EVALUATION OF CLASSIFIER MODELS FOR THE DETECTION OF DIABETES DISEASE

BY
KABIRU ABDULLAHI
Department of Computer Science, School of Science and Technology
Hussaini Adamu Federal Polytechnic, Kazaure, Jigawa state
Correspondence E-mail: kabjah@yahoo.com

ABSTRACT
Diabetes is one of the major diseases which is commonly found among all age groups and people of different origins. Diabetes is a disease which may lead to the failure of different organs, and cause high risk of blindness, kidney failure, heart disease and problems in the nervous system. Data mining algorithms could be used as alternative way for diagnosis by discovering patterns from the history of patient data and also by capturing the experience of experts. In this research different classifier models ware designed and implemented to predict type one and type two diabetes diseases. Different performances measure was evaluated to identify an optimal classifier accuracy models. The classifiers used in the experimental approaches are Decision Tree (C4.5), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The optimal models identification was done using performance evaluation matrices include accuracy (Acc), specificity (Spe.), sensitivity (Sen) and precision (Pre.). The models was tested with Pima Indian Heritage diabetes database from University California Irvine (UCI) Machine learning repository and Virginia Commonwealth University (VCU) database collected across 139 hospitals in United states of America (USA).
Keywords: *Data Mining Classification, Decision tree, Artificial neural network, Support vector machine*

1.0 Introduction

Diabetes is one of the most prevalent and quickly spreading disease in the world, and it is one of the main health issues in the majority of countries, commonly found among all age groups and people of different origin. Is a condition when the body's organs refuse to regulate the release of insulin., the substance that unblock the body cell by allowing glucose to speed up its secretion, (Anand, Pratap, Kirar, & Burse, 2012) stated that diabetes disease is also characterized to cause failure of different organs, which result in any of the following ailment such as high risk of blindness, kidney failure, heart disease and sometimes nervous system damage which may even lead to death. Statistical analysis by international diabetes federation (IDF) (Guidelines & Force, 2012), and world health organization (WHO), indicated that approximately 366 million people have diabetes worldwide in the year 2012, adding to that about 183 million people are unaware they had the disease.

For prevalence, area like North America and Caribbean is the region encountered with higher prevalence, followed by Middle East and North Africa then Western pacific region. If immediate action is not taken, the IDF estimates that the number of people with diabetes worldwide would

reach 552 million by 2030. In addition the centers for disease control and prevention (CDC) claim that in 2014 about 29.1 million people almost 9.3% of United States population (including children and adults) were affected with diabetes (United, 2014), regarding this figure, 21.0 million were diagnosed, of which approximately 8.1 million people were not. And about 4.3 million people aged 20 years or older was identified and found with diabetes disease.

Moreover the Association of American diabetes society stated that, about 11.2 million people age between 65 years and above have diabetes mellitus while 25.6 million peoples age between 20 years or older have also diabetes mellitus and according to UK National Health Services("Key statistics on diabetes Contents," 2012), indicated that In Northern Ireland, a total of 72,693 adults, or 3.8% of the population, had been diagnosed with diabetes. An additional 10,000 people have diabetes but are unaware of it. The Coordinator for Strategies for Improving Diabetes Care in Nigeria (SIDCAIN), Christopher Alebiosu, was quoted. According to the News Agency of Nigeria Newspaper (NAN), (2014), Nigeria had more cases of diabetes than any other African nation, with 3 million cases, followed by South Africa with 1.9 million cases and Ethiopia and Kenya with 1.4 million and 769,000 cases, respectively.

Because of this, there have been numerous attempts to uncover significant knowledge that can be acquired from data generated based on patient diagnostic information such as symptoms, treatment history, and novel therapy patterns that may be analyzed and advised. Because of this, even if data mining is important, using the right data mining techniques and methodologies to find relevant information that can aid in decision-making is a challenge for medical diabetes data diseases.

1.1 Problem statement

Based on the earlier stated fact on problem associated with "Diabetes disease as one of the leading causes of death in the entire world and also associated with significant medical complication such as failure of nervous system like kidney and heart disease, blindness and, hypertension, amputation and lot of others chronic ailment, the cost of diabetic disease is not only physical and psychological but also economical.

Hence a lot of growing concern on how to get knowledge out of patients diagnostic and treatments data. with traditional data analysis cannot discover complex relationships among it, and so query might not be able to retrieve relevant information due to user bias or lack of experience. As a result of this researchers turned to data mining and knowledge discoveries as a general approach for knowledge "hidden" in a raw data especially in medical domain, if it is to be deemed the output of a knowledge discovery tool should be accurate, stable and interpretable. Fulfilling such requirement involves incorporating concepts and methods from various disciplines such as machine learning, artificial intelligence and data mining, for effective decision-making and clinical reasoning.

Based on the aforementioned reasons, this study employ the usage of three different classifiers, by identifying the following issues.

1. How to identify the optimal models for providing accurate diagnosis using diverse classifiers such as Decision tree, Support vector machine (SVM) and Neural Network

2. How to use appropriate knowledge acquisition tool to incorporate domain knowledge into Models
3. How to identify reliable predictive features from patient's clinical records in order to evaluate the performance models.

1.2 Research questions:

Based on this problem associated with diabetes disease the following questions were formulated as follows

1. What data driven techniques are currently being used to predict diabetics disease?
2. What is the best classifier architecture that can be used for predicting type1 and type2 diabetes disease?
3. What are the optimal learning parameters that can be used in the classification ?

2.0 LITERATURE REVIEW

2.1 Overview:

This paper reviews a related literature pertaining to the study, it highlights on several studies about the topic and identify the merit and demerits of approaches proposed by different researchers which provides further directions on our proposed design approaches. It also discusses about the concept of data mining and classification and also on WEKA and MATLAB the software tools that will be used for identifying the optimal model.

2.2 Related Work

In recent years use of predictive classification and evaluating the performance of medical diagnosis has receive different techniques involves in classifying the data, many studies was conducted by researchers classification and prediction for finding appropriate and best performing tools and algorithm for diagnosing of diabetes disease was conducted in this area to assist physicians in getting appropriate diagnosis tool for decision making.

Moreover many work related to data mining and machine learning algorithm in the domain of diabetes has been concentrated on the extensive studies of Pima Indian diabetes dataset, many classification algorithm have been applied to the dataset and yet most of the algorithm have performed moderately, (Shanker & Hu, 1999), used Neural network to predict onset of diabetes mellitus among Pima Indian female population near Phoenix, Arizona using network with hidden layer and obtained an overall accuracy of 81.25%. Also, (Zolfaghari, 2012), research on diagnosis of diabetes in Female population of Pima Indian heritage using ensemble classification of Neural network and support vector machine obtained a predicted accuracy of 88.02% which was compared with previous literature studies using logistic and regression model with 80.2% and neural network with 77%

Other Studies on data mining techniques with Pima Indian Diabetes dataset using various classification model such as Naïve Bayes, fuzzy logic, JRip, MLP Probabilistic neural network ANFIS and many more which I did not mention was conducted . Smith et al. used a neural network with ADAP algorithm using Hebbian learning to build associative models for diabetes

diagnosis with 768 instances, they used 576 randomly selected cases for training and the remaining 192 test cases showed an accuracy of 76% was obtained. (Temurtas, H. et al, 2009), in there study on "comparative studies on diabetes disease with PIDD", the authors applied MLNN and LM algorithm on the first stage the LLNN and LM used 10-fold cross validation to compare the classification accuracy. And on the second stages they applied PNN on training and testing and also used 10-fold cross validation for comparing the accuracy of the result which at end the methodology adopted obtained better estimation compared to the previous study in his literature, but based on the observation I made the research MLNN is black box in nature and required greater computational burden which will cost more time to finish so we need faster enabling work instead of the one taken time.

In another studies conducted by (Patil, Joshi, & Toshniwal, 2010), using C4.5 and K-means clustering algorithm on PIDD with 768 instance and having eight attributes to build the model for predicting type2 diabetic patient, the decision tree was constructed on J48 and 10-fold cross validation was applied to evaluate the performance using specificity and sensitivity and obtained classification accuracy of 92% which is higher above other previous studies of in his literature, but the problem of the studies is that data used on identifying type 1 Diabetes is same with his data which put a question mark on how he model and applied to type2 diabetes prediction

Furthermore study conducted by (Rahman & Afroz, 2013), on "comparison of various classification using different data mining tools for diabetes" conducted on (PIDD) from UCI machine learning repository, the main objective of their study is to evaluate and investigate nine selected classification algorithm which are MLP, Bayes net, J48graft, C4.5, Fuzzy Lattice Reasoning (FLR), ANFIS, Fuzzy Inference system (FIS), JRip, and Adaptive Neuro fuzzy Inference system, using WEKA, MATLAB and TANAGRA software, with best algorithm obtained in the three of the software tools are , WEKA is J48graft having 81.33% accuracy and take 0.135 second in training while in TANAGRA Naïve Bayes classifier provide 100% accuracy with training time of 0.001 second and using MATLAB was ANFIA having 78.79%accuracy. Based on the outcomes of the result Bayes network classification and J48graft has significant impression in the use of medical domain classification. But the demerit of this classifiers is that total time required to build this model is crucial in comparing the classification algorithm using Bayes network and J48graft will take longer time.

Therefore any approach for achieving diabetes control need to rely on some kinds of developed model in order to provide a framework, the depth of the model should corresponds to the level of my understanding of what is going on, quality and quantity of data obtained should determine the guide for control action of getting appropriate prediction model as well as classification accuracy of the evaluated parameters

2.3 Data mining

The term data mining sometimes known as (knowledge discovery) is defined as kind of pattern involves in analyzing data from different perspective and summarizes it into useful information". (Fayyad, et al, 1996). Data mining is generally a representation of the real world situation about how and which kind of data going to be collected and store in the database. Therefore the overall objectives of data mining task is extraction of information from data set and then transform the

data into useful form and understanding structure with the intention of future use. Data mining analysis involves different level of analysis such as Artificial Neural network, genetic algorithm, decision tree, rule induction and data visualization, the importance of using data mining concept in medical domain has been proven to be effective methods in describing specific patterns such as dependencies, interrelation and many irregularities. Which may be present in the data?

2.4 Data Mining and Diagnosing

(Kumari,2013),(Koklu&Unal,2013), stated that a diagnosis is define as "process of conducting several experiment which describes analyzed a symptoms in patient body" performed by and expert or medical specialist and obtain appropriate and reliable result for medical prescription. Therefore the most effective way of detecting symptoms of diabetes disease is earlier diagnosis in order to reduce the occurrence of high risk of several ailmentwhich are associated with the disease, such as blindness, heart disease or nervous system damage. Therefore data mining is one of the major techniques when apply to medical domain can assist in diagnosis process.

2.5 Classification of diabetes

Basically diabetes is disease which occur There are three basic types of diabetes that are caused by either the pancreas not creating enough insulin or the body's cells not responding to the produced insulin adequately as follows:

2.5.1 Type 1

Type 1 Diabetes It is the body's inability to create enough insulin, also known as diabetes mellitus. Referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes" and adolescent but may present at any age. The cause is unknown(Kumari,2013).

2.5.2 Type 2

Failure to produce enough insulin and insulin resistance are the causes of type 2 diabetes. Reducing calorie intake and increasing physical activity can help manage high blood sugar levels. Absolute deficiency is a possibility over time. Due to the fact that type 2 diabetes is frequently asymptomatic in the beginning, it frequently goes misdiagnosed for years. The risk of having macro- and micro-vascular problems is higher in people with untreated type 2 diabetes. (Kumari,2013),(Polat&Gunes, 2007).

3.0 METHODOLOGY

3.1 Overview

This paper discuss the methods and the approach adopted in this study and will explain how the classification method is being selected and the protocol in developing the model for performance evaluation of the classifier, therefore it is very important to examine the data thoroughly before undertaking any further steps in analyzing it.

3.2 Method of developing classification model

Classifier models are developed to identify presence or absence of diabetes and also to differentiate between type 1 and type 2 diabetes. Classifier model one is developed to identify the presence or absence of diabetes disease within a group of persons. Classifier model two is developed to differentiate the presence of type 1 or type 2 diabetes within 'yes' group identified with model one as indicated in fig 3.1 below:

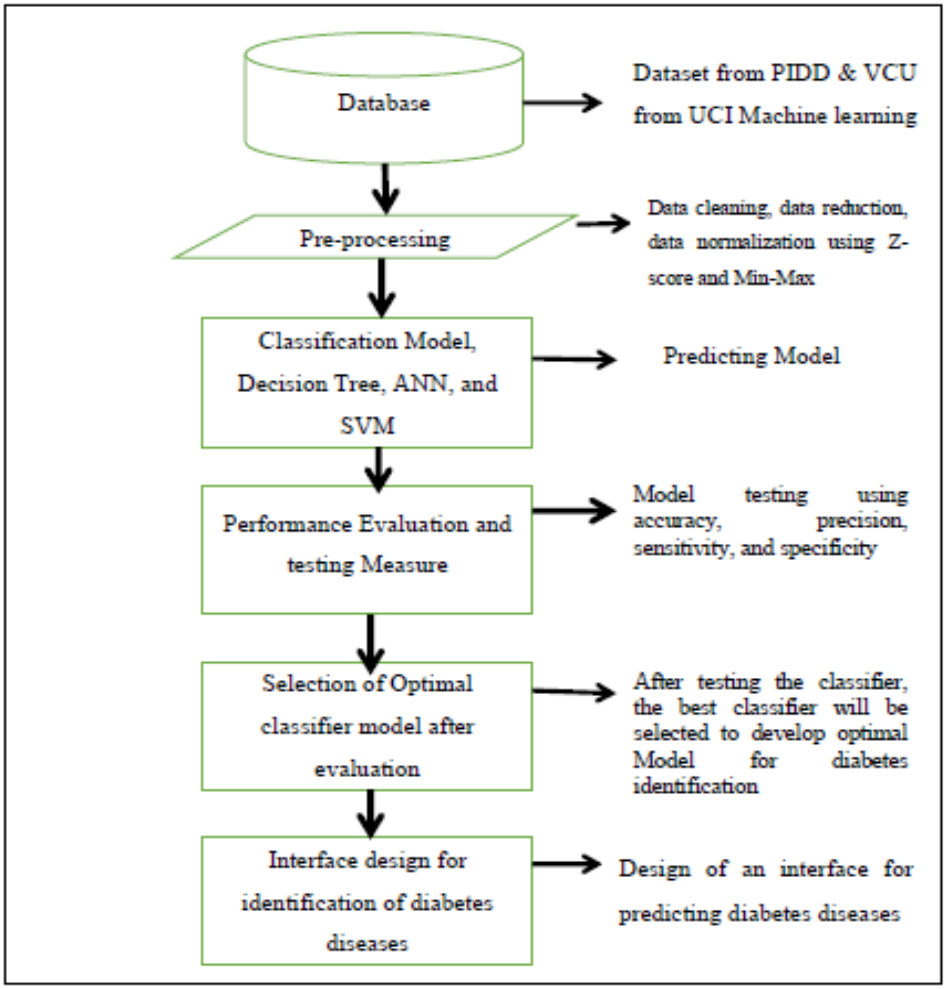

Fig. 3.1. Proposed Methodology Framework

3.3 Data Source

In their study, Han and Kamber (2006), revealed that, there are considerable amount of records and data information available in the database which makes information retrieval and manipulation complex, therefore based on this assertion the study employ the use of two data sources in order to achieve stated objectives for successful development of the model, the data set are source from "The Pima Indians heritage diabetes database" from the UCI machine learning repository The national institute of diabetes and digestive of kidney diseases is the original owner of thes data set. Vincent Sigillito of the John Hopkins University's applied physics laboratory donated the database.

and data collected from the Virginia Commonwealth University Center for Clinical and Translational Research (VCU), obtained from public data repository in UCI machine learning repository. Some attributes ware selected for type 2 model development, several constraints before selecting the data set was encountered which caused a lot of hinderance in this methodology ranging from extracting the data, to analysis of the result. A total of 768 instances of the diagnostic, binary valued variable was examined in order to determine if the patient exhibited symptoms of diabetes according to the World Health Organization guidelines. The table 3.1 below for PIMA attributes lists eight (8) attributes.

Table 3.1 PIMA Attributes and Instances list with missing values

S/No.	Attribute	Description	Missing Value
1	Pregnant	a record of how many times the woman has been pregnant	111
2	blood glucose	Utilizing a two-hour oral glucose tolerance test, plasma glucose concentration is determined (mm Hg)	5
3	Diastolic BP	Diastolic blood pressure of patient arteries measure (mm Hg)	35
4	Triceps SFT	Triceps skin fold thickness body measurement in (mm)	227
5	Serum-Insulin	Two hours serum insulin (MuU/ml).	374
6	BMI	Body mass index of patient measured in (weight Kg/height in (mm) ^{2})	11
7	DPF	Diabetes pedigree function (likelihood)	0
8	Age	Age of diabetes patient in (year)	0
9	Class	Diabetes on set within five year	0

While the second data source used was collected from Virginia Common wealth university (VCU)with original database contains 1087 instance with incomplete, redundant and noisy information, the data set contains fifty attributes but many of them in carrying out the experiment ware deleted because are not important in this study hence only ten (10) attributes was selected for this experiment listed below as

1 Race
2. Gender
3. Age
4. Body mass index
5. Triceps skin fold thickness
6. Diabetes Pedigree function
7. Plasma glucose concentration
8. Diastolic blood pressure
9. Serum insulin
10. Class

3.4 Data Pre-processing

Large databases in the real world and in practice frequently contain data that is susceptible to noise and inconsistent due to their size and originated from several sources. Additionally, there are numerous record discrepancies that might easily exacerbate analytical issues in the medical field, thus it is crucial to eliminate all the noisy data and data inconsistencies. hence missing or incomplete data can invalidate the result obtained with standard analysis, procedure that has not indicate wrong on the data and one of the common problems of data quality is missing data therefore reliable decision making rely heavenly on high quality data, In order to create high-quality mining results, preprocessing approaches have a substantial impact on the performance of machine learning algorithms. Because of this, it is clear to use pre-processing for the raw data input in this situation because it has a significant impact on getting the data ready for training and testing and generating correct and trustworthy data. In addition to increasing accuracy, data pre-processing can also reduce bias in a classifier by commencing the training process for each feature within the same scale and in the same range of values for each input feature.

Various data preprocessing methods are categorized into the following

1. Data cleaning method
2. Data reduction
2. Data transformation (Normalization)
4. Data Integration

Prior to the clustering process, any of these techniques can be used to successfully analyze data, removing noisy data and correcting any inconsistencies., data cleaning process could be apply to fill in the gap of missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies, while for data reduction method involves reduction data size (volume) by process aggregation and eliminating redundant features, or by grouping data in to cluster, with same attributes, but produces the same or similar analytical results and data integration is a process of Integration of multiple databases, data cubes, or files while Normalization, also known as data transformation, is a crucial preprocessing step in data mining that standardizes the values of all variables from the dynamic range into a certain range. Normalized values make it possible to compare corresponding normalized values across various datasets while removing the impact of some significant external variables.

3.5 Classification and prediction Approaches

According to(Han & Kamber, 2006), define “classification as a process of assigning an object to a certain class based on its similarities, classification predict categorical and involves identifying a model that distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class level is not known”, Classification and prediction methods involves two techniques in analysis of data which can be used to extract models and describing the relevance of the data classes in and predict future data trends. This analysis can help to provide experts with a better understanding of the data classified for prediction purpose.

To analyze the effectiveness of the three classifiers chosen which are C4.5 algorithm, the ANN, and the SVM algorithm and to develop the best model for predicting patients with type 2 diabetes, the study employ to apply classification techniques..

3.6 Decision tree

One of the most popular classification techniques used in data analysis in the medical field is the decision tree. (Nong, 2013), because of its numerous advantages over statistical approaches, it is very easy to visualize and understand using software tools obtained from Waikato Environment for Knowledge Analysis (WEKA) developed at the university Waikato, New Zealand. This software is free under the GNU license, which contains visualizing tools and algorithm for data analysis and predicting modelling, normally resistance to noisy data. In the user interface of the software the decision tree construction can be used to classify records to a proper class. Furthermore, to classify the data items, a decision tree classifier constructs an attribute tree structure. The decision tree divides the test data set into many classifications based on the value of the characteristics and conditions.

These characteristics, which resemble a tree structure with nodes that specify conditional attributes of symptoms $symbol\ of\ \ S = \{S1, S2, \ldots \ldots Sn\}$, with branches that show value of Vi,n for example h-th ranges for i-th symptoms and leaves which present decision $D = \{di, \ldots \ldots \ldots dk\}$ in which their vinary value Wdk ={0,1}

3.6.1 C4.5 Algorithm

Developed by Ross Quinlan, C4.5 is an algorithm that creates a decision tree. The core ID3 method has been improved with the help of C4.5, a software extension. (Mitchell, 1997). The C4.5 technique also frequently handles missing data and values and can build decision trees that can be used for categorization and handling continuous properties. Moreover C4.5 can easily generates rules from a single tree which can as well convert to generate into multiple tree construction and create classification rules, for this reason, C4.5 is often called as statistical classifier and is widely proven in medical diagnosis study.. Based on this advantages as well as important of this classifier the experiment on decision tree was induce using C4.5 algorithm. The data set was loaded in WEKA explorer and the data was saved in arffformat as shown below on fig.3.2 interface of WEKA graphic explorer. With decision tree structure also shown on fig 3.3 at the same page indicating C4.5 algorithm.

Fig. 3.2. WEKA Explorer Interface (Source WEKA Interface Apps)

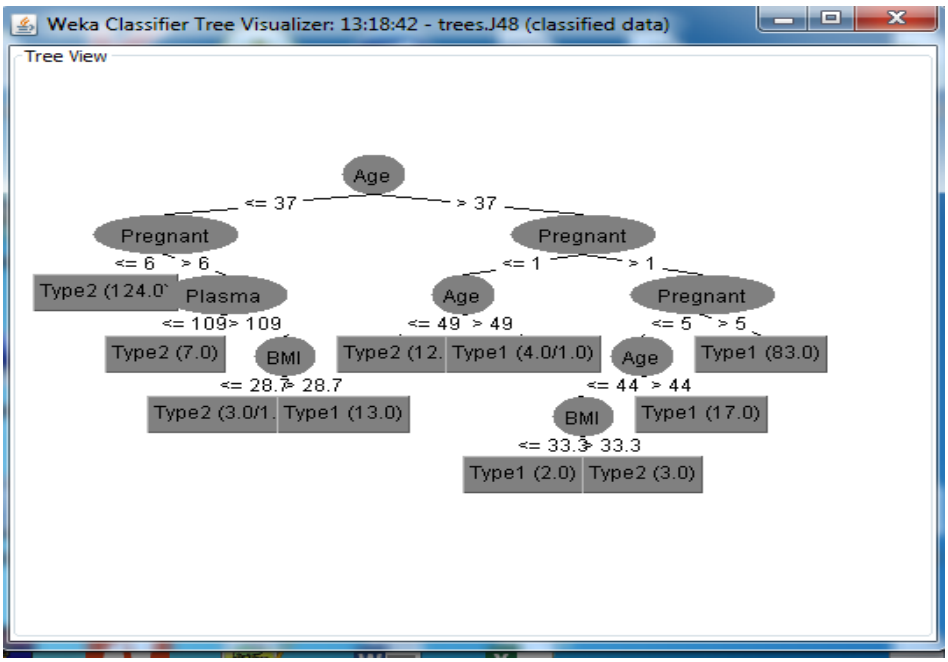

Fig. 3.3. Visualize Decision tree structure after the experiment conducted in WEKA

3.7 Neural Network Classification

Neural networks are a relatively recent technique in data mining and medical domain compare to decision tree and Bayes network, but with a widely used class of mathematical model which can be applied to problems such as time series prediction, classification, and function or functional approximation. The reasons of selecting Neural Network is that because it performed task that linear program cannot when an element of neuron falls, it can continue without problem. Also it has been used to analyzed blood and Urine samples and can also truck glucose level, while this study is based on the diagnosing the diabetes which is the presence of glucose on patient. Fig 3.4 shown how Neural network experiment was conducted based on loaded data in the MatLab work space with target output after preprocessing was performed.

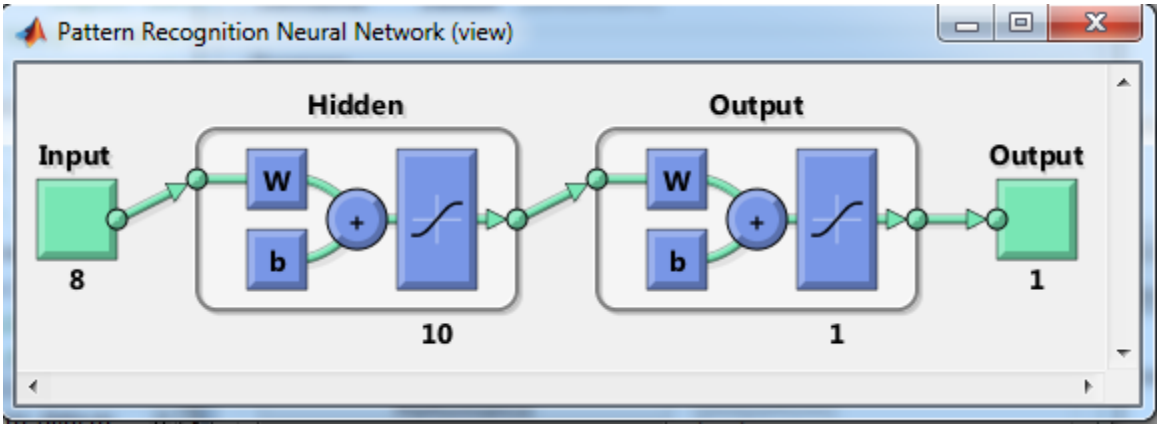

Fig. 3.4. Neural Network Design input/ output obtained after visualizing in the MATLAB Apps

3.7.1 Neural Network classification experiment with MATLAB

Experiment was conducted on MATLAB to develop the model, the reason of using MATLAB is because it has higher performance language for technical computing by integrating the computation and visualization.

Step one:

Data was loaded in to workspace of the MATLAB environment,
Original input and target output was given in excel format usually called matrices
By assigning normalize inputs and target output, preprocessing was done in order to normalize the data collection of input and output as Pn and Tn
Where P= normalize input and
T= target output
Mean standard deviation of original input and target where define before network was trained

Step two:

In order to determine the input, principal component analysis (PCA) was used. Comparison was made between eight (8) back propagation (BP) algorithms which are:-
Resilient (BP) (Rprog) RP
Fletcher Reeves conjugate Gradient BP
Polale-Ribiere Conjugate Gradient BP
Powell –Beale Conjugate Gradient BP
Lavenberg Marquardt BP
Scaled Conjugate Gradient BP
BF GS Quasi-Newton BP
One-step secout BP
Comparison was made based on R-values (accuracy) and mean square errors (MSE).
The one with highest R-values (accuracy) and least MSE was selected as the NN algorithm
Data training validation and testing
The data was divided in to training, validation and testing subset in which
¼ of the data was taken for Validation
½ of the data for training
¼ of the data for testing
Meaning that out of 768 data set for diabetes 192 was used for Validation, 384 for training and 192 for testing

3.7.2 Identification of the Optimal Neural Network Model

After conducting an optimization between neuron number and mean square error (MSE), the best BP algorithm selected a two-layer NN as the best training method among the eight BP algorithms. In optimizing the NN, five (5) neurons was first used in the hidden layer as an initial guess, followed with increased number of neurons in order to get an optimum neuron values obtained for the training set for the selected BP algorithm, the Neural Network training experiment was carried out on MATLAB and train as shown in figure 3.5 below.

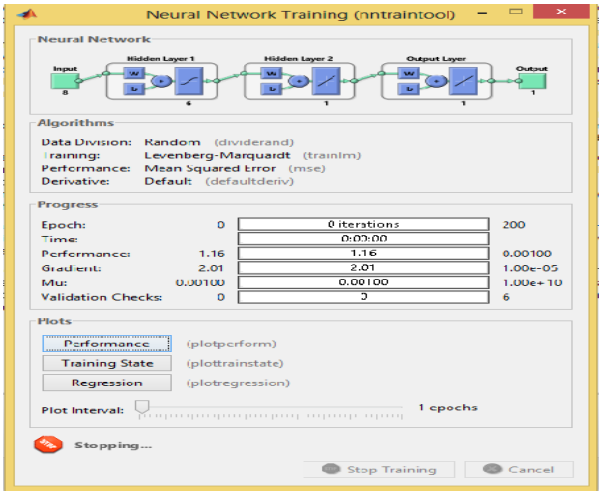

Fig. 3.5. Neural Network Training experiment on MATLAB

3.8 **Support vector machines (SVM) Classification**

SVM is a supervised learning technique that can classify both linear and nonlinear data. It was once considered the best option for classification problems in many biomedical fields, including bioinformatics, (Yu, Liu, Valdez, Gwinn, & Khoury, 2010). SVM was developed by Vepnik (1995) and has been studied increasingly in recent years. It was applied to the problem of diagnosis of diabetes diseases in several research works due to outstanding characteristics and excellent generalization performance, (Priya, &Rajalaxmi, 2012), 2011; Polat et al., 2008).

3.8.1 SVM experiment

The experiment conducted using WEKA tool environment to illustrate methods on detecting person with presence of diabetes based on this classification. We used data set from (PIDD) from UCI Machine learning repository.

The data set is stored in excel sheet save as CSV format and preprocessing procedure was also done similar to C4.5 algorithm in which data was clean. The data was further examined using normalization procedure to determine appropriate value of some attributes whose feature a non-numeric and convert them to numeric format such as sex which determine person is either male or female in which the value will be change to male as 1 while female as 0 value and any other non-numeric variable features as shown below diagram an example of Support vector machine (SVM) process experiment as shown in fig 3.6 below with input being scattered to obtain the value (sources scikit-learn.org)

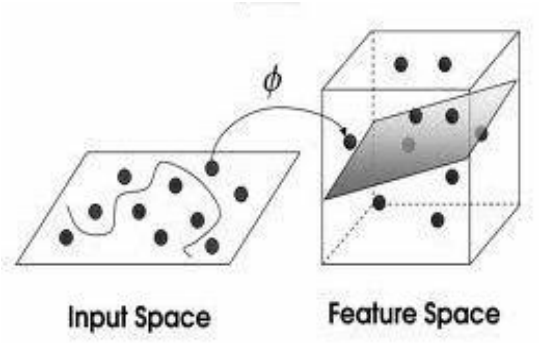

Fig. 3.6. SVM process experiment (source scikit-learn.org)

3.9 Data split into training and testing

Based on the outcomes of the three different classification experiment conducted, was evaluated and compare the result of the classifier method for new model classification in order to obtained best classification accuracy. The performance of this model is measured using Accuracy, Precision, Sensitivity, and Specificity.(Meng, Huang, Rao, Zhang, & Liu, 2013),(Polat, Güneş, & Arslan, 2008). Which refer to as

Accuracy. Is the measurement tolerance that define the errors limits made when the tools is used to determine classified accurate result. Which denotes: Number of correctly classified variables in the text set for diabetes disease data set divide by the total number of variable in the text set which is given in the formula

Accuracy = (TP+TN)/ (TP+TN+FP+FN

The **Table 3.2** below shows the result obtained for accuracy measure for the normalized value with Min =0 and Max=1

Table 3.2 Accuracy measure for Normalize value of Min=0 and Max=1

Accuracy		
10 fold C.V	C4.5(J48)	SVM (RBF)
	75.27%	78.18%

Sensitivity: also known as the true positive (recognition) rate (that is, the proportion of positive tuples that are correctly identified) while

specificity represents the genuine negative rate (that is, the proportion of negative tuples that are correctly identified).

Sensitivity = TP/(TP + FN)

Specificity =TN/(FP + TN)

Where TP=Represent True positive

TN = True negative

FN= False negative and

FP= False positive

Example TP = a /a + b

False Positive (FP) is the percentage of instances among all instances that are not of the same class C that are categorized as class C but actually belong to a different class**.**
Example. FP= C/ c + d

Precision is the percentage of instances correctly classified as belonging to class C over all instances correctly classified as belonging to class C.. therefore precision measure characterized the system probability by providing equal result. The value of precision is calculated by using average (mean) value of the data and also use average as well to represent the best estimate of the result in order to obtain an optimal result after evaluation as shown in Table 3.3 with data split into training and testing to obtain the value also table 3.4 enumerate performance accuracy measure for the normalize testing data after conducting the experiment to identify C4.5 algorithm with SVM accuracy measure using 10 fold cross validation shown in table 3.5

Table 3.3: data split into training and testing in order to evaluate the accuracy

Data set	**No. of training data**		**No. of Test data**		**Total**
768	537 (70%)		230(30%0		768
	+ve	-ve	+ve	-ve	
	183	354	88	144	

Table 3.4: Performance accuracy measure for Normalize Testing data

Accuracy		
10 fold C.V	C4.5(J48)	SVM (RBF)
	74.19%	74.19%

Table 3.5 Accuracy measure of Type1&type2 data set

	Accuracy
	C4.5
	J48
10 fold C.V	96.28%

Table 3.6 confusion Matrix for prediction type 1 and type2

		PREDICTION	
	P	POSITIVE	NEGATIVE
		TP=a	FN=b
	N	FP=c	TN=d

Example
In the above table 3.6 outcomes of the confusion matrix the table show prediction outcomes of Type 1 and Type 2 diabetes identification using true positive and false positive measure
Therefore for a single prediction there are possibilities of four outcomes namely: true positive, true negative, false positive and false negative. With these values the accuracy, sensitivity and specificity can be calculated.

3.10 Performance evaluation Accuracy measure

The data presented below represent our ideal model for classifying Type 1 and Type 2 diabetes in table 3.7 below to predict the outcomes of Accuracy, Sensitivity, Precision, and S. After the data set was classified and the result of the accuracy measure was obtained, the data set was further evaluated to obtain the performance accuracy using the Accuracy, Sensitivity, Specificity, and Precision.

Table 3.7: Evaluation Result of classified Model Two for Type 1 and Type 2 classification

	C4.5 (J48) Classifier
Accuracy	96%
Sensitivity	91%
Specificity	1%
Precision	1%

3.10.1 Comparison between the classifier

Though both the classifier have shown moderate accuracy model for the classification which was developed for classification of diabetes data set. The experiment was conducted in WEKA explorer as well as MATLAB R2012a, the data set are stored in MS excel document and read directly to MATLAB and WEKA. The diagnostic of the Performance of developed models is evaluated using true positive, false positive, true negative, false negative with accuracy sensitivity, specificity and precision.
The first model developed was observed between decision tree, SVM and Neural Network, data was split in two for testing and training in order to validated the accuracy the decision tree C4.5(J48) was found to be higher accuracy with 85% (testing accuracy measure) compare to

SVM(RBF) and NN(FFBP) with 75 and 67 respectively, while for the data set of VCU on 10 fold cross validation shows the accuracy of J48 and SVM on RBF to be 98% and 97% respectively, but when it comes to true positive false positive and true negative the accuracy of specificity , sensitivity and precision was bad which indicate 2%, , 3% and 1% respectively. Model for the classification, decision tree was build based on the highest classifier that give accuracy result which is J48 algorithm, can be find in figure 4.2 and the table represent Model one of our work can find below on table 3.8 as shown

Table 3.8 Model one The second model after clustering to find the class of type1 and type 2, classification was performed with cluster zero (0) identify as Type1 diabetes and cluster (1) was identify as Type 2 diabetes based on expert opinion as shown below on the

Table 3.9.

	J48	**RBF**	**NN(FFBP)**
Accuracy	85%	75%	67
Sensitivity	72%	55%	73
Specificity	93%	88%	78
Precision	88%	74%	47

4.0 CONCLUSION

This research work provide some of the new approaches about type1 and type2 diabetes among their classes. Within this research empirical study of different learning approaches have been considered and investigated, decision tree using C4.5 algorithm shows it advantages over others in terms of generalize performance. The performance of classifier is quite problem dependents, influential factors may be able to influence many factors which include data set itself, the way to split the data, selected features, result of Pima Indian data set and decision tree perform well consider to other approaches, this indicates use of many complex classifiers may not be necessary for better classifier. Henceforth it is suggested to use decision tree for prediction of similar data set like that of Pima Indian

5.0 Recommendation

Further investigation based on large data set should be carried out, since data mining shows it power on large amount of data respectively, work found based on this research should be reviewed further by collaborating with diabetes expert and machine learning professional for the sake of efficiency

REFERENCES

Anand, R., Pratap, V., Kirar, S., & Burse, K. (2012). Data Pre-processing and Neural Network Algorithms for Diagnosis of Type II Diabetes : A Survey, (1), 49–52.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., &Uthurusamy, R. G. R. (1996). Advances in knowledge discovery and data mining. Menlo Park, CA: AAAI Press/The MIT Press.

Guidelines, C., & Force, T. (2012). Global Guideline for Type 2 Diabetes
.

Han, J., & Kamber, M. (2006). Data Mining : Concepts and Techniques (2nd edition) Bibliographic Notes for Chapter 6 Classification and Prediction.

Key statistics on diabetes Contents. (2012)

Koklu, M., &Unal, Y.(2013).Analysisof aPopulation of DiabeticPatientsDatabases with Classifiers, (8), 222–224.

Kumari, V. A. (2013). Classification Of Diabetes Disease Using Support Vector Machine, 3(2), 1797–1801.

Meng, X., Huang, Y., Rao, D., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung Journal of Medical Sciences, 29(2), 93–99. doi:10.1016/j.kjms.2012.08.016

Mitchell, T. M. (1997). Machine learning. Redmond McGraw-Hill

Nong, Y. (2013). Handbook of data mining. Lawrence Earlbain associate.

Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications, 37(12), 8102–8108. doi:10.1016/j.eswa.2010.05.078

Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine. Expert Systems with Applications, 34(1), 482–487. doi:10.1016/j.eswa.2006.09.012

Priya, S., Rajalaxmi, R. R. (2012).An Improved Data Mining Model to Predict the Occurrenceof Type-2 Diabetes using Neural Network, (Icon3c), 26–29.

Rahman, R. M., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, 2013(March), 85–97. doi:10.4236/jsea.2013.6301

Shanker, M., & Hu, M.Y. (1999).Estimating Probabilities of Diabetes Mellitus.

Temurtas, H., et al.(2009). A comparative study on diabetes disease diagnosis using neural networks.Expert Systems with Applications,36(4),8610–8615.

doi:10.1016/j.eswa.2008.10.032

United, S. (2014). National Diabetes Statistics Report, 2014 Estimates of Diabetes and Its Burden in the Epidemiologic estimation methods, 2009–2012. Retrieved from http://www.cdc.gov/diabetes/pub/references/4.htm

Yu, W.,Liu, T., Valdez,R., Gwinn, M., &Khoury,M. J. (2010).Application ofsupport vector machinemodeling forprediction ofcommon diseases : the caseof diabetes and pre-diabetes.

Zolfaghari, R. (2012).Diagnosis ofDiabetes inFemalePopulation of Pima Indian Heritagewith Ensembleof BPNeural Network and SVM,15(4), 11