# Prediction of Potential-Diabetic Obese-Patients using Machine Learning Techniques

Raghda Essam Ali[1], Hatem El-Kadi[2], Soha Safwat Labib[3], Yasmine Ibrahim Saad[4]
Teaching Assistant at Faculty of Computer Science, MSA University
Faculty of Computers and Artificial Intelligence, Giza, Egypt[1]
Associate Professor at Faculty of Computers and Artificial Intelligence, Giza, Egypt[2]
Associate Professor in Computer Science, Cairo University, Cairo, Egypt[3]
Associate Professor at Endemic Medicine, Hepatogastroenterology and Clinical Nutrition
Faculty of Medicine, Cairo University, Giza, Egypt[4]

***Abstract*—Diabetes is a disease that is chronic. Improper blood glucose control may cause serious complications in diabetic patients as heart and kidney disease, strokes, and blindness. Obesity is considered to be a massive risk factor of type 2 diabetes. Machine Learning has been applied to many medical health aspects. In this paper, two machine learning techniques were applied; Support Vector Machine (SVM) and Artificial Neural Network (ANN) to predict diabetes mellitus. The proposed techniques were applied on a real dataset from Al-Kasr Al-Aini Hospital in Giza, Egypt. The models were examined using four-fold cross validation. The results were conducted from two phases in which forecasting patients with fatty liver disease using Support Vector Machine in the first phase reached the highest accuracy of 95% when applied on 8 attributes. Then, Artificial Neural Network technique to predict diabetic patients were applied on the output of phase 1 and another different 8 attributes to predict non-diabetic, pre-diabetic and diabetic patients with accuracy of 86.6%.**

***Keywords—Obesity; diabetes; nonalcoholic fatty liver disease; artificial neural network; support vector machine***

## I. Introduction

Applying Machine Learning (ML) and Data Mining (DM) techniques in data mining studies are a main approach for using big quantities of accessible knowledge-based diabetes information. DM is one of the top priorities in science and medicine studies, this inevitably produces enormous quantities of data, due to the specific social impact of the effect of the severe disease. Consequently, without a doubt, for elements of clinical administration, diagnosis and management ML and DM techniques are of excellent interest. As a result, attempts were made to review the current literature on machine learning and approaches of data mining in diabetes research as part of this study.

Globally, obesity and diabetes have become huge public health problems, both associated multifactorial, complicated diseases [1]. However, many conditions can actually be avoided. Obesity is a notable increasing health issue; some call this the New World Syndrome [2]. It is described as an unusual or excessive accumulation of fat that poses a health danger.

Over 1.9 billion teenagers, 18 years of age and older, were overweight, more than 650 million of these were obese, in 2016 [3]. For non-communicable diseases as: cardiovascular illnesses, diabetes, musculoskeletal disorders, and types of cancers [4], it is a significant risk factor. Weight gain and body mass are essential to type 1 and type 2 diabetes development and increased incidence.

The definition of overweight and obesity is an extraordinary or excessive accumulation of fat that poses a health danger. The latest CDC (Center for Disease Control and Prevention) study demonstrates that the age-adjusted incidence of diagnosed diabetes increased dramatically from 3.5 to 6.6 for every 1000 population from 1980 and 2014 [5].

The Body Mass Index (BMI) [3] is a straightforward height to weight index usually used for adult's classification to be either underweight, overweight or obese. *'Overweight*' means a body mass index (BMI) of 25-29.9 kg/m² and *'Obese'* means a BMI of greater than 30 kg/m².

Overweight and obesity are powerful risk factors for type 2 diabetes and contribute significantly to precocious death. These metabolic disorders in the Eastern Mediterranean region are growing rapidly among adolescents. Adult data all over 16 countries in the Mediterranean region from age of 15 years and older show the highest levels of overweight and obesity such as in Egypt, Bahrain, Jordan, Kuwait, Saudi Arabia and the United Arab Emirates [5].

Diabetes mellitus is a one of the chronic diseases that is characterized by hyperglycemia. It can trigger a lot of complications [6]. As a result of the increasing mortality in the latest years, in 2040, the world's diabetic patients will reach 642 million [7], this means that there will be one adult per each ten adults suffers from diabetes in the future according to the WHO (World Health Organization) statistics.

This frightening estimate must undoubtedly be faced. Diabetes can cause chronic harm and abnormality or impairment in the function of a specified bodily organ or different tissues including eyes, heart, nerves, kidneys and blood vessels [8].

Diabetes is subdivided into two classifications, Type one Diabetes (T1D) and Type two Diabetes (T2D) [9]. T1D patients usually are younger, mainly under the age of 30 years. The common symptoms for these patients may include accelerated thirst, frequently urinated, high levels of glucose in blood [10]. This kind of diabetes cannot really be efficiently healed by using only oral drugs but also by using insulin treatment which considered to be very necessary. T2D which are usually linked with obesity, hypertension, fatty liver, dyslipidemia, arteriosclerotic and other diseases, is more frequently present in the mid-aged and elderly humans.

Diabetes can be diagnosed by evaluating glycated hemoglobin (HbA1c) taken from a blood sample. If the HbA1c reveals ⩾ 48 mmol/mol (6.5%) diagnosis may be alleged. HbA1c glycation is a measure of the plasma glucose concentration level and is used for both diagnosis and diabetes surveillance. HbA1c represents the mean plasma glucose of a patient in such a long period of time "*about three months*".

Nonalcoholic fatty liver disease (NAFLD) is frequently recorded in patients T2D [11], which has been proposed as a leading cause for NAFLD progression, or nonalcoholic steatohepatitis, probably reflect the quick succession of obesity and resistance to insulin in T2D.

Metabolic syndrome becomes more and more prevalent. It happens when there are a combination of a number of metabolic risk variables, such as obesity and insulin resistance. The risk of developing T2D increased by metabolic syndrome. Most usually, overweight individuals who are pre-diabetic or T2D produces much more insulin than nondiabetic individuals due to the greater bodily fat-muscle proportion. The possible explanation is that the body cannot use its insulin efficiently enough, which results in insulin resistance. It is therefore logical for the body to generate more insulin to offset. Furthermore, the growing quantity of insulin in the body progressively makes the body more resistant, it may also be seen as a comparable method of developing tolerances to drugs for drug users.

In order to diagnose an individual to be metabolism, Three out of Five requirements must be fulfilled: The elements of the Metabolic Syndrome, according to the WHO proposal [12] are: (1) *in abdominal obesity*: waist circumference of men ⩾ 102 cm and ⩾ 88 cm in women, (2) hypertriglyceridemia is greater than or equal to 150 mg/dl (1.695 mmol/L), (3) low HDL-C in men is less than 40 mg/dL (1.04 mmol/dL) and less than 50 mg/dL (1.30 mmol/dL) in women, (4) high blood pressure (BP) of greater than 130/85 mmHg and (5) high fasting glucose of more than 110 mg/dl (6.1 mmol/L).

Nonalcoholic Fatty Liver (NAFLD) could be classified as an added metabolic syndrome feature [13] with a certain resistance to hepatic insulin. The presence of fatty liver in patients with T2D and obesity has long been reported. Fatty liver has long been recorded in patients with type 2 diabetes and obesity [14]. It is generally regarded an incidental discovery, with little or no clinical significance. Sedentary lifestyle and bad nutritional habits contribute to weight gain and the chance of developing the metabolic syndrome and nonalcoholic fatty liver will increase eventually.

The importance of applying the proposed method for prediction of the diabetic patients who are already affected with both nonalcoholic fatty liver disease and obesity will help in minimizing the huge complications that results in an enormous health problems as an early diagnosis is the starting point for a successful living without the disease, it will also encourage and promote efficient interventions to monitor, prevent and manage diabetes mellitus disease and its complications in low and middle income nations, in particular.

The paper is organized as follows: Section 2 provides the required machine learning background understanding, Section 3 is divided into 3 subsections present the proposed system, Machine Learning techniques used and the methodological approach adopted, Section 4 provides the evaluation methods followed for valuation, Section 5 showed the proposed system results, Section 6 provides the conclusions.

## II. Background

### A. Machine Learning

The term "Machine Learning" is for many scientists identical to that of "Artificial Intelligence" [15], as the possibility of learning is the principal feature of an entity called intelligence. Machine learning is designed to build computer systems that can accommodate and benefit from their knowledge.

Knowledge discovery in database (KDD) [16] is an area that includes theories, methods and techniques, which attempt to create sense of information and derive helpful and valuable knowledge from it. The most significant stage in the KDD method is data mining, which idealize the implementation of machine learning algorithms in the analysis of data. It is regarded a multi-stepping process (selection, preprocessing, conversion, data mining interpretation and evaluation).

### B. Machine Learning Types

The mining of data utilizes a variety of machine learning techniques to find hidden data patterns. These techniques are classified into three major categories: supervised learning techniques, semi-supervised and unsupervised learning techniques [16]. Physicians can use expert systems that are developed through machine learning techniques to help them easily diagnose and predict diseases given the importance of diseases diagnosis for humans, various studies on the classification methodologies have been conducted.

### C. ML Applications on Medical Data

Medical diagnosis is an optimal field for algorithms of ML [17]. Many of them are recognized by patterns recognition on big quantities of data. To be effective in the field, an algorithm must be prepared, on comparatively few medical tests, to manage noisy and empty records of data.

Many studies were conducted in the area of machine learning in healthcare. Healthcare machine learning becomes one of the most researchers' priority. Insights can be obtained using different DM techniques and methods in hidden patterns recognition. These insights can also be used for forecasting of diseases and epidemics.

Kumar [18], showed that the goal of the different data mining techniques in health care systems is to highlight applications of data mining in healthcare depending on the nature of the dataset; as Artificial Neural Network and Support Vector Machine were applied in predicting

Parkinson’s disease with accuracy of 95%. And using statistical Neural Network in diagnosing breast cancer disease to improve the detection rate by 98.8%, and applied ANN, Multiple Association rule and immature bayed in predicting the heart disease.

Measuring the performance of DM techniques in healthcare prediction by applying multiple learning techniques (Basma Boukenze Hajar Mausannif & Abdelkrim Haqiq [19]); Decision tree, SVM and ANN, simulation of results showed that decision tree proved its performance in predicting chronic kidney failure disease than other learning techniques.

Also, when applying the same data mining techniques to specify the anemia type for anemic patients (M. Abdullah and S. Al-Asmari [20]), Decision Tree performs the best with accuracy result 93.75%. While using only SVM in classifying diabetes disease (Kumari and Chitra [21]), using Matlab 2010a tool to detect diabetes disease with accuracy of 78%.

Building decision tree and classification data mining methods help health care providers making better clinical decisions to identify chronic diabetes in early phases [22].

El-Halees and Shurrab [23], generated a model that can distinguish patients with normal blood disease from those who have blood tumor by using Multiple Association rules, classification techniques and ANN that resulted in accuracy of 79.45%.

## III. Proposed System

### A. Data Set

The dataset utilized were acquired form Al-Kasr Al-Aini, Faculty of Medicine, Cairo University. The dataset consist of 30 attributes, it were divided into two phases, the first phase consists of 8 attributes which are; Age, Sex, Schistosomiasis (Shisto), Alanine Aminotransferase (ALT), An aspartate aminotransferase (AST), Alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT) and nonalcoholic fatty Liver disease. The second phase consists of 8 attributes which are; Nonalcoholic Fatty Liver disease attribute (output of phase one), Weight, Height, Waist Circumference (WC), Fasting Blood Sugar (FBS), History of Hypertension, History of Diabetes and Hemoglobin A1C (HBA1C).

### B. System Model

The suggested model in this paper comprises of two phases; the model starts from preprocessing step of filtering data then estimating the missing values, standardize data, normalize data after that handling the imbalanced data then verifying data to finally be ready for feature selection and extraction (Fig. 1).

Then, the first phase is developing Support Vector Machine algorithm to classify patients with nonalcoholic fatty liver disease (NAFLD). The learning algorithm was applied on 8 attributes and number of patients with NAFLD were detected and patients with other reasons of liver illness (alcohol, medication, etc.) were excluded to give off the results to either be 0; does not have liver disease or 1; affected with liver disease.

The second phase is developing a back propagation neural network that takes the output from phase 1 as an input in phase 2 in addition to another 7 attributes then train the artificial neural network algorithm with distinct topologies and range of epochs to achieve weights that give the optimum outcomes to categorize patients to three different classes; pre-diabetic, diabetic or nondiabetic patients “Fig. 2”.

*1) Preprocessing:* After reading the dataset file, preprocessing steps start in order to remove unwanted records that are presented in the dataset file by applying the following steps:

*a) Filtering:* by removing noise or unwanted data to display useful feature for prediction.

*b) Estimating* missing values: by calculating missing values as a weighted sum of linear interpolations from the closest accessible points. A total of five estimates from column-wise and five from row-wise, linear interpolation estimates for one-d are calculated. The best case was weighing; such that interpolation is equivalent to the average Lager 4-points of the nearest points in rows and columns (separated missing points far from the border).

*c) Standardize* and normalize: by rescaling attributes to the range of 0 to 1.

*d) Handling* imbalanced data: by adopting k-fold cross-validation steps in which data are randomly sorted, and then splited into k folds. after that, dividing the data using a common value of k=4 (four folds). The validation set consists of one fold, while the three remaining folds together are used for training. For each one of the validation sets, the validation accuracy is calculated and the final cross validation precision is averaged.

When you run 'k' cross validation rounds, one of the validation folds were used every round and the remaining folds were used for training. Its precision on the validation data was evaluated after being trained by the classifier. Average precision throughout the k round to obtain the final precision of cross-validation. Verify data by checking the dataset accuracy and inconsistency after data migration and proofreading data involving checking the data entered against the original document.

*e) Export*: release data to be ready for data mining.

*2) Feature selection and extraction:*There are number of characteristics for health information used for the system instruction. Noise, unauthorized or irrelevant information may also be present. The training dataset must be preprocessed in order to clean the data.

The Correlation Feature Selection (CFS) measure makes an evaluation for the subsets of features according to the following basics: “Good feature subsets contains features which are extremely associated with classification, but are not associated with each other” [24].

The primary goal of the feature selection method is to remove the redundancy and the non-relevant information. The result of improving classifier effectiveness and improving the accuracy is an increasing percentage of true positive predictive values. Decrease in accuracy is a result of nonrelevant features.

Feature Selection is considered to be one of the most important steps in the transformation phase of the KDD [16]. It is defined as the selection method of features from the area of study, which is more related and informative for model building. Feature selection [25] has many advantages that are relative to various elements of data analysis like improved data visualization and data realizing, reduced computational time and analysis time, and improved prediction precision.

### 3) *Machine learning techniques*

*a) Support Vector Machine (SVM):* Support Vector Machine (SVM) is considered to be one of the popular linear discrimination methods on the basis of a straightforward but strong powerful concept.

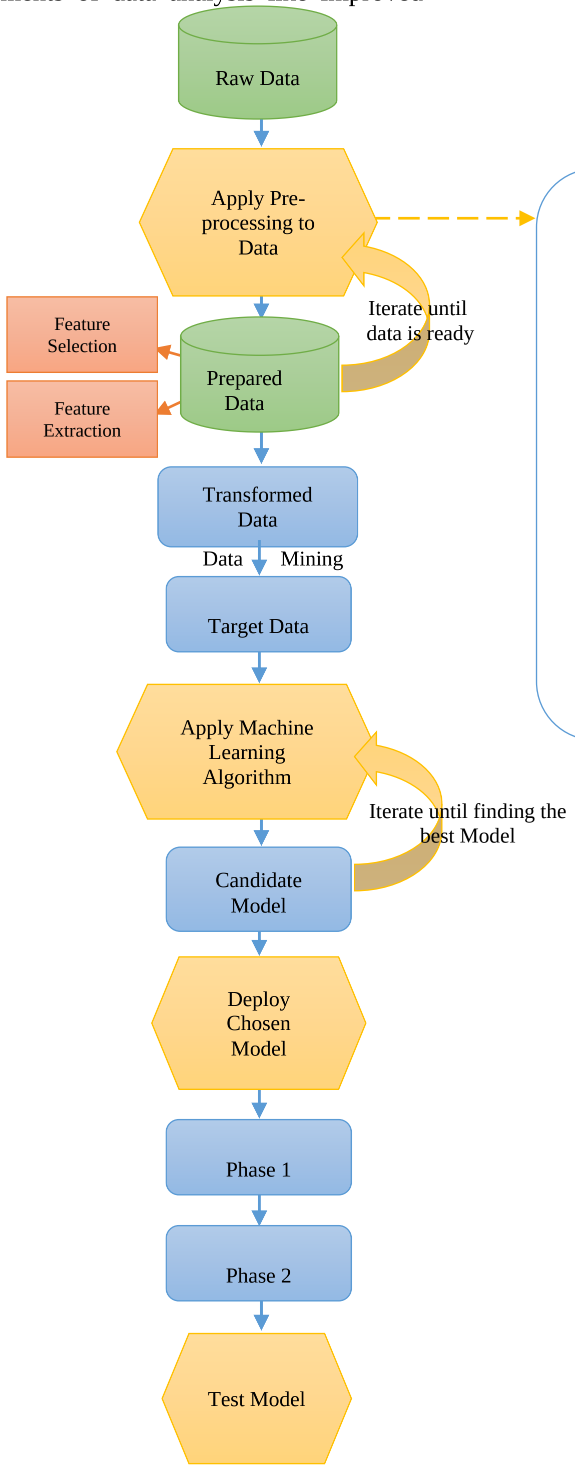

Fig. 1. Data Preprocessing.

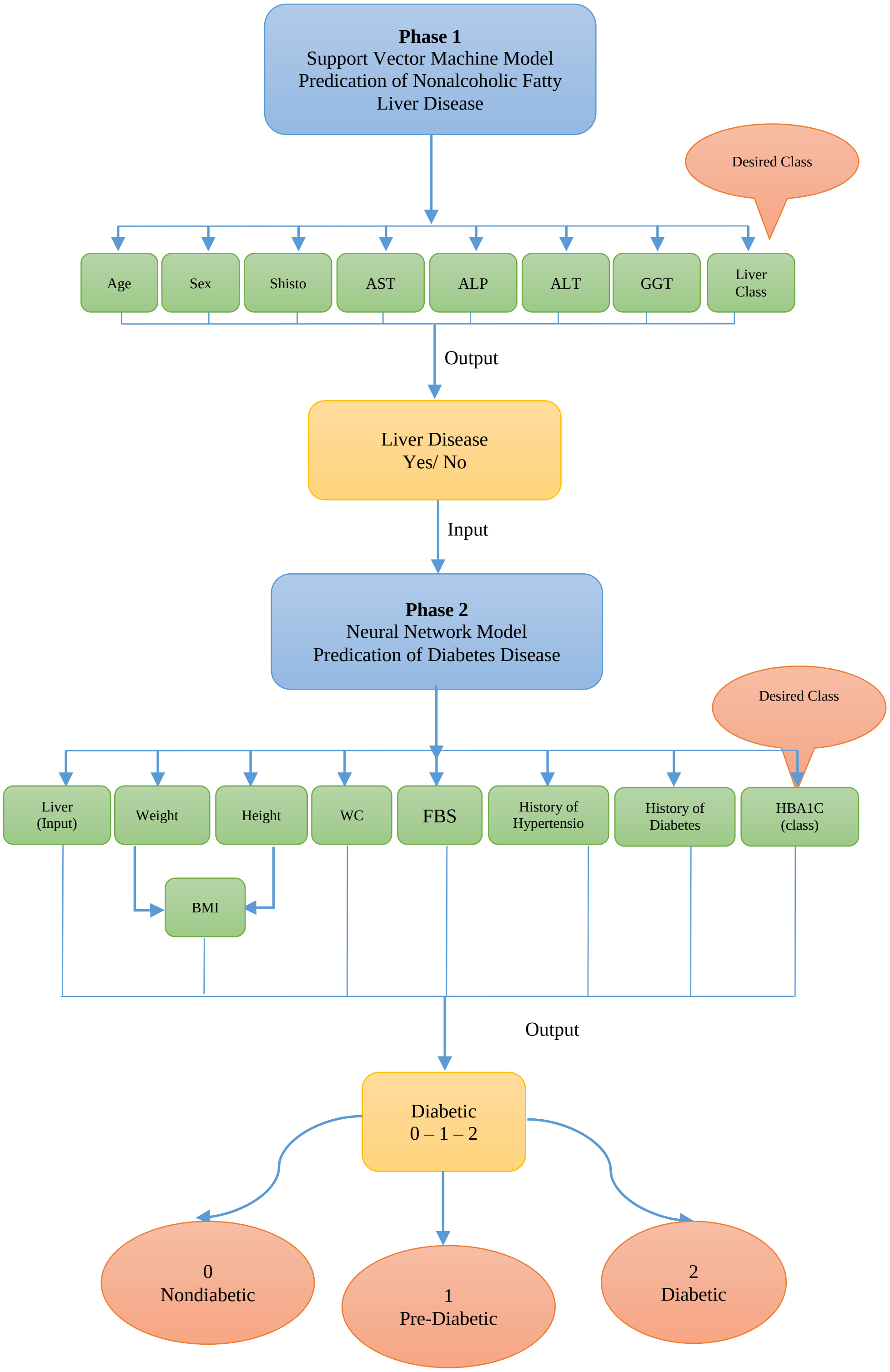

Fig. 2. Proposed System.

The first stage is to map the samples from the original entry to a high feature space so that the best way to separate samples is got. If its margin is largest then a hyperplane separating H is considered to be the top. The margin is the largest distance between two parallel hyperplanes to H on both sides that have no sample points between them [26, 27]. It comes from the concept of risk minimization (the expected loss assessment function as a miss-classification of samples) that the higher the margin, the higher the generalization error of the classifier.

Numerous kernels can be used in the Support Vector Machine models, including linear, polynomial, radial-based function (RBF) and Gaussian function:

$$K(X_i, X_j) = \begin{cases} X_i \,.\, X_j & Linear \\ (\gamma X_i \,.\, X_j + C)^d & Polynomial \\ exp(-\gamma \; |X_i - X_j|^2) & RBF \\ \tanh(\gamma X_i \,.\, X_j + C)^d & Sigmoid \end{cases} \qquad (1)$$

Where $K(X_i, X_j) = \emptyset(X_i) \,.\, \emptyset(X_j)$ that is, the kernel function shows a dot product of input data that is mapped by transformation into the higher level feature space ϕ. Gamma is an adaptive parameter of some kernel functions.

Eventually, the RBF is the most common option in Support Vector Machines for Kernel types. This is primarily due to their localized and finite reactions throughout the true x-axis.

*Algorithm 1 " Support Vector Machine (Phase 1)"*

*Detecting Nonalcoholic Fatty Liver Disease*

```
Input: set of samples (input and output) for training pair samples; the input samples are x1, x2…xn, and the output class is y.
    1. Finding Pair of samples in the training samples that are closed
        candidateSV = {closest pair from classes that are opposite}
        do
        Find a violator sample
    2. Adding this sample to the Support Vector data samples
    Candidate_Vector = candidateSV_ Vector ∪ violator
    3. Pruning
        if any $\alpha_p < 0$ as a result of adding c to S then
        candidateSV = candidateSV / p
    4. Repeat till all points are pruned
         end if
```

*b) Artificial Neural Network:* Artificial Neural Network (ANN) method was used in the classification phase, as it utilizes complexity issues to be solved. The artificial neural network adapts itself by sequential training algorithm and its architecture and linked weights [24]. This paper utilized the learning algorithm for back propagation.

The Artificial Neural Network (ANN) is defined as a computational model made up of interconnected nodes that are called neurons arranged in layers; input, hidden and output. Each interconnection has a weight that changes during the training phase till adequate outcomes are achieved. ANN is used to model complex/nonlinear inter-relationships between inputs and outputs, for extracting significant patterns. In [26], ANN based classifier is used to model Diabetes dataset. The proposed ANN classifier has $i$- $h$ - $o$ configuration, where $i = 8$ (the number of attributes to the model inputs), $h$ is the number of neurons in the hidden layer, where $h = 7$ (using one hidden layer), and $o$ is the number of outputs that is equal one.

*Algorithm 2 " Artificial Neural Network (Phase 2)"*

*Detecting Diabetes Disease*

```
1. Initialization step: set all weights equal to small random Values.
    Do while (from step 2 : 9)
2. Iterate steps from 3 to 8: for each sample in the training set,
    Forward Phase:
3. Each feature vector in the input are forward to the above
layer (the hidden layer)
4. Each hidden unit (Zj) sums its weighted i/p signals,
        $Z - i_{nj} = V_{aj} + \sum_{i=1}^{n} x_i v_{ij}$ ,$Where\ V_{aj}\ is\ a\ bias$
    Apply the transfer function
        $Z_j = 1/(1 + e^{-(Z-inj)})$
    send this value to all units that present in the above layer
5. Compute the output:
        $Y - i_{nk} = W_{ok} + \sum_{i=1}^{n} Z_j w_{jk}$ ,$Where\ W_{ok}\ is\ a\ bias$
        $Y_k = 1/(1 + e^{-(Y-i_{nk})})$
    Backward Phase:
6. Calculate the error in the output layer
        $\delta_{2k} = Y_k(1 - Y_k) * (T_k - Y_k), T_k\ is\ the\ target$
7. Computes its error information in hidden layers
        $\delta_{1j} = Z_j(1 - Z_j) * \sum_{k=1}^{m} \delta_{2k}\, w_{lk},$
    Update Phase:
8. Update weights in all layers and bias
        $W_{jk}(new) = \eta * \delta_{2k} * Z_j + \alpha * W_{jk}(old)$
        $V_{ij}(new) = \eta * \delta_{1j} * x_i + \alpha * V_{ij} old$
9. Test stopping condition.
```

## IV. Evaluation Methods

### A. Precision and Recall

Precision and recall are both common metrics for evaluating classifier efficiency and will be used widely in this dissertation. Precision is the proportion that when making a choice, the model properly predicts positive. To be more specific, precision is the number of positive instances properly identified divided by all number of positive examples “(2)”. Recall is the percentage of identified correct positive from all the current positives; it is the number of the correct positive classified exampled divided by the total number of true positive examples in the tested set.

Both high recall and precision are considered to be an ideal model. The F-measure “(5)” is the harmonic measure of precision and recall in a single measure [28]. The F- measure varies from 0 to 1, as a classified is a measure of 1 that completely captures precision and recall.

$$Precision \quad = \frac{TP}{TP+FP} \tag{2}$$

$$Sensitivity \quad = \frac{TP}{TP+FN} \tag{3}$$

$$Specificinty = \frac{TN}{TN+FP} \tag{4}$$

$$F-measure = \frac{2(Precision)(Sensitivity)}{(Precision)+Sensitivity} \tag{5}$$

Where

TN (True Negative): Negative case truly expected,

TP (True Positive): Positive case truly expected,

FN (False Negative): Negative case was positive but negatively expected,

FP (False Positive): Positive case was negative but positively expected.

## *B. Kappa Coefficient*

Cohen's kappa statistics provide the second approach for datasets evaluation, which are 1 for an ideal classifier that usually classifies the right ones and 0 for a random classifier. The value of the kappa coefficient can be calculated using the following equation:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{6}$$

Where, $p_0$ is the classification accuracy and $p_e$ is the hypothetical accuracy of a random classifier on the same data.

# V. Results

## *A. Support Vector Machine*

Algorithm 1 " Support Vector Machine (Phase 1)"

*Detecting Fatty Liver Disease*

As shown in “Fig. 3”, it is obvious that the SVM with Gaussian and RBF kernel function give the best accuracy results. Thus, as both function return the same results RBF function was chosen to measure its precision and recall as shown in Table I.

## *B. Artificial Neural Network*

Algorithm 2 "Artificial Neural Network (Phase 2)"

*Detecting Diabetes Disease*

The best performance was achieved using the primary dataset with overall 16 attributes, scaling each feature to a value between 0 and 1. Then, the classifier is trained as showed in Table II and demonstrates that the optimal accuracy results achieved in 50 iterations with 3 layers; 8 input nodes, 7 nodes in the hidden layer and 1 node in the output layer was 86.6% as shown in “Fig. 4”.

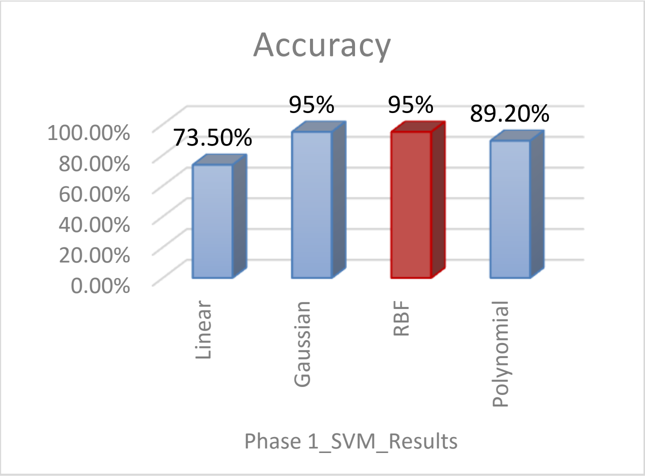

Fig. 3. SVM Accuracy Results.

TABLE. I. Phase 1_RBF Function Accuracy

| **RBF Function** | |
|---|---|
| ***Measure*** | ***Value*** |
| Sensitivity | *0.90* |
| Specificity | *1.00* |
| Precision | *1.00* |
| Negative Predictive Value | *0.90* |
| False Positive | *0.00* |
| False Negative | *0.09* |
| Accuracy | *0.95* |
| F1 Score | *0.05* |

TABLE. II. Phase 2_ANN Accuracy

| **Phase 2** | **Artificial Neural Network** | | | |
|---|---|---|---|---|
| *Iterations* | *Input* | *Hidden* | *Output* | *Accuracy* |
| 50 | 5 | 5 | 1 | 82.69% |
| 100 | 5 | 5 | 1 | 82.69% |
| 150 | 5 | 5 | 1 | 83.07% |
| 200 | 5 | 5 | 1 | 82.30% |
| 50 | 7 | 5 | 1 | 83.84% |
| 100 | 7 | 5 | 1 | 84.61% |
| 150 | 7 | 5 | 1 | 84.61% |
| 200 | 7 | 5 | 1 | 84.61% |
| 50 | 3 | 3 | 1 | 81.53% |
| 100 | 3 | 3 | 1 | 82.30% |
| 150 | 3 | 3 | 1 | 83.07% |
| 200 | 3 | 3 | 1 | 83.07% |
| **50** | **8** | **7** | **1** | **86.56%** |
| 100 | 8 | 7 | 1 | 85.38% |
| 150 | 8 | 7 | 1 | 85.38% |
| 200 | 8 | 7 | 1 | 85.38% |

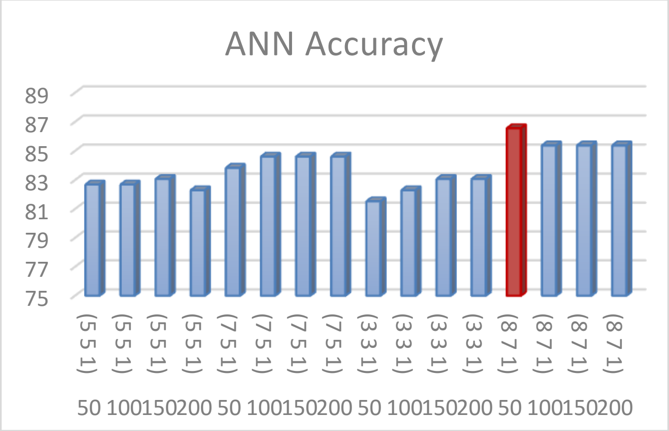

Fig. 4. ANN Accuracy Results.

In the experiments, when investigating the effect of the training data size on the classification accuracy, it has been noted that the size of the training set improves the classification accuracy. After analyzing the results, it can be concluded that the use of the hybrid system that combines SVM and ANN in one system is clearly preferable than using each classifier individually. Primarily, because SVM classifier is more efficient with binary class problem and very sensitive to the dimensionality of the feature vectors. In addition to the ANN algorithm which supposed to be a very flexible classifier. These combination leads to a powerful technique for classification problems.

## VI. Conclusion and Future Work

Metabolic syndrome, non-alcoholic fatty liver and diabetes mellitus patients are at a growth of a very dangerous outcome like cirrhosis and morals for patients. In this study, a model for predicting chronic Diabetes mellitus was proposed.

The proposed model combines two machine learning techniques which are Support Vector Machine and Artificial Neural Network. The accuracy results showed that predicting nonalcoholic fatty liver disease by using the RBF kernel function in Support Vector Machine was 95% and by applying ANN classifier the findings obtained for optimal accuracy was 86.6% in 50 iterations with 3 layers; 8 input nodes, 7 nodes in the hidden layer and 1 node in the output layer.

Accordingly, good results on the obtained dataset showed that the proposed model performed out exemplary of the existing classifiers.

This analysis implies that patients with obesity, nonalcoholic fatty liver disease can lead to diabetes mellitus disease and more violent illnesses such as cirrhosis and mortality are expected to occur.

The results of this study exhibited the need of some further work to be done in the future. Firstly, more experiments will be required, since the imbalance of the dataset likely had a detrimental effect on the performance and the data processing was limited by the size of the dataset. Thus, more data should be involved to create a balanced dataset that would probably lead to a very important improvement in the performance for various learners, in order to make the research more universal. Secondly, more research is required to improve the quality of experimental information in preprocessing phase for data cleaning and estimation of the missing values. However, there are still many challenges in the medical research, and further work should be carried out to really advance this technology beyond laboratory demonstrations and disease prediction in order to restrict disease propagation.

## Acknowledgment

I would like to express my great appreciation to Dr. Farid A. Mousa for his assistance, my gratitude to my family and my best friends.

## References

[1] Bhupathiraju, S.N. and F.B. Hu, Epidemiology of obesity and diabetes and their cardiovascular complications. Circulation research, 2016. 118(11): p. 1723-1735.

[2] Ng, M., et al., Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The lancet, 2014. 384(9945): p. 766-781.

[3] Al-Goblan, A.S., M.A. Al-Alfi, and M.Z. Khan, Mechanism linking diabetes mellitus and obesity. Diabetes, metabolic syndrome and obesity: targets and therapy, 2014. 7: p. 587.

[4] Iyer, A., S. Jeyalatha, and R. Sumbaly, Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774, 2015.

[5] Control, C.f.D. and Prevention, National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services, 2014. 2014.

[6] Cichosz, S.L., Predictive models in diabetes: Early prediction and detecting of type 2 diabetes and related complications. 2016, Aalborg Universitetsforlag.

[7] Zou, Q., et al., Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 2018. 9.

[8] Krasteva, A., et al., Oral cavity and systemic diseases—diabetes mellitus. Biotechnology & Biotechnological Equipment, 2011. 25(1): p. 2183-2186.

[9] Devi, M.R. and J.M. Shyla, Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research, 2016. 11(1): p. 727-730.

[10] Iancu, I., M. Mota, and E. Iancu. Method for the analysing of blood glucose dynamics in diabetes mellitus patients. in 2008 IEEE International Conference on Automation, Quality and Testing, Robotics. 2008. IEEE.

[11] Mills, E.P., et al., Treating nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus: a review of efficacy and safety. Therapeutic advances in endocrinology and metabolism, 2018. 9(1): p. 15-28.

[12] Isomaa, B., et al., Cardiovascular morbidity and mortality associated with the metabolic syndrome. Diabetes care, 2001. 24(4): p. 683-689.

[13] Marchesini, G., et al., Nonalcoholic fatty liver disease: a feature of the metabolic syndrome. Diabetes, 2001. 50(8): p. 1844-1850.

[14] Ballestri, S., et al., Nonalcoholic fatty liver disease is associated with an almost twofold increased risk of incident type 2 diabetes and metabolic syndrome. Evidence from a systematic review and meta‐analysis. Journal of gastroenterology and hepatology, 2016. 31(5): p. 936-944.

[15] Peterson, D.M., The mind's new labels?: Review of RA Wilson and FC Keil (Eds.), The MIT Encyclopedia of the Cognitive Sciences. 2001, Elsevier.

[16] Kavakiotis, I., et al., Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 2017. 15: p. 104-116.

[17] Nilashi, M., et al., An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 2017. 106: p. 212-223.

[18] Kumar, R.N. and M.A. Kumar, Medical Data Mining Techniques for Health Care Systems. International Journal of Engineering Science, 2016. 3498.

[19] Boukenze, B., H. Mousannif, and A. Haqiq, Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease. Int. Journal of Database Managment systems, 2016. 8(30): p. 1-9.

[20] Abdullah, M. and S. Al-Asmari, Anemia types prediction based on data mining classification algorithms. Communication, Management and Information Technology–Sampaio de Alencar (Ed.), 2017.

[21] Kumari, V.A. and R. Chitra, Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 2013. 3(2): p. 1797-1801.

[22] Daghistani, T. and R. Alshammari, Diagnosis of diabetes by applying data mining classification techniques. International Journal of Advanced Computer Science and Applications (IJACSA), 2016. 7(7): p. 329-332.

[23] El-Halees, A.M. and A.H. Shurrab, Blood tumor prediction using data mining techniques. Blood tumor prediction using data mining techniques, 2017. 6.

[24] Wiley, M.T., Machine learning for diabetes decision support. 2011, Ohio University.

[25] Jahankhani, P., V. Kodogiannis, and K. Revett. EEG signal classification using wavelet feature extraction and neural networks. in IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06). 2006. IEEE.

[26] Eljil, K.A.A.S., Predicting Hypoglycemia in Diabetic Patients using Machine Learning Techniques. 2014.

[27] Hashmi, S.F., A Machine Learning Approach to Diagnosis of Parkinson’s Disease. 2013.

[28] Frutuoso, D.G., SMITH-Smart MonITor Health system. 2015.