in cooperation with the Moh because no studies were done of such kind from before in Palestine, so no ready data was there, and though worked side by side with the MoH staff to pick up the right patient profile with the right fields, though the works were done on that basis. For the international standardization of datasets, there was no public dataset dedicated for the Cholesterol studies, since Cholesterol is usually considered a risk factor for other disease studies such as hypertension. But extracted the data from a common dataset called KNHANES for disease studies, though utilized it for the ML experiments that is similar to what other studies have done of prediction and classification for the Cholesterol [16] [18].

Table 2.1: Classification of lipid profile measurements as defined in NCEP [19].

Range	Diagnosis
LDL-C:	
< 100	Optimal
100 - 129	Near or above optimal
130 - 159	Borderline high
160 - 189	High
≥ 190	Very high
Total Cholesterol TCH:	
< 200	Desirable
200 - 239	Borderline high
≥ 240	High
HDL-C:	
< 40	Low
≥ 60	High
Triglyceride:	
< 150	Normal
150 - 199	Borderline high
200 - 499	High
≥ 500	Very high

The fields of the used datasets to be explained with details in the below section, both datasets included the lipid profile levels LDL, HDL, TCG, and TG. These Cholesterol levels have standard ranges for the classification and diagnosis per what is standardized in the National Cholesterol Education Panel NCEP [19], per illustrated in Table 2.1 all levels concentrations expressed in mg/dL. For the gender related features, in this thesis works didn't consider the effect of gender differences on the used features because the ranges of the used features are very close in the both genders.

2.2.1 National Data

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 hospitals are the Governmental Palestine Medical Complex Hospital, Hebron Governmental Public Hospital, Rafedia Governmental Public Hospital, and Jenin Governmental Public Hospital. The extracted patient profiles are 5484 records that include the features illustrated in Table 2.2, this number of records extracted after a preprocessing, 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 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.

Table 2.2: Features list with their values ranges, Palestine data

Feature Name	Value Range	Average	Normal Ranges
LDL	2.1 - 727.1	116	< 160 mg/dL
HDL	7 - 403	42	40-60 mg/dL
Total Cholesterol TCH	33.7 - 928.2	179	< 240 mg/dL
Triglyceride TG	4.3 - 5990	186	< 200 mg/dL
Age	0.5 – 104.33	53	--
Gender	Male - Female	--	--

The medical records profiles are distributed between 3076 males and 2408 females, the selection was based on the profile consists the minimal common features as explained in the above table. So, the sample is limited, randomly selected, and filtered per the completeness of the profile. By all means, if the features would be more that would optimize the results. Since the LDL-C is the major goal for the study so the figure 2.2.1 illustrates the dataset distributions according to the LDL-C diagnosis classes.

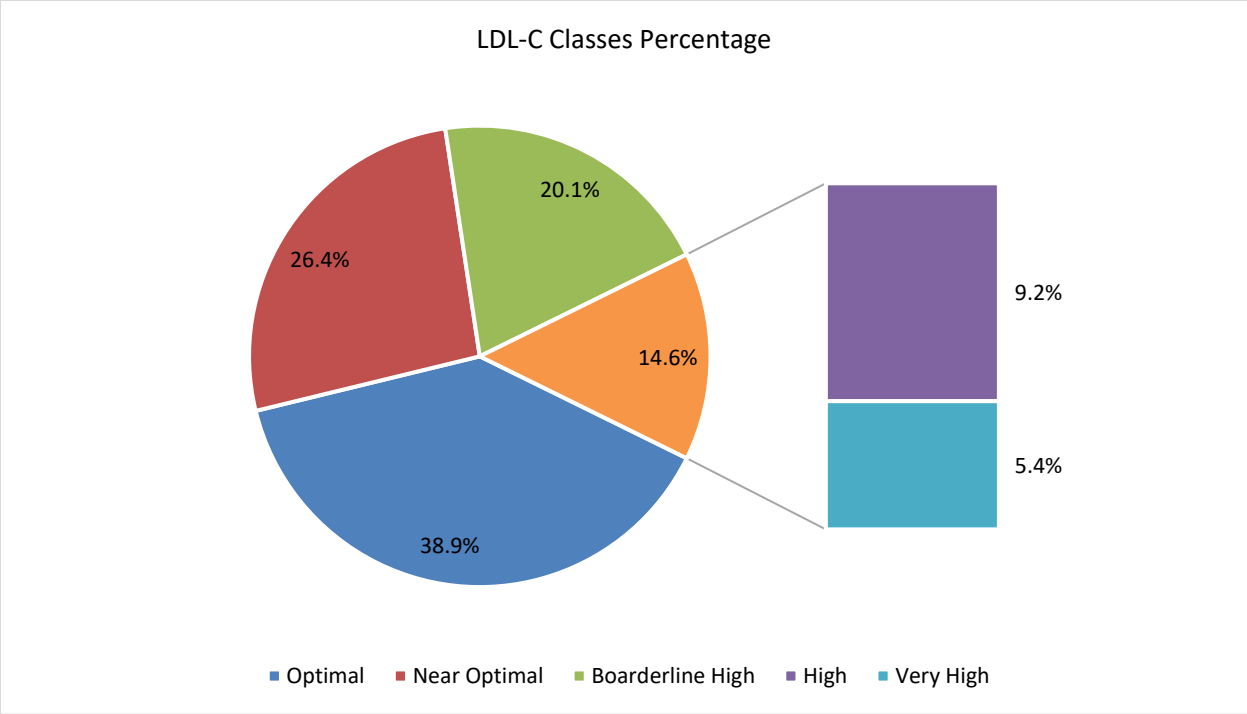


Figure 2.2.1: LDL-C diagnosis classes distribution among the Palestinian dataset records

2.2.2 International Data

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 selected based on the data availability, as the records with missing values 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, because the original data files had too much invalid content missing values, corrupted data, then once

excluded the records with the invalid data features the sample reduces. So used the most 10 related fields per the reviewed works, when tried to include more fields then the filtration process caused to decrease the sample size less than the 910 available records, which would not satisfy the experiments and real evaluation for the accuracy.

Table 2.3: The Korean dataset selected features and their ranges.

Feature Name	Value Range	Average	Normal Ranges
LDL	20 - 252	115	< 160 mg/dL
HDL	8 - 84	42	40 - 60 mg/dL
Total Cholesterol TCH	92 - 372	208	< 240 mg/dL
Triglyceride TG	200 - 2115	310	< 200 mg/dL
Age	10 - 80	51	--
Gender	Male - Female	--	--
BMI	16.72 - 42.88	26	< 30
WHtR	0.37 - 0.75	0.5	0.4 - 0.5
SBP	84 - 204	124	90 - 120 mmHg
DBP	49 - 135	80	60 - 80 mmHg

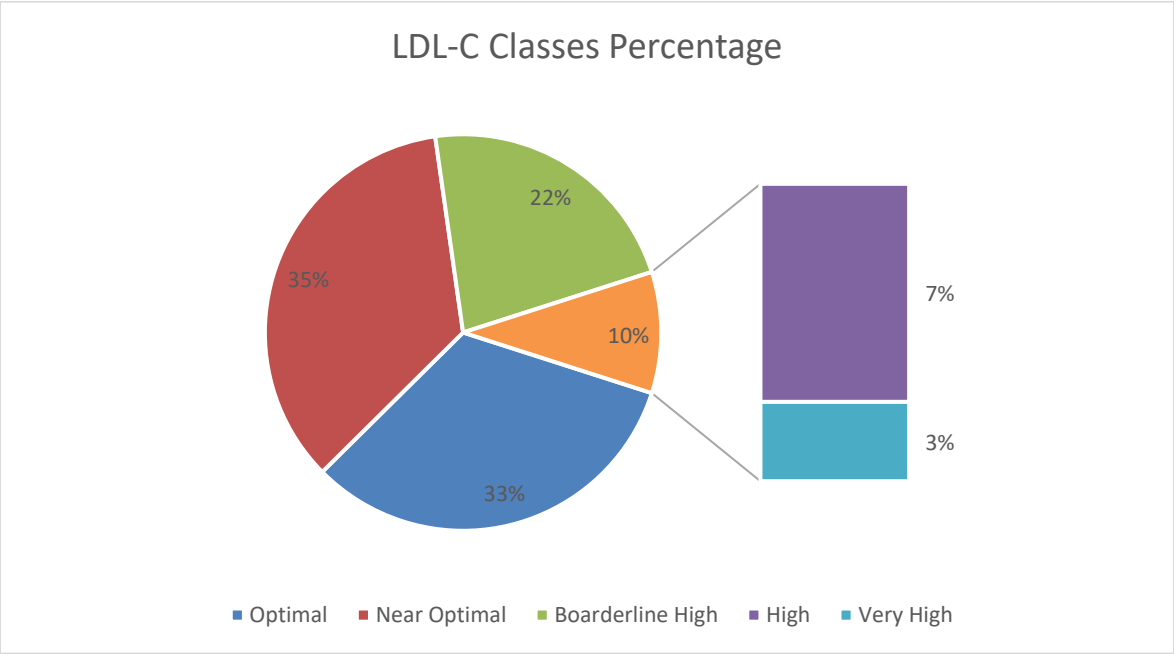


Figure 2.2.2: LDL-C classes diagnosis distribution among the Korean dataset records.

After extracting the data and analyzing the statistics for the LDL-C in the Korean dataset, figure 2.2.2 illustrated the LDL-C diagnosis classes distribution among the dataset records, as introduced, the main Cholesterol level that affects diagnosing the disease is the LDL-C value, hence need to have an initial impact about its diagnosis distribution among the data.

The Features Definition:

1. BMI: the Body Mass Index, and usually calculated using the equation 2.2.1, represents a method to describe the patient's general obesity, as it is directly linked as a risk factor for Cholesterol disease, where its ranges are explained in Table 2.4 [20].

Table 2.4: Body Mass Index (BMI) categories ranges.

Category	Range
Non-Obesity	BMI < 30
Obesity	BMI ≥ 30

$$BMI = \frac{Weight}{Height^2} \qquad [21] \ (2.2.1)$$

Table 2.5: Waist-To-Height ratio ranges in % percentage.

Category	Male	Female
Extremely Slim	≤ 34	≤ 34
Slim	35 - 42	35 - 41
Healthy	43 - 52	42 - 48
Overweight	53 - 57	49 - 53
Very Overweight	58 - 62	54 - 57
Obesity	≥ 63	≥ 58

2. WHtR: the waist to height ratio, it's an index to measure abdominal obesity by dividing the waist circumference by the height, and the same BMI it's a major risk factor for Cholesterol disease. Its categories explained in Table 2.5 [21] [22].
3. SBP and DBP: Systolic Blood Pressure SBP, and Diastolic Blood Pressure DBP, respectively. The unit of measurement is the millimeter-high mmHg. They measure the pressure caused by the heart pumping work for the blood through arteries to the body forth and back. The ranges are classified in Table 2.6 [22].

Table 2.6: The blood pressure classification ranges per SBP/DBP, their measurement unit is mmHg.

Category	**SBP**	**DBP**
Hypotension	< 90	< 60
Normal	90 - 119	60 - 79
Prehypertension	120 - 129	60 - 79
Hypertension Stage 1	130 - 139	80 - 89
Hypertension Stage 2	> 140	> 90
Hypertensive Crisis	≥ 180	≥ 120
Isolated Systolic	≥ 140	< 90
Isolated Diastolic	< 140	≥ 90

4. The lipid profile values: LDL, HDL, TCH, and TG are already introduced at the beginning of section 2.2 with details and ranges.
5. Gender: the gender of the patient if male or female, because the diagnosis for the females is different than males, and thought needs pattern recognition ability to distinguish between them.
6. Age: the patient's age affects because the child's circumstances are different than the adult than the old people, each age has its condition.

2.3 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 thesis for the Palestinian Patients Cholesterol. Here would have a look at the closest and related efforts to the thesis goal, even it's very few efforts being done on the same idea. Milan et al. [4] have worked on predicting the LDL and other lipid profile levels using ANNs, one-layered feed-forward ANNs with a backpropagation learning algorithm, and a varied number of hidden neurons started by 3 and change repeatedly with the multiplication of the 3 until used 30 neurons, each test with 100 repetitions. Their goal is to predict, mainly based on age, gender, blood pressure, and obesity indicators such as BMI. So, all the parameters are noninvasive which need no blood or any clinical laboratory tests. this thesis 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. Such critical and sensitive diseases would need very accurate evaluations because significant decisions have to be made based on the prediction results, the decision would guide the treatment actions including changing the lifestyle, pharmaceutical and clinical medications, or even a surgical one [2].

The efforts done in [23] went toward comparing the correlation coefficient of the traditional equations methods in estimating the LDL-C with a novel ML model. Both views used the lipid profile test measurements as inputs for predicting the LDL-C value. The data sample size 17,500 profiles for 10,936 unique patients collected from New York-Presbyterian Hospital/Weill Cornell Medicine (NYP-WCM), they aged from 1 year to 103 years, with around 40% female's percent.

The used ML model is the Random Forests and specifically the Weill Cornell model. The ML model results outperformed the used traditional equations of Friedewald and Martin-Hopkins, the correlation coefficient between the estimated and the measured LDL-C using the ML Weill Cornell model is 0.982, for the of Friedewald and Martin-Hopkins are 0.950 and 0.962, respectively. The ML model outperformed because has the ability of learning based on the data trends, while the traditional equations has limitation because they depend on the lipid profile constant values, that would lead to mis-estimation specially the cases when high triglycerides and very low LDL-C as proved in the reviewed work.

Similar to the above reviewed work, in [24] the researchers compared performance of five ML models to the of Friedewald and Martin-Hopkins traditional equations in predicting the LDL-C. The inputs to the utilized methods and models are the laboratory lipid profile test levels, which are the Total Cholesterol, Triglycerides, and High-Density Lipoprotein Cholesterol HDL-C. The collected data sample size is 7397 directly measured LDL-C. So, used 80% of the data to train the used ML models which are Linear Regression, Random Forests, Gradient Boosting Machine, Support Vector Machine SVM, and Artificial Neural Networks ANNs. The tested them using the rest 20% of the collected data. Using the Mean Square Error MSE for the performance comparison, so the SVM achieved the lowest MSE in comparison to the traditional equations of Friedewald and Martin-Hopkins.

Additional efforts were done in [8] 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. They used data collected from a Thai hospital called "Srinagarind Hospital", their experiment was done with 10,949 participants, collected their lipid profile clinically, and applied the mentioned techniques to measure how accurate their prediction

abilities to the goal, the study had three inputs which are HDL-C, TCH, and TG; the output is the LDL-C value. The used ANNs structure is a single hidden layer, that is validated by a ten-fold cross. Their study analyzed the correlation between the fields of the lipid profile with the LDL-C which showed a high correlation of average equals 95%, which proves and supports the continued efforts toward the tries of accurate prediction for the LDL-C.

More efforts are done in [10] 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. They constructed DNN with six hidden layers and 30 neurons for each layer, the activation function of the hidden neurons is the "g" function as explained in equation 2.3.1.

$$z = bias + \sum_{k=0}^{n}(X_i W_i), \text{and } g(z) = \max(0, z) \qquad (2.3.1)$$

Where, X_i: the neuron i input, W_i: the neuron i weight to the next layer, z: the neuron response, and $g(z)$: the neurons activation function to be calculated based on the neuron response value z.

Also, they used the five-fold cross-validation. The datasets, relied on two sources, one is the same dataset used in this thesis which is the KNHANES, and the other one collected from Wonju Severance Christian Hospital, with approximately 20,000 samples collected and included in the experiment, all the samples contained the lipid profile measurements. 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 for the constructed model in comparison with the traditional equations such as Friedewald and Novel, the DNN won with the lowest MSE around 6%, and the highest accuracy among their studied options.

The back-propagation technique of the ANNs (BP-ANN) is used in [13] for accurate prediction of LDL-C, more specifically in overweight people. 757 subjects covered in their experiment and divided into 302 normal weights, and 273 overweighted. In their research, they included the liver function and kidney function, in addition to the TG and TCH as inputs for the analysis of correlation with the Cholesterol level. The data samples were collected from the First Affiliated Hospital in the Wenzhou Medical University, China. The results showed a correlation mapping between the inputs and output around 94% with an ending condition of 1000 epochs. It worth telling, the full inputs they included are weight, height, age, BMI, liver function, kidney function, and fasting blood glucose, so they got started to add more inputs than just the lipid profile and obesity. The actual goal of the researchers to increase the accuracy by the lowest cost, though including the kidney and liver test in the Cholesterol prediction means more costive and complicated labs, which go a little bit far from the original goal of simplicity and accuracy in the prediction.

The Genetic Algorithm GA used as an optimization technique for the BP-ANN inputs and structure construction in [25] by efforts from Tone et al., a single hidden layer is used with 50 neurons, and the inputs are 12 of the blood plasma contents, including the main interested value which is the LDL-C. the inputs are VLDL, IDL, LDL, and HDL, these values with all corresponding versions of the cholesterol, triglyceride, apolipoprotein A1 ApoA1, and apolipoprotein B ApoB. The dataset was collected in a study approved by the ethics committee, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway. 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.

Another view for the Cholesterol abnormalities detections has been tried in [26], it utilized another type of inputs which is the medical imaging for the Cholesterol i.e. X-Ray Computed Tomography CT, Ultrasound Imaging US, and Magnetic Resonance Imaging MRI. In their works, the first level is the wavelet filtration for the data collected from the Cholesterol MRI and used the ANN technique from the machine learning side. Their efforts focused more on the detection of the Cholesterol distribution in the human body, which reached the accuracy of 98.8% per their experiments on MATLAB. The used tool is the multilayer perceptron with back-propagation BP-ANN. The wavelet images are transformed by extracting the pixels and storing them as strings, and this is done by phases of image pre-processing, image segmentation, wavelet processing and so to apply the wavelet filtration then to be the ANN inputs for the classification. The proposed method of combining the wavelet filtration and BP-ANN outperformed the traditional ways for image processing as input for the ANNs.

The Cholesterol excess amounts in the body cause major harm to the human body system, so the tries to predict it continues from many viewpoints, and not the last idea was to work on the iridology of eye iris imaging per the works done by Louis et al. [27], the efforts put on to detect by the eye iris as easy and fast way, but the accuracy was not that much high and reached with them to 86%, they had lacked in the data as just collected a sample of size 35 objects, which is not enough to have a clear decision about the accuracy. They used the Gabor filter technique for the image recognition, so to extract the features and compare with database images defined as classification standards, and make graph matching with them then decide the results.

More ideas went to use the genetic expressions for predicting Cholesterol levels, similar to the works done in [28], where they used meta-dimensional data of gene Expression Variables EVs and Single Nucleotide Polymorphisms SNPs to predict the HDL-C. They have designed and used the

Analysis Tool for Heritable and Environmental Network Associations ATHENA in a biology view system, they integrated the massive dimensional genome data with the analysis and prediction tools to predict the HDL-C levels. Their tool has filtration and modeling modules, the module is called Grammatical Evolution Neural Networks (GENN) that combine the genetic programming GA to optimize the ANNs for achieving the highest outcome accuracy. Their module accuracy reached have r-squared of 16% that outperformed the other analysis techniques such as the linear regression with an r-squared 23%. Their tool ATHENA has more options of modules such as Grammatical Evolution Symbolic Regression GESR to try but their study focused on the results of GENN.

A lot of works and ideas have shown the light for the sake of high performance of Cholesterol predictions, the ideas moved between using non-clinical and clinical risk factors evaluation considerations, also used the laboratory and noninvasive considerations, then moved to reach more areas that used the medical imaging, iridology science and finally the genome world.

In this thesis, 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. This thesis works proved very high accuracy reached 97.10% in the LDL-C and 97.80% in the HDL-C, that was using the BP-ANN in the classification, which outperformed the works done in [4], their works are the closest and most related to the this thesis works, so would compare with their results in the accuracy specifically, they reached the accuracy of 80% for LDL-C and 81% for HDL-C, their low accuracy because they included only the lipid profile inputs in their experiments and that led the low accuracy in their works. The local data also showed very acceptable prediction performance with MSE equals to 0.0003, and 0.0001 in the Korean dataset, that was using the Neuro-Fuzzy model which outperformed the other prediction techniques mentioned for the prediction. By all means, it worth telling, the utilization of the ML techniques on the Palestinian dataset for recognizing and diagnosing the LDL-C is done for the first time on this thesis, which makes it novel work. It is an honor for this thesis that its main goal to support the health sector by utilizing the ML techniques, which is a leading idea. The thesis detailed results will be introduced and discussed in the next sections related to the analysis and discussion of the outcomes in Chapter 4.

Chapter 3

Methodology

Chapter 3

3.1The Proposed Method

This chapter will describe the ML techniques and modules that would utilize, combine and propose to support the thesis goals, which is to provide accurate and precise classification and prediction abilities for the Cholesterol levels of LDL-C and HDL-C, in the Palestinian dataset specially and Korean dataset generally. So, the methods would apply to the datasets to evaluate the local results based on the international dataset research results. The Palestinian dataset is the first time be collected for Cholesterol experiments, the Korean dataset collected from South Korea and supplied for research purposes by Korean National Health and Nutritional Examination Survey KNHANES, it is from the rare international resources that can be utilized to support researches of applying ML in the prediction and classification of the Cholesterol disease.

For the classification, purposes will use Logistic Regressions LR, Support Vector Machines SVM, Decision Trees (Fine Trees) DT, and the Two-Layered Feed-Forward Back-Propagation Artificial Neural Networks BP-ANN, also built a Fuzzy Inference System FIS for the classification, specially that the Cholesterol levels has uncertain ranges of measurements. In the prediction, would again use the BP-ANN but with non-linear output, Recurrent Neural Networks RNN, Radial Basis Function Neural Networks RBFNN, then combined the ANNs with Fuzzy Log in the Neuro-Fuzzy model which to enhance on the performance accuracy and error, the Neuro-Fuzzy prediction outputs could supply into Fuzzy System to decide the ranges of the Cholesterol, the Fuzzy Inference System FIS would be built base on rules and feedback collected from expert physicians in the Cholesterol disease, the Fuzzy Logic used for the diagnosis as usually no sharp values to describe the harm level of the disease, but depends on the case of the patient diagnosis by the expert experience, which produce the case of uncertainty that require the fuzzy classification. At

the end of the chapter will describe the results evaluation techniques and measurement metrics that would be followed in the results of the work in the discussion part.

3.2Preprocessing Phase

The data sources for the thesis experimental part are from national Palestinian hospitals and from the international source the KNHANES, the data targeted and collected after reviewing the field works of Cholesterol estimation techniques acted from the ML area. The works majorly relied on data and features described in two main references [4] [10]. Even though, the clear related features mentioned in Tables 2.2 and 2.3 per the reviewed related works and references, but the available data especially the Palestinian one didn't cover all the required features, so no hard works in the feature selection needed as there is a shortage, the thesis works much tried to extract and collect more risk factors as possible form the datasets as to optimize on the results. The original files of the data were not that well-formatted to meet the experimental works on the MATLAB, so works for the data transformation, filtration, cleaning, extraction, standardization, and finally normalization to meet the goals of the data processing and experimentation in the used tools.

3.2.1 Feature Selection

Both local and international data files had extra fields, invalid records, non-formatted well to meet the requirements, so the first steps were to format them well, the Palestinian data records collected from four local hospitals and were divided into 12 excel files with one risk factor in each line so the patient has multiple distributed lines, also not all the required inputs from the data are available and correct, for that used the python language to transform the data into formatted excel files, then filtered the invalid data and fields to collect the required inputs per the main references [4] [10] that relied on them in collecting the required data and risk factors. After the preprocessing 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 using the equation 3.2.1, 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.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (3.2.1)$$

Where, r: correlation coefficient, x_i: inputs values, representing the risk factors, $\bar{x}$: the average of the input, y_i: the output value, representing the targets which are LDL-C and HDL-C and $\bar{y}$: the average of the output.

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, and the considered risk factors after extraction are 10 fields out of 785 which are the lipid profile levels, age, gender, BMI, WHtR, SBP, and DBP. All the extracted final data records were explained and listed in section 2.2, including the fields lists and definitions.

3.2.2 Data Normalization

The classification experiments used the extracted data in its original values and scales as the goal to define between classes i.e. 0 or 1, that reflect if the patient is diagnosed with the Cholesterol disease or not, so majorly used these two classes for the classification target. The classification didn't include the normalization because the experiments showed no worthy difference in the results with the normalized data. But in the prediction methods the non-normalized data showed overfitting values which made it non-acceptable results. The prediction output would approximate

as much as possible to the actual value, so the inputs must be scaled to a balanced level to each other, because if they are not scaled well then the features and inputs would have a different effect than it's real weight since they could be distributed across many levels of numerical values, which cause the higher to dominate if didn't normalize [29]. The below table 3.1 shows a sample of the Palestinian data before and after normalization.

Table 3.1: Sample of the Palestinian before and after the Min-Max normalization

Gender	Age	Total Cholesterol	HDL-C	Triglycerides	LDL-C
Original Data					
Male	30.45	157.4	31.8	293	101.8
Male	58.77	305.8	21.1	771.2	60.6
Male	46.94	176.1	43.9	115.5	114.9
Male	35.98	185.2	43.1	99.9	145.8
Male	13.98	130.7	61	132.3	58.7
Female	66.24	188.9	41.5	395.5	107.9
Male	71.31	141.5	40.2	164.4	84.6
Female	41.70	127.5	35.5	158.6	77.7
Female	9.48	187	55.3	106	139
Normalized Data					
1	0.288472068	0.138289547	0.062626263	0.048231619	0.137517241
1	0.561204498	0.304192286	0.035606061	0.128122024	0.080689655
1	0.447293323	0.159195081	0.093181818	0.01857761	0.155586207
1	0.341720331	0.169368362	0.091161616	0.015971398	0.198206897
1	0.129861905	0.10844047	0.136363636	0.021384299	0.078068966
0	0.633160897	0.173504751	0.087121212	0.065355765	0.145931034
1	0.682002589	0.120514254	0.083838384	0.026747081	0.113793103
0	0.396842046	0.104863052	0.071969697	0.025778104	0.104275862
0	0.08648225	0.17138066	0.121969697	0.016990494	0.188827586

There are many normalization techniques for the numerical data that scale the values to be in the interval [0, 1], so all inputs has the same scale, and the one utilized in this thesis works is the Min-Max method [29] per explained in equation 3.2.2, that is applied on all the inputs-outputs data of both datasets before importing them in the prediction techniques.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (3.2.2)$$

Where, y: the normalized value of the input x, x: the input x -numerical value, x_{min}: the minimum value of the input x, and x_{max}: the maximum value of the input x.

Any input that has no numerical value then needs to be translated into a corresponding numerical value for the experimental purposes, to have unified inputs for the machine learning techniques, so for example the gender value will be reflected as "1" for males and "0" for females, which needs no normalization since it's already in the scaling range of the values.

3.3Building Models Phase

The used ML techniques and hybrid models in this thesis will be introduced in this section with all implementation details. The models and methods are applied on the local Palestinian dataset which is named PDS, and the international South Korean dataset which is named KNHANES. There are two purposes of the techniques, one for classification, another for prediction of the Cholesterol levels of both LDL-C and HDL-C. The classification will utilize the BP-ANNs with linear output, Logistic Regressions LR, Support Vector Machine SVM, Decision Trees (Fine Trees) DT. Also, the Fuzzy Inference System FIS using Fuzzy Logic with rules collected from experts in Cholesterol that is for the purpose of ranges classification. On the other side, the prediction will use the BP-ANNs with non-linear output, Recurrent ANN called RNN, Radial Basis Function RBF of ANNs which is commonly known as RBFNN, then the hybrid model of Neuro-Fuzzy that is called Adaptive Network-Based Fuzzy Inference System ANFIS, and finally the output will go through Fuzzy Logic FIS to classify the Cholesterol in the shallow levels thresholds.

Figure 3.3.1 illustrates the general procedure flowchart for building the models. After deciding the idea in cooperation and guidance from the supervisors, and reviewing the idea of applying ML

techniques in the recognition and diagnosis of the disease, the journey started by looking for collecting the required data to utilize in the selected ML models, then processing the data, after that start applying and trying the different techniques for getting the best results. So here will summarize the flow in the following points for better sequencing and review:

- Data collection: it was on two tracks, the Palestinian data because of the study about the Cholesterol in Palestinian patients, and another dataset for generalization and comparisons, both tracks had hard times until reached fulfillment, because it's very few the available datasets that support the Cholesterol studies in the field of ML, also the Palestinian dataset was collected in the times of Covid-19 pandemic which slowed the process.
- Data preprocessing: it passed through many stages, started by the data extraction from the original files, then transformation to a format suitable for the experimental environment, which is the MATLAB, after that the cleaning process for excluding any invalid data, and finally the scaling stage for normalizing the data to be suitable for the measurement as one level of precision.
- Experimental tests: once the data prepared well then selected the models of Machine Learning that would be applied and suited for the thesis goals, which is done based on the related reviews and studies. So, for the prediction, the chosen techniques are the Neuro-Fuzzy, BP-ANN, RNN, RBFNN, and for classification, purposes selected the BP-ANN, SVM, LR, DT, and FIS of the Fuzzy Logic.

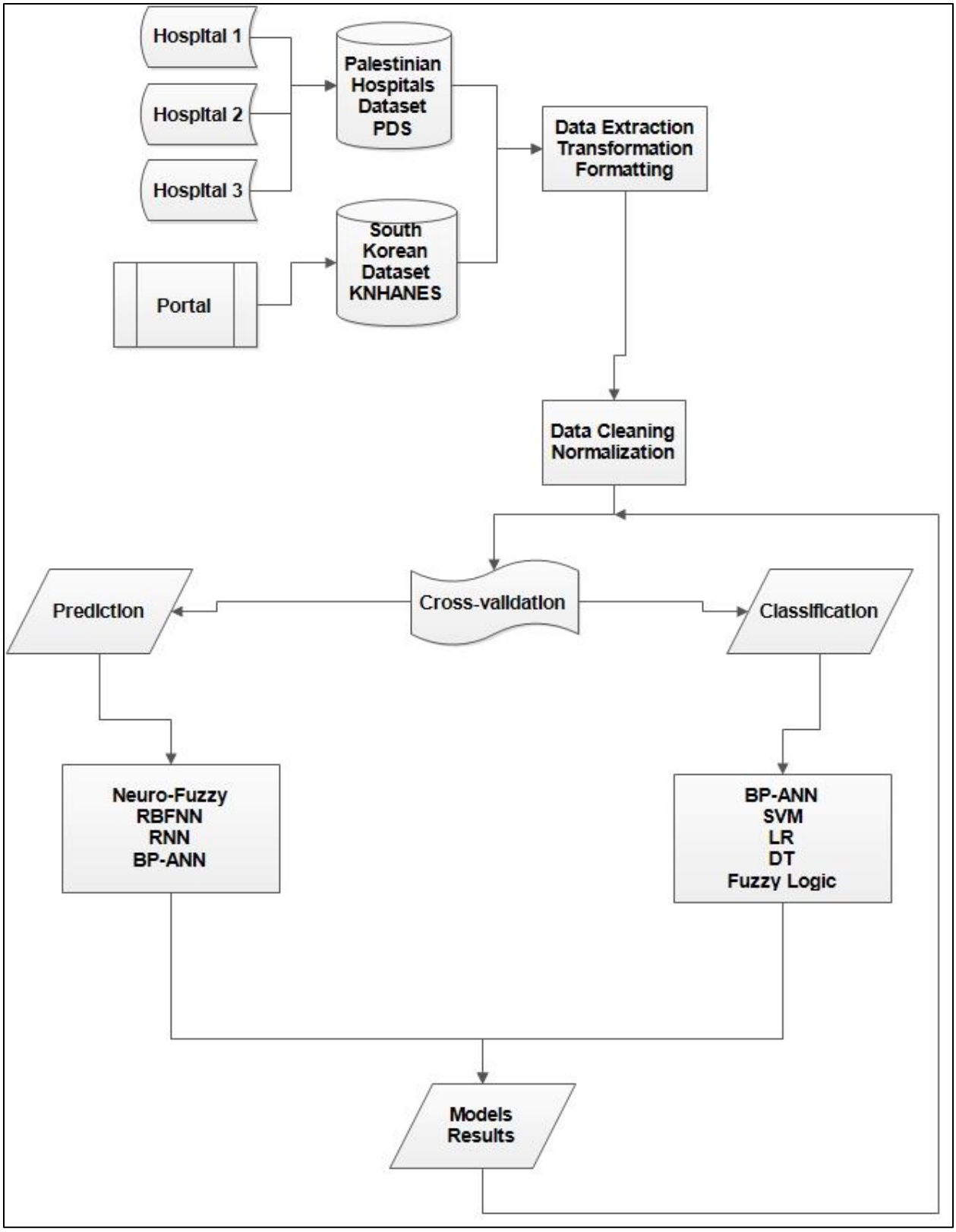


Figure 3.3.1: the general method procedures flowchart

- Results collection: the results were the judge for keep trying or ending the test, that is based on the different possible optimization, enhancements, accuracy results, and error percentage.
- Analyzing the results: once all the tests are finalized with the best results, then made the comparison and reviews for deciding the best techniques achievements reached in the experiments, and supplied the recommendations for the future works.

3.3.1 Multi-Layer Feed-Forward Back-Propagation ANNs BP-ANNs

The Artificial Neural Networks ANNs algorithm is from the most common techniques used in the machine learning field, where it emulates the brain biological work of exactly the learning feature, it is a successful trial of describing the way of human learning, as it is a powerful tool in the mapping between inputs-outputs of complex and non-linear data relations, similar to how the brain makes recognize the complex patterns and create weighted relations between the inputs and outputs, similar to many needs exist in the human life such as objects classification and values predictions [30].

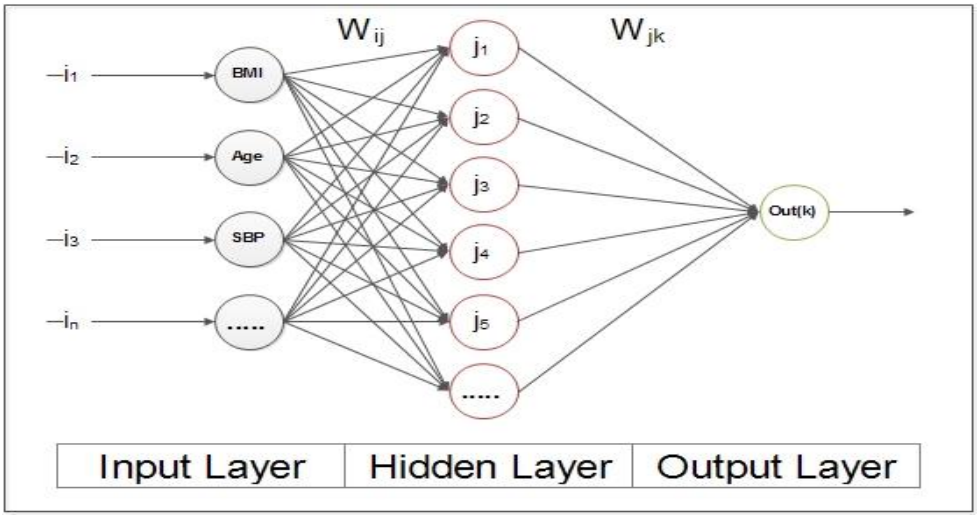


Figure 3.3.2: Two-Layer Feed-Forward Back-Propagation ANNs (BP-ANNs) – Structure

So, as the brain consists of networked layers of biological neurons that have input signals from other connected neurons, processing phase, and output signals to other connected neurons, and there are features or parameter inputs that have a higher impact on the decisions than others. The BP-ANNs is an artificial multilayer perceptron with a structure consists of the input layer, hidden layers, and output layer per illustrated in figure 3.3.2, the hidden layers could be one or more per the problem complexity [31].

As shown in the above figure, the risk factors that are considered for the Cholesterol prediction and classification are the inputs layer, the used scheme BP-ANNs has only one hidden layer with activation function for each neuron, the used output neuron is Softmax function for the classification problem and Sigmoidal function for the prediction problem. These functions selected after many tries to varied number of the available function in the MATLAB, and the results showed similar performance for the most of the function but these functions had the highest.

In the light of figure 3.2.2, the ANNs with feed-forward and back-propagation steps can be mathematically explained and described as per the following sequenced points with the mathematical correspondence [31] [32]:

- Input layer represent by the input value X_i, i counts from 1 till number of inputs n
- Hidden layer: each X_i is connected to the hidden layer neuron j with weight W_{ij}
- Output layer: represented by the neurons that calculate the network output Y_k
- the feed-forward stage to find the output is calculated per equation 3.3.1, the used function for the prediction is the Hyperbolic tangent sigmoid function per equation 3.3.2, and Softmax for the classification output as in the equation 3.3.3

$$Output = f(\sum_{i=1}^{n} X_i.W_{ij} + \theta_j) \quad (3.3.1)$$

Where: X_i: the *i*th input, W_{ij}: the *i*th corresponding weight between the *i*th input and the *j*th neuron, θ_i : the neuron bias value and takes the content value 1, f: the activation function that could be linear or non-linear per the case complexity.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.3.2)$$

Where: x: the input and e: Euler's number which is a mathematical constant equal 2.71828.

$$f(x) = \frac{e^x}{\sum_{i=1}^{n} e^{xi}} \quad (3.3.3)$$

Where: x: the input and e: Euler's number which is a mathematical constant equal 2.71828.

- The back-propagation stage used to update the network weights based on the error difference between the actual and calculated outputs, it usually uses the gradient decent algorithm and relies on the Sigmoid activation function per equation 3.3.4

$$Y_k = f(X_j) = \frac{1}{1+e^{-X_j}} \quad (3.3.4)$$

Where, Y_k: the network output, X_j: the input to the next layer, e: Euler's number which is a mathematical constant equal 2.71828

- The calculated output is Y_k, and considering the actual target Y_t value, then the error difference calculated as the equation 3.3.5

$$\Delta e = Y_t - Y_k \quad (3.3.5)$$

- The output layer error to be considered for the propagation can be found per equation 3.3.6, the is called the error signal δ_k:

$$\delta_k = \Delta e * Y_k * \ (1 - Y_k) \qquad (3.3.6)$$

- The next step to finding the weight needed to be changed based on the calculated error signal δ_k:, so need to find it using the delta rule per equation 3.3.7

$$\Delta w_{jk} = \ l * \ \delta_k * X_k \qquad (3.3.7)$$

Where, l: the positive real number represents the learning rate and could be tuned to reach the best results

- Since The inputs can have different weights based on their importance to determine output, this technique can learn how to find the best weights for reaching the most accurate output, so the initial weights usually selected randomly which mostly not leading to reach the best accurate results of the goal output, so the back-propagation phase is needed and would rely on the calculated output error to update back the weights and fine-tune them into the layers back using the equation 3.3.8 until reaching the goal accuracy

$$w_{jk\ new} = w_{jk} + \ \Delta w_{jk} \qquad (3.3.8)$$

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.

The used training function in the prediction is the Levenberg–Marquardt and it calculates the delta in the weight using the equation 3.3.9 as explained below.

$$\Delta w = (J^T J + \mu I)^{-1} J^T e \qquad (3.3.9)$$

Where: w: the weighs vector, I: the identity matrix, μ: combination coefficient, J: Jacobian matrix, e: error vector.

Generally, the output error is important in the training algorithm and learning phase as it will be used to balance the weights until reach the optimal or goal results, form the most common equations and that will be used in this thesis for calculating the output error is the Mean Square Error MSE, as per the below equation.

$$MSE = \frac{1}{2}\left(\sum_{i=1}^{n} Output_d - Output_i\right) \quad (3.3.10)$$

Where: $Output_d$: the actual output, and $Output_i$: the calculated output.

The training epochs will continue until reach a specific number of epochs or a minimal target error value, then test the resulted model on new testing data. The general algorithm pseudocode is illustrated in the below Algorithm 1.

Algorithm 1: BP-ANNs algorithmic steps

Input: Dataset, Number of Neurons, Learning Rate

Output: Trained model (Levenberg–Marquardt)

Main:

```
For row in Dataset:
    Y ← SumofMultiply (field.inputs, field.weights, Number of Neurons)
    Predicted ← Levenberg–Marquardt (Y)
    Error ← Error (Predicted, Target)
    Signal_Error δ ← Signal (Error, Learning Rate)
    Weights_Update ← update (δ, field.weights)
    If Error <= Threshold
        Stop
    Else
```

Continue

Return Model (Weights_Update)

3.3.2 Logistic Regression LR

It's one of the machine learning techniques that is used in the classification problems, especially the binary ones, like when classifying into two classes such as in the thesis 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 equation 3.3.11 [33].

$$f(x) = \frac{1}{1+e^{-x}} \qquad (3.3.11)$$

Where: x: the input to be classified that is defined per the below equation 3.3.12, and e: Euler's number which is a mathematical constant equal 2.71828.

$$X = a + bx \qquad (3.3.12)$$

Where: a: is a parameter that its value yields the X when x is zero, and b: adjust how $f(x)$ is quick changed when X is changed, this value considered as the weight.

The used loss function is the least square to measure the fit between the actual and the models results, where used the equation 3.3.13 for minimizing the error using the weight b [33].

$$b = (X'X)^{-1}X'y \qquad (3.3.13)$$

Where X: is already defined in the above equation, and y: is the output that will be used to re-update the weight b.

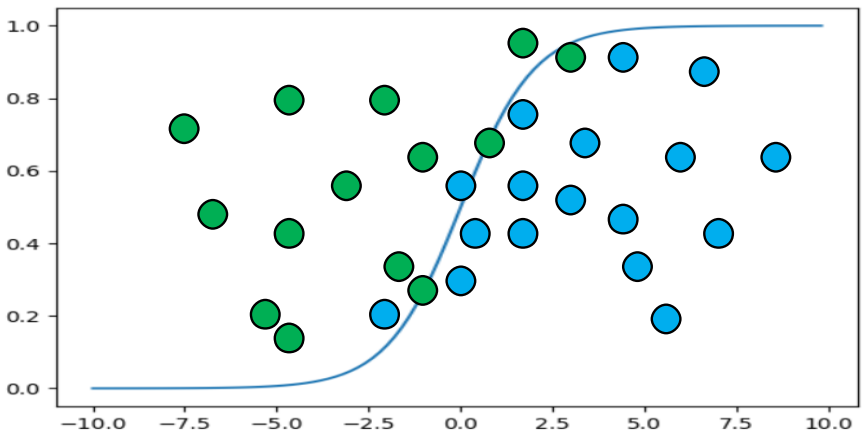


Figure 3.3.3: The S-Shape of the Logistic Regression – Sigmoid Function

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 as shown in figure 3.3.3. Another interval for the output can be used which is [-1, 1]. The above figure can see it reflects the S-Shape of the logistic function. The technique output has two classes, one for values closer to zero which can consider zero, and the other for values closer to 1, that can reflect in the following conditional equation number 3.3.14 [34]:

$$f(x) = \begin{cases} 0, & x < 0.5 \\ 1, & x \geq 0.5 \end{cases} \qquad (3.3.14)$$

The input data of the risk factors for classifying the cholesterol would be supplied paired with the output class of the LDL-C and HDL-C, so the class to be either 1 or 0, where 1 will mean the high and 0 the normal. The below Algorithm 2 illustrated the algorithmic steps for the Logistic Regression model.

Algorithm 2: Logistic Regression algorithmic steps

Input: Dataset

Output: Trained model

Main:

For row **in** Dataset:

// Weight: is the regression coefficient

Weight ← Calculate_Weight (row)

Cost ← Calculate_Cost (Gradient)

Weight ← Update_Weight (Weight, Cost)

Output ← Predict ()

Accuracy ← Calculate_Accuracy (Output, Targets)

Return Model (Weight)

3.3.3 Support Vector Machine SVM

The Support Vector Machine 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. It has been used in many applications such as image recognition, speech recognition, and face recognition, this thesis 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 [35]. As shown in figure 3.3.4, the general idea of the SVM 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 [36].

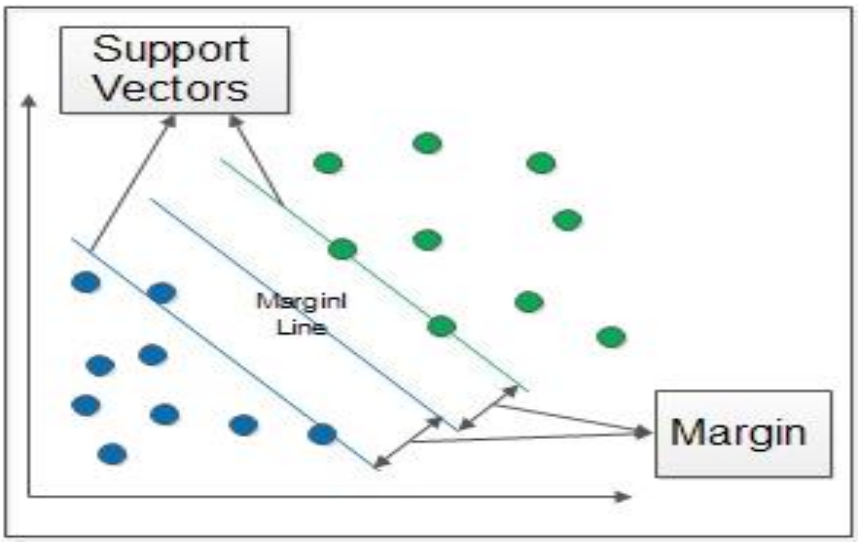


Figure 3.3.4: The SVM Hyperplanes, margins, and support vectors

As the target data is linearly separable so the marginal line of separation and the other two support vectors approximation with their margins, would rely on the mentioned below equations 3.3.15, 3.3.16, and 3.3.17, respectively [37].

$$w.x + b = 0 \qquad (3.3.15)$$

$$w.x + b = 1 \qquad (3.3.16)$$

$$w.x + b = -1 \qquad (3.3.17)$$

Where: w: the vertical vector on the marginal vector of the hyperplane separation, x: the input vectors, and b: a flexible scalar constant to maximize the margin as much as possible.

So, the distance |w| has to be divided into two equal parts per the cross with the hyperplane of separation, such that all the points belong to the area fail between the two support vectors with the interval $w.x + b \leq 1$ or $w.x + b \geq -1$ [37].

The SVM algorithmic steps can be found in Algorithm 3, which is illustrated below.

Algorithm 3: SVM algorithmic steps

Input: Dataset

Output: Trained model

Main:

```
Margine_Line ← Construct_Initial_Margin_Line ()
Vector_Point ← Find_Initial_Support_Vectors ()
For row in Dataset:
    Margine_Line ← Calculate_Distance (row, Vector_Point)
    Vector_Point ← Checck_If_Vector_Point(Y)
    Error ← Calculate_Error (Margine_Line, Vector_Point, row)
    If Error <= Threshold
        Stop
    Else
        Continue
Return Model (Margine_Line, Vector_Point)
```

3.3.4 Decision Trees

The Decision Tree DT is an effective classification technique with high performance and accuracy, where it became very common in fields related to marketing, fraud detection, image recognition, network attacks detection, and many clinical-related decisions such as the goal of the thesis to apply on the Cholesterol. While it requires simple structure and construction methods but has fast computation and powerful classification abilities. For the simplicity of DT and high abilities so it is very suitable in processing the big data in the data mining fields, though it would show worthy experiments in the PDS experiments as it has several records in thousands [38].

The main idea behind the 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 shown

in figure 3.3.5 below. 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 [39].

It recursively divides the data into two subsets and each sub-set split forward until reaching the end node, that is the leaf node which represents the category and class. It will be applied on all features to have weighting for the features across the path starting from the root ending with the class [40].

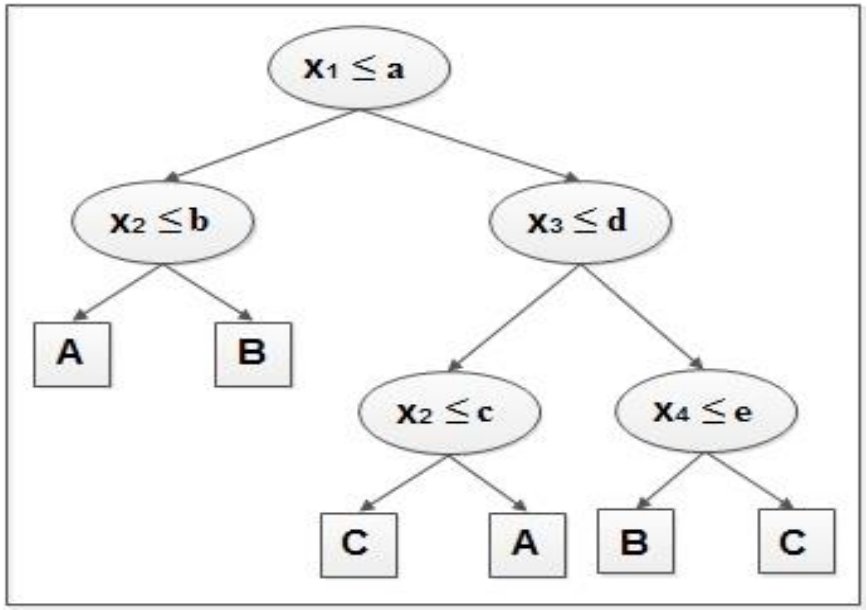


Figure 3.3.5: This Decision Tree has three classes A, B, and C, with x_i inputs, and five thresholds

The used splitting criterion is the CART (Classification and Regression Trees) algorithm that uses the Gini Diversity Index formula, it splits the data recursively into two equal subsets of data, that makes the simplicity in the trees because has easy decision to which direction go from the two possible sub-trees, as the targeted problem is binary classification [41].

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

3.3.5 Recurrent Neural Network RNN

On the opposite side of the Feed-Forward ANNs that transfer the activation outputs as inputs to the next layer, there are the Recurrent Neural Networks 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 the figure 3.3.6, 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 [42].

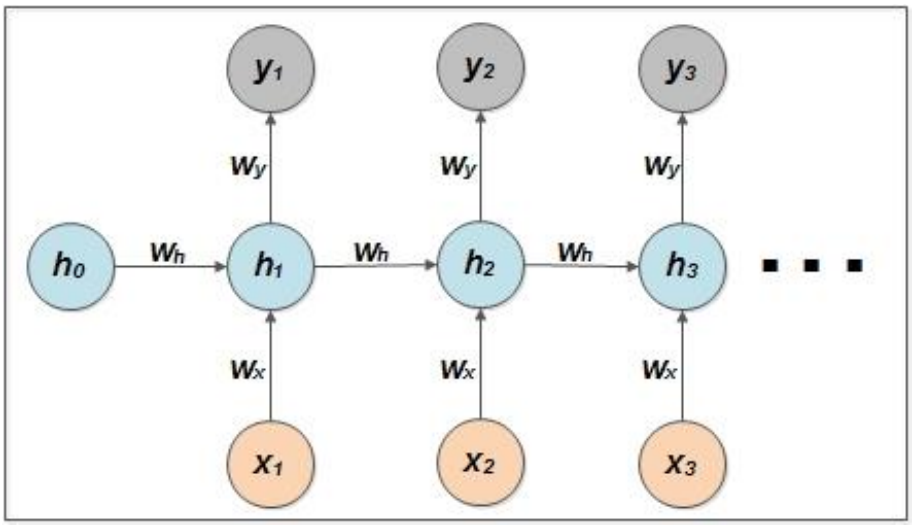


Figure 3.3.6: Recurrent Neural Network RNN Structure, the prior output is the current input

It has powerful abilities than the other versions of the ANNs in the prediction, especially the cases when the training data has a sequential effect or cumulative meaning, such as Natural Language Processing NLP such as word-level language modeling where it can suggest what is the expected next word in the sentence. That is in addition to many other applications, the case when the data follow ordinal sequencing by space or time similar to the mentioned fields of recognition and modeling [43]. To represent the transition of the previous hidden state as input to the next or current hidden state, will translate it to the equation 3.3.18, which use time sequencing to reflect how the prior output affects the new output as an addition to the current input [44], per show in the following.

$$h_t = \alpha\,(Wx_t + Uh_{t-1} + b) \qquad (3.3.18)$$

Where, x_t: the input at time t, h_t: the hidden state activation at time t, W: the weights for the current input, U: the weights for the recurrent input, b: the bias value of the neurons, and α: the activation function of the neurons.

Many activation functions are used for the hidden state and of the most common ones is the sigmoid as illustrated and explained in figure 3.3.3 and equation 3.3.8, respectively. A major weakness in this technique is the repeated multiplication, which causes exploding in the size of weights and inputs matrices multiply [44].

The algorithmic steps are very similar to the one mentioned for BP-ANNs in the Algorithm 1 illustration, but the idea here that the backpropagation step could happen at the level of one neuron, and that is why called recurrent because consists of cycles for at least one neuron.

3.3.6 Radial Basis Function Neural Network RBFNN

This kind of the Artificial Neural Networks ANNs that is using the Radial Basis Function RBF as its activation function, is similar to the Feed-Forward ANNs and has only a single hidden layer for the estimation and prediction goals, where it relies on the data inputs distribution along with clusters with centers, so each center will be represented by a neuron that uses the RBF activations for transferring to the output layer. As it works using clustering centers, then the hidden layer neurons usually fewer than the high dimensionality of the input data, and also there is the option of the exact number of neurons to the inputs, but that cause overfitting and computation complexity for no purpose which degrades the performance and accuracy, specifically when the data mapping has some significance in the error [45].

As in the below figure 3.3.7 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 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 [46].

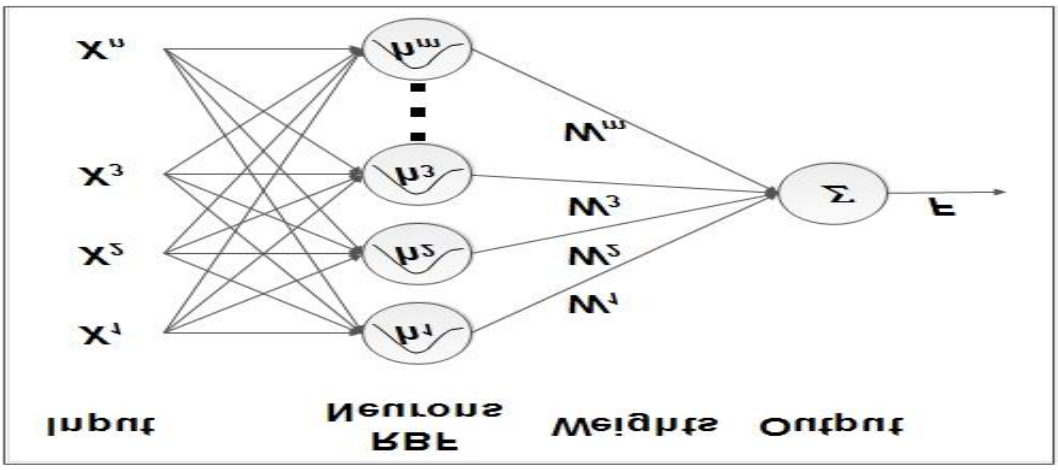


Figure 3.3.7: Radial Basis Function Neural Network RBFNN, structure, and flow

The commonly used hidden layer activation function is the Gaussian function as the initial data mapping is non-linear, it takes the shape as illustrated in figure 3.3.8, 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.

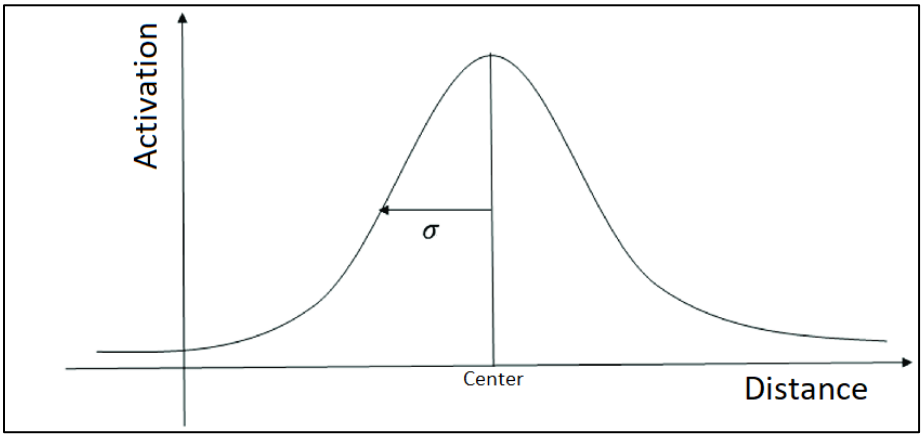


Figure 3.3.8: Gaussian Function Distribution, center and spread constantly

The above figure 3.3.8 reflects the Gaussian function distribution that is defined in equation 3.3.19, which shows how each parameter is usually calculated, the width of the neuron, the cluster

centering distance of maximum from center to the input, and the final activation transferring value itself using the exponential functioning [47].

$$h_j = \exp\left(\frac{v_j^2}{2\sigma_j^2}\right) \quad (3.3.19)$$

Where: h_j: the calculated Gaussian function value of the transfer, σ_j: the width of the neuron which is the function spread constant, that affects the clusters count, and v_j: is the maximum radial distance between the neuron center the data points of the input vector, this one usually calculated using the Euclidean distance equation 3.3.20 below [47].

$$v_j = \sqrt{\sum_{i=1}^{m}(x_i - c_{ij})} \quad (3.3.20)$$

Where, x_i: is the input values counted between 1-m, c_{ij}: the center of the *jth* RBF neuron It is clear that this model very useful in the cases where the data is corrupted and non-linear, as it utilizes the clustering to re-arrange the data and apply the activation as if it's linear problem, also need no much activation units especially when the data is high dimensional, because it re-distribute the high dimensional data into centers, and need only one stage of training as forward instead of the repetition as in the normal Multilayer Perceptron Network MLP models, with high performance and accuracy when the case of the neuron is fewer [45].

The algorithmic steps share the same part of Algorithm 1 in the BP-ANNs, more specifically the feed-forward part because it's one direction algorithm, and utilizes the RBF as its activation and transfer function, this function detail is already explained in the above discussions.

3.3.7 Fuzzy Inference System FIS

The disease diagnosis usually has kind of uncertainty, because the symptoms and clinical test have sliding scales of values that relied on the diagnostic decisions, similar to what happens in the

Cholesterol diagnosis, as the physicians and experts don't depend on the sharp or crisp value of LDL-C in the treatment prescription [48], and that is the case in what got with the discussion had with experts in Ibn Sina hospital, Jenin city, Palestine. As they don't use exact scales for the diagnosis because they would be treated proactively, so that is why they put ranges for the prescribing. And here comes the job of the Fuzzy Logic systems, where they process the scenarios when the input and output data is uncertain and ambiguous, as no exact degree of the illness or disease, but average or ranged descriptive diagnostic. Similar to what happens in the physician's mind when deciding the diagnosis of the disease that consider many factors to figure out the acuteness degree of the pain and disease effects. Though the expert experience will be transformed into a rule-based inference system of the form If-Rules, that is to rely on for the machine-based decisions of the diagnosis and treatment, in a similar way of human-like capabilities in the Artificial Intelligence AI [49].

For building the FIS a discussion session was done with the hospital experts in Cholesterol diagnosing in the view of the medication consideration wise. The laboratory test usually cover the whole lipid profile levels in the results for the patient who visit the physician to check the Cholesterol status, the lipid profile test has four values the TCH, TG, HDL-C, and LDL-C, but the physician majorly rely on the LDL-C only, especially if the patient has no other risk factors or diseases that could affect the Cholesterol, and that is the claim in this experiment to consider the patient is free from the other affecting diagnostics, so per their feedback would rely majorly on the LDL-C. The values of LDL-C ranges per the discussion with the experts are illustrated in table 3.2, with a descriptive translation to the fuzziness terms.

Table 3.2: The experts' feedback on the considered ranges in the LDL-C diagnosis, it's expressed in the measurement unit of mg/dL

Range	Diagnosis
< 130	Optimal
100-160	Borderline High
130-190	High
> 190	Risky

Based on the above values will build the Fuzzy Inference System FIS as one of the techniques studied in this thesis. The FIS usually consists of four phases which followed to reach the targeted output as in the following points [50]:

- Fuzzification: the fuzzy sets created as per the corresponding ranges in table 3.1, and illustrated in figure 3.3.9.
- Rules: collected from the experts in Ibn Sina hospital, local hospital, and utilized Mamdani FIS, because it's the most suitable for reflecting the experts' feedback into fuzzy sets, as this technique utilizes the sets to represent the inputs and output. So would use the AND operation to compose the collected rules if the inputs more than the LDL, below listed the initial collected rules that can be considered as a start for the new efforts:
    - If (LDL is Low) then (LDL-Cholesterol is Optimal)
    - (LDL is Moderate) then (LDL-Cholesterol is Borderline-High)
    - (LDL is High) then (LDL-Cholesterol is High)
    - (LDL is Very-High) then (LDL-Cholesterol is Risky)
- Membership Functions MF Aggregation: used the most common MF which is the triangular per illustrated in figure 3.3.9, then aggregated the inputs MFs to unify and combine their results into a single output fuzzy set.
- Defuzzification: where the crisp output value to be calculated, that find a pointed slice the aggregate output set into two equal areas, that to use for the final decision recommendation

corresponding to the inputs. The most common technique for the defuzzification to the crisp output value is the Centroid technique, the equation 3.3.21 explain how to calculate the Centroid of Area technique COA [50].

$$Crisp = \frac{\sum_{i=1}^{n} x_i * \mu(x_i)}{\sum_{i=1}^{n} \mu(x_i)} \qquad (3.3.21)$$

Where, x_i: the x-axis elements that represent the aggregate set value of the output, which in the thesis represent the output LDL-C crisp value, $\mu(x_i)$: the MF degree to the output set elements

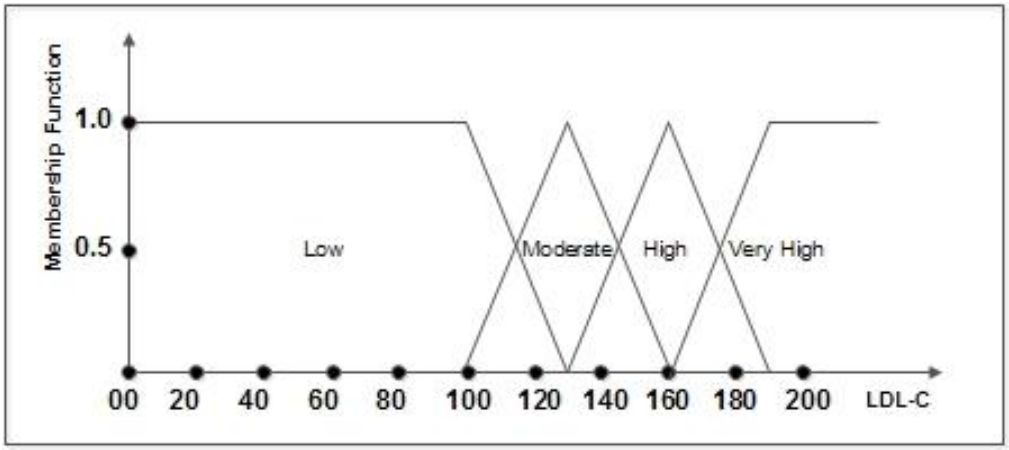


Figure 3.3.9: Fuzzy Logic, Triangular Membership Function of LDL-C

The experiment was done using the FIS toolbox in MATLAB, where the sets, membership function, rules, and defuzzification is done using a graphical user interface. The final results to be shown and analyzed in the next chapter, Chapter 4.

3.3.8 Neuro-Fuzzy Model: Adaptive-Network-Based Fuzzy Inference System ANFIS

The Adaptive-Network-Based Fuzzy Inference System ANFIS is a prediction model produced by Jang, around twenty years back, where it benefits from both Fuzzy logic and Artificial Neural Networks ANNs, the 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 [51]. This thesis utilized the input-output pairs collected from the 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 uses Sugeno FIS because represents the sets membership as a calculation summation of membership degrees but not using the defined fuzzy sets as in Mamdani, which suits the case how the ANNs generate the rules based on the pattern recognition in the provided training data. Its 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, it starts by defining the inputs and fuzzy set, passing through the membership degree, and end with the de-fuzzification. Each step represents one layer of the ANNs, the following explanation for each layer considering the model structure in the figures 3.3.10 & 3.3.11 [52]:

- The input layer (layer 1): represents the input variables to the ANNs, where each neuron at this layer transfer the input crisp value to the next layer, as defined in the below equation 3.3.22

$$O_i = \mu_{Ai}(X), for\ i = 1,2\ \ldots. \qquad (3.3.22)$$

Where, O_i: the output at the *ith* node neuron, X: the input at node neuron i, A: linguistic variable of fuzzy set.

- Fuzzification layer (layer 2): it 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. Equation 3.3.23 denotes the product operation of previous layer outputs that represent the input to this layer.

$$W_i = \mu_{Ai}(X) * \mu_{Bi}(X), for\ i = 1,2\ ....\quad (3.3.23)$$

Where, W_i: the outputs of the second layer nodes, X: the input at node neuron i, A, B: linguistic variable of fuzzy sets, could be more sets.

- Fuzzy rules (layer 3): it receives the membership degrees defined for the input in the fuzzification layer, and each neuron in this layer represents a fuzzy rule that is generated based on the fuzzy sets' membership of the input crisp values. So, each node at this layer represents the rule strength to the overall rule's strengths as per equation 3.3.24.

$$S_i = \frac{W_i}{\sum W_i}, for\ i = 1,2\ ....\quad (3.3.24)$$

Where, S_i: the outputs of the third layer nodes that represent the ratio strength of each rule to the overall rules' strength, W_i: the outputs of the second layer that represents the multiplication of the input's memberships.

- Output Membership (layer 4): this layer is the 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 probabilistic OR operation for all received inputs belonging to the same fuzzy set, as per illustrated in the equation 3.3.25

$$M_i = W_i * f_i = W_i * (p_i x + q_i y + r_i)\quad (3.3.25)$$

Where, M_i: the outputs of the fourth layer nodes that represents the rule contribution, W_i: the outputs of the third layer, and p_i, q_i, r_i: the parameters set of the layer 4 nodes.

- Defuzzification Layer (Layer 5, Output): the output layer in the network represents the defuzzification layer, where 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. though this layer combines all the inputs into a single value as per the equation 3.3.26 below.

$$D_5 = \frac{\sum w_i}{\sum w_i f_i}, for\ i = 1,2\ ....\qquad (3.3.26)$$

Where, D_5: the final output of the last layer 5, and the rest of the parameters already explained in the previous equations.

The structures of the Neuro-Fuzzy models are two types, the first is the Grid Partitioning, it generates the rules based on the relation between the number of inputs and the number of membership function for each input per the equation 3.3.27, also the below figure 3.3.10 illustrates the complexity caused from the Grid Partitioning structure, as it's exponential relation the more the inputs and functions increase the more complexity would reach.

And that leads to a large number of rules which requires long computation time, which could be many days or long hours if the inputs number and data size are large. That is on the contrary in the Subtractive Clustering structure per the below figure 3.3.11, where each rule is represented using one cluster only, and that makes the experiments of this structure faster with lower complexity in the rules and generation.

$$Rules\ Number = numbe\ of\ Membership\ Functions^{Number\ of\ Inputs}\qquad (3.3.27)$$

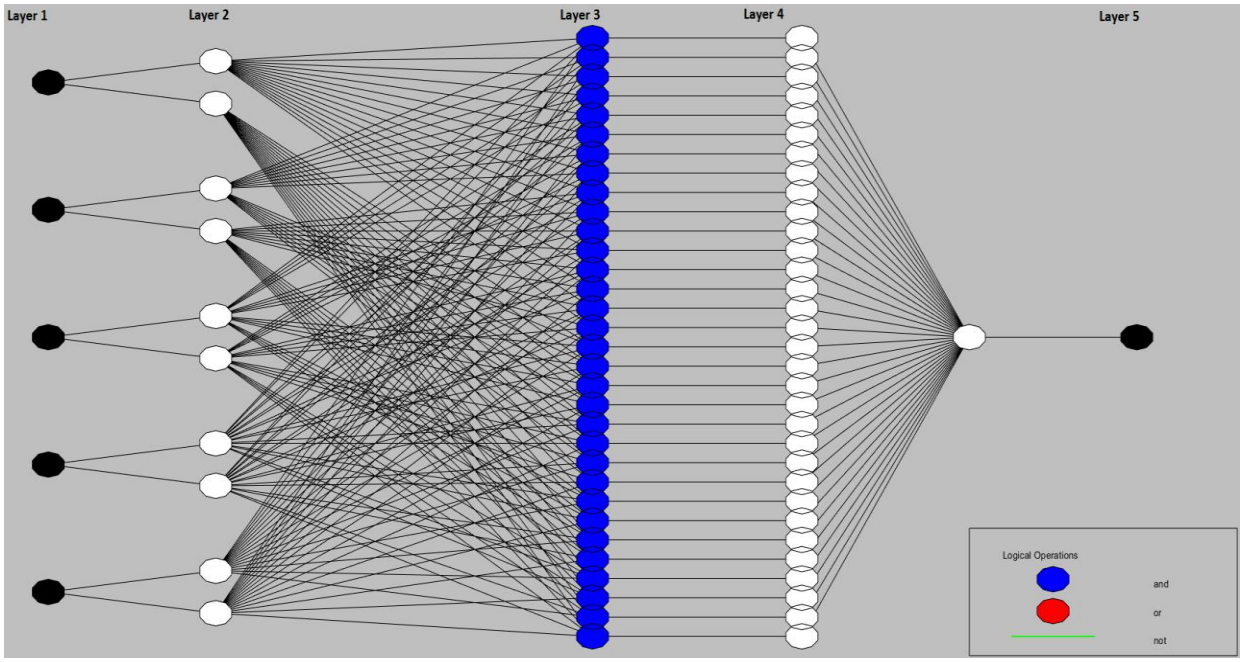


Figure 3.3.10: Neuro-Fuzzy structure using Grid Partitioning structure

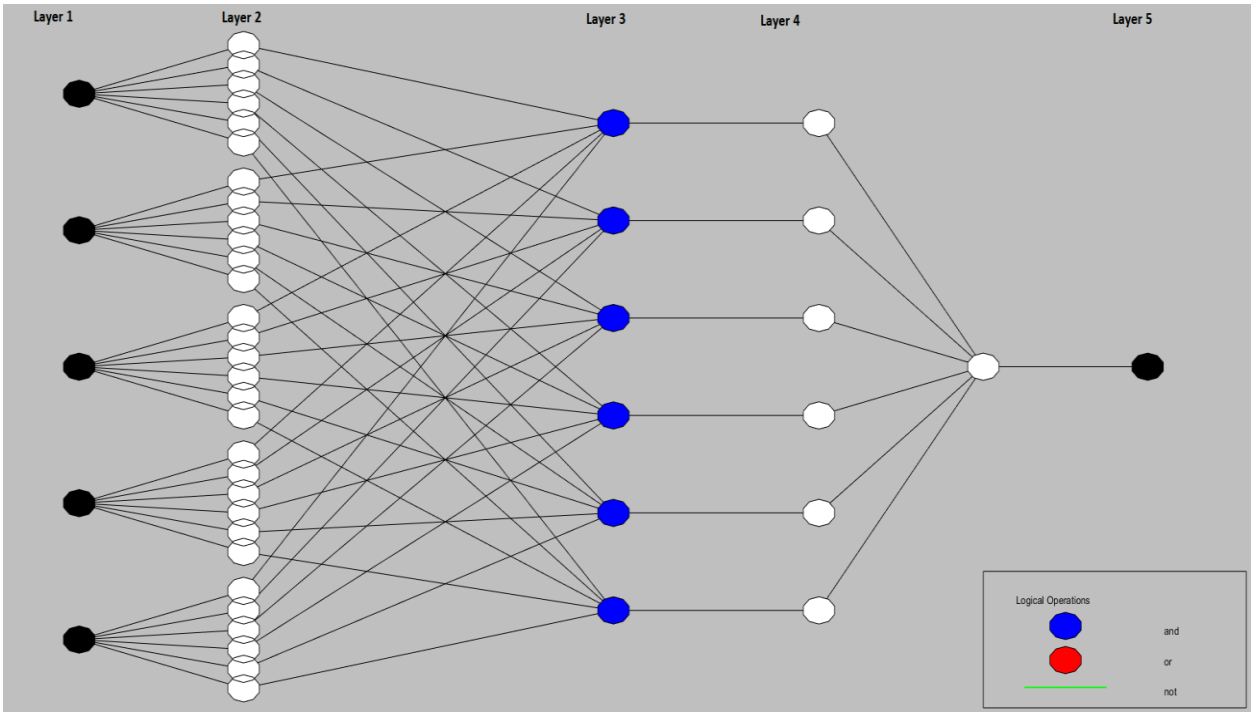


Figure 3.3.11: Neuro-Fuzzy structure using Subtractive Clustering structure

The datasets will be utilized in the ANFIS model to generate the rules from the data pairs of input factors with output diagnosis for the LDL-C and HDL-C. This ANFIS model can recognize from the data mapping for generating the rules and predicting the diagnosis, that is what is applied in this thesis, it proved promising results in predicting the Cholesterol levels. The outcomes would be discussed in the results section Chapter 4.

Algorithm 4: Neuro-Fuzzy (ANFIS) algorithmic steps

- Data normalization: inputs scaling
- Data loading: both datasets
- Data partitioning: training and testing
- ANFIS training and testing
    - Structure
        - Grid Partitioning
        - Subtractive Clustering
    - Optimization
        - Back Propagation
        - Hybrid: uses a combination of backpropagation and least-squares regression
- Evaluate results
- Measure the performance
    - Mean Square Error MSE
- Optimize the structure
    - By changing the membership functions types and counts
- Re-run the model

3.4Metrics Selection

This thesis 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 per illustrated in figure 3.4.1, those classes are the "**True Positive**", "**False Positive**", "**True Negative**", "**False Negative**". The definitions of these terms are listed in the following points [53] [54] in the light of the LDL-C classification:

- True Positive **TP**: the number of data instances that are correctly identified as high LDL-C
- False Positive **FP**: the number of data instances that are incorrectly identified as high LDL-C
- True Negative **TN**: the number of data instances that are correctly identified as normal LDL-C
- False Negative **FN**: the number of data instances that are incorrectly identified as normal LDL-C

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:

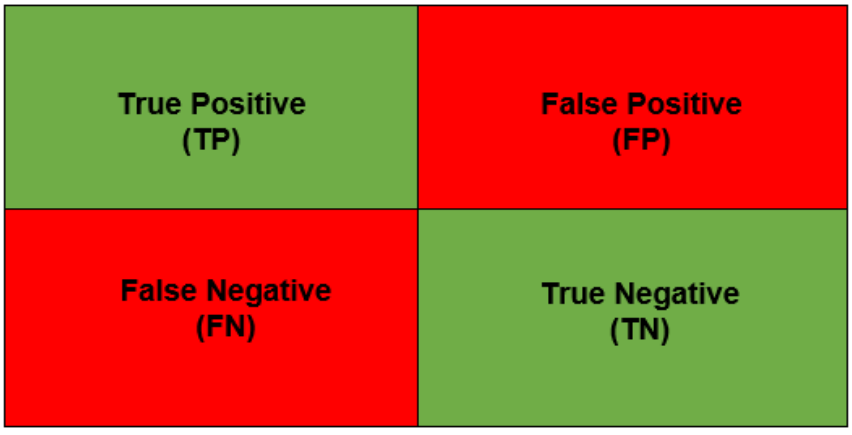


Figure 3.4.1: The Confusion Matrix

- Accuracy: it is the general used metric to measure the overall results of the model, regardless of the positive and negative recognition
  - $Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$ (4.4.1)
- F-Measuring (F-Score): the mean balance of precision and sensitivity
  - $F - Measuring = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision}$ (4.4.2)
- G-Mean: it measures the geometric mean of normal classes specificity and hostile classes sensitivity
  - $G - Mean = \sqrt{(Sensitivity * Specificity)}$ (4.4.3)
- Sensitivity: the percent of the TP instances
  - $Sensitivity = \frac{TP}{TP + FN}$ (4.4.4)
- Specificity: the percent of the TN instances
  - $Specificity = \frac{TN}{TN + FP}$ (4.4.5)
- Precision: the percent of TP instances to all instances of positive class
  - $Precision = \frac{TP}{TP + FP}$ (4.4.6)

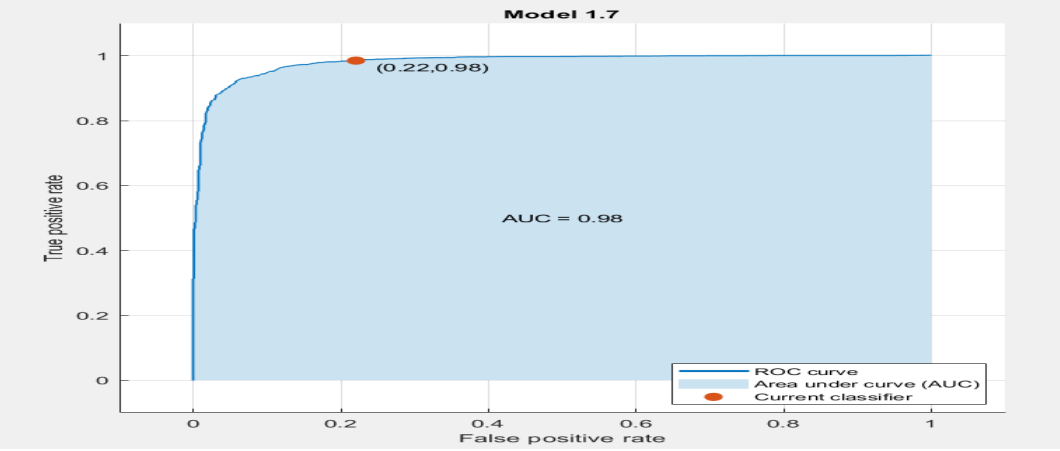


Figure 3.4.2: The Receiver Operative Characteristic ROC curve and the Area Under the Curve AUC

Another method for evaluating the accuracy of the classification techniques is the Receiver Operating Characteristic ROC curve, it's commonly used to represent the medical recognition and classification between the disease and normal diagnosed subjects, that is similar to what did in this thesis of classifying into high and normal cholesterol level of LDL-C.

The ROC curve as in figure 3.4.2 plots the relation of True Positive Rate TPR y-axis versus the False Positive Rate FPR x-axis. The TPR represents the ratio of the positively correctly classified to the all positive classifications, and the FPR represents the ratio of the incorrectly positively classified to the all negatively classified instances. The Area Under the Curve AUC is used to be an effective measure for the model's accuracy. This curve uses the logistic regression function that is already discussed and explained in the previous sections, it has a value of 0.5 as the threshold in the interval [0.0 – 1.0], the values below the 0.5 are considered as the zero class, and the values above or equal the 0.5 are considered as the 1 class, so can use the zero and one to represent the two classes of normal and high in the Cholesterol disease diagnosis. The curve is useful in graphically representing the cut-off points in the sensitive and specificity, because it describes each of them per the curve axis of x and y, as the vertical one shows the cut-off in the disease classification sensitivity which is the TPR, and the horizontal one shows the cut-off in the FPR which is the complement of the specificity.

The "Classification Learner" application in the MATLAB tool automatically generates the curve for all the classification models per figure 3.4.2, which is already extracted from the application through the thesis experimental parts.

3.5Executive Summary

This chapter covers the methodology and techniques that are followed to reach the accepted results in the experimental parts. It started with the preprocessing phase of the data, where there were two utilized datasets of local Palestinian one and international Korean one, the original sources of the data needed cleaning, feature extraction and normalization. So, used the correlation for the selection and Min-Max normalization.

It introduced the general methodology of data collection, data pre-processing, experimentation and analysis. As have many goals and techniques to achieve it. The thesis works focus on classification and prediction for the both datasets, specifically the LDL-C and HDL-C values of the lipid profile. That is based on the input fields of the lipid profile itself, in addition to the obesity measurements such as BMI, also the age and gender, and the blood hypertension of systolic and diastolic.

The chapter illustrated the structure and the implementation of the utilized techniques for the classification and the prediction. In the classification used the SVM, LR, DT, Fuzzy Logic FIS, and BP-ANNS. In the prediction the used techniques are the Neuro-Fuzzy, RNN, RBFNN, and non-linear BP-ANNs.

Finally, described and explained the metrics used to measure the performance and accuracy of the prediction and classification results. The mainly used metric is the confusion matrix values of True Positive, False Positive, True Negative, False Negative, with their relying metric calculations of accuracy, F-measuring, G-mean, sensitivity, specificity, and precision. That is in addition to the Receiver Operating Characteristic ROC and the Area Under Curve AUC.

Chapter 4

Experiments and Results

Chapter 4

4.1 Experiments and Results

The experiments were applied on two datasets of local and international, the local one collected from the Palestinian hospital, called PDS, and the international one is retrieved from a standard source created for the research purposes called KNHANES. Both datasets were gathered to serve the purpose of utilizing the ML techniques on the prediction and classification for the Cholesterol levels of LDL-C and HDL-C. The data fields contain laboratory data such as lipid profile and non-laboratory data such as gender and BMI, that collected fields represent the risk factors of the disease that help to recognize and diagnosis, the more include related risk factors would help to exploit and increase the accuracy of input-output mapping and pattern recognition using the ML.

This thesis covered many 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, in addition to using they Fuzzy Inference System FIS that is built using the AI technique of Fuzzy Logic. On the other side, the techniques used for the prediction purposes are BP-ANN non-linear output (Sigmoid), Recurrent Neural Networks RNN, Radial Basis Function Neural Networks 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. Figure 4.1.1 illustrates how was the computer's performance while running the MATLAB tools and applications.

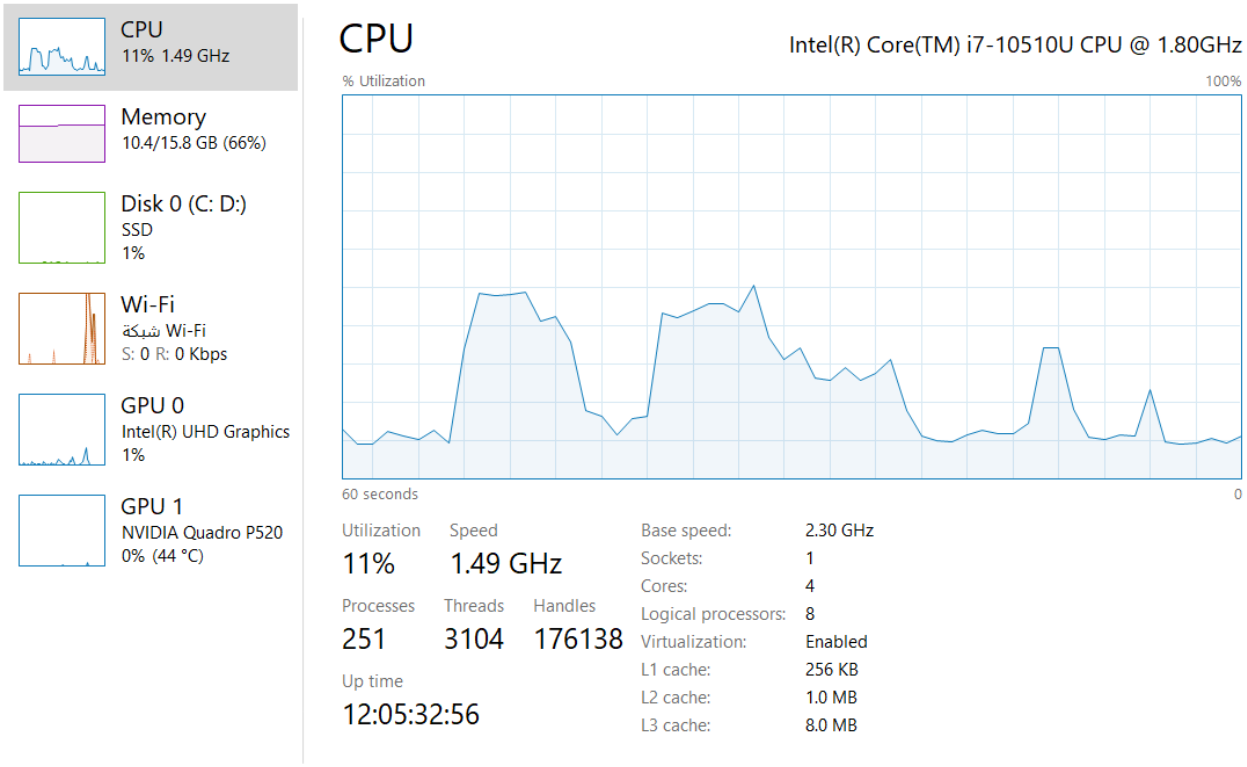


Figure 4.1.1: The computer performance while running MATLAB tools and application for the tests

The tool framework that was used is MATLAB, which provides many facilities to make the experiments faster and easier for repeating the tests. The MATLAB has many ready-made toolboxes and Graphical User Interfaces GUIs to represents and facilitate using each of the mentioned models. Each of the methods and techniques used has an affecting parameter - per the done experiments - as listed in below table 4.1, those listed parameters that affect to change the results toward try reaching the optimal accurate results. There are more features that can be included but the time limitation and data availability caused to be limited for the listed, as all this thesis works done under the coronary of Covid-19.

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. 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.

Table 4.1: the parameters and attributes that affect the accuracy of the utilized machine learning techniques.

Technique	Parameters
Neuro-Fuzzy	• Membership Function MF Input Type • MF Output Type • MF Count • Optimization Method • Number of Iterations • Cluster Range of Influence • Data Size
Radial Basis Function Neural Network RBFNN	• The Radial Spread Constant • Number of Neurons • Data Size
Recurrent Neural Network RNN	• Number of Neurons • Data Size
BP-ANN Linear and Non-Linear	• Number of Neurons • Data Size
Fuzzy Logic	• MF Type • MF Count • Input Classes • Data Size
Support Vector Machine SVM **Logistic Regression LR** **Decision Tres DT (Fine Trees)**	• Feature Selection • Data Size

The accuracy and output errors were measured and represented using the confusion matrix values and calculating the mean square error MSE. The confusion matrix helped to calculate the classification accuracy, sensitivity (recall), precision, G-mean, F-measure. Also, the prediction outputs performance is measured by how far from the actual ones, which is done using the error difference between the actual and the predicted values using the MSE. The data was divided into three parts, training got 70% of the data size, the testing 15%, and validation 15%, that division is per the commonly used standards in the reviewed related researches, and also got the recommendations in the MATLAB environment as it relies on the most accurate standards that lead to the optimal results.

4.2 The national and International Dataset Experiments

- **National Dataset**: The PDS dataset consisted of the fields introduced per table 2.2, it clarifies the data fields that consist of the lipid profile measurements (LDL-C, HDL-C, TCH, TG), in addition to the age and gender. It contains about 5484 patient profiles, that is for the sake to apply experiments for predicting and classifying the LDL-C and also HDL-C. Many ML techniques were applied that would introduce and discuss their results in the next sections.
- **International Dataset**: this dataset grabbed from the Korea National Health and Nutrition Examination Survey KNHANES, where the original dataset consists of thousands of records, with around 750 fields, but the filtration and preprocessing phase extracted data suitable for the needed experiments inputs, so a 910 patient profiles got available with 10 fields, consists of the lipid profile measurements (LDL-C, HDL-C, TCH, TG), age, gender, BMI, WHtR, SBP, and DBP.

The classification experiments used Logistic Regression LR, Support Vector Machine SVM, Decision Trees DT (Fine Tree), BP-ANN with linear output, and then built Fuzzy Logic based on feedback from the experts. On the other side, the prediction experiments used BP-ANN with non-linear output, Recurrent Neural Networks RNN, Radial Basis Function RBFNN, and finally utilized a hybrid combination of ANN and Fuzzy Logic in the model called Neuro-Fuzzy. The results would show high accuracy and acceptable results per the following sections.

4.2.1 Two-Layer Linear Output BP-ANN Experiments on the Datasets Classification

The linear output BP-ANN applied on the PDS data to classify the patient profiles into LDL-C with normal or high measurements, the used threshold to define the normal and high is the threshold value of 159 mg/dL, so the level above the threshold would consider as danger area and a healing treatment has to be considered for lowering the LDL-C cholesterol. Additionally, applied the classification on the HDL-C with threshold 59 mg/dL, it should not be lower than this level, as it represents the good cholesterol that helps to get rid of the excess bad cholesterol in the body.

For the mentioned purposes the BP-ANN experiment tried multiple times using a different number of neurons, that is to find the best accuracy with the lowest complexity and cost, per the showed parameters in table 4.1 the major parameter used in the BP-ANN is the number of neurons, it can have many other parameters, but there were many experiments on around nine algorithms and two datasets, so a time limitation would force us to focus on the significant affecting parameters only to reach the best accuracy for the binary classification of LDL-C and HDL-C. And though the following test is tried to get the needed results:

Test: the BP-ANN was 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-validation to overcome the overfitting possibility. The results as shown in the below tables 4.2 & 4.3 for the PDS dataset, and 4.4 & 4.5 for the KNHANES dataset, the best accuracy happened as explained in the followings:

- **PDS** dataset:
    - **LDL-C**: table 4.2 shows that the best classification accuracy with the percent of 95.55%, was at using 35 neurons. Each number of neurons experiment was tried 3 times to take the average accuracy results because the used model relied on the randomization to initialize the ANN weights, which need repeating the experiment to not have biased results.
    - **HDL-C**: table 4.3 represents the results of the HDL-C classification of the PDS dataset, the highest accuracy proved at using 10 neurons was in the percent of 91.80%. The same experimental conditions were applied as per the PDS LDL-C classification, to have an equal chance for the accuracy comparison.

Table 4.2: BP-ANN model results, PDS dataset for LDL-C classification

Neurons	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
5	94.77	98.10	74.90	95.90	85.72	96.99
10	94.53	97.63	75.93	96.07	86.10	96.84
15	94.03	97.27	74.80	95.83	85.30	96.54
20	94.23	97.20	77.43	96.10	86.76	96.65
25	94.37	97.60	76.23	95.83	86.26	96.71
30	94.60	97.50	87.87	96.17	92.56	96.83
35	**95.55**	98.40	77.90	96.17	87.55	97.27
40	93.47	98.03	68.63	94.50	82.03	96.23
45	94.67	98.13	75.57	95.67	86.11	96.88
50	95.43	97.90	79.70	96.83	88.33	97.36

Table 4.3: BP-ANN model results, PDS dataset for HDL-C classification

Neurons	Accuracy	Sensitivity	Specificity	Precision	G-Mean	F-Measuring
5	89.70	99.37	10.57	90.10	32.40	94.51
10	**91.80**	98.43	29.23	92.87	53.64	95.57
15	90.57	98.83	21.93	91.37	46.56	94.95
20	89.73	99.33	9.87	90.20	31.31	94.55
25	89.50	99.33	10.70	89.93	32.60	94.40
30	90.73	99.10	23.30	91.23	48.05	95.00
35	91.53	98.67	25.37	92.47	50.03	95.47
40	91.60	98.93	23.43	92.30	48.15	95.50
45	90.97	99.33	17.33	91.40	41.49	95.20
50	90.07	99.63	5.33	90.33	23.05	94.76

- **KNHANES** dataset:
    - **LDL-C**: table 4.4 shows the results of applying the LDL-C on the KNHANES dataset using the BP-ANN, the best accuracy proved at using 15 neurons, with an average accuracy of 97.10%, the experiments for each number of neurons was repeated 3 times to get the average.

Table 4.4: BP-ANN model results, KNHANES dataset for LDL-C classification

Neurons	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
5	94.90	97.33	68.80	97.07	81.83	97.20
10	95.13	98.13	71.47	96.53	83.75	97.33
15	**97.10**	98.40	81.20	98.43	89.39	98.42
20	95.17	99.47	58.20	95.33	76.09	97.36
25	92.23	97.53	47.10	93.97	67.78	95.72
30	96.37	99.20	50.70	96.97	70.92	98.07
35	93.17	97.53	61.30	94.90	77.32	96.20
40	96.10	98.90	71.30	96.80	83.97	97.84
45	93.93	96.77	68.67	96.50	81.51	96.63
50	93.43	98.13	53.33	94.80	72.34	96.44

- **HDL**-C: table 4.5 contains the results of applying the classification technique on the Korean dataset that is targeting the HDL-C, while this Cholesterol Level is not the study goal and added as additional works, but the results are also promising and showed high accuracy especially in the international dataset because it had more required fields, so the accuracy of this experiment in the BP-ANN is 97.80% at using 10 neurons.

Table 4.5: BP-ANN model results, KNHANES dataset for HDL-C classification

Neurons	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
5	94.90	100.00	0.00	94.90	0.00	97.38
10	**97.80**	99.77	0.00	98.03	0.00	98.89
15	96.60	100.00	0.00	96.60	0.00	98.27
20	95.63	99.77	0.00	95.87	0.00	97.78
25	95.87	99.27	0.00	96.60	0.00	97.92
30	95.40	99.73	0.00	95.63	0.00	97.64
35	95.40	100.00	11.13	95.33	33.37	97.61
40	95.13	100.00	0.00	95.13	0.00	97.51
45	95.13	99.73	4.77	95.37	21.80	97.50
50	94.67	99.23	0.00	95.37	0.00	97.26

4.2.1.1 Summary:

The accuracy metrics measurements are summarized on the level of both datasets national and international, for the LDL-C and HDL-C as per in the following:

- **PDS LDL-C Classification**: The BP-ANN model proved the highest classification average accuracy of 95.55% at 35 neurons number, with the average Sensitivity, Specificity, Precision, G-Mean, and F-Measuring as 98.40%, 77.90%, 96.17%, 87.55%, and 97.27%, respectively.
- **PDS HDL-C Classification**: in the PDS dataset the BP-ANN model proved the highest classification average accuracy of 91.80% at 10 neurons number, with the average

Sensitivity, Specificity, Precision, G-Mean, and F-Measuring as 98.43%, 29.23%, 92.87%, 53.64%, and 95.57%, respectively.

- **KNHANES LDL-C Classification**: in the Korean dataset the BP-ANN model proved the highest LDL-C classification average accuracy of 97.10% at 15 neurons number, with the average Sensitivity, Specificity, Precision, G-Mean, and F-Measuring as 98.40%, 81.20%, 98.43%, 89.39%, and 98.42%, respectively.
- **KNHANES HDL-C Classification**: and for the Korean dataset HDL-C classification, the BP-ANN model proved the highest classification average accuracy of 97.80% at 10 neurons number, with the average Sensitivity, Specificity, Precision, G-Mean, and F-Measuring as 99.77%, 0.00%, 98.03%, 0.00%, and 98.89%, respectively.

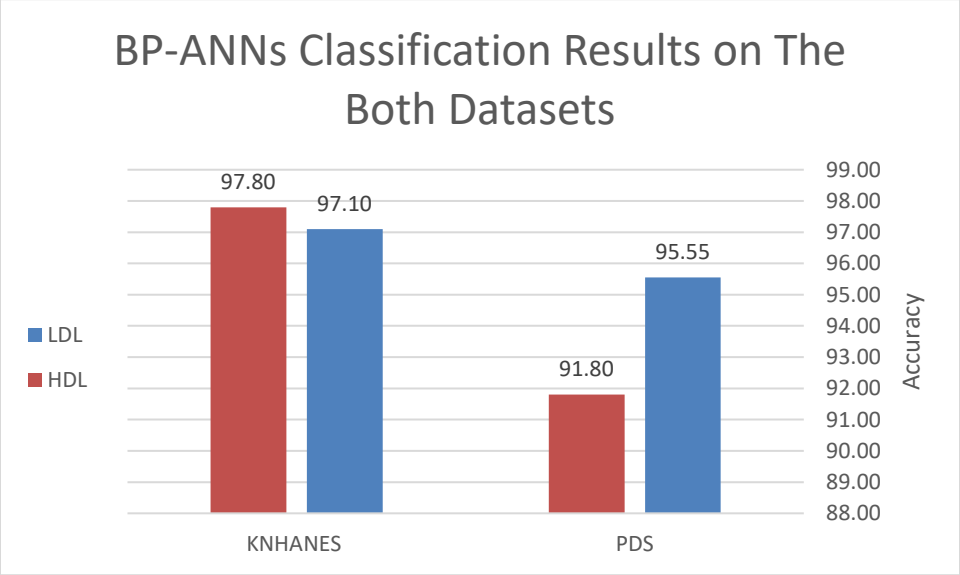


Figure 4.2.1: The classification accuracy results of using the BP-ANN on the datasets

As per above figure 4.2.1 can see that the accuracy in classifying the LDL-C is close to each other in both datasets, while there is a worthy difference reaches 6% in the classification accuracy of the HDL-C, and the results proved the outperforming in the results of the Korean dataset, and that is related to the availability of the required risk factors, but it's lacked in the Palestinian set.

4.2.2 LR, SVM, and DT Experiments on the Datasets Classification

The same LDL-C and HDL-C threshold used in the BP-ANN experiment are re-used again in this experiment of LDL-C and HDL-C classification, which would make a scaled comparison between the models' accuracy. The used techniques for this classification are the Logistic Regression LR, Support Vector Machine SVM, and the Decision Trees DT. All of them and many more combined in one application in the MATLAB called the "Classification Learner".

Both datasets experimented on the application using cross-validation of 5 folds, that is for avoiding the overfitting, so the accuracy verification is done on the level of each fold. Per the results would notice a difference in the classification accuracy results of the datasets. The PDS dataset has a smaller number of features than the KNHANES dataset that has more required fields similar to BMI and patient hypertension. To utilize the classification app so the datasets were transformed into data inputs that suit the following experiment:

Test: The "Classification Learner" application allows to select using the cross-validation, and the experiments were done using the five cross-validation folds, each technique of the three has its calculated confusion matrix for the targeted goal values which are the LDL-C and HDL-C, the experiment is done on both datasets as per illustrated in tables 4.6, 4.7, 4.8, 4.9 and explained in the following:

- **PDS** dataset:
    - **LDL-C** classification accuracy: per table 4.6 the highest accuracy is achieved using the Support Vector Machine SVM technique which got a percent of 95.50%, while the rest are the Logistic Regression LR is 95.30% and the Decision Trees DT is 94.20%, it's clear how close to each other these techniques.

Table 4.6: LR, SVM, and DT models result, PDS dataset for LDL-C classification

Method	LR	SVM	DT
TP	4597	4601	4533
FP	75	71	139
FN	181	178	177
TN	631	634	635
Accuracy%	95.30%	**95.50%**	94.20%
Recall	96.21%	96.28%	96.24%
Specificity	89.38%	89.93%	82.04%
Precision	98.39%	98.48%	97.02%
G-Mean	92.73%	93.05%	88.86%
F-Measuring	97.29%	97.37%	96.63%

- **HDL-C** classification accuracy: per table 4.7 the highest accuracy is reached using the Logistic Regression technique which got percent of 91.70%, while the rest are the Support Vector Machine SVM is 90.00% and the Decision Trees DT is 90.40%, it's lower than the LDL-C accuracy as the correlation between the input fields and output fields is less than what in the LDL-C.

Table 4.7: LR, SVM, and DT models result, PDS dataset for HDL-C classification

Method	LR	SVM	DT
TP	4841	4873	4750
FP	73	41	164
FN	384	506	365
TN	186	64	205
Accuracy %	**91.70%**	90%	90.40%
Recall	92.65%	90.59%	92.86%
Specificity	71.81%	60.95%	55.56%
Precision	98.51%	99.17%	96.66%
G-Mean	81.57%	74.31%	71.83%
F-Measuring	95.49%	94.69%	94.73%

- **KNHANES** dataset:
    - **LDL-C** classification accuracy: below table 4.8 shows the highest accuracy happened using the Support Vector Machines SVM technique which got a percent of 95.20%, while the rest are the Logistic Regression LR is 95.10% and the Decision Trees DT is 93.00%.

Table 4.8: LR, SVM, and DT models result, KNHANES dataset for LDL-C classification

Value	LR	SVM	DT
TP	803	812	789
FP	17	8	31
FN	28	36	33
TN	62	54	57
Accuracy %	95.10%	**95.20%**	93.00%
Recall	96.63%	95.75%	95.99%
Specificity	78.48%	87.10%	64.77%
Precision	97.93%	99.02%	96.22%
G-Mean	87.08%	91.32%	78.85%
F-Measuring	97.27%	97.36%	96.10%

Table 4.9: LR, SVM, and DT models result, KNHANES dataset for HDL-C classification

Method	LR	SVM	DT
TP	866	871	846
FP	5	1	25
FN	33	39	35
TN	6	1	4
Accuracy %	**95.80%**	95.70%	93.40%
Recall	96.33%	95.71%	96.03%
Specificity	54.55%	50.00%	13.79%
Precision	99.43%	99.89%	97.13%
G-Mean	72.49%	69.18%	36.39%
F-Measuring	97.85%	97.76%	96.58%

    - **HDL-C** classification accuracy: classification accuracy: above table 4.9 shows the highest accuracy is recognized using the Logistic Regression technique which got

a percent of 95.80%, while the rest are the Support Vector Machines SVM is 95.70% and the Decision Trees DT is 93.40%. It's shown that the accuracy in the KNHANES dataset is higher than in the PDS dataset because of the extra fields included in the international dataset, and also a higher correlation between its inputs and outputs.

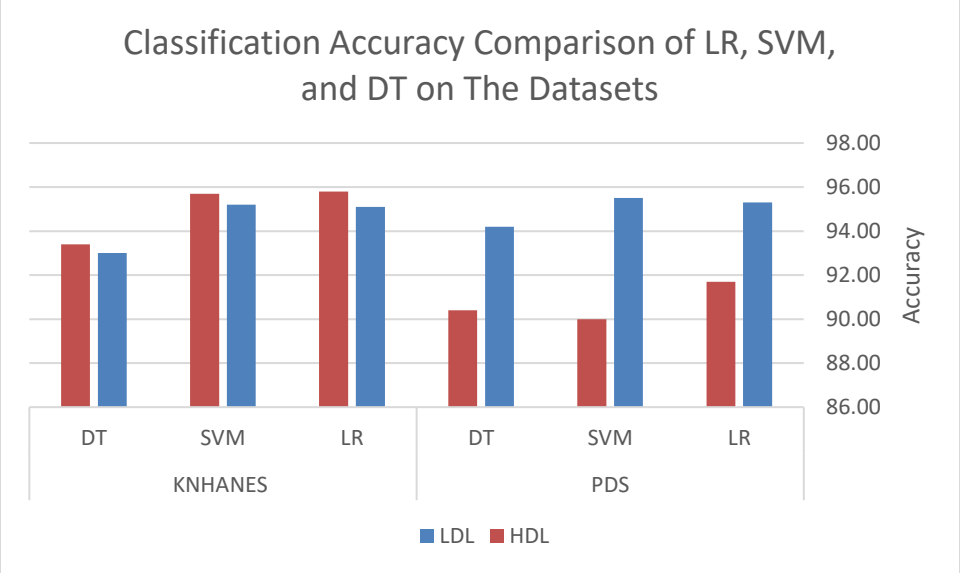


Figure 4.2.2: The classification accuracy results of applying LR, SVM, and DT on both datasets

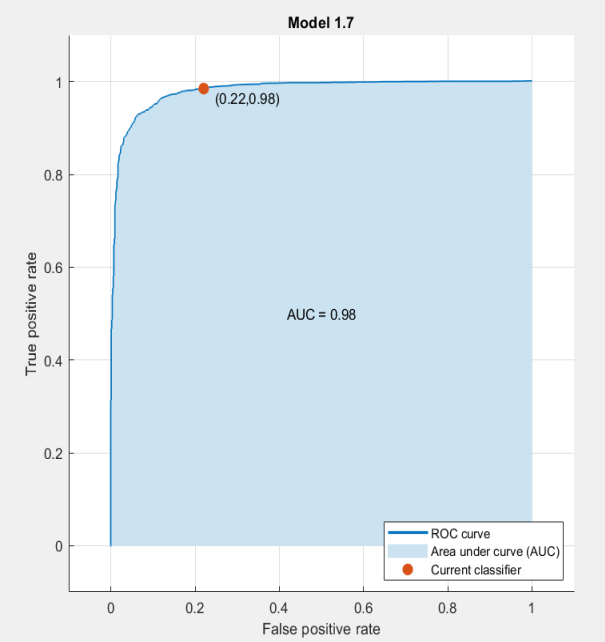
4.2.2.1 Summary:

The three models' results are very close to each other in every dataset, while the international data set provided higher accuracy for the major introduced reason, which is the available correlated features are more than what is in the local dataset. Based on the confusion matrices presented in table 4.6-4.9, the concluded accuracy measurements results are summarized as Accuracy, Sensitivity, Specificity, Precision, G-Mean, and F-measuring. For each table the highest accuracy

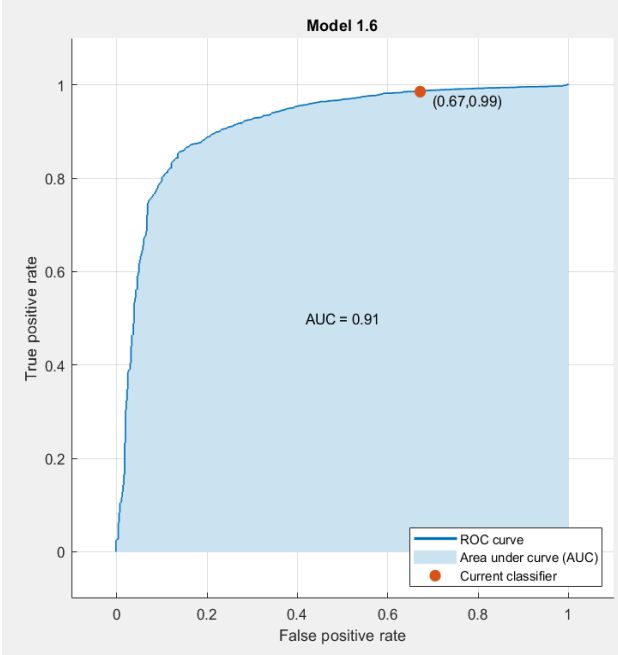
row is highlighted in yellow with the bolded red font for the accuracy value, though, the full would represent the rest of the accuracy mentioned metrics based on the table headers.

It's noticed from above figure 4.2.2 how the accuracy for the LDL-C classification is close to each other in each corresponding technique on both datasets, and still, the results higher accuracy in predicting the LDL-C than in the HDL-C in both datasets, and that the accuracy results in the international datasets are generally high for all modules for both LDL-C and HDL-C, that is even if sometimes the local data achieve higher accuracy as in the LDL-C for 95.30% in LR and 95.50% in SVM. The difference in the HDL-C between the dataset classification is 4.10%, 50.70%, and 3% in the LR, SVM, and DT respectively, and that difference was to the right of the Korean data.

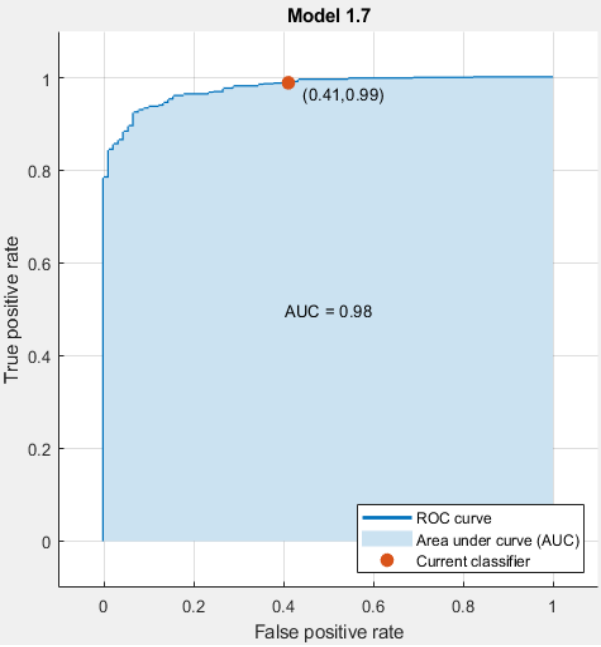
Also used the ROC curves to measure the accuracy among the different used models, which indicates the level of accuracy and actuality of the results, the below figure 4.2.3 illustrates the corresponding ROC curve with its AUC for each of the models that achieved the highest accuracy, that is in classifying the LDL-C and HDL-C in both datasets. The higher AUC means more accurate and correct classification since the high AUC means more realistic separation and classification between the real diagnostic of the normal and high values of LDL-C and HDL-C. The above figure shows how the PDS dataset got the highest LDL-C accuracy with the highest AUC of value 98% using the SVM model.



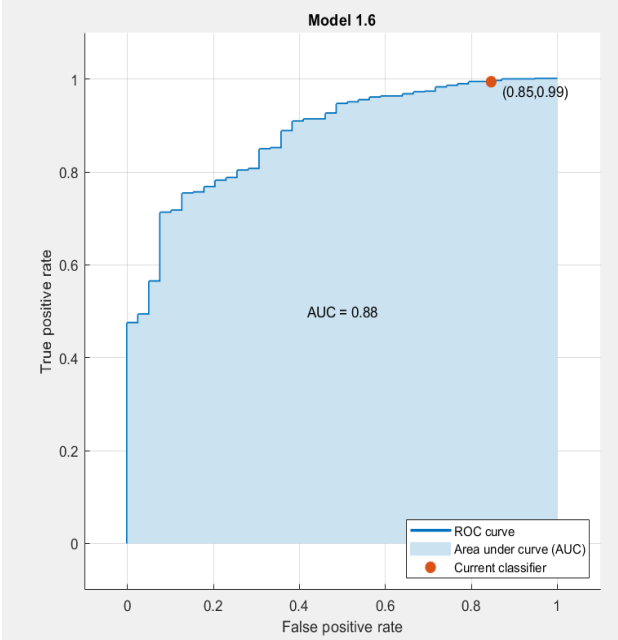
A. SVM ROC Curve: 98%, PDS LDL-C: 95.5%



B. LR ROC Curve: 91.00%, PDS HDL-C: 91.70%



C. SVM ROC Curve: 98%, Korean LDL-C: 95.20%



D. LR ROC Curve: 95.8%, Korean HDL-C:95.8%

Figure 4.2.3: The ROC curve and AUC of the highest accuracy classifiers on the datasets

4.2.3 Fuzzy Logic (FIS) Experiments on the Datasets Classification of the LDL-C

As introduced in section 3.3.3, the Fuzzy Logic idea of classification is represented to be an initial root for a Fuzzy Inference System to classify the LDL-C, as it's a chronic disease and a major risk factor for the CVDs. The practical ranges of LDL-C that are considered for diagnosing are collected from experts in the local hospital "Ibn Sina Hospital", Jenin City, Palestine. The discussion was limited to claim that the patient has no other diseases or symptoms except measuring the LDL-C, the current Cholesterol diagnostics are considering only the laboratory value of the LDL-C for the treatment decisions, ignoring the other lipid levels.

Table 4.10: LDL-C ranges that are collected from the experts

Range	Class	Diagnosis
< 130	Low	Optimal
100-160	Moderate	Borderline High
130-190	High	High
> 190	Very High	Risky

So, per the above introductory, there would be only one input which is the LDL-C level, and one output which is the LDL-C Cholesterol classification, the input-output ranges and classification are illustrated in table 4.10, it includes the input descriptive ranges and a corresponding output class, which can rely on them to define the fuzzy sets.

The main collected rules for the seed idea are based on the above table, it covers one input and one output, where can consider this initial formulation as the root for expanding the idea, the rules are listed below, which are mainly four rules collected from the experts:

- If (LDL is Low) then (LDL-Cholesterol is Optimal)
- (LDL is Moderate) then (LDL-Cholesterol is Borderline-High)

- (LDL is High) then (LDL-Cholesterol is High)
- (LDL is Very-High) then (LDL-Cholesterol is Risky)

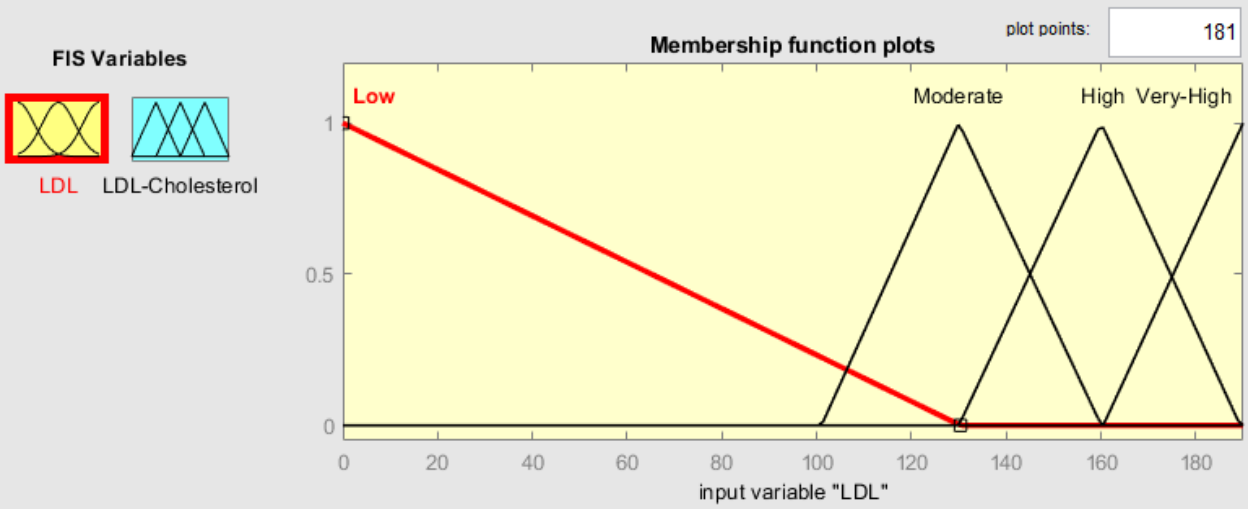


Figure 4.2.4: MATLAB FIS Fuzzy Sets and Triangular Membership Function

The fuzzy sets created based on table 4.10, and applied the MATLAB Fuzzy Inference System FIS as showing the figure 4.2.2, it used Mamdani FIS which depends on the fuzzy sets for representing the inputs and outputs, it is selected because the inputs and output ranges would be used to represent the fuzzy sets and that is most suitable to reflect the expert's collected ranges and rules. The defuzzification method used in the output is the also most recommended option in MATLAB which is the centroid, and the membership function is triangular, a simulation for the input and output can be found in Figure 4.2.3, where the Mamdani FIS utilized the triangular membership and centroid method to represent the sets and de-fuzzified the result.

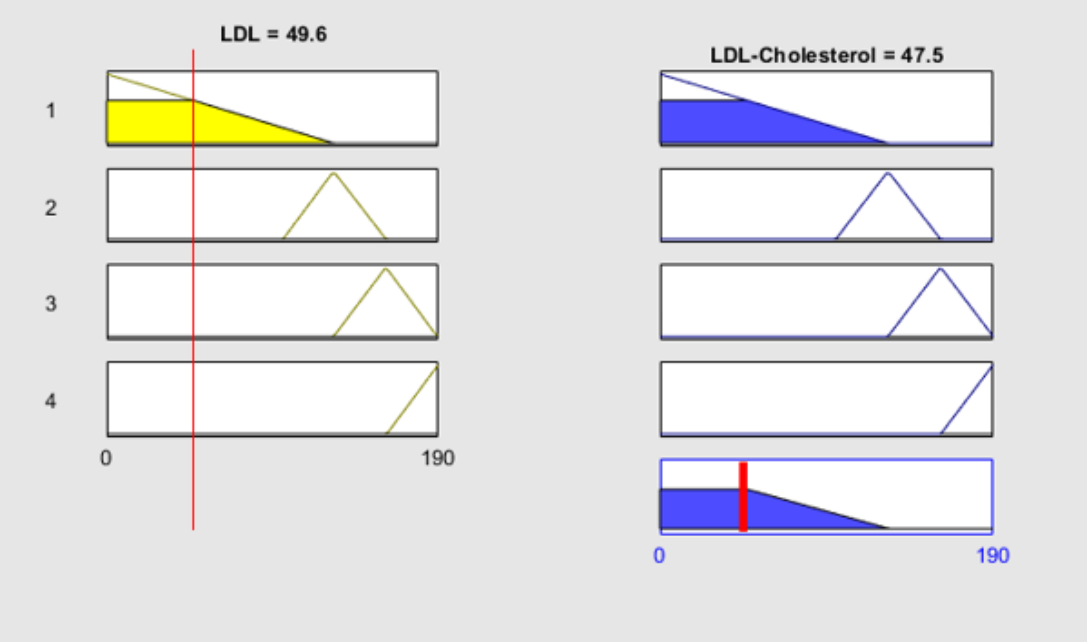


Figure 4.2.5: MATLAB Fuzzy Logic, rules output screen of Mamdani FIS

Since the input is one factor, and the output is a corresponding class to the same input field, so the accuracy would be optimal and theoretically as 100%, but be fair, not biased, and honest in collecting the expert's feedback, they claimed all other clinical measurements are normal, and the goal is to classify the LDL-C ignoring any other disease. It can build based on this idea to make some exceptions and add more factors. In this thesis works, were able to collect only those rules and inputs because the experts are working under high pressure especially for the covid-19 and the general situation in Palestine.

4.2.4 Non-linear Output BP-ANN and RNN Experiments on the Datasets Prediction

Both RNN and BP-ANN models were applied using the neural network toolbox that is available in MATLAB, they are used for the prediction of LDL-C and HDL-C, both models have the concept of backpropagation with a little bit of difference. The RNN can make multiple loops on the same layer and neuron, instead of back-propagating to the whole network starting from the output layer and backward in the hidden layers. The used version in this experiment of BP-ANN is the one with

non-linear output that is using the sigmoidal function for the prediction goals. the main attribute that is used to change for reaching better accuracy was the number of neurons. Similar to all other experiments it started per the following test details:

Test: constructing five 5 neurons at the hidden layer for each model of RNN and BP-ANN, then increasing fives 5s until reached fifty 50 neurons, then selected the structure of neurons with the minimal MSE, as the MSE is used to measure the prediction performance closure to the actual values of both LDL-C and HDL-C. Tables 4.11 - 4.14 shows the results of both models' prediction in the average MSE. Each neuron set was tried 3 runs to avoid the bias of random weights initialization, so the prediction results on both datasets using the two models are as the following:

- **PDS** dataset:
    - **LDL-C**:
        - **RNN** model: Table 4.11 illustrates the different neuron numbers that were used in the RNN model and shows how the minimal MSE was achieved using 20 neurons with the recurrent loops. While it's complex computing to use the recurrent layering but the prediction precision in the medical field worth it, so the MSE at that structure is **0.00039**, which is a very acceptable level of the prediction for the LDL-C, which is the major risk factor as mentioned for the dangerous diseases of CVD.
        - **BP-ANN** model: Table 4.11 also shows that the BP-ANN model reached a smaller MSE than the RNN but it needed a greater number of neurons, so can see that BP-ANN reached **0.00036** MSE at the number of neurons equal to 35, with a computation difference of 15 neurons.

Table 4.11: RNN and BP-ANN models result, PDS dataset for LDL-C prediction

Neurons #	Average Mean Square Error (MSE)	
	BP-ANN	**RNN**
5	0.00052	0.00044
10	0.00049	0.00045
15	0.00043	0.00045
20	0.00037	**0.00039**
25	0.00040	0.00051
30	0.00039	0.00040
35	**0.00036**	0.00049
40	0.00038	0.00045
45	0.00045	0.00045
50	0.00040	0.00053

Table 4.12: RNN and BP-ANN models result, PDS dataset for HDL-C prediction

Neurons #	Average Mean Square Error (MSE)	
	BP-ANN	**RNN**
5	0.00063	0.00074
10	0.00054	0.00050
15	0.00069	0.00056
20	0.00063	0.00060
25	0.00060	0.00067
30	0.00064	0.00067
35	0.00067	0.00066
40	**0.00048**	0.00050
45	0.00061	**0.00047**
50	0.00059	0.00069

- **HDL-C**:
  - **RNN** Model: for the HDL-C prediction, table 4.12 illustrates the runs that were applied for the HDL-C prediction until reached the 50 neurons. It found the minimal MSE as **0.00047**, it was by using 45 neurons in the RNN model.

- **BP-ANN** model: as per above table 4.12, the experiments of BP-ANN on the Palestinian dataset PDS, show an average of MSE equals **0.00048** at 40 neurons. In the HDL-C the picture is reflected and the BP-ANN outperformed the RNN model by using a smaller number of neurons. That reflects the correlation level between the inputs and outputs that supported the BP-ANN accuracy, and also that less sequencing values in the HDL-C than what is in the LDL-C, which led to lower accuracy in the RNN.

Table 4.13: RNN and BP-ANN models result, KNHANES dataset for LDL-C prediction

Neurons #	Average Mean Square Error (MSE)	
	BP-ANN	RNN
5	0.0015	0.0027
10	0.0015	0.0014
15	0.0013	0.0015
20	0.0013	**0.0014**
25	0.0011	0.0014
30	0.0014	0.0015
35	0.0012	0.0015
40	**0.0009**	0.0015
45	0.0012	0.0015
50	0.0011	0.0014

- **KNHANES** dataset:
  - **LDL-C**:
    - **RNN** model: above table 4.13 breaks down the MSE results for using the RNN model in predicting the LDL-C in the South Korean dataset. The used neurons range from 5 to 50 neurons. The minimal MSE was **0.0014** at using 20 neurons. The level of MSE is varied in the international dataset than

what is in the national because the data size in the local set is 5 times larger than the Korean one.

- **BP-ANN** model: per above table 4.13, the BP-ANN model outperformed the RNN model in predicting the LDL-C in the KNHANES dataset, but that done using a larger number of neurons that reached to 20 neurons difference with MSE value as **0.0009**, that value is the minimal in both models tries.

- **HDL-C**:
  - **RNN** model: as cleared in table 4.14, the minimal MSE in the RNN model tries was at using 35 neurons, with MSE equals to **0.0058**. The MSE value has raised in the HDL-C prediction than in the LDL-C because the patterns have a higher correlation in the LDL-C.
  - **BP-ANN** model: the BP-ANN model reached a smaller MSE than what in RNN per table 4.14, but also needs a larger number of neurons to achieve that difference, as it used 50 neurons to achieve **0.00485** MSE.

Table 4.14: RNN and BP-ANN models result, KNHANES dataset for HDL-C prediction

Neurons #	**Average Mean Square Error (MSE)**	
	BP-ANN	**RNN**
5	0.00818	0.0087
10	0.00688	0.0082
15	0.00647	0.0072
20	0.00614	0.0076
25	0.00580	0.0079
30	0.00566	0.0070
35	0.00519	**0.0058**
40	0.00574	0.0067
45	0.00488	0.0064
50	**0.00485**	0.0066

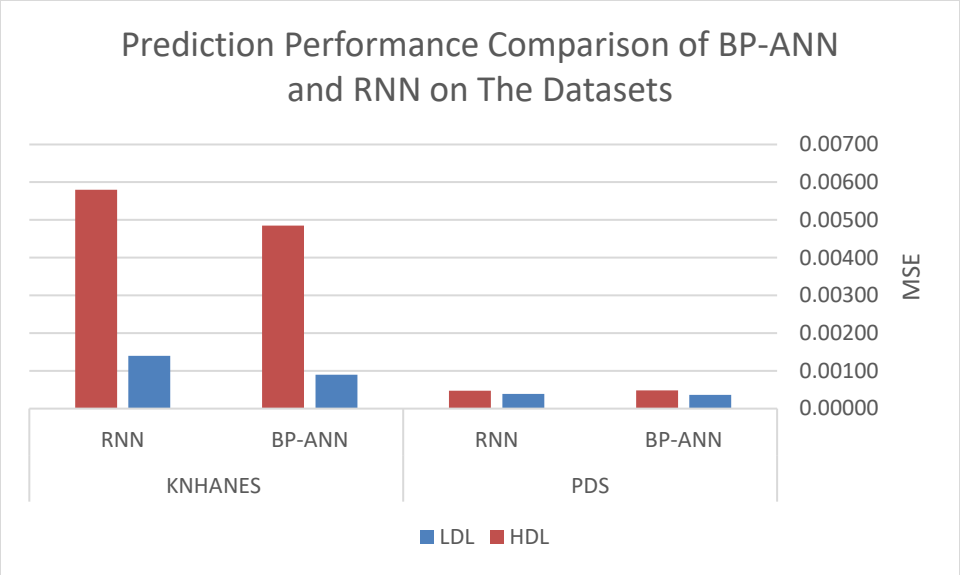


Figure 4.2.6: The prediction results of using the BP-ANN and RNN on both datasets

4.2.4.1 Summary:

As shown in figure 4.2.6, the results of RNN and BP-ANN as in the introduced tables, the model BP-ANN has usually outperformed, or very close to the RNN MSE if didn't outperform. In the PDS dataset, the BP-ANN has LDL-C prediction MSE 0.00036. In the HDL-C prediction the BP-ANN MSE is 0.00048 while the RNN 0.00047 that is a very close value with a small difference, even though, the BP-ANN compensated the small MSE difference with a fewer number of neurons for the BP-ANN that makes it less complexity, and can choose in between based on the need trade-offs if accuracy or performance. In the KNHANES dataset, the BP-ANN model outperformed the RNN in both LDL-C prediction with MSE of 0.0009 and HDL-C prediction with MSE of 0.00485, which both are smaller than what RNN achieved.

4.2.5 Radial Basis Function RBFNN Experiments on the Datasets Prediction

The RBFNN technique relies mainly on re-distributing the inputs into clusters, the cluster has a center and width defined by the radial, the distance from the center to the far input point is defined

using Euclidian distance, each cluster represents a neuron that uses the Gaussian function with the spread constant as the main attribute, to determine how much each neuron will cover points, and though how much neurons there will be, that is because the larger cluster size using the spread constant will mean fewer number of neurons, so the fewer number of neurons could lead to higher error if the data distribution has higher error mapping caused by the distance from the centers.

This experiment relied mainly on two attributes to minimize the error as much as possible, which are the spread constant and the number of neurons, the spread constant is from its name a constant real number, where we can select any number but for sure not all numbers would have a positive effect, since the larger numbers mean larger clusters, and though weaker relation between the inputs and cluster centers, that leads to higher error. To practice reaching the lowest error of MSE so tried many options by changing the attributes as in the following test:

Test: the used tool in MATLAB allows changing the targeted attributes, so the spread constant value started the experiment with 0.5 then started increasing the value with halves until reached the value 5.0, at every increase in the constant it was tried three runs by changing in the number of neurons, starting by 50 neurons, then increase the 50s until reached 150 neurons to measure the minimal possible MSE. Tables, 4.15 – 4.18 proved that the higher the spread constant the higher MSE which is contrary to the goal to find the lowest possible error. The following points summarize the results of applying the RBFNN on both utilized datasets for predicting the values of the LDL-C and HDL-C:

- **PDS** dataset
    - **LDL-C**: per below table 4.15 shows how RBFNN achieved the minimal MSE at using a spread constant of value 1.5, that is by utilizing 150 neurons, that MSE was

0.00034, where also a close value to the mentioned MSE which is **0.00036** can be achieved at a smaller spread constant and fewer number of neurons, so that is related to the judgment of the situation if the goal is the error minimization or the performance and complexity, can choose between them based on the trade-offs.

Table 4.15: RBFNN model result, PDS dataset for LDL-C prediction

Spread Constant	Average MSE		
	Fifty Neurons	**Hundred Neurons**	**One Hundred Fifty Neurons**
0.5	0.00042	0.00036	0.00034
1.0	0.00041	0.00036	**0.00034**
1.5	0.00042	0.00037	0.00036
2.0	0.00042	0.00038	0.00038
2.5	0.00042	0.00039	0.00038
3.0	0.00041	0.00040	0.00040
3.5	0.00043	0.00041	0.00041
4.0	0.00044	0.00042	0.00042
4.5	0.00045	0.00042	0.00042
5.0	0.00045	0.00043	0.00043

Table 4.16: RBFNN model result, PDS dataset for HDL-C prediction

Spread Constant	Average MSE		
	Fifty Neurons	**Hundred Neurons**	**One Hundred Fifty Neurons**
0.5	0.000676	0.000612	0.000573
1.0	0.000547	0.000596	0.000562
1.5	0.000637	0.000608	0.000610
2.0	0.000649	0.000615	0.000615
2.5	0.000641	0.000648	0.000640
3.0	0.000674	0.000658	0.000650
3.5	0.000663	0.000655	0.000651
4.0	0.000664	0.000640	0.000640
4.5	0.000680	0.000654	0.000648
5.0	0.000682	0.000658	0.000659

- **HDL-C**: also, above table 4.16 illustrates the results of RBFNN on predicting the HDL-C in the Palestinian data, it showed higher error but fewer neurons than what is the LDL-C, as at try of using 1.0 as spread constant with 50 neurons, the MSE achieved is **0.00055**, which still in the acceptable area, and with less complexity.

- **KNHANES** dataset:
  - **LDL-C**: per below table 4.17, the Korean data produced a higher error in the prediction of LDL-C than what is in the Palestinian dataset, that is caused by the high dimensionality in the PDS dataset, as the fewer data in the Korean data with the high number of used neurons can lead to the overfitting which causes higher prediction error, so the RBFNN in this experiment achieved lowest MSE at spread constant value 1.0 with using 150 neurons, and the minimal achieved value of MSE is **0.00121** per what presented in the listed experiments in the mentioned table.

Table 4.17: RBFNN model result, KNHANES dataset for LDL-C prediction

Spread Constant	**Average MSE**		
	Fifty Neurons	**Hundred Neurons**	**One Hundred Fifty Neurons**
0.5	0.00218	0.00153	0.00125
1.0	0.00166	0.00136	**0.00121**
1.5	0.00162	0.00138	0.00125
2.0	0.00156	0.00139	0.00125
2.5	0.00158	0.00138	0.00129
3.0	0.00157	0.00139	0.00131
3.5	0.00159	0.00141	0.00127
4.0	0.00157	0.00139	0.00131
4.5	0.00157	0.00139	0.00129
5.0	0.00157	0.00138	0.00131

Table 4.18: RBFNN model result, KNHANES dataset for HDL-C prediction

Spread Constant	Average MSE		
	Fifty Neurons	**Hundred Neurons**	**One Hundred Fifty Neurons**
0.5	0.00872	0.00729	0.00628
1.0	0.00831	0.00718	0.00643
1.5	0.00816	0.00724	0.00660
2.0	0.00811	0.00735	0.00655
2.5	0.00817	0.00732	0.00660
3.0	0.00821	0.00730	0.00654
3.5	0.00819	0.00729	0.00665
4.0	0.00819	0.00733	0.00658
4.5	0.00825	0.00729	0.00681
5.0	0.00822	0.00734	0.00682

- **HDL-C**: the same in LDL-C, the Palestinian dataset outperformed the Korean data in the RBFNN, as per illustrated in above table 4.18, the minimal achieved MSE is 0.0063 which is worse than the PDS results which are 0.00055, and that caused by the larger data and dimensionality, as the RBFNN is suitable for the scenarios of dimensional data, this experiment of predicting the HDL-C in the Korean dataset utilized 150 neurons and used 0.5 as the spread constant.

4.2.5.1 Summary:

as per below figure 4.2.7 the PDS dataset prediction results outperformed the KNHANES dataset, which emerged from the higher dimensionality and larger data size because the RBFNN utilized usually for the large dataset size and dimensionality. Per the produced results it's shown how the PDS dataset had MSE of 0.00034 for LDL-C and 0.00055 for the HDL-C, which proved lower MSE than what is for the KNHANES that has 0.0012 for the LDL-C and 0.0063 for the HDL-C values. The figure illustrates the high prediction ability difference to the right of the Palestinian dataset.

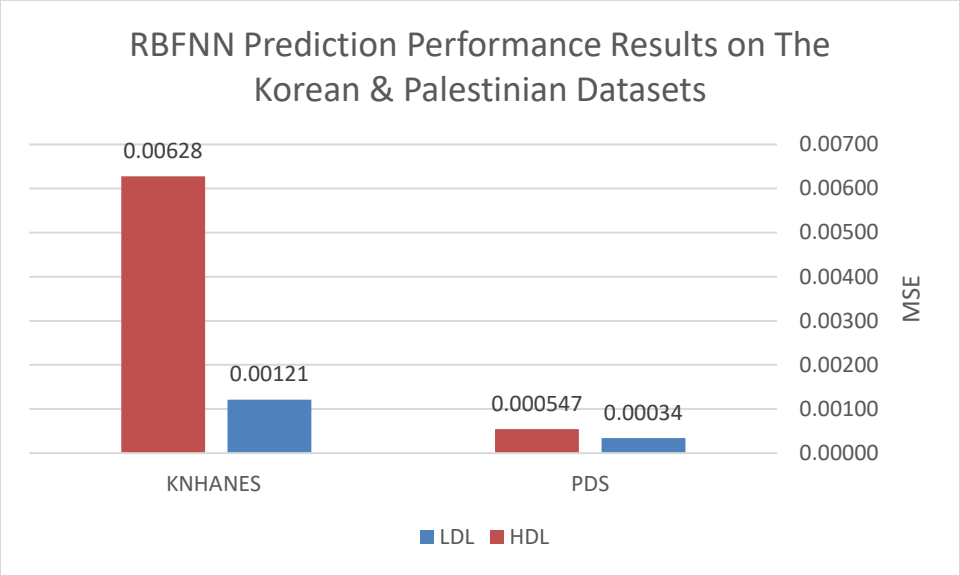


Figure 4.2.7: The RBFNN model prediction performance results

4.2.6 Neuro-Fuzzy (ANFIS) Experiments on the Datasets Prediction

The ANFIS (Adaptive Neuro-Fuzzy Inference System) model has become widely used in medical classification and prediction problems, especially in disease recognition and diagnosis, which is similar to the goal of this thesis. The main target of this work is to recognize and diagnose if the LDL-C level is high or normal in the blood, that is for the dangerous effects of having high LDL-C, which requires urgent treatment action. Though the ANFIS utilizes the strong power of ANNs in the recognition and classification and the decision-making ability in the Fuzzy Logic specifically in the cases of uncertainty, that is the same as what happens in the disease diagnosis as the harm has levels. By utilizing the medical data pairs into this model of the form input-output, the system will be automatically able to generate the If-Rule from the data itself and representing the Fuzzy decisions using the ANNs structure through the introduced layers in the proposed methods section. The model can generate the rules and also digesting if any additional supportive rules from outside

of the data, like if get the rules from the matter experts then can utilize them manually in the system rules.

As shown in figure 4.2.8, MATLAB has provided an application for ANFIS designing user interface screen, where it offers many options for choosing the model structure as Grid Partitioning or Subtractive Clustering, the optimization method if Hybrid or Back Propagation, selecting the membership function and it has many available functions such Gaussian and Triangular, picking the type of output membership if linear or constant, and flexible to choose a different number of membership functions for each input value. There is a chunk of options and combinations to try for finding the lowest possible error in the Cholesterol levels prediction. Before starting the test there was a normalization process that is done on the data because they had different scaling and that would affect the model results because would not have a specified scale for measuring the error. A different combination of the parameters and attributes available in the designer screen were tried to predict the LDL-C and HDLC, as per the following experiments:

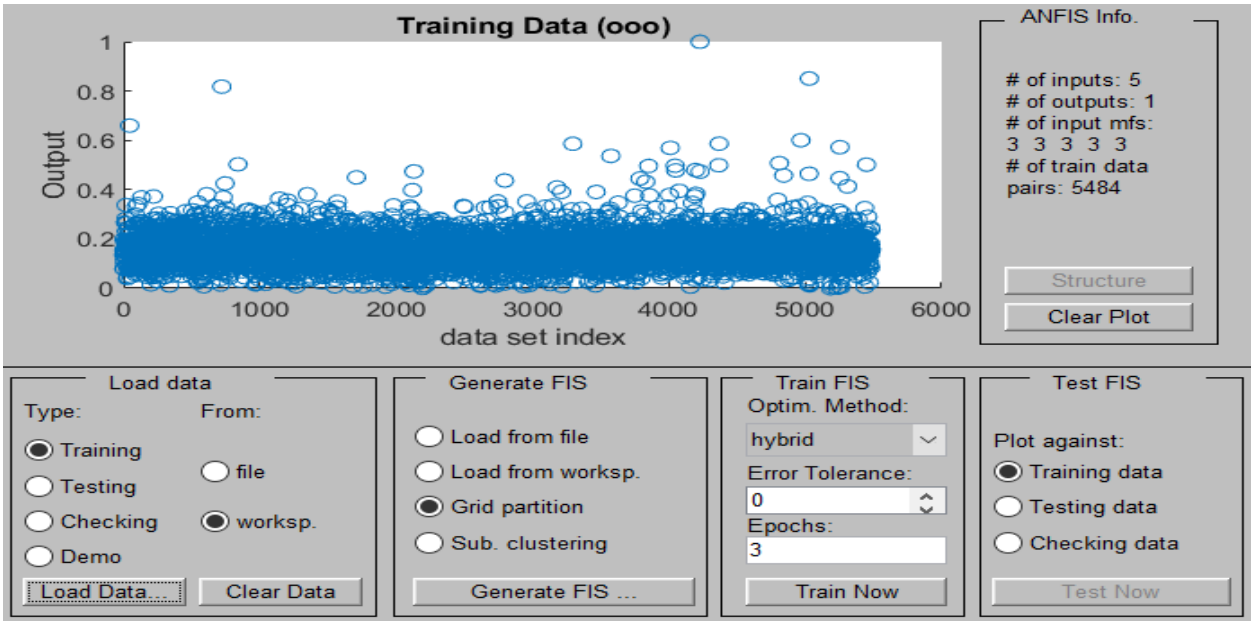


Figure 4.2.8: The MATLAB application interface of the ANFIS designer screen

Test: there are options to select many different membership functions, so tried the most common ones for the comparison, which are the Triangular, Trapezoidal, Gaussian, and Sigmodal. Each of them tried changing the optimization method and output type, that was in the case of the Grid Partitioning structure. For the Subtractive Clustering structure, the main attribute that affected finding the lowest MSE was the Range of Influence, in addition to the change in the optimization method. The MSE results are illustrated in tables 4.19 – 4.26, which summarizes the error in predicting the LDL-C and HDL-C in both datasets of Palestine and Korea as per the following:

- **PDS** dataset
  - **LDL-C**:
    - **ANFIS – Grid Partitioning**: as per below table 4.19 the Grid Partitioning structure of the ANFIS model was tried with four types of membership functions, two optimization techniques, and two types of output, the least error in predicting the LDL-C of the Palestinian dataset was for the Gaussian function, it achieved MSE value of **0.0003** using the hybrid optimization technique and linear output type.

Table 4.19: ANFIS model result using Grid Partitioning, PDS dataset for LDL-C prediction

Membership Function	**Average MSE**			
	Linear Output		**Constant Output**	
	Hybrid Optimization	**Back Propagation Optimization**	**Hybrid Optimization**	**Back Propagation Optimization**
Triangular	0.00040	0.00076	0.00054	0.00055
Trapezoidal	0.00045	0.00046	0.00347	0.00294
Gaussian	**0.00031**	0.00598	0.00073	0.00084
Sigmoidal	0.00042	0.00042	0.00336	0.00333

- **ANFIS – Subtractive Clustering**: in the case of Subtractive Clustering structure, below table 4.20 showed that the least error was at using Range of Influence equals to 0.25 and using the Back Propagation optimization technique, that reached to MSE value of **0.00055**, which is higher than the error in the Grid Partitioning, and though the Grid structure achieved the minimal MSE in predicting LDL-C in the Palestinian dataset.

Table 4.20: ANFIS model result using Subtractive Clustering, PDS dataset for LDL-C prediction

Range of Influence	Average MSE	
	Hybrid Optimization	Back Propagation Optimization
0.25	0.00057	**0.00055**
0.50	0.00060	0.00061
0.75	0.00058	0.00067
1.00	0.00058	0.00064

- **HDL-C**:

Table 4.21: ANFIS model result using Grid Partitioning, PDS dataset for HDL-C prediction

Membership Function	Average MSE			
	Linear Output		Constant Output	
	Hybrid Optimization	Back Propagation Optimization	Hybrid Optimization	Back Propagation Optimization
Triangular	0.00063	0.00073	0.00079	0.00080
Trapezoidal	0.00074	0.00075	0.00122	0.00123
Gaussian	**0.00052**	0.01225	0.00087	0.00090
Sigmoidal	0.00071	0.00072	0.00119	0.00120

- **ANFIS – Grid Partitioning**: in the HDL-C prediction for the Palestinian dataset, the lowest error in the Grid Partitioning structure achieved using the Gaussian membership function, that was by depending on the hybrid

optimization technique and linear output type, that reached value **0.0005** as the MSE minimal in the used structure per what illustrated in the above table 4.21.

- **ANFIS – Subtractive Clustering**: as per the below table 4.22, the Subtractive Clustering structure reached to lowest MSE using Range of Influence equals to 0.25 and Back Propagation optimization technique, the MSE value is **0.0007**, which proves that the Grid Partitioning also outperformed in the prediction of HDL-C in the Palestinian dataset. From both results, the LDL-C prediction in the Palestinian dataset prediction proved a lower error than the HDL-C prediction.

Table 4.22: ANFIS model result using Subtractive Clustering, PDS dataset for HDL-C prediction

Range of Influence	Average MSE	
	Hybrid Optimization	Back Propagation Optimization
0.25	0.0008	**0.0007**
0.50	0.0009	0.0009
0.75	0.0008	0.0012
1.00	0.0009	0.0026

- **KNHANES**:
  - **LDL-C**:
    - **ANFIS – Grid Partitioning**: the lowest MSE for applying Grid Partitioning on the Korean dataset for predicting the LDL-C was **0.0001**, and as per the below table 4.23 results, it illustrates the experiment details, the lowest MSE is done using Sigmoidal membership function with linear output and hybrid optimization.

Table 4.23: ANFIS result using Grid Partitioning, KNHANES dataset for LDL-C prediction

Membership Function	Average MSE			
	Linear Output		Constant Output	
	Hybrid Optimization	Back Propagation Optimization	Hybrid Optimization	Back Propagation Optimization
Triangular	0.00051	0.00052	0.00111	0.18664
Trapezoidal	0.0005	0.00011	0.00351	0.18429
Gaussian	0.00037	0.00041	0.00135	0.00145
Sigmoidal	**0.00010**	0.00015	0.00408	0.00456

- **ANFIS – Subtractive Clustering**: below table 4.24 shows lowest achieved error in the Subtractive Clustering structure was at using hybrid optimization at Range of Influence equals to 0.25, that MSE is **0.0002**, which is higher than what achieved in the other structure, and though again, the Grid Partitioning produces better prediction results than Subtractive Clustering, it's a small difference between them and the decision on which to select depends on the goal of the prediction, is the performance or the accuracy more important, as the Subtractive Clustering performance is better.

Table 4.24: ANFIS result - Subtractive Clustering, KNHANES dataset for LDL-C prediction

Range of Influence	Average MSE	
	Hybrid Optimization	Back Propagation Optimization
0.25	**0.0002**	0.0004
0.50	0.0015	0.0017
0.75	0.0019	0.0020
1.00	0.0019	0.0020

- HDL-C:
  - **ANFIS – Grid Partitioning**: for the HDL-C prediction in the Korean dataset, the Trapezoidal function achieved the best prediction results, it had

an MSE value equals 0.0004 at using the linear output and hybrid optimization technique, the details illustrated in below table 4.25.

- **ANFIS – Subtractive Clustering**: below table 4.26 presents the experiment details of predicting the HDL-C in the Korean dataset using the Subtractive Clustering, where the minimal achieved MSE equals 0.0007, that got at using the hybrid optimization using the Range of Influence value of 0.25, still the Grid Partitioning outperforming the Subtractive Clustering in all prediction results as per reviewed in the all previous experiments of ANFIS results.

Table 4.25: ANFIS result using Grid Partitioning, KNHANES dataset for HDL-C prediction

Membership Function	Average MSE			
	Linear Output		Constant Output	
	Hybrid Optimization	Back Propagation Optimization	Hybrid Optimization	Back Propagation Optimization
Triangular	0.00272	0.00274	0.00545	0.00549
Trapezoidal	**0.00041**	0.00062	0.00806	0.00809
Gaussian	0.00203	0.00228	0.00578	0.00616
Sigmoidal	0.00099	0.00129	0.00809	0.00898

Table 4.26: ANFIS result - Subtractive Clustering, KNHANES dataset for HDL-C prediction

Range of Influence	Average MSE	
	Hybrid Optimization	Back Propagation Optimization
0.25	**0.0007**	0.0010
0.50	0.0078	0.0076
0.75	0.0095	0.0097
1.00	0.0094	0.0095

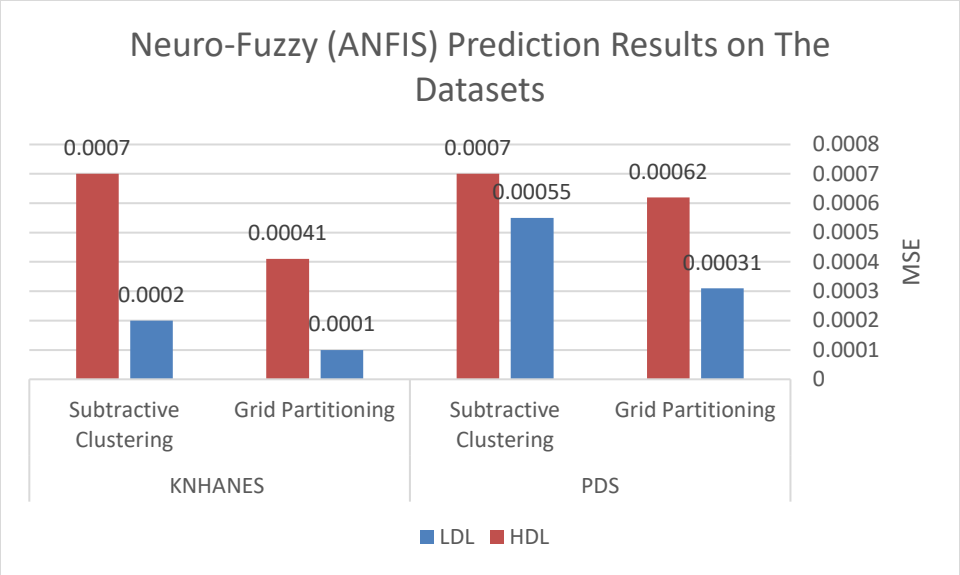


Figure 4.2.9: The prediction results of utilizing the Neuro-Fuzzy model on the PDS and KNHANES datasets

4.2.6.1 Summary:

As per above figure 4.2.9 visualized, the LDL-C usually has better results in both datasets in comparison to HDLC, because the MSE is lower which means a closer result between the predicted and actual values. But if want to compare the results on the level of the datasets, the Korean data results have outperformed the PDS data results, because the prediction training requires more inputs that are relative and affecting the output values, where the case in the Korean dataset as it has more input fields represent the risk factors. On the level of the Neuro-Fuzzy structure, the Grid partitioning has reached better prediction performance results in both datasets, while it's more complex but achieved better results, on the other side, the Subtractive Clustering structure achieved very close results to the Grid Partitioning and with less complexity, and the choice depends on which is more before select between them, is it the error or performance.

The same MATLAB ANFIS designer application has the option to graphically test the generated model and manually modify the rules as below shown in figures 4.2.10 & 4.2.11, respectively.

Figure 4.2.10: The ANFIS designer screen for testing the generated model rules

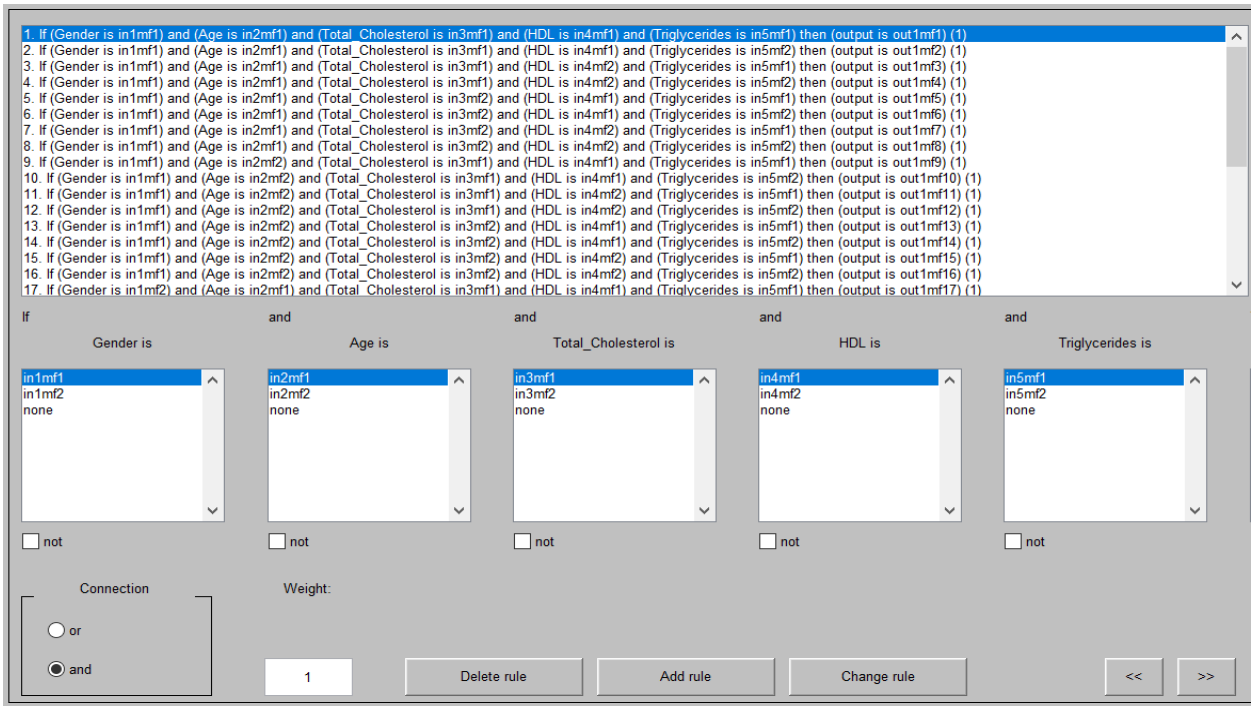


Figure 4.2.11: The ANFIS designer screen for editing the generated model rules manually

The below table 4.27 illustrates a sample of the rules that get generated in the Neuro-Fuzzy model, those roles that are generated based on the training data using the ANNs, so the rules will be representing the Fuzzy Logic knowledge base. That experiment utilized 5 inputs with 2 membership function for each using the Gaussian function, it resulted in 32 rules as base.

Table 4.27: Sample of the rules generated using the Neuro-Fuzzy model in ANFIS tool

Rule #	Neuro-Fuzzy Rule
1	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf1) (1)
2	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf2) (1)
3	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf3) (1)
4	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf4) (1)
5	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf5) (1)
6	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf6) (1)
7	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf7) (1)
8	If (Gender is in1mf1) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf8) (1)
9	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf9) (1)
10	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf10) (1)
11	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf11) (1)
12	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf12) (1)
13	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf13) (1)
14	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf14) (1)
15	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf15) (1)
16	If (Gender is in1mf1) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf16) (1)
17	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf17) (1)
18	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf18) (1)

19	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf19) (1)
20	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf20) (1)
21	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf21) (1)
22	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf22) (1)
23	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf23) (1)
24	If (Gender is in1mf2) and (Age is in2mf1) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf24) (1)
25	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf25) (1)
26	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf26) (1)
27	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf27) (1)
28	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf1) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf28) (1)
29	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf1) then (LDL is out1mf29) (1)
30	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf1) and (HDL is in5mf2) then (LDL is out1mf30) (1)
31	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf1) then (LDL is out1mf31) (1)
32	If (Gender is in1mf2) and (Age is in2mf2) and (TCH is in3mf2) and (TG is in4mf2) and (HDL is in5mf2) then (LDL is out1mf32) (1)

4.3 Discussion of the Results

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 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 4.28 to 4.35, in each table will illustrate the best result done using which technique for which dataset and to which Cholesterol level if LDL or HDL.

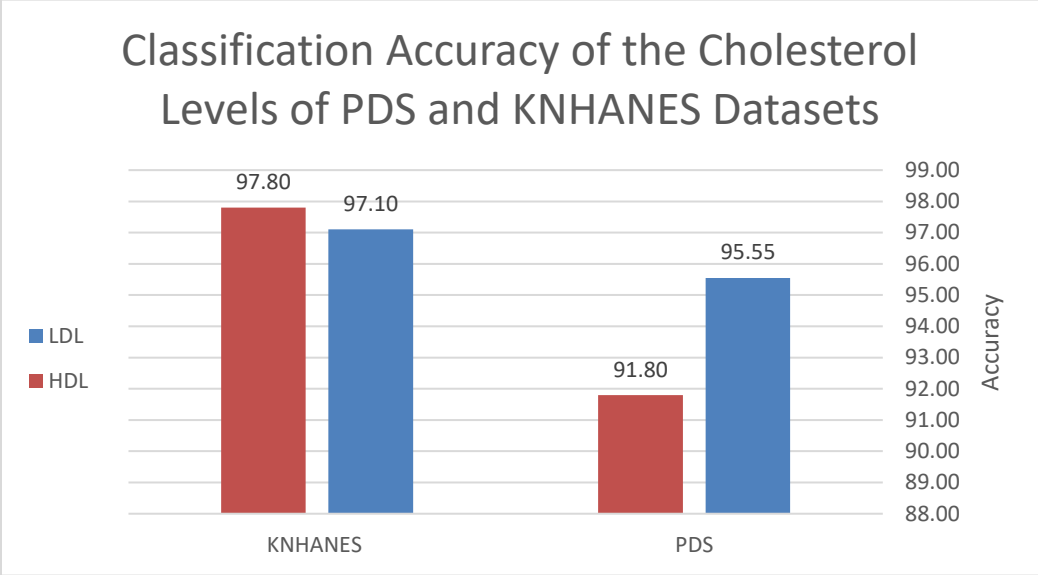


Figure 4.3.1: The accuracy results of using BP-ANN for classifying LDL-C and HDL-C on the PDS and KNHANES datasets

Table 4.28: All Techniques Results Comparison, PDS dataset for LDL-C classification

Technique	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
BP-ANN	**95.55**	97.50	87.87	96.17	92.56	96.83
LR	95.30	96.21	89.38	98.39	92.73	97.29
SVM	95.50	96.28	89.93	98.48	93.05	97.37
DT	94.20	96.24	82.04	97.02	88.86	96.63

Table 4.29: All Techniques Results Comparison, PDS dataset for HDL-C classification

Technique	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
BP-ANN	**91.80**	98.43	29.23	92.87	53.64	95.57
LR	91.70	0.93	0.72	0.99	0.82	0.95
SVM	90.00	0.91	0.61	0.99	0.74	0.95
DT	90.40	0.93	0.56	0.97	0.72	0.95

Table 4.30: All Techniques Results Comparison, KNHANES dataset for LDL-C classification

Technique	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
BP-ANN	**97.10**	98.40	81.20	98.43	89.39	98.42
LR	95.10	96.63	78.48	97.93	87.08	97.27
SVM	95.20	95.75	87.10	99.02	91.32	97.36
DT	93.00	95.99	64.77	96.22	78.85	96.10

Table 4.31: All Techniques Results Comparison, KNHANES dataset for HDL-C classification

Technique	Accuracy %	Sensitivity %	Specificity %	Precision %	G-Mean %	F-Measuring %
BP-ANN	**97.80**	99.77	0.00	98.03	0.00	98.89
LR	95.80	96.33	54.55	99.43	72.49	97.85
SVM	95.70	95.71	50.00	99.89	69.18	97.76
DT	93.40	96.03	13.79	97.13	36.39	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 4.28 to 4.31, 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 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-ANN, but the BP-ANN proved the highest ability in the classification. Figure 4.3.1 illustrates a comparison between the accuracy in the two different local and foreign datasets using the BP-ANN technique, that reached the best results.

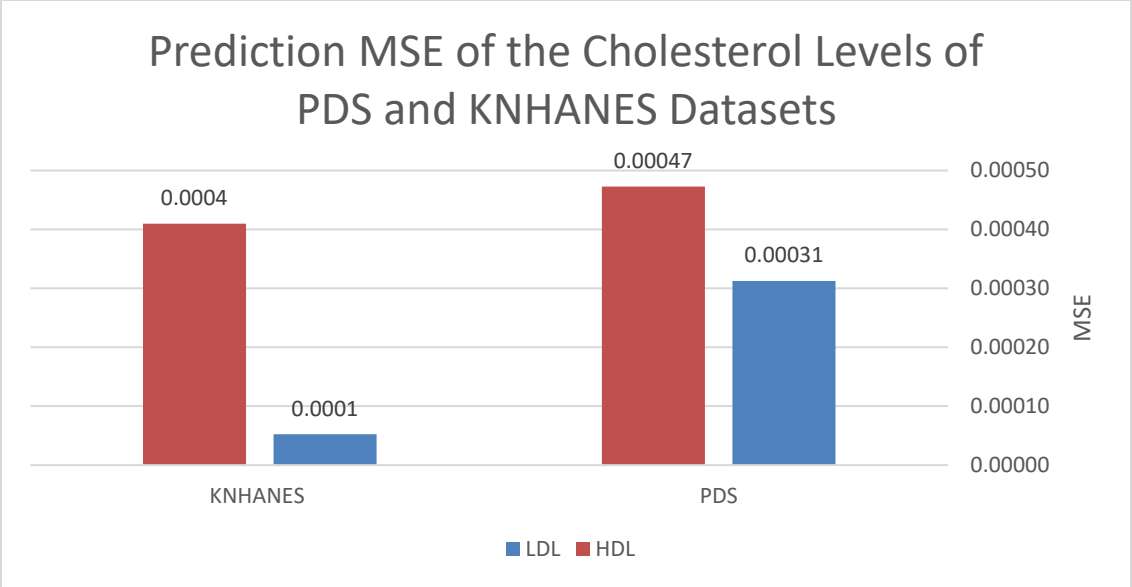


Figure 4.3.2: The MSE of predicting LDL-C and HDL-C on the PDS and KNHANES datasets

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 RNN outperformed the Neuro-Fuzzy in the HDL-C prediction of the PDS dataset per shown in table 4.33, 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 4.32: All Techniques Results Comparison, PDS dataset for LDL-C prediction

Average MSE

Technique	**BP-ANN**	**RNN**	**RBFNN**	**Neuro-Fuzzy**
Minimal	0.00036	0.00039	0.00034	**0.00031**
Condition	35 Neurons	20 Neurons	Spread Constant = 1.0 150 Neurons	Gaussian Function Hybrid Optimization Linear Output

Table 4.33: All Techniques Results Comparison, PDS dataset for HDL-C prediction

Average MSE

Technique	BP-ANN	RNN	RBFNN	Neuro-Fuzzy
Minimal	0.00048	**0.00047**	0.00055	0.00052
Condition	50 Neurons	35 Neurons	Spread Constant = 0.5 150 Neurons	Gaussian Function Hybrid Optimization Linear Output

Table 4.34: All Techniques Results Comparison, KNHANES dataset for LDL-C prediction

Average MSE

Technique	BP-ANN	RNN	RBFNN	Neuro-Fuzzy
Minimal	0.0009	0.0014	0.0012	**0.0001**
Condition	40 Neurons	20 Neurons	Spread Constant = 1.0 150 Neurons	Sigmoidal Function Hybrid Optimization Linear Output

Table 4.35: All Techniques Results Comparison, KNHANES dataset for HDL-C prediction

Average MSE

Technique	BP-ANN	RNN	RBFNN	Neuro-Fuzzy
Minimal	0.0049	0.0058	0.0063	**0.0004**
Condition	50 Neurons	35 Neurons	Spread Constant = 0.5 150 Neurons	Trapezoidal Function Hybrid Optimization Linear Output

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 4.32 to 4.35, 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-ANN, RNN, and Radial Basis Function Neural Networks RBFNN.

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 figure 4.3.2 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.

4.4 Limitation

Within the year of the thesis, there were many challenges to complete the thesis requirements, many of them related to the new situation forced by the Covid-19 pandemic. One side is the data collection of the Palestine dataset which is the focus of the study, the data is received within the last two months, because the accessibility was hard at that time of the pandemic, and all organizations staffs were under high pressure, especially the Health sector. The data received from the Palestinian Ministry of Health MoH didn't include all the required fields. Additionally, the experiments needed a long time to run especially the Neuro-Fuzzy model because some experiments took a day while others took hours, in addition to the other nine models needed to experiment. Also the running machine for the experiments had limited ability of computation because the used is normal laptop, while if have better machine performance would include more experiments and parameters.

Since the local data received within the past short time, so all the previous time was tried to utilize for international datasets such as the KNHANES. Not only but also, most of the thesis times passed under the online communications and follow up, which compensated the physical meetings with the supervisors and experts by conference calls and meetings.

4.5 Executive Summary

This chapter 4 previews the experimental part of the thesis, including the technical environment, the tools, and the techniques test results. Also review the local and international dataset expectation and contents.

The results parts covered all the detailed scenarios for each used technique, and illustrated the effect of each parameter change on the targeted technique. Some techniques changed the number of neurons and iterations to optimize the results that is similar to what used in the BP-ANN, RNN, and RBFNN. Others affected by the inputs count and the data size similar to the classification techniques in the SVM, LR, and DT. Additionally the Neuro-Fuzzy model had many parameters to change such as the membership functions count, the structure if Subtractive Clustering or Grid Partitioning, also the membership functions types such Gaussian, Triangular, Sigmoidal, and Trapezoidal. The Neuro-Fuzzy model also had options for optimization method if hybrid or back propagation, and the type of the output if linear or constant, all of these parameters changed in the results toward reaching the highest performance and accuracy. The results of the all utilized techniques are discussed in details in section 4.2, and summarized in section 4.3 that concluded the final comparison to figure out the best techniques achieved the highest accuracy and performance, in both the prediction and classification purposes.

The targeted output values were the LDL-C and the HDL-C of Cholesterol, these values that are usually results of the laboratory test called "Lipid Profile". In the previous results discussion part showed how the BP-ANNs model achieved the highest accuracy in the classification goals, and the Neuro-Fuzzy mostly outperformed in the prediction performance. All these ideas and techniques were applied on the Palestinian and Korean datasets, the Korean dataset had better result in the classification techniques because consisted of more input fields to the ML techniques, while the Palestinian one achieved better results in the prediction because contained more data samples.

Chapter 5

Conclusion and Future work

Chapter 5

5.1 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 thesis 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, nine 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 proved 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.

5.2 Future Work

It is reviewed how dangerous the effect of the LDL-C in causing the cardiovascular diseases CVDs, so the studies have to continue for utilizing the Machine Learning ML to more accurately with lower errors to classify and predict the LDL-C value, also toward being proactive in discovering the Cholesterol disease at early stages, since diagnosing it at early stages can cost zero because the treatment could be only a change in the lifestyle-related to the food and exercises while recognizing it at late stages could cost human life. Also, very recommended to search toward diagnosing the disease using noninvasive risk factors that need no blood samples or lab tests, as that would avoid the patient waiting for the results or supply blood samples for the tests, because the laboratory works require devices and specialized practitioners. Also, collecting the non-clinical factors would push the people toward checking their Cholesterol status, and be proactive for any actions toward avoiding failing in the acute and chronic diseases such as CVDs.

From the ML wise, there is great opportunity to optimize the accuracy and performance in all the utilized technique, specifically if included more related risk factor especially the ones that were collected for this thesis 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. Since already the patients have medical history so the maximum grab from those fields would support more. It's recommended to utilize the Genetic Algorithm for the ANNs optimizations because that would recommend selecting the best structure and wights, especially in the initial phases. The Fuzzy Logic system that is built in cooperation with the experts would require to continue the efforts in collecting more inputs and rules, in addition to supporting the Fuzzy Logic rules with the classification phase using the Support Vector Machines SVM. If there would be more time it's very priority to try Deep Learning through trying multiple hidden layers with a different number of neurons to achieve better outcomes.

The reached results support utilizing the utilized techniques specially the Neuro-Fuzzy in the market, as it achieved high accuracy with that few fields, so if go toward cooperation the Hospitals that would lead to more practical and realistic benefits. For sure the hospital would offer more generous data and experience, and though would enrich the expert system of the Fuzzy Logic system. And the returns would on two levels, the proactive recognition and diagnosis, and the automated checkups that would lower the operation cost.

Bibliography

[1] B. A. Ference *et al.*, “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,” *Eur. Heart J.*, vol. 38, no. 32, pp. 2459–2472, 2017.

[2] 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,” *J. Clin. Lipidol.*, vol. 10, no. 3, pp. 472–489, 2016.

[3] S. Berger, G. Raman, R. Vishwanathan, P. F. Jacques, and E. J. Johnson, “Dietary cholesterol and cardiovascular disease: a systematic review and meta-analysis,” *Am. J. Clin. Nutr.*, vol. 102, no. 2, pp. 276–294, 2015.

[4] 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.

[5] 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.

[6] L. Badimon and G. Vilahur, “LDL-cholesterol versus HDL-cholesterol in the atherosclerotic plaque: inflammatory resolution versus thrombotic chaos,” *Ann. N. Y. Acad. Sci.*, vol. 1254, no. 1, pp. 18–32, 2012.

[7] M. Nauck, G. R. Warnick, and N. Rifai, “Methods for measurement of LDL-cholesterol: a critical assessment of direct measurement by homogeneous assays versus calculation,” *Clin. Chem.*, vol. 48, no. 2, pp. 236–254, 2002.

[8] 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.

[9] J. Knopfholz *et al.*, “Validation of the friedewald formula in patients with metabolic syndrome,” *Cholesterol*, vol. 2014, 2014.

[10] T. Lee, J. Kim, Y. Uh, and H. Lee, “Deep neural network for estimating low density lipoprotein cholesterol,” *Clin. Chim. Acta*, vol. 489, pp. 35–40, 2019.

[11] K. H. Cho *et al.*, “Low-density lipoprotein cholesterol level in patients with acute myocardial infarction having percutaneous coronary intervention (the cholesterol paradox),” *Am. J. Cardiol.*, vol. 106, no. 8, pp. 1061–1068, 2010.

[12] 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.

[13] J. Ma *et al.*, “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.

[14] B. Abu-Nasser, “Medical expert systems survey,” *Int. J. Eng. Inf. Syst.*, vol. 1, no. 7, pp. 218–224, 2017.

[15] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[16] 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.

[17] I. Ali *et al.*, “Prevalence of Dyslipidemia in Undiagnosed Palestinian Men: A Cross-Sectional Study,” *J. Lipids*, vol. 2019, 2019.

[18] B. J. Lee and J. Y. Kim, “Identification of the best anthropometric predictors of serum high-and low-density lipoproteins using machine learning,” *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 5, pp. 1747–1756, 2014.

[19] R. H. Nelson, “Hyperlipidemia as a risk factor for cardiovascular disease,” *Prim. Care Clin. Off. Pract.*, vol. 40, no. 1, pp. 195–211, 2013.

[20] A. Hattori and R. Sturm, “The obesity epidemic and changes in self-report biases in BMI,” *Obesity*, vol. 21, no. 4, pp. 856–860, 2013.

[21] K. M. Flegal, M. D. Carroll, B. K. Kit, and C. L. Ogden, “Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010,” *Jama*, vol. 307, no. 5, pp. 491–497, 2012.

[22] P. K. Whelton *et al.*, “2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr,” *J. Am. Coll. Cardiol.*, vol. 71, no. 19, pp. e127–e248, 2018.

[23] G. Singh *et al.*, “Comparing a novel machine learning method to the Friedewald formula and Martin-Hopkins equation for low-density lipoprotein estimation,” *PLoS One*, vol. 15, no. 9, p. e0239934, 2020.

[24] M. N. Rana, C. Kim, C. T. Longenecker, and S. Al-Kindi, “Machine Learning for Prediction of Low Density Lipoprotein Cholesterol in Fasting and Non-fasting Patients With and Without HIV Infection,” *Circulation*, vol. 142, no. Suppl_3, pp. A14313–A14313, 2020.

[25] 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.

[26] K. G. Adi, P. V Rao, and V. K. Adi, “Analysis and Detection of Cholesterol by Wavelets based and ANN Classification,” *Procedia Mater. Sci.*, vol. 10, pp. 409–418, 2015.

[27] L. P. D. Simangunsong, I. N. Napitupulu, R. E. Lumbantoruan, S. Zalukhu, and M. Zebua, “The Expert System of Cholesterol Detection Based on Iris Using the Gabor Filter,” *Sink. J. dan Penelit. Tek. Inform.*, vol. 4, no. 1, pp. 13–18, 2019.

[28] E. R. Holzinger, S. M. Dudek, A. T. Frase, R. M. Krauss, M. W. Medina, and M. D. Ritchie, “ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels,” in *Biocomputing 2013*, World Scientific, 2013, pp. 385–396.

[29] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, p. 105524, 2020.

[30] J. Liu *et al.*, “Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network,” *Eur. Radiol.*, vol. 28, no. 8, pp. 3268–3275, 2018.

[31] N. Saravanan and A. Rathinam, “A comparitive study on ANN based fault location and classification technique for double circuit transmission line,” in *2012 Fourth International Conference on Computational Intelligence and Communication Networks*, 2012, pp. 824–830.

[32] D. A. Cirovic, “Feed-forward artificial neural networks: applications to spectroscopy,” *TrAC Trends Anal. Chem.*, vol. 16, no. 3, pp. 148–155, 1997.

[33] A. I. Schein and L. H. Ungar, “Active learning for logistic regression: an evaluation,” *Mach. Learn.*, vol. 68, no. 3, pp. 235–265, 2007.

[34] I. Kurt, M. Ture, and A. T. Kurum, “Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease,” *Expert Syst. Appl.*, vol. 34, no. 1, pp. 366–374, 2008.

[35] A. Pradhan, “Support vector machine-a survey,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, 2012.

[36] S. Suthaharan, “Support vector machine,” in *Machine learning models and algorithms for big data classification*, Springer, 2016, pp. 207–235.

[37] K. S. Durgesh and B. Lekha, “Data classification using support vector machine,” *J. Theor. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 1–7, 2010.

[38] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, “A survey on decision tree algorithm for classification,” 2014.

[39] 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.

[40] M. Pal and P. M. Mather, “An assessment of the effectiveness of decision tree methods for land cover classification,” *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, 2003.

[41] H. Sharma and S. Kumar, “A survey on decision tree algorithms of classification in data mining,” *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, 2016.

[42] H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” 2014.

[43] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, “Hierarchical recurrent neural network for document modeling,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 899–907.

[44] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, “Independently recurrent neural network (indrnn): Building a longer and deeper rnn,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5457–5466.

[45] P. Venkatesan and S. Anitha, “Application of a radial basis function neural network for diagnosis of diabetes mellitus,” *Curr. Sci.*, vol. 91, no. 9, pp. 1195–1199, 2006.

[46] T. Xie, H. Yu, and B. Wilamowski, “Comparison between traditional neural networks and radial basis function networks,” in *2011 IEEE International Symposium on Industrial Electronics*, 2011, pp. 1194–1199.

[47] J.-D. Wu and J.-C. Liu, “A forecasting system for car fuel consumption using a radial basis function neural network,” *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1883–1888, 2012.

[48] A. Adeli and M. Neshat, “A fuzzy expert system for heart disease diagnosis,” in *Proceedings of international multi conference of engineers and computer scientists, Hong Kong*, 2010, vol. 1, pp. 28–30.

[49] M. A. Kadhim, M. A. Alam, and H. Kaur, “Design and implementation of fuzzy expert system for back pain diagnosis,” *Int. J. Innov. Technol. Creat. Eng.*, vol. 1, no. 9, pp. 16–22, 2011.

[50] P. Singhala, D. Shah, and B. Patel, “Temperature control using fuzzy logic,” *arXiv Prepr. arXiv1402.3654*, 2014.

[51] J.-S. Jang, “ANFIS: adaptive-network-based fuzzy inference system,” *IEEE Trans. Syst. Man. Cybern.*, vol. 23, no. 3, pp. 665–685, 1993.

[52] E. Joelianto and B. Rahmat, “Adaptive Neuro Fuzzy Inference System (ANFIS) with Error Back propagation Algorithm using Mapping Function,” *Int. J. Artif. Intell.*, vol. 1, no. A08, 2008.

[53] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.

[54] D. Chicco, N. Tötsch, and G. Jurman, “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” *BioData Min.*, vol. 14, no. 1, pp. 1–22, 2021.

Appendix

This part includes the results of the experiments that applied on both datasets, the national one is the Palestinian dataset PDS, and the international one is the Korean National Health and Nutritional Examination Survey KNHANES dataset. In addition to the supportive documents and official letters.

Appendix A

Appendix A section includes all the detailed experiments results that were performed on the Palestinian dataset PDS, including all used techniques such that Logistic Regression LR, Support Vector Machine SVM, Decision Tree DT, BP-ANN, RNN, RBFNN, Fuzzy Logic, and Neuro-Fuzzy model.

Table A.1: result of LDL-C classification using the BP-ANN

Accuracy1	Accuracy2	Accuracy3	Avg. Accuracy	Neurons	Iterations1	Iterations2	Iterations3
95.70	95.30	93.30	94.77	5	33	30	20
95.30	94.40	93.90	94.53	10	41	24	29
93.60	94.50	94.00	94.03	15	36	25	20
93.90	94.00	94.80	94.23	20	26	22	42
95.10	93.20	94.80	94.37	25	19	21	38
96.00	94.70	93.10	94.60	30	23	32	30
94.50	96.85	95.30	95.55	35	23	21	16
93.40	95.00	92.00	93.47	40	38	28	26
93.80	94.30	95.90	94.67	45	30	23	19
95.00	96.00	95.30	95.43	50	27	17	40

Table A.2: Confusion matrix of BP-ANN results in classifying LDL-C

Neurons	5	10	15	20	25	30	35	40	45	50
Sensitivity 1	98.60	98.40	97.40	97.00	97.30	97.50	99.30	98.30	98.10	97.50
Sensitivity 2	97.60	97.00	97.30	97.40	97.70	97.30	97.90	97.40	97.90	98.40
Sensitivity 3	98.10	97.50	97.10	97.20	97.80	97.70	98.00	98.40	98.40	97.80
Average Sensitivity	98.10	97.63	97.27	97.20	97.60	97.50	98.40	98.03	98.13	97.90
Specificity 1	75.00	77.80	72.00	77.00	82.40	86.70	68.80	75.40	70.80	80.30
Specificity 2	82.50	78.30	78.30	76.20	67.70	80.00	84.70	71.90	74.40	80.90
Specificity 3	67.20	71.70	74.10	79.10	78.60	96.90	80.20	58.60	81.50	77.90
Average Specificity	74.90	75.93	74.80	77.43	76.23	87.87	77.90	68.63	75.57	79.70
Precision 1	96.60	96.10	95.10	95.90	97.00	97.90	94.50	96.00	94.70	96.80
Precision 2	96.90	96.50	96.30	95.60	94.50	96.40	97.50	95.00	95.50	96.90
Precision 3	94.20	95.60	96.10	96.80	96.00	94.20	96.50	92.50	96.80	96.80
Average Precision	95.90	96.07	95.83	96.10	95.83	96.17	96.17	94.50	95.67	96.83
G-Mean	85.72	86.10	85.30	86.76	86.26	92.56	87.55	82.03	86.11	88.33
F-Measuring	96.99	96.84	96.54	96.65	96.71	96.83	97.27	96.23	96.88	97.36

Table A.3: result of HDL-C classification using the BP-ANN

Accuracy1	Accuracy2	Accuracy3	Avg. Accuracy	Neurons	Iterations1	Iterations2	Iterations3
88.50	89.30	91.30	89.70	5	95	33	84
92.50	92.00	90.90	91.80	10	74	63	27
87.80	91.90	92.00	90.57	15	83	89	26
90.00	90.30	88.90	89.73	20	39	62	80
89.30	89.80	89.40	89.50	25	57	22	73
92.20	89.20	90.80	90.73	30	28	58	70
90.80	92.20	91.60	91.53	35	85	90	58
90.80	91.40	92.60	91.60	40	69	68	50
91.60	92.60	88.70	90.97	45	58	57	22
89.90	90.00	90.30	90.07	50	71	77	20

Table A.4: Confusion matrix of BP-ANN results in classifying HDL-C

Neurons	5	10	15	20	25	30	35	40	45	50
Sensitivity 1	99.90	98.10	99.20	99.60	98.80	98.90	98.80	98.10	99.30	99.60
Sensitivity 2	99.60	98.80	98.50	98.90	99.20	99.60	98.40	99.50	98.80	99.60
Sensitivity 3	98.60	98.40	98.80	99.50	100.00	98.80	98.80	99.20	99.90	99.70
Average Sensitivity	99.37	98.43	98.83	99.33	99.33	99.10	98.67	98.93	99.33	99.63
Specificity 1	2.10	37.70	3.10	13.20	15.10	37.80	25.60	28.70	19.00	3.60
Specificity 2	3.40	25.00	34.90	15.30	17.00	6.50	24.60	13.00	28.80	0.00
Specificity 3	26.20	25.00	27.80	1.10	0.00	25.60	25.90	28.60	4.20	12.40
Average Specificity	10.57	29.23	21.93	9.87	10.70	23.30	25.37	23.43	17.33	5.33
Precision 1	88.50	93.80	88.50	90.20	90.10	92.80	91.50	92.10	92.00	90.20
Precision 2	89.60	92.80	92.80	91.00	90.30	89.40	93.50	91.70	93.40	90.40
Precision 3	92.20	92.00	92.80	89.40	89.40	91.50	92.40	93.10	88.80	90.40
Average Precision	90.10	92.87	91.37	90.20	89.93	91.23	92.47	92.30	91.40	90.33
G-Mean	32.40	53.64	46.56	31.31	32.60	48.05	50.03	48.15	41.49	23.05
F-Measuring	94.51	95.57	94.95	94.55	94.40	95.00	95.47	95.50	95.20	94.76

Table A.5: result of LDL-C prediction using the BP-ANN

MSE 1	MSE 2	MSE 3	Avg. MSE	Neurons	Iterations1	Iterations2	Iterations3
0.00052	0.00052	0.00052	0.00052	5	20	6	6
0.00048	0.00048	0.00052	0.00049	10	17	8	6
0.00047	0.00041	0.00040	0.00043	15	14	34	6
0.00036	0.00039	0.00035	0.00037	20	40	6	6
0.00038	0.00042	0.00042	0.00040	25	26	6	6
0.00036	0.00044	0.00038	0.00039	30	33	6	9
0.00035	0.00036	0.00038	0.00036	35	26	9	6
0.00037	0.00041	0.00037	0.00038	40	42	6	6
0.00032	0.00051	0.00054	0.00045	45	6	6	6
0.00042	0.00043	0.00036	0.00040	50	11	6	22

Table A.6: result of LDL-C prediction using the RNN

MSE 1	MSE 2	MSE 3	Avg. MSE	Layer 1 Neurons	Iterations1	Iterations2	Iterations3
0.000445	0.000412	0.000459	0.00044	5	85	6	6
0.000573	0.000388	0.000381	0.00045	10	14	57	20
0.000435	0.000455	0.000471	0.00045	15	16	6	13
0.000396	0.000376	0.000389	0.00039	20	56	6	6
0.000497	0.000509	0.000524	0.00051	25	22	6	6
0.000399	0.000394	0.000397	0.00040	30	62	6	6
0.000430	0.000522	0.000521	0.00049	35	19	7	6
0.000457	0.000473	0.000434	0.00045	40	15	6	20
0.000437	0.000509	0.000394	0.00045	45	13	6	46
0.000536	0.000536	0.000516	0.00053	50	28	6	6

Table A.7: result of LDL-C prediction using the Neuro-Fuzzy Grid Partitioning

Input Membership Function Type	Output Type	Optimization Technique	Iterations	MSE
Triangular	Constant	Hybrid	3	0.00054
Triangular	Constant	Back Propagation	3	0.00055
Triangular	Linear	Hybrid	3	0.00040
Triangular	Linear	Back Propagation	3	0.00076
Trapezoidal	Constant	Hybrid	3	0.00347
Trapezoidal	Constant	Back Propagation	3	0.00294
Trapezoidal	Linear	Hybrid	3	0.00045
Trapezoidal	Linear	Back Propagation	3	0.00046
Gaussian	Constant	Hybrid	3	0.00073
Gaussian	Constant	Back Propagation	3	0.00084
Gaussian	Linear	Hybrid	3	0.00031
Gaussian	Linear	Back Propagation	3	0.00598
Sigmoidal	Constant	Hybrid	3	0.00336
Sigmoidal	Constant	Back Propagation	3	0.00333
Sigmoidal	Linear	Hybrid	3	0.00042
Sigmoidal	Linear	Back Propagation	3	0.00042

Table A.8: result of LDL-C prediction using the Neuro-Fuzzy Sub Clustering

Range of Influence	Squash Factor	Accept Ratio	Reject Ratio	MSE	Iteration	Optimization Technique
0.25	1.25	0.5	0.15	0.000568	100	Hybrid
0.25	1.25	0.5	0.15	0.000549	100	Back Propagation
0.5	1.25	0.5	0.15	0.000598	100	Hybrid
0.5	1.25	0.5	0.15	0.000606	100	Back Propagation
0.75	1.25	0.5	0.15	0.000579	100	Hybrid
0.75	1.25	0.5	0.15	0.000667	100	Back Propagation
1	1.25	0.5	0.15	0.000581	100	Hybrid
1	1.25	0.5	0.15	0.000644	100	Back Propagation

Table A.9: result of HDL-C prediction using the BP-ANN

MSE 1	MSE 2	MSE 3	Avg. MSE	Neurons	Iterations1	Iterations2	Iterations3
0.00044	0.00059	0.00085	0.00063	5	98	6	6
0.00059	0.00060	0.00045	0.00054	10	15	6	14
0.00069	0.00070	0.00069	0.00069	15	20	7	6
0.00069	0.00048	0.00072	0.00063	20	19	6	7
0.00067	0.00066	0.00048	0.00060	25	43	16	6
0.00073	0.00050	0.00070	0.00064	30	19	6	25
0.00070	0.00069	0.00063	0.00067	35	18	6	37
0.00062	0.00041	0.00041	0.00048	40	52	6	6
0.00067	0.00067	0.00048	0.00061	45	23	6	6
0.00074	0.00064	0.00040	0.00059	50	8	42	6

Table A.10: result of LDL-C prediction using the RNN

MSE 1	MSE 2	MSE 3	Avg. MSE	Layer 1 Neurons	Iterations1	Iterations2	Iterations3
0.00092	0.00063	0.00066	0.00074	5	16	72	10
0.00067	0.00042	0.00040	0.00050	10	61	6	6
0.00044	0.00060	0.00065	0.00056	15	36	76	6
0.00067	0.00046	0.00068	0.00060	20	63	6	6
0.00065	0.00070	0.00067	0.00067	25	25	6	7
0.00078	0.00062	0.00062	0.00067	30	9	96	6
0.00069	0.00064	0.00064	0.00066	35	12	55	6
0.00053	0.00046	0.00050	0.00050	40	12	20	6
0.00041	0.00063	0.00039	0.00047	45	35	6	6
0.00068	0.00069	0.00071	0.00069	50	13	6	6

Table A.11: result of HDL-C prediction using the Neuro-Fuzzy Grid Partitioning

Input Membership Function Type	Output Type	Optimization Technique	Iterations	MSE
Triangular	Constant	Hybrid	3	0.00079
Triangular	Constant	Back Propagation	3	0.00080
Triangular	Linear	Hybrid	3	0.00063
Triangular	Linear	Back Propagation	3	0.00073
Trapezoidal	Constant	Hybrid	3	0.00122
Trapezoidal	Constant	Back Propagation	3	0.00123
Trapezoidal	Linear	Hybrid	3	0.00052
Trapezoidal	Linear	Back Propagation	3	0.00075
Gaussian	Constant	Hybrid	3	0.00087
Gaussian	Constant	Back Propagation	3	0.00090
Gaussian	Linear	Hybrid	3	0.00062
Gaussian	Linear	Back Propagation	3	0.01225
Sigmoidal	Constant	Hybrid	3	0.00119
Sigmoidal	Constant	Back Propagation	3	0.00120
Sigmoidal	Linear	Hybrid	3	0.00071
Sigmoidal	Linear	Back Propagation	3	0.00072

Table A.12: result of HDL-C prediction using the Neuro-Fuzzy Sub Clustering

Range of Influence	Squash Factor	Accept Ratio	Reject Ratio	MSE	Iteration	Optimization Technique
0.25	1.25	0.5	0.15	0.000776	100	Hybrid
0.25	1.25	0.5	0.15	0.000733	100	Back Propagation
0.5	1.25	0.5	0.15	0.000891	100	Hybrid
0.5	1.25	0.5	0.15	0.000911	100	Back Propagation
0.75	1.25	0.5	0.15	0.000845	100	Hybrid
0.75	1.25	0.5	0.15	0.001209	100	Back Propagation
1	1.25	0.5	0.15	0.000865	100	Hybrid
1	1.25	0.5	0.15	0.002580	100	Back Propagation

Appendix B

Appendix B section includes all the experiments detailed results that performed on the dataset of Korean National Health and Nutritional Examination Survey KNHANES, including all used techniques such that Logistic Regression LR, Support Vector Machine SVM, Decision Tree DT, BP-ANN, RNN, RBFNN, Fuzzy Logic, and Neuro-Fuzzy model.

Table B.1: result of LDL-C classification using the BP-ANN

Accuracy1	Accuracy2	Accuracy3	Avg. Accuracy	Neurons	Iterations1	Iterations2	Iterations3
97.10	93.40	94.20	94.90	5	17	30	20
93.40	96.40	95.60	95.13	10	41	24	29
97.10	97.10	97.10	97.10	15	36	25	20
94.90	96.40	94.20	95.17	20	26	22	42
94.90	94.20	87.60	92.23	25	19	21	38
97.10	97.80	94.20	96.37	30	23	32	30
92.70	91.20	95.60	93.17	35	23	21	16
95.60	97.80	94.90	96.10	40	38	28	26
94.20	92.70	94.90	93.93	45	30	23	19
92.70	92.00	95.60	93.43	50	27	17	40

Table B.2: Confusion matrix of BP-ANN results in classifying LDL-C

Neurons	5	10	15	20	25	30	35	40	45	50
Sensitivity 1	99.20	98.40	100.00	99.20	99.20	99.20	99.20	99.20	96.70	95.20
Sensitivity 2	96.80	99.20	97.60	100.00	98.40	100.00	95.00	99.20	94.40	99.20
Sensitivity 3	96.00	96.80	97.60	99.20	95.00	98.40	98.40	98.30	99.20	100.00
Average Sensitivity	97.33	98.13	98.40	99.47	97.53	99.20	97.53	98.90	96.77	98.13
Specificity 1	75.00	53.30	63.60	60.00	50.00	57.10	50.00	61.50	73.30	66.70
Specificity 2	54.50	76.50	90.00	58.30	60.00	70.00	64.70	81.80	72.70	33.30
Specificity 3	76.90	84.60	90.00	56.30	31.30	25.00	69.20	70.60	60.00	60.00
Average Specificity	68.80	71.47	81.20	58.20	47.10	50.70	61.30	71.30	68.67	53.33
Precision 1	97.60	94.50	96.90	95.30	95.40	97.70	92.90	96.10	96.70	96.70
Precision 2	96.10	96.70	99.20	96.20	95.20	97.70	95.00	98.40	97.50	92.40
Precision 3	97.50	98.40	99.20	94.50	91.30	95.50	96.80	95.90	95.30	95.30
Average Precision	97.07	96.53	98.43	95.33	93.97	96.97	94.90	96.80	96.50	94.80
G-Mean	81.83	83.75	89.39	76.09	67.78	70.92	77.32	83.97	81.51	72.34
F-Measuring	97.20	97.33	98.42	97.36	95.72	98.07	96.20	97.84	96.63	96.44

Table B.3: result of HDL-C classification using the BP-ANN

Accuracy1	Accuracy2	Accuracy3	Avg. Accuracy	Neurons	Iterations1	Iterations2	Iterations3
96.40	91.20	97.10	94.90	5	9	13	62
98.50	97.10	97.80	97.80	10	17	8	8
95.60	97.10	97.10	96.60	15	28	13	35
94.20	97.10	95.60	95.63	20	21	17	46
97.10	94.90	95.60	95.87	25	18	13	28
94.20	95.60	96.40	95.40	30	35	42	24
96.40	93.40	96.40	95.40	35	13	9	15
95.60	93.40	96.40	95.13	40	21	16	34
95.60	94.90	94.90	95.13	45	24	26	26
96.40	94.20	93.40	94.67	50	36	28	19

Table B.4: Confusion matrix of BP-ANN results in classifying HDL-C

Neurons	5	10	15	20	25	30	35	40	45	50
Sensitivity 1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Sensitivity 2	100.00	100.00	100.00	99.30	100.00	99.20	100.00	100.00	99.20	100.00
Sensitivity 3	100.00	99.30	100.00	100.00	97.80	100.00	100.00	100.00	100.00	97.70
Average Sensitivity	100.00	99.77	100.00	99.77	99.27	99.73	100.00	100.00	99.73	99.23
Specificity 1	0.00	0.00	0.00	0.00	0.00	0.00	16.70	0.00	0.00	0.00
Specificity 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.30	0.00
Specificity 3	0.00	0.00	0.00	0.00	0.00	0.00	16.70	0.00	0.00	0.00
Average Specificity	0.00	0.00	0.00	0.00	0.00	0.00	11.13	0.00	4.77	0.00
Precision 1	96.40	98.50	95.60	94.20	97.10	94.20	96.30	95.60	95.60	96.40
Precision 2	91.20	97.10	97.10	97.80	94.90	96.30	93.40	93.40	95.60	94.20
Precision 3	97.10	98.50	97.10	95.60	97.80	96.40	96.30	96.40	94.90	95.50
Average Precision	94.90	98.03	96.60	95.87	96.60	95.63	95.33	95.13	95.37	95.37
G-Mean	0.00	0.00	0.00	0.00	0.00	0.00	33.37	0.00	21.80	0.00
F-Measuring	97.38	98.89	98.27	97.78	97.92	97.64	97.61	97.51	97.50	97.26

Table B.5: result of LDL-C prediction using the BP-ANN

MSE 1	MSE 2	MSE 3	Avg. MSE	Neurons	Iterations1	Iterations2	Iterations3
0.00143	0.00160	0.00158	0.00154	5	31	6	6
0.00146	0.00149	0.00161	0.00152	10	18	6	6
0.00120	0.00130	0.00135	0.00128	15	14	6	7
0.00122	0.00135	0.00140	0.00132	20	16	6	6
0.00104	0.00122	0.00116	0.00114	25	23	7	6
0.00140	0.00133	0.00147	0.00140	30	12	6	6
0.00113	0.00120	0.00128	0.00120	35	12	6	6
0.00081	0.00085	0.00096	0.00087	40	15	7	6
0.00129	0.00090	0.00130	0.00116	45	9	7	6
0.00104	0.00113	0.00112	0.00110	50	12	6	6

Table B.6: result of LDL-C prediction using the RNN

MSE 1	MSE 2	MSE 3	Avg. MSE	Layer 1 Neurons	Iterations1	Iterations2	Iterations3
0.00307	0.00320	0.00191	0.00273	5	23	6	6
0.00148	0.00143	0.00133	0.00141	10	23	6	6
0.00160	0.00152	0.00152	0.00155	15	11	6	6
0.00136	0.00141	0.00137	0.00138	20	12	6	7
0.00133	0.00152	0.00137	0.00141	25	13	6	6
0.00147	0.00146	0.00143	0.00145	30	12	7	6
0.00146	0.00151	0.00158	0.00152	35	10	6	6
0.00139	0.00162	0.00154	0.00152	40	9	6	9
0.00148	0.00150	0.00148	0.00149	45	10	6	6
0.00131	0.00146	0.00138	0.00138	50	12	6	6

Table B.7: result of LDL-C prediction using the Neuro-Fuzzy Grid Partitioning

Input Membership Function Type	Output Type	Optimization Technique	Iterations	MSE
Triangular	Constant	Hybrid	3	0.00111
Triangular	Constant	Back Propagation	3	0.18664
Triangular	Linear	Hybrid	3	0.00051
Triangular	Linear	Back Propagation	3	0.00052
Trapezoidal	Constant	Hybrid	3	0.00351
Trapezoidal	Constant	Back Propagation	3	0.18429
Trapezoidal	Linear	Hybrid	3	0.00005
Trapezoidal	Linear	Back Propagation	3	0.00011
Gaussian	Constant	Hybrid	3	0.00135
Gaussian	Constant	Back Propagation	3	0.00145
Gaussian	Linear	Hybrid	3	0.00037
Gaussian	Linear	Back Propagation	3	0.00041
Sigmodal	Constant	Hybrid	3	0.00408
Sigmodal	Constant	Back Propagation	3	0.00456
Sigmodal	Linear	Hybrid	3	0.00010
Sigmodal	Linear	Back Propagation	3	0.00015

Table B.8: result of LDL-C prediction using the Neuro-Fuzzy Sub Clustering

Range of Influence	Squash Factor	Accept Ratio	Reject Ratio	MSE	Iteration	Optimization Technique
0.25	1.25	0.5	0.15	0.00023	100	Hybrid
0.25	1.25	0.5	0.15	0.00041	100	Back Propagation
0.5	1.25	0.5	0.15	0.00149	100	Hybrid
0.5	1.25	0.5	0.15	0.00171	100	Back Propagation
0.75	1.25	0.5	0.15	0.00192	100	Hybrid
0.75	1.25	0.5	0.15	0.00205	100	Back Propagation
1	1.25	0.5	0.15	0.00191	100	Hybrid
1	1.25	0.5	0.15	0.00204	100	Back Propagation

Table B.9: result of HDL-C prediction using the BP-ANN

MSE 1	MSE 2	MSE 3	Avg. MSE	Neurons	Iterations1	Iterations2	Iterations3
0.0079	0.0081	0.0085	0.0082	5	15	14	16
0.0079	0.0070	0.0057	0.0069	10	15	19	20
0.0063	0.0062	0.0069	0.0065	15	17	11	11
0.0058	0.0066	0.0060	0.0061	20	11	9	9
0.0053	0.0058	0.0062	0.0058	25	12	11	10
0.0057	0.0055	0.0058	0.0057	30	10	9	8
0.0053	0.0050	0.0053	0.0052	35	11	11	8
0.0054	0.0048	0.0071	0.0057	40	9	10	10
0.0046	0.0051	0.0049	0.0049	45	9	10	10
0.0051	0.0051	0.0044	0.0049	50	9	9	10

Table B.10: result of LDL-C prediction using the RNN

MSE 1	MSE 2	MSE 3	Avg. MSE	Neurons	Iterations1	Iterations2	Iterations3
0.0086	0.0089	0.0086	0.0087	5	14	9	6
0.0080	0.0082	0.0083	0.0082	10	15	7	6
0.0066	0.0082	0.0069	0.0072	15	11	6	6
0.0072	0.0084	0.0071	0.0076	20	18	6	6
0.0078	0.0081	0.0077	0.0079	25	12	6	6
0.0080	0.0069	0.0063	0.0070	30	12	6	12
0.0060	0.0056	0.0058	0.0058	35	17	7	6
0.0062	0.0074	0.0064	0.0067	40	20	6	6
0.0063	0.0058	0.0072	0.0064	45	11	6	6
0.0069	0.0068	0.0061	0.0066	50	14	6	6

Table B.11: result of HDL-C prediction using the Neuro-Fuzzy Grid Partitioning

Input Membership Function Type	Output Type	Optimization Technique	Iteration	MSE
Triangular	Constant	Hybrid	3	0.0054
Triangular	Constant	Back Propagation	3	0.0055
Triangular	Linear	Hybrid	3	0.0027
Triangular	Linear	Back Propagation	3	0.0027
Trapezoidal	Constant	Hybrid	3	0.0081
Trapezoidal	Constant	Back Propagation	3	0.0081
Trapezoidal	Linear	Hybrid	3	0.0004
Trapezoidal	Linear	Back Propagation	3	0.0006
Gaussian	Constant	Hybrid	3	0.0058
Gaussian	Constant	Back Propagation	3	0.0062
Gaussian	Linear	Hybrid	3	0.0020
Gaussian	Linear	Back Propagation	3	0.0023
Sigmoidal	Constant	Hybrid	3	0.0081
Sigmoidal	Constant	Back Propagation	3	0.0090
Sigmoidal	Linear	Hybrid	3	0.0010
Sigmoidal	Linear	Back Propagation	3	0.0013

Table B.12: result of HDL-C prediction using the Neuro-Fuzzy Sub Clustering

Range of Influence	Squash Factor	Accept Ratio	Reject Ratio	MSE	Iteration	Optimization Technique
0.25	1.25	0.5	0.15	0.0007	100	Hybrid
0.25	1.25	0.5	0.15	0.0010	100	Back Propagation
0.50	1.25	0.5	0.15	0.0078	100	Hybrid
0.50	1.25	0.5	0.15	0.0076	100	Back Propagation
0.75	1.25	0.5	0.15	0.0095	100	Hybrid
0.75	1.25	0.5	0.15	0.0097	100	Back Propagation
1.00	1.25	0.5	0.15	0.0094	100	Hybrid
1.00	1.25	0.5	0.15	0.0095	100	Back Propagation

Appendix C

Appendix C section includes all supportive documents such as the official letters submitted to the Palestinian Ministry of Health, the Local Hospitals, and for the data collection.

Arab American University
Faculty of Graduate Studies
الجامعة العربية الأمريكية
كلية الدراسات العليا

2020-12-3

الى من يهمه الامر

تسهيل مهمة بحثية

تحية طيبة وبعد،

تهديكم كلية الدراسات العليا في الجامعة العربية الامريكية اطيب التحيات، وبالإشارة الى الموضوع أعلاه، تشهد كلية الدراسات العليا في الجامعة أن الطالب **سند احمد انيس ملايشة** والذي يحمل الرقم الجامعي **20710492** هو طالب ماجستير في برنامج علم الحاسوب ويعمل حالياً على دراسة علمية بعنوان:

"التصنيفات والتنبوءات للدهون (LDL) للمرضى الفلسطينين باستخدام تقنية التعليم الالي "تحت اشراف الدكتور احمد حدرب، نأمل من حضرتكم الايعاز لمن يلزم لمساعدته للحصول على المعلومات اللازمة للدراسة علماً ان المعلومات ستستخدم لغاية البحث فقط وسيتم التعامل معها بغاية السرية، وقد أعطي هذه الرسالة بناءً على طلبه.

وتفضلوا بقبول فائق الاحترام

عميد كلية الدراسات العليا

د. اشرف الميمي

كلية الدراسات العليا
FACULTY OF GRADUATE STUDIES

Jenin Tel: +970-4-2418888 Ext.:1471,1472 Fax: +970-4-2510810 P.O. Box:240
Ramallah Tel: +970-2-2941999 Fax: +970-2-2941979 Abu Qash - Near Alrehan
E-mail: FGS@aaup.edu ; PGS@aaup.edu Website: www.aaup.edu

State of Palestine
Ministry of Health - Nablus
General Directorate of Education in Health

دولة فلسطين
وزارة الصحة– نابلس
الإدارة العامة للتعليم الصحي

Ref.: .....
Date:.....

الرقـــم: ٢٠٢١/ ١٢٥ /١٢٤
التـــاريخ: ٢٠٢١/١١/٢٥

الأخ مدير عام الادارة العامة للمستشفيات المحترم،،،
الأخ مدير عام وحدة الهندسة والحاسوب المحترم،،،
تحية واحترام،،،

الموضـــوع: تسهيل مهمة بحث

لاحقاً لموافقة معالي وزيرة الصحة، يرجى تسهيل مهمة الطالب: **سند احمد انيس ملايشة،** ماجستير علم الحاسوب، الجامعة العربية الامريكية، لعمل بحث رسالة الماجستير بعنوان:

"**التصنيفات والتنبوءات للدهون (LDL) للمرضى الفلسطينيين باستخدام تقنية التعليم الآلي**"، وذلك بالسماح له بالحصول على المعلومات المطلوبة من ملفات المرضى وذلك بالتنسيق مع مسؤولي قسم الحاسوب بحيث لا يتم التعرض للمعلومات التعريفية للمرضى، وذلك في:

- **مستشفى جنين**

على ان يتم الالتزام بجميع تعليمات واجراءات الوقاية والسلامة الصادرة عن وزارة الصحة بخصوص جائحة كورونا، وتحت طائلة المسؤولية.

وان يتم تزويدنا بنسخة من نتائج الدراسة والتعهد بعدم النشر.

وتقبلوا فائق الاحترام،،،

د. عبد الله القواسمي
مدير التعليم الصحي

دولة فلسطين وزارة الصحة
الإدارة العامة للتعليم الصحي

نسخة: مشرفة الدراسة المحترم/ الجامعة العربية الامريكية

Data Availability Check Form

Note:

Please use the third column named "**Available (Yes\No)**" to indicate if the field exists and has values in the patient medical record with **"Yes"**, and "**No**" if not exists.

Data Field Name	Importance	Available (Yes\No)
Height	**Required**	Yes
Weight	**Required**	yes
Age	**Required**	yes
Gender	**Required**	yes
Body Mass Index BMI	**Required**	no
Sagittal abdominal diameter to height ratio SADHtR	**Required**	no
Waist to height ratio WHtR	**Required**	no
Body fat percentage BF%	**Required**	no
Diastolic blood pressure DBP	**Required**	yes
Systolic blood pressure SBP	**Required**	yes
Total Cholesterol TCH	**Required**	yes
Low Density Lipoprotein LDL	**Required**	yes
High Density Lipoprotein HDL	**Required**	yes
Triglyceride TG	**Required**	yes
Very Low-Density Lipoprotein VLDL	*Optional*	yes
Alanine transaminase ALT	*Optional*	yes
Aspartate transaminase AST	*Optional*	yes
Alkaline phosphatase ALP	*Optional*	yes
Albumin	*Optional*	yes
Bilirubin	*Optional*	yes
Gamma-glutamyl transferase GGT	*Optional*	yes
L-lactate dehydrogenase LD	*Optional*	yes
Prothrombin time PT	*Optional*	yes
Cardio-ankle vascular index CAVI	*Optional*	yes
Heart Rate	*Optional*	yes
Asthma diagnosis	*Optional*	yes
Diabetes mellitus	*Optional*	yes
Fasting glucose	*Optional*	yes
Chronic kidney disease	*Optional*	no
SBP CAVI measurement	*Optional*	yes
Marital Status	*Optional*	yes
Smoking cessation	*Optional*	yes
Alcohol use	*Optional*	yes

الملخص

يعتبر مرض الكوليسترول واحد من الأسباب الرئيسية في الوفيات محليا في فلسطين وعلى مستوى العالم، وذلك بنسبة تصل الى 31.5% محليا و31.4% عالميا. يتكون الكوليسترول من أربعة قيم أساسية وهي الكوليسترول الكلي TCH، الدهون الثلاثية TG، كوليسترول البروتين الدهني منخفض الكثافة LDL-C، وكوليسترول البروتين الدهني مرتفع الكثافة HDL-C. القيمة الرئيسية المؤثرة والمسببة لأمراض القلب والاوعية الدموية هي LDL-C التي تسمى الكوليسترول السيء، لأنه يتراكم على الاسطح الداخلية للشرايين مسببا تضيقها وبطئ وصول الدم المغذي الى القلب والدماغ، بالتالي هذا يؤدي الى النوبات القلبية والسكتات الدماغية.

التعلم الالي ML أصبح لاعب رئيسي في دعم كشف وتشخيص الكوليسترول منخفض الكثافة LDL-C، يتم ذلك بالاعتماد على بيانات الملفات الطبية للمرضى المشخصين بارتفاع الكوليسترول، ويتم توصيف هذا الحالي عندما تتخطى قيمة LDL-C المستوى المقبول والصحي الذي هو قياس 160 mg/dL. تصنيف والتنبؤ بقيمة LDL-C باستخدام أدوات التعلم الالي ML سوف يضمن تشخيص وتقدير دقيق لهذا المرض، متجنبا الأخطاء البشرية، والتكاليف المخبرية والوقت في انتظار نتائج الفحص، لذلك التعلم الالي لديه تأثير إيجابي كبير على تشخيص المرض وبالتالي العلاج. ولهذا الهدف استخدمت هذه الاطروحة أدوات التعلم الالي في التصنيف والتنبؤ لقيمة LDL-C، بالإضافة الى تطبيقها على قيمة HDL-C. من هذه الأدوات المستخدمة الشبكات العصبونية الاصطناعية ANNs بنسخها المختلفة مثل RNN وRBFNN، المنطق الضبابي Fuzzy Logic، شجرة اتخاذ القرار Decision Tree، Support Vector Machine SVM، اقتران الانحدار اللوجستي Logistic Regression، ونماذج تدمج بين ANNs والمنطق الضبابي Fuzzy Logic فيما يسمى نموذج Neuro-Fuzzy لتحسين دقة النتائج وتقليل نسبة الخطأ. هذه الأدوات المذكورة تحتاج الى مجموعة بيانات عينة من المرضى لتدريبها واختبارها، وبما ان الاطروحة تستهدف المجتمع الفلسطيني فانه تم التعاون مع وزارة الصحة الفلسطينية للحصول على البيانات المطلوبة من حقول وسجلات للمرضى والتي قاموا بتزويدها من عوامل مؤثرة وبيانات مرضى كوليسترول فلسطينيين. كما تم تدعيم الاطروحة بمجموعة بيانات أخرى عالمية تم جمعها في كوريا الشمالية بترتيب من استبيان الفحوصات الغذائية والصحية الكورية الوطنية KNHANES، ذلك لتعميم النتائج ومقارنتها بالجهود العالمية في مجال البحث. الأدوات والبيانات المستخدمة في هذه الاطروحة اثبتت تفوق في الدقة على الجهود الأخرى السابقة

التي كانت نسبتها 80% في البحث المعنون " Lipid Profile Prediction Based on Artificial Neural Networks" حيث انه استخدم حقول بيانات اقل متعلقة فقط بقياسات فحص الكوليسترول، بينما في هذه الاطروحة وصلت نسبة الدقة الى 97.10% في البينات الكورية و95.55% في البيانات الفلسطينية. هذه الاطروحة تقترح وتدعم باتجاه استخدام حقول أكثر تأثيرا والتي لا تشمل إجراءات جراحية للحصول على نتائج ادق، اقل تكلفة ووقت.