A Hybrid Intelligent System Design for Diabetes Risk Classification

Thirumalaimuthu Thirumalaiappan Ramanathan, B.E (AU)

Faculty of Education, Science, Technology and Mathematics

University of Canberra

November 2015

A Dissertation Submitted in partial fulfillment of the requirements for the degree of Master of Information Sciences (Research)

Abstract

Risk classification is a major technical challenge in medical diagnosis and chronic illness management. Various computational techniques have been developed for risk classification in recent years with improvements in outcomes. This thesis investigates a novel approach combining support vector machine (SVM) and fuzzy modelling. The proposed hybrid model (called SVM-Fuzzy) is designed, implemented and evaluated on an available benchmark dataset.

Diagnosis and management of diabetes mellitus (also known as type 2 diabetes) are the motivating problem for the current investigation of risk classification. Type 2 diabetes is a chronic condition marked by elevated levels of blood glucose. The prevalence of diabetes is increasing at a fast pace due to obesity, in particular, central obesity, physical inactivity, and unhealthy dietary habits. In type 2 diabetes patients, a range of different organs are under stress and are influenced by the altered metabolic condition. Hence the ailment has to be properly managed within stipulated boundaries to minimize other health complications caused by type 2 diabetes and to assure longevity of life. In the proposed model, fuzzy reasoning is used to classify the level of risks from data. SVM is used to design the fuzzy rules. The well-known Pima Indian diabetes dataset (Frank, 2010) is used to train the SVM and for testing the fuzzy system. The goal is to evaluate the proposed design for better accuracy in risk classification and to investigate training the machine learning algorithm using sample real world data. Another goal is also to investigate efficiency in classification by optimizing selection of right sized datasets through appropriate dataset size selection from experiments. The experiments from the SVM-Fuzzy model show promising results on the benchmark Pima Indian diabetes dataset.

Early in the thesis the various soft computing approaches, such as artificial neural network, SVM, Bayesian network and fuzzy logic models which have been previously reported as relevant to diabetes risk classification are reviewed. The drawbacks in these models are analyzed and identified gaps from the previous computational models on risk classification are targeted in the research questions.

The main objective in risk classification is to handle uncertain and incomplete input data for risk classification and to make use of sample datasets using fuzzy reasoning. The experimental results from the proposed model (SVM-Fuzzy) from the type 2 diabetes risk classification problem demonstrates that it is able to handle uncertain input data by making use of sample datasets for fuzzy reasoning. The outcomes from the proposed

model on type 2 diabetes risk classification are summarized and compared with the previous computational models. These experimental results are encouraging. The SVM-Fuzzy model can be further extended to manage type 2 diabetes risks using its output risk values in planning lifestyle and diet for responsive management of type 2 diabetes conditions. The model can further be extended to a multi-agent learning and planning environment for solving classification problems in complex applications, such as extracting patterns in structure of proteins in bioinformatics.

Acknowledgement

I thank my God for giving me strength and knowledge to conduct the research and write this thesis. I would like to convey my gratitude to my supervisor Professor Dharmendra Sharma for his valuable contribution and expertise. I discussed all my ideas with him and he gave me specific direction to follow on. His guidance, motivation and encouragement helped in every step of this research work. My thanks to Professor Xu Huang, Associate Professor Dat Tran and Dr Chris Barnes provided for their guidance and support. I would also like to extend my appreciation to the Faculty of Education, Sciences, Technology and Mathematics for their support for my research. My special thanks to my parents and my uncle Professor Kaliappa Kalirajan (from the Australian National University) for keeping me motivated and for all their love and understanding throughout the journey.

CONTENTS

List of Tables and Figures

Chapter 1

Introduction

This chapter describes what this thesis is about: why diabetes management is taken as the motivational problem; the research problem in diabetes management; the objectives of the thesis; the research question that arises in diabetes management; the methodology used for the research; and the organization of the thesis.

1.1 Motivation

Diabetes management is taken as the motivating problem for developing a risk classification system. The motivation for this thesis is to develop an algorithm to improve the classification of diabetes risks.

As reported in (Health Direct Australia, 2013), diabetes is a group of metabolic diseases in which a person has high blood sugar, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced. There are two main types of diabetes – type 1 and type 2. Type 1 diabetes results from the body's failure to produce insulin, and currently requires the person to inject insulin or wear an insulin pump. Type 2 diabetes results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. The prevalence of diabetes is increasing at a fast pace due to obesity, in particular, central obesity, physical inactivity, and unhealthy dietary habits (Uusitupa, 2002).

As reported in (Government Must Act Now to Stop Diabetes, 2012), type 2 diabetes is a chronic condition marked by elevated levels of blood glucose. In addition to health and social issues, diabetes management also costs individuals significantly. Type 2 diabetes costs the Australian Government and Australian taxpayers more than an estimated $6 billion per year in direct costs and is the fastest growing major chronic illness in Australia, with numbers set to more than double from 1.5 million to 3.5 million in the next 20 years if nothing is done to address the burgeoning pandemic. Efforts have been made in recent years to develop computational techniques for aiding the identification of risks in diabetic patients. Type 2 diabetes is a condition where there is too much glucose, a type of sugar, in the blood. In type 2 diabetes patients different organs are under stress and are influenced by the altered metabolic condition. Poor management of type 2 diabetes control may result in organ failure, low quality of life (higher morbidity) and a significant reduction in

longevity. Hence the ailment has to be properly managed within stipulated guidelines to minimize other health complications caused by type 2 diabetes and to assure longevity of life. Classification systems have been widely utilized in the medical domain to investigate and model patient's data and extract a predictive model. They help physicians to improve their prognosis, diagnosis or treatment planning procedures.

1.2 Research problem

Type 2 diabetes can be managed by following given lifestyle guidelines (Government Must Act Now to Stop Diabetes, 2012). The goal of this research is to provide lifestyle guidelines to type 2 diabetes patients according to their perceived risk level. The whole project contains two phases: the first phase is to develop a computational model that is able to classify the diabetic risks efficiently and the second phase is to develop an intelligent planning system that is able to interrogate the patient's data to derive guidelines according to their risk level, in order to optimally manage their health conditions. This thesis contributes to the first phase which involves describing a computational model that is able to classify the level of risks with high accuracy. The outcomes from this thesis can be extended to give guidelines to patients through the development of an intelligent planner.

Computational techniques can help type 2 diabetes patients by classifying and modelling their diabetic risks and advising diabetic patients on how to manage them. This research involves developing a "soft" computing system design for classifying type 2 diabetes risk with better accuracy than available at present. There are many computational techniques applied to classify type 2 diabetes risk which will be described in Chapter 2. The problem with these techniques is that they are not able to classify the risks clearly. Most of the computational techniques used in the recent years provide binary classification of whether the patients are diabetic or non-diabetic. The computational challenge in this thesis is to classify the risk more accurately such as showing the level of risks from the input data. Instead of classifying whether the person is diabetic or non-diabetic showing their level of risks in diabetic will be more useful to people as this may help the type 2 diabetes patients to manage their health conditions by following proper life style according to their risks. The computational system design reported in this thesis is about investigating risk classification for people with type 2 diabetes but the model can also be applied to determine the levels of risk for non-diabetics.

1.3 Objectives

The main objective of this thesis is to investigate a novel, hybrid computational model for risk classification and to evaluate its accuracy and efficiency. The proposed model combines support vector machine (SVM) (Morik et al., 1998) and fuzzy modeling (Zadeh, 1965). An optimal sized data set is determined by the SVM-Fuzzy model for improved efficiency.

1.4 Research questions

There are various problems found in solving type 2 diabetes risks classification and diabetes management. The major research questions include

1) *Can the level of risk be modelled as a classification problem?*
2) *Can an improved algorithm be developed to provide optimal classification?*
3) *Can mining a smaller subset from the data set would result in same outcomes as bigger data sets?*
4) *Can the risk classification results be used for developing an intelligent planning system for providing guidelines?*
5) *Can the risk classification algorithm be improved using multi-agent learning?*

1.5 Methodology

A build research methodology is used where a new system design is proposed which is a combined approach of SVM (Morik et al., 1998) and fuzzy logic (Novák et al., 1999) for classifying risks in type 2 diabetes. SVM is used to extract pattern from a diabetes dataset. The extracted patterns are used in the rule base of a fuzzy expert system to reason and classify the risk level. The main focus of this research is to classify the level of risks in type 2 diabetes by using fuzzy logic with extracting information from the proved cases in type 2 diabetes for the fuzzy reasoning. In addition an experimental methodology is used to determine the efficiency of the proposed system in risk classification by testing with different sized diabetes datasets. The goal is to determine whether a subset of large dataset is sufficient to train the machine learning algorithm. As reported in (Elio et al., 2011), a build research methodology consists of building an artifact either a physical artifact or a software system to demonstrate that it is possible. An experimental methodology is about taking measurements that will help identify what are the questions that should be asked

about the system and evaluating the system to answer these questions.

1.6 Thesis organization

The remainder of the thesis is organized as follows: Chapter 2 is dedicated for literature review on soft computing techniques for risk classification in diabetes. This chapter describes the algorithm used for risks classification. It also identifies gaps in the soft computing approaches in classification. Chapter 3 describes the data sets, risk classification and the proposed algorithm for risk classification. Chapter 4 elaborates the proposed system design for risk classification. Chapter 5 describes the performance of the proposed model. Chapter 6 concludes the thesis with some suggestions for future work. Appendix A gives details of a published paper as the outcome of this research work. Appendix B presents a snapshot of Pima Indian diabetes dataset from (Frank, 2010) and gives its description.

Chapter 2

Advances in Computational Classification: A Review

2.1 Introduction

Computation classification is the problem of identifying to which of a set of categories (sub-population) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known (Alpaydin, 2010). Risk classification is an instance of the classification problem. Risk classification is the formulation of different premiums for the same coverage based on group characteristics (A Practical Guide to Risk Assessment, 2008). Many computational models have been developed and applied for risk classification in recent years. This chapter is dedicated to a literature review on computational approaches for risk classification. It describes various algorithm proposed for risk classification; how they are used; their inputs, outputs; and summarizes their limitations. The classification accuracy of the reviewed models is listed and their drawbacks are analyzed.

2.2 Artificial Neural Network

Stergiou et al (2003) defines artificial neural network (ANN) as an information processing paradigm that is inspired by the way of biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.

Dowla et al (1995) described that the ANN learns to solve problems by example rather than following a set of heuristics or theoretical mechanisms. It has the ability to abstract or to respond appropriately to input patterns different from those involved in the training of the network. A particular strength of the ANN approach is its ability to identify nonlinearities of phenomena. A well designed ANN can be relatively tolerant to noisy, incomplete or even spurious data. Once trained neural network have synthesized understanding into a compact system of weights, they become portable and easy to insert

into other systems. They can be easily retrained with new data and inserted to upgrade existing systems.

Tu (1996) describes ANN as a "black box" and have limited ability to explicitly identify possible casual relationships. ANN modelling requires greater computational resources. ANN models may be more difficult to use in the field and are prone to over fitting. ANN model development is empirical, and many methodological issues remain to be resolved.

2.2.1 Artificial Neural Network for Classification

Dey et al (2008) applied backpropagation algorithm of ANN for the classification problem of diagnosis of diabetes. For the dataset, data of 530 patients from Sikkim Manipal Institute of Medical Science hospital, Sikkim, India were collected out of which 249 were suffering from diabetes and rest were non-diabetic. The inputs to these architectures are the parameters random blood sugar test result, fasting blood sugar test result, post plasma blood sugar test, age, sex and occupation.

The output is either 0 or 1, where 0 indicates non-diabetic and 1 indicates diabetic. The results presented for the diabetes classification problem validates the fact that, the network is able to classify diabetic and non-diabetic patients with the network performance of 92.5%.

2.2.2 Drawbacks of Artificial Neural Network in Classification

The ANN classifies the risk into diabetic and non-diabetic. The output of risk classification using ANN limits binary classification. The level of risks in diabetes needs to be modeled which is done in this thesis.

2.3 Support Vector Machine

Morik et al (1998) describes SVM that it operates by finding a linear hyper plane that separates the positive and negative examples with a maximum interclass distance or margin d. In the case of unequal misclassification costs, a cost factor J (C+ /C−) is introduced by which training errors on positive examples outweigh errors on negative examples. Therefore, the optimization problem becomes

$$\text{minimize } \frac{1}{2}\|w\|^2 + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{j:y_j=-1} \xi_j$$

$$\text{Subject to } y_k(wx_k + b) \geq 1 - \xi_k, \qquad \xi_k \geq 0.$$

where y_i is the class label, w is normal to the hyper-plane, |b| ||w|| is the perpendicular distance from the hyper-plane to the origin, ||w|| is the Euclidean norm of w, C is a regularization parameter, which defines the tradeoff between the training error and the margin d, and ξ_i is a slack variable to allow errors in classification. For handling nonlinearly separable data, kernel functions are used, including kernel functions and a Lagrange multiplier α_i, the dual optimization problem becomes

$$\text{maximize } w(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{l} \alpha_i y_i \alpha_j y_j K(x_i . x_j)$$

$$C \geq \alpha_i \geq 0 \qquad \forall_i, \qquad \sum_{i=1}^{l} \alpha_i y_i = 0.$$

Solving for α, training examples with a nonzero α are called support vectors (SVs) and the hyper-plane is completely defined by the SVs alone.

Auria et al (2008) comment that SVM that by introducing the kernel, SVMs gain flexibility in the choice of the form of the threshold separating solvent from insolvent companies, which needs not be linear and even needs not have the same functional form for all data, since its function is non-parametric and operates locally. Since the kernel implicitly contains a non-linear transformation, no assumption about the functional form of the transformation, which makes data linearly separable, is necessary. The transformation occurs implicitly on a robust theoretical basis and human expertise judgment beforehand is not needed. SVMs provide a good out of sample generalization, if the parameters (in case of a Gaussian kernel) are approximately chosen. SVMs deliver a unique solution, since the optimality problem is convex.

The disadvantage of SVM is embodied in determining the appropriate kernel and its parameters to optimize the overall performance of the machine (Sammany and Zaghloul, 2006). Another limitation as mentioned in (Bebis, 2010) is speed and size. For large training sets, it typically selects a small number of support vectors, thereby minimizing the computational requirements during testing. Below describes two SVM models which has been developed and applied for diabetes risk classification.

2.3.1 Intelligible SVM Model for Classification

Barakat et al (2010) employed SVM for diabetes diagnosis where an additional rule-based explanation component is utilized to provide comprehensibility. The SVM and the rules extracted from it are intended to work as a second opinion for diagnosis of and as a tool to predict diabetes through identifying people at high risk. For the dataset data from

3014 subjects of age 20 years and above was collected. The inputs are age, sex, family history of diabetes, body mass index (BMI), waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, cholesterol, fasting blood sugar, 2 hour post glucose load. The output class is diabetic or non-diabetic. The training procedure was used as follows.

1) *A number of SVM models were generated by varying the misclassification cost factor J, starting with a small value and increasing J until no change in true positive or false positive rates was observed.*
2) *Each of the generated models was then used to classify the independent test set and accuracy, true positive and false positive rates calculated.*
3) *Rules were then extracted from each of the models using SQRex-SVM and eclectic methods.*
4) *Each of the extracted rulesets were then used to classify the same independent test set and again accuracy, true positive and false positive rates, as well as fidelity were computed.*

Results on a real-life diabetes dataset show that intelligible SVMs provide a promising tool for the prediction of diabetes, where comprehensible ruleset have been generated with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. The predictive results obtained by the proposed rule needs to further refined.

2.3.2 Region based SVM algorithm for Classification

Karatsiolis et al (2012) handled the problem of diagnosing Pima Indian diabetes dataset with a modified SVM strategy. The proposed algorithm divides the training set into two subsets, one that arises from the joining of coherent data regions and the second subset comprises of data portions that is difficult to be clustered. The first subset is used to train the SVM with radial basis function (RBF) kernel and the second subset is used to train the SVM with polynomial kernel. The inputs are the parameters plasma concentration, blood pressure, insulin, BMI, diabetes pedigree function, age, triceps thickness (mm), no. of times pregnant. The output is either positive (diabetic) or negative (non-diabetic). The following steps define the algorithm's major functionalities:

1) *Initialize the genetic algorithm (GA) with a population of N individuals with random genes and train initial SVMs accordingly. For each population individual there are two sets of information: the training sets that are used by the RBF SVMs to create combined regions of classification and the actual SVM set.*
2) *Fitness function reflects the search for solely positive regions. Each individual's*

fitness function is the number of positive cases that are classified as positive by all of its SVM models over the number of negative cases that are classified as positive by all of its SVM models.

3) *The population is evolved through crossover and mutation by maximizing fitness function. After a genetic operator is applied to an individual its corresponding SVMs must be trained again.*
4) *If after an epoch an individual has a high fitness function (meaning zero negative class cases) and the number of the positive cases is more than or equal to 5% of the total positive data set examples, then the region is saved an the included cases are removed from the total training set. The process is repeated from step (1).*

The presented algorithm has limitation that a small dataset is not eligible for solving with the presented algorithm because of the algorithm's natural approach to divide the dataset to clusters which in turn reduce the size of the test sets making the test phase unreliable, prone to over fitting or even unfeasible when just a bunch of test examples are available.

2.3.3 Drawbacks of SVM Models in Classification

The classification accuracy of Intelligible SVM and Regional SVM are only 94% and 82.2% respectively in diabetes risk classification. This motivates this thesis to improve the accuracy in diabetes risk classification by using the SVM in a better way which is investigated in this thesis.

2.4. Fuzzy logic

The term fuzzy logic was introduced with the 1965 proposal of fuzzy set theory by (Zadeh, 1965). Fuzzy logic (Novak et al., 1999) is a form of many-valued logic that deals with approximate, rather than fixed and exact reasoning. Compared to traditional binary logic (where variables may take on true or false values), fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions (Ahlawat et al., 2014).

The generic architecture of the fuzzy expert system is shown in the Figure 2.1 (Mendel, 2001). In the fuzzy expert system, the crisp inputs are fuzzified into fuzzy sets. In fuzzification, the degree to which each of the crisp inputs belongs to the appropriate fuzzy

sets is determined. The rules are used to reason the fuzzy sets and give inference. The output fuzzy sets are defuzzified into crisp outputs. In defuzzification, the output fuzzy set is aggregated into a single number.

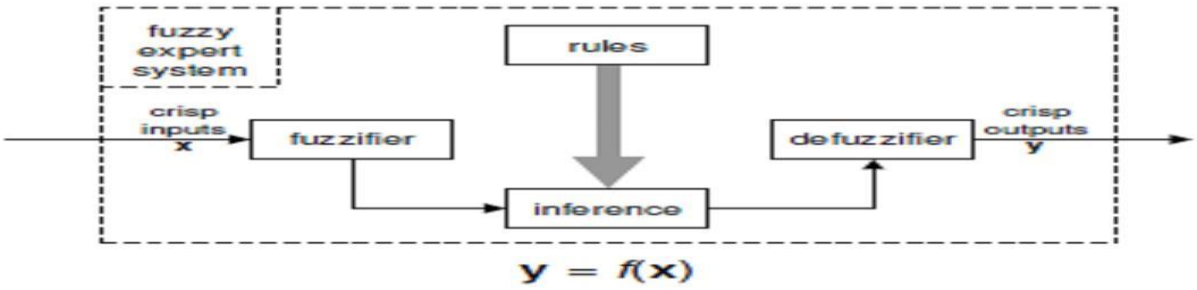

Figure 2.1: Generic architecture of a fuzzy expert system (Mendel, 2001)

Below describes a fuzzy inference technique.

2.4.1 Mamdani-Style Fuzzy Inference

Fuzzy inference can be defined as a process of mapping from a given input to an output, using the theory of fuzzy sets (Negnevitsky, 2011). The most commonly used fuzzy inference technique is the so-called Mamdani method. In 1975, Mamdani built one of the first fuzzy systems to control a steam engine and boiler combination (Mamdani and Assilian, 1975). He applied a set of fuzzy rules supplied by experienced human operators.

The Mamdani-style fuzzy inference process is performed in four steps (Mamdani and Assilian, 1975): fuzzification of the input variables, rule evaluation, aggregation of the rule outputs, and finally defuzzification.

Step1: ***Fuzzification***

The first step is to take the crisp inputs and determine the degree to which these inputs belong to each of the appropriate fuzzy sets.

Step2: ***Rule evaluation***

The second step is to take the fuzzified inputs and apply them to the antecedents of the fuzzy rules. If a given fuzzy rule has multiple antecedents, the fuzzy operator (AND or OR) is used to obtain a single number that represents the result of the antecedent evaluation. This number is then applied to the consequent membership function (MF). To evaluate the disjunction of the rule antecedents, the OR fuzzy operation will be used.

Similarly, in order to evaluate the conjunction o the rule antecedents, the AND fuzzy operation is applied.

Step3: ***Aggregation of the rule outputs***

Aggregation is the process of unification of the outputs of all rules. In other words, the MFs of all rule consequents previously clipped or scaled is chosen and combining them into a single fuzzy set. Thus the input of the aggregation process is the list of clipped or scaled consequent MFs, and the output is one fuzzy set for each output variable.

Step4: ***Defuzzification***

The last step in the fuzzy inference process is defuzzification. Fuzziness helps us to evaluate the rules, but the final output of a fuzzy system has to be a crisp number. The input for the defuzzification process is the aggregate output fuzzy set and the output is a single number.

2.4.2 Fuzzy Logic for Classification

Narashiman et al (2014) used fuzzy logic for risk classification of diabetic nephropathy. The input parameters are plasma concentration, diastolic blood pressure, BMI and age. Mamdani type fuzzy inference system is used. 25 numbers of rules are given for the risk prediction. The Pima women diabetes dataset is taken for simulation.

2.4.3 Drawbacks of Fuzzy System in Classification

The fuzzy system depends on the fuzzy set and rules designed by a human expert in a particular field. The fuzzy system should make use of the patterns derived from available datasets. The fuzzy system is investigated in this thesis whether it is able to perform better through using sample dataset.

2.5. Hybrid Models

Various hybrid models have been proposed in the recent years for classifying risks in diabetes. The relevant ones are summarized and analyzed in this section below.

2.5.1 Hybrid Model using Modified PSO and Least Squares SVM for Classification

Soliman et al (2014) proposed a hybrid model using Modified Particle Swarm Optimization **(**PSO) and Least Squares SVM (LS-SVM) for diabetes risk classification. Below describes about PSO, Modified-PSO, LS-SVM and their hybrid model.

2.5.1.1 Particle Swarm Optimization

Blondin (2009) describes PSO as an algorithm inspired from the nature social behavior and dynamic movements and communications of insects, birds and fish. The main strength of PSO is its fast convergence, comparing with many global optimization algorithms like GA, Simulated Annealing and other global optimization algorithms. PSO is a technique used to explore the search space of a given problem to find the settings or parameters required to maximize a particular objective. The PSO algorithm works by simultaneously maintaining several candidate solutions in the search space. During each iteration of the algorithm, each candidate solution is evaluated by the objective function being optimized, determining the fitness of that solution. Each candidate solution can be thought of as a particle "flying" through the fitness landscape finding the maximum or minimum of the objective function.

2.5.1.2 Modified-PSO

Wang et al (2013) proposed a modified version of PSO. The main idea of this modified version is illustrated in the following algorithm.

1) *Initialize population of particles X(t) which consists of random positions $x_1, x_2, \dots, x_n$ and velocities V(t) are made up of the particle's initial velocity $v_1, v_2, \dots, v_n$ on n dimensions.*
2) *Evaluate the fitness for each particle.*
3) *For each particle, find the maximum fitness and compare it to the best found so far (pbest), if f(x_i) < f(pbestk), then f(pbestk) = x_i.*
4) *Set i^{th} position of the particles, P_i equals to the location of the maximum fitness value X_i.*
5) *Compare fitness evaluation with the population' overall previous best. If current value is better than gbest, then other gbest to the current particle's array index and value.*
6) *Calculate the convergence factor.*
7) *Calculate the inertia weight wid.*
8) *Update the positon of the particle and the new population X(t+1) will be generated.*
9) *Adjust the acceleration of the particles.*
10) *Loop to step (2) until stopping criterion is satisfied. (Reach a maximum number of iteration Tmax).*

2.5.1.3 Least Squares-SVM

Suykens et al (1999) describe that LS-SVM classifier is one particular sample of SVM and is used for finding an optimal hyper plane, which separates various classes. It obtains this optimal hyper-plane by using maximum Euclidean distance to the nearest point. It is a parametric algorithm that is popular with its sensitivity to the changes in the values of its parameters. Below is the algorithm of LS-SVM.

1) *Load the training data set of n data points, $\{x_k, y_k\}_{k=1}^{N}$ where x_i is the i^{th} input vector and $y_i \epsilon R$ is the corresponding i^{th}target with values {-1,+1}.*
2) *Generate random weights for each input data point.*
3) *Determine the value of the bias term b and initialize the error e for each point randomly.*
4) *Initialize γ and σ using random values.*
5) *Search for values that minimize the objectives function.*
6) *Construct the Lagrangian function with the solution that must satisfy the Karush-Kuhn Tucker conditions in the set.*
7) *Calculate number of support vectors.*
8) *Training data for LS-SVM model could be classified with RBF kernel function.*
9) *Classify any new point using RBF kernel function.*
10) *Loop until stopping criteria is met, usually until reach the maximum number of iterations.*

2.5.1.4 Integrating Modified-PSO and Least Squares-SVM

Below is the algorithm of hybrid model using Modified-PSO and LS-SVM for diabetes risk classification (Soliman et al., 2014).

1) *Load the dataset of n data points, $\{xk, yk\}_{k=1}^{N}$ where xi is the i^{th} input vector and yi ϵR is the corresponding i^{th}target with values{-1,+1}.*
2) *Generate random weights for each input data point.*
3) *Initialize the bias term b and the error e for each point randomly.*
4) *Find the optimal value for using algorithm 1.*
5) *Find the optimal values for the objective function.*
6) *Count number of support vectors.*
7) *Classify any new point using RBF function.*
8) *Loop until stopping criteria is met, usually until reach the maximum number of iterations.*

The hybrid algorithm was worked on Pima Indians diabetes dataset (Frank, 2010). The input to the Modified-PSO is total of 768 records. About 768 random individuals in the search space for

100 iterations. The output from the Modified-PSO is the optimal values for $\gamma\ and\ \sigma$, which are 100 and 0.5 respectively. LS-SVM is run with the optimized parameters and RBF kernel function seeking to find the optimal hyper plane that separates the search space into two classes. In the proposed algorithm, Modified-PSO is used as parameter optimization technique to improve the sitting of the parameter values of LS-SVM. LS-SVM is used for classification which consists of training and testing phase. The performance of the proposed algorithm was evaluated by calculating the classification accuracy (TP+TN/TP+TN+FP+FN) where TP,TN, FP, FN stand for true positive, true negative, false positive, false negative. The average classification accuracy of proposed algorithm which was worked on Pima diabetes dataset (Frank, 2010) showed 97.833%.

2.5.2 A Soft Intelligent Binary Model for Classification

A hybrid model of multilayer perceptron (MLP) is proposed using fuzzy logic in (Khashei et al., 2012). Below describes the MLP and the hybrid algorithm using MLP and fuzzy logic.

2.5.2.1 Multi-Layer Perceptron

Delashmit et al (2005) describe about MLP neural networks. MLP consist of units arranged in layers. Each layer is composed of nodes and in the fully connected networks considered in this paper each node connects to every node in subsequent layers. Each MLP is composed of a minimum of three layers consisting of an input layer, one or more hidden layer(s) and an output layer. The above definition ignores the degenerate linear multilayer perceptron consisting of only an input layer and an output layer. The input layer distributes the inputs to subsequent layers. Input nodes have linear activation functions and no thresholds. Each hidden unit node and each output node have thresholds associated with them in addition to the weights. The hidden unit nodes have nonlinear activation functions and the outputs have linear activation functions. Hence, each signal feeding into a node in a subsequent layer has the original input multiplied by a weight with a threshold added and then is passed through an activation function that may be linear or nonlinear (hidden units).

Watts (2013) describes that MLP with one hidden layer of sufficient size can approximate any continuous function to any desired accuracy. MLP can learn conditional probabilities. MLP are multivariate non-linear regression models can accurately model previously unseen examples. Some of the problems in using MLP is choosing the number of hidden layers, initialization of weights, the network forgets what it learned about the old data, it only knows about the new data problems with MLP.

2.5.2.2 Multi-Layer Perceptron using Fuzzy Logic

Khashei et al (2012) enhances MLP using fuzzy logic. The neural network is trained using the available information observations. Using the obtained weights the minimal fuzziness is determined. The membership probability of the output in each class is calculated. The output is assigned to appropriate class by the largest probability. The Pima diabetes dataset is used for training. The proposed model consists of five phases as follows:

1) *Training the neural network using the available information from observations.*
2) *Determining the minimal fuzziness using the obtained weights and same criterion.*
3) *Deleting the outliers in accordance with Ishibuchi's recommendations.*
4) *Calculating the membership probability of the output in each class.*
5) *Assigning the output to appropriate class by the largest probability.*

The MLP model using fuzzy logic does not require experimentation and final selection of a kernel function and a penalty parameter as is required by SVM. The model solely relies on a training process in order to identify the final classifier model. Finally, the model does not need large amount of data in order to yield accurate results, as traditional MLP.

2.5.3 A Hybrid Genetic Algorithm for Classification

Dalakleidi et al (2013) proposed a hybrid approach based on the combined use of GA and a nearest classifier for the selection for classification in patients with type 2 diabetes. Below describes the GA, K-nearest-neighbor (KNN).

2.5.3.1 Genetic Algorithm

Dulay (2001) defines GA as an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomized, GA are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate process in natural systems necessary for evolution, especially those follow the principles of survival of the fittest. Since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker one.

Navlakha et al (2013) describe that benefits of GA are its concept is easy to understand, it is modular, separate from application, it supports multi-objective optimization, it always gives an answer, the answer gets better with time, it is easy to exploit previous or alternate solutions and it is flexible for building blocks for hybrid applications.

Safaric (2011) comments that some of the disadvantages in using GA are certain optimization problems cannot be solved due to poorly known fitness functions, there is no absolute assurance that GA will find a global optimum and the GA cannot assure constant optimization response time.

2.5.3.2 K-Nearest-Neighbor

Peterson (2009) defines about KNN that KNN classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. KNN classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. Below describes the different version of KNN.

2.5.3.3 The Weighted KNN Algorithm

To weigh the closer neighbors more heavily than the farther ones, a Weighted KNN **(**WKNN) algorithm (Marinakis et al., 2009) is applied in which the distance-weighted function $w$to the i-th nearest neighbor is defined as,

$$w_i = k + 1 - i / \sum m$$

where m is an integer in the interval (l,k) and k is the total number of the neighbors. Thus all weights are in the interval ($1/\sum$m, k/$\sum$m), and a neighbor with smaller distance is weighed more heavily than one with greater distance.

2.5.3.4 The Distance-WKNN Algorithm

In order to address the effect of the number of neighbors on the classification performance, a Distance-WKNN (DWKNN) algorithm (Gou et al., 2011) has been proposed. The DWKNN algorithm gives different weights to the k nearest neighbors depending on distances between them and their ranking according to their distance from the query object. The distance-weighted function $w_i$of the i-th nearest neighbor is

computed as

$$w_i = ((d_k^{NN} - d_i^{NN})/(d_k^{NN} - d_i^{NN})) * 1/i \text{ , } d_k^{NN} \neq d_i^{NN} \text{ or } 1, d_k^{NN} = d_i^{NN}$$

where d_i^{NN} is the distance of the i-th nearest neighbor from the query object, d_1^{NN} is the distance of the nearest neighbor, and d_k^{NN} is the distance of the k-furthest neighbor. Thus the weight of the nearest neighbor is 1, and the weight of the furthest k-th neighbor is 0, whereas other weights are distributed between 0 and 1.

2.5.3.5 Combined use of Genetic Algorithm and K-Nearest-Neighbor

Dalakleidi et al (2013) applies a hybrid model using GA and a KNN classifier for the selection for the critical clinical features which are strongly related with the incidence of fatal and non-fatal Cardiovascular Disease (CVD) in patients with type 2 diabetes Mellitus. The feature selection task is performed by a hybrid GA, where the proposed subsets of features are ranked using different versions of a KNN classifier described above. For evaluation data from the medical records of 560 patients with type 2 diabetes are used. The dataset comprises 32 features providing information related to demographics, lifestyle, laboratory examinations, complications/comorbidities and treatment. The output class is either fatal or non-fatal CVD.

2.5.4 Drawbacks of Hybrid Models in Classification

The hybrid models reviewed above are not able to model the actual level of risks in diabetes. They give a result of yes or no but not the levels in between. This motivates this thesis to design an alternative hybrid model for showing the level of risks in diabetes patients.

2.6 Bayesian Network

Ben-Gal et al (2007) describes Bayesian network (BN) which is also known as belief network belonging to the family of probabilistic graphical models. These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, BN combine principles from graph theory, probability

theory, computer science, and statistics.

Machine Learning (2005) describes that Bayesian learning methods interpolate all the way to pure engineering. Bayesian methods interpolate to an extreme because the Bayesian prior can be a delta function on one model of the world. Bayesian and near-Bayesian methods have an associated language for specifying priors and posteriors. Bayesian learning involves specifying a prior and integration, two activities which seem to be universally useful. The limitations of Bayesian learning are it is information theoretically infeasible, computationally infeasible and non-automatic.

2.6.1 Bayesian Network for Classification

Kumari et al (2014) applies BN classifier to predict the persons whether diabetic or not using data of 206 persons with and without diabetes collected from a hospital. Weka tool is used for experiment and analysis. The techniques of pre-processing applied are attributes identification and selection, data normalization, and numerical discretization. The inputs are the parameters plasma concentration, blood pressure, insulin, BMI, diabetes pedigree function, age, triceps thickness (mm), no. of times pregnant, glucose tolerance test. The outputs are predicted as non – diabetic or pre-diabetic or diabetic.

2.6.2 Drawbacks of Bayesian Learning in Classification

Kumari et al (2014) comments that BN needs to be improved considering uncertain factors of some diabetes attributes. In the proposed model of this thesis, fuzzy reasoning is used which is able to overcome this drawback.

2.7 A Neuro-Fuzzy System

Nauck et al (1997) defines NEFCLASS model neuro-fuzzy system as a fuzzy system that is trained by a learning algorithm (usually) derived from neural network theory. The (heuristically) learning procedure operates on local information, and causes only local modifications in the underlying fuzzy system. The learning process is not knowledge based, but data driven. A neuro-fuzzy system can be viewed as a special three-layer feed forward neural network. The first layer represents input variables, the middle (hidden) layer represents fuzzy rules, and the third layer represents output variables. Some of the advantages of NEFCLASS model are vague knowledge can be used, the classier is

interpretable in form of linguistic rules and from an application view the classier is easy to implement, to use and to understand.

2.7.1 Neuro-Fuzzy for Classification

Selami et al (2004) applies neuro-fuzzy system for the effect of diabetes on blood flow of hemodynamics of the ophthalmic arteries. Blood flow hemodynamics was obtained from 80 ophthalmic arteries of 20 healthy patients and 20 diabetic patients. Inputs are peak systole, peak diastole, resistive index, pulsatile index and systole/diastole rate (SDR). These values were applied to neuro-fuzzy system using NEFCLASS model. The outputs are determined as three classes which are Class1: diabetic neuropathy, Class2: proliferative retinopathy. Class3: Control group.

2.7.2 Drawbacks of Neuro-Fuzzy in Classification

The classification accuracy of neuro-fuzzy system is only 87.5% **(**Selami et al., 2004). This motivated the research reported in this thesis to find an improved model on Pima diabetes dataset (Frank, 2010) that is able to give better accuracy in diabetes risk classification.

2.8 Summary

This chapter has described various computing techniques in diabetes risk classification. Table 2.1 shows classification accuracy of the reviewed algorithms above. Although some of the reviewed models showed best accuracy in classification listed in Table 2.1, it is based on binary classification. The research gaps found from the above reviewed models are their lack of ability in dealing with uncertain data and the level of risks is not classified. These drawbacks have motivated the current research. The proposed model in this thesis fills these gaps by using fuzzy reasoning that is able to deal with incomplete data and also showing the level of risks from the input. The proposed model described in this thesis enhances the previous fuzzy models by using patterns from the diabetes datasets in fuzzy rule base through a machine learning algorithm.

Algorithm	Accuracy	Testing dataset
Neuro-fuzzy	87.5%	Dataset of 20 healthy patients and 20 diabetic patients
ANN	92.5 %	Dataset of 530 patients from Sikkim Manipal Institute, India
Intelligible SVM	94 %	Dataset of 3014 subjects
Region based SVM algorithm	82.2 %	Pima Indian dataset
Hybrid model of MLP's and fuzzy logic	82.4%	Pima Indian dataset
Hybrid Modified-PSO and LS-SVM	97.833 %	Pima Indian dataset
Hybrid GA	96 %	Dataset of 560 patients with type 2 diabetes
Fuzzy Logic	98.88%	Pima women diabetes dataset
BN	99.51 %	Pima Indian dataset

Table 2.1 Classification accuracy of reviewed models

Chapter 3

Risk Classification: Proposed SVM-Fuzzy Model

3.1 Introduction

In the previous chapter various soft computing models developed and applied for risk classification were discussed. Most of the reviewed models drawbacks are their lack of ability in dealing with incomplete data. The algorithm is proposed in this chapter uses fuzzy reasoning that makes use of datasets through SVM which fills the gaps in previous models. For managing type 2 diabetes, the patient risks need to be known accurately to give suitable medication and advise. Sample datasets play important role in medical risk classification. This chapter provides background information about risk classification, the importance of datasets in classifying type 2 diabetes risks and describes the proposed algorithm for risk classification.

3.2 Foundation of Risk

As reported in (Fischhoff et al., 2011), the foundations of risk lie in decision theory, which articulates concepts whose emergence must have begun with the first human thought about uncertain choices. Applications of decision theory have led to unique collaborations among disciplines. Natural scientists have assessed probabilities for outcomes identified by ethicists reflecting on tradition politics and policy dilemmas. Social scientists have devised ways to explain these prospects and help individuals decide what they want, given what they might be able to have. Mathematicians and philosophers have formulated questions about uncertainties that computer scientists and psychologists have helped to answer. Sociologists and political scientists have shown how selecting experts and defining 'risks can highlight some issues and obscure others. These collaborations have also enriched the participating disciplines, by confronting them with issued outside their normal sphere. As a result, risk has changed sciences, as well as societies. As reported in (A Practical Guide to Risk Assessment, 2008), risk assessment is a systematic process for identifying and evaluating events (i.e., possible risks and opportunities) that could affect the achievement of objectives, positively or negatively. Such events can be identified in the external environment (e.g., economic trends, regulatory landscape, and competition) and within an organization's internal environment (e.g., people, process, and infrastructure). When these events intersect with an organization's objectives or can be predicted to do so, they become

risks.

3.3 Risk Classification

Risk is an uncertain event or condition that, if it occurs, has an effect on losing something of value (A Practical Guide to Risk Assessment, 2008). Computational classification is an example of the more general problem of pattern recognition, which is the assignment of some sort of output value to a given input value. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available (Alpaydin and Ethem, 2010). Risk classification is defined as grouping of different risks according to their estimated cost or likely impact, likelihood of occurrence, countermeasures required, etc. (Business Dictionary, 2000). As reported in (Damodaran, 2007), risk is incorporated into many different disciplines and is defined in different ways by each discipline.

Risk versus Probability: While some definition of risk focus only on the probability of an event occurring, more comprehensive definitions incorporate both the probability of the event occurring and the consequences of the event. Thus, the probability of a severe earthquake may be very small but the consequences are so catastrophic that it would be categorized as a high-risk event.

Risk versus Threat: In some disciplines, a contrast is drawn between risk and a threat. A threat is a low probability event with very large negative consequences, where analysts may be unable to assess the probability. A risk, on the other hand, is defines to be higher probability event, where there is enough information to make assessments of both the probability and the consequences.

All outcomes versus Negative outcomes: Some definitions of risk tend to focus only on the downside scenarios, whereas others are more expansive and consider all variability as risk. The engineering definition of risk is defined as the product of the probability of an event occurring, that is viewed as undesirable, and an assessment of the expected harm from the event occurring as below.

*Risk = Probability of an accident*Consequence in lost money/deaths.*

In contrast, risk in finance is defined in terms of variability of actual returns on an investment around an expected return, even when those returns represent positive outcomes. In case of diabetes, it is affected by many factors (Diagnosis of Diabetes, 2014) such as BMI, blood pressure, glucose conc. Diabetes risk is level of factors such as BMI, blood pressure, etc., in patients.

3.4 Importance of Dataset in Risk Classification

The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Several characteristics define a dataset's structure and properties. These include the number and types of the attributed or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis (Żytkow and Jan Rauch, 1999). The values may be numbers, such as real numbers or integers, for example representing a person's height in centimetres, but may also be nominal data (i.e., not consisting of numerical values). As an example for a diabetes dataset, we use Pima Indian diabetes dataset (Frank, 2010). The dataset contains 768 people data belonging to nine attributes: glucose concentration, blood pressure, insulin, BMI, diabetes pedigree fn., age, triceps thickness (mm), no. of times pregnant, class (0 or 1). For some kind of risk classification problem, we can't simply develop a computational model without a proof. Datasets are useful for these problems which assist the computational techniques for solution. When starting with the risk classification problem, the application should be observed as in what type of risk does it sits in. The Rumsfeld approach (Rumsfeld, 2011) characterized risks into:

1) Known/known: i.e. we know the risk exists and we know how to model the outcomes.
2) Known/unknown: i.e. we know the risk exists, but we don't know how to model it with any reliability.
3) Unknown/unknown: i.e. we have no idea what risks might exist and by definition, no idea how to model the risks.

The diabetes risks falls into second category according to Rumsfeld approach. We know that the diabetes risk exists, but we don't know how to model it mathematically. The sample dataset of proved cases helps this kind of situation, where we make use of this dataset through artificial intelligence approach to model the risk level in new patients. Dataset of sample cases is used in medical diagnosis for assisting in decision making for new cases through learning insights from the dataset. Data driven computational algorithms provide effective modeling for ill-structured problems like the diabetes risk classification. There are few questions in using very large datasets that whether to use whole dataset for training the machine learning algorithm or a small proportion of them is enough.

Data mining refers to extracting or mining knowledge from large amounts of data. It is an essential process where intelligent methods are applied in order to extract data patterns (Jiawei Han et al., 2011).

3.5 Proposed Algorithm: SVM-Fuzzy

The proposed algorithm, called SVM-Fuzzy, is a combined computational approach of SVM and fuzzy logic.

The following top level algorithm captures the SVM-Fuzzy design

SVM-Fuzzy algorithm

{Inputs: Glucose concentration (GC), serum insulin (SI), blood pressure (BP), BMI

Outputs: Points between 0 and 85, Diabetic (D): (0-50), Non-Diabetic (ND) (35-85)}

For each input and output {Input: GL, SI, BP, BMI; Output: D or ND}

Create fuzzy sets (); *{Create fuzzy sets(f_1, f_2,….., f_x) using triangular and trapezoidal membership function}*

For each input fuzzy set

Evaluate Median_F_i ; {Median_F_i: Median value of range in fuzzy set f_i}

IP_i = Median_F_i; *{IP_i = Input for SVM classifier assigned from fuzzy set f_i}*

Train SVM (); *{Train SVM using Pima Indian diabetes dataset}*

For each possible combination of binary inputs

SVM-Classify = (IP_i , IP_j); {SVM classification for binary inputs (GC, SI), (GC,BP),(GC, BMI), (SI, BP), (SI, BMI), (BP,BMI)}

Add SVM results to fuzzy rule base ();

Fuzzy rules (); *{Apply fuzzy rules to fuzzy sets (f_1, f_2,….., f_x)}*

Defuzzification (); *{Defuzzify output fuzzy sets to crisp output using center of gravity (COG) algorithm}*

End

The inputs to the system are 2-hour serum insulin (muU/ml), BMI (kg/m^2), plasma glucose concentration, diastolic blood pressure (mmHg). For each input fuzzy sets are created BMI (VLow, Low, Medium, High), plasma glucose level (VLow, Low, Medium, High), blood presure (VLow, Low, Medium, High), serum insulin (VLow, Low, Medium, High)). The output fuzzy sets are diabetic and non-diabetic. The fuzzy

sets are calculated using triangular and trapezoidal MFs from SVM. SVM is trained by Pima dataset. The SVM classification is used for each fuzzy rule as per the above algorithm. The fuzzy rules are used to give inferences. Defuzzification is done to the fuzzified output to crisp output value using Centre of Gravity (COG) method which is described in Section 4.5. The output value is in terms of points between 0 and 85 showing the level of risks from the input data.

3.6 Pima Indian Diabetes Dataset

The Pima Indian diabetes dataset summarized in Appendix B (from Frank, 2010) is used to train the SVM and to test the performance of the system against risk classification. The dataset contains 706 people data belonging to nine attributes: glucose concentration, blood pressure, insulin, BMI, diabetes pedigree fn., age, triceps thickness (mm), number of times pregnant, class (0 or 1). The proposed system is for type 2 diabetes patients, so the attributes influencing type 2 diabetes which are glucose concentration, blood pressure, insulin, BMI, output class are used (NIDDK, 2014). The Pima dataset (Frank, 2010) was grouped into smaller, bigger, random and testing datasets. The smaller dataset contains first 100 people records from Pima dataset, the bigger dataset contains first 400 people records from Pima dataset, the random dataset contains 100 people records that were randomly picked from the Pima dataset and the testing dataset contains 50 people data from the Pima dataset which are different from the other three groups. SVM was trained using smaller, bigger and random datasets individually and the testing dataset was used to test the performance of the whole system. The sample Pima Indian diabetes dataset is presented in the appendix.

3.7 Summary

This chapter has described the definition of risk, the importance of datasets in risk classification. The algorithm of combined approach of SVM and fuzzy reasoning is proposed. The inputs and outputs of the proposed model are described. The proposed SVM-Fuzzy algorithm in this chapter is able to overcome the drawbacks in the reviewed model descried in Chapter 2. It is also shown how the diabetes dataset is grouped to investigate the different sized datasets in risk classification. The next chapter describes the architecture and their components of the proposed system.

Chapter 4

System Design for SVM-Fuzzy and testing results

4.1 Introduction

Fuzzy expert system differs from machine learning techniques in risk classification by giving approximate reasoning rather than exact. Both attempt to derive insights from data to inform future action or response. The fuzzy sets and rules decide the performance of the fuzzy system. For designing the fuzzy rules of any classification problem, using the data of sample cases would make the fuzzy system to reason efficiently. The prosed model make use of SVM for learning a diabetes dataset and use the SVM classification results in the fuzzy rule base as a fuzzy expert system. The SVM-Fuzzy algorithm described in the previous chapter is capable of overcoming the drawbacks found in the reviewed models for risk classification. This chapter explains how the SVM-Fuzzy algorithm is implemented describing its architecture, the fuzzy sets, using of SVM in fuzzy rule base and defuzzification technique in the SVM-Fuzzy model.

4.2 Architecture of SVM-Fuzzy

The proposed system is the enhancement to the fuzzy expert system architecture described in the Chapter 2. A novel design shown in the Figure 4.1 is proposed for the SVM-Fuzzy model. The Mamdani type fuzzy expert system design is followed. SVM is used to design rules in the fuzzy system by using the SVM classification in the rule inference. SVM is trained by Pima Indian diabetes dataset accessible from (Frank, 2010). Figure 4.1 show the top level design of the various components of the proposed system. The inputs are fuzzified into fuzzy sets followed by reasoning with the fuzzy sets using rules which are designed using SVM classification results. Then the output fuzzy sets are defuzzified into crisp value.

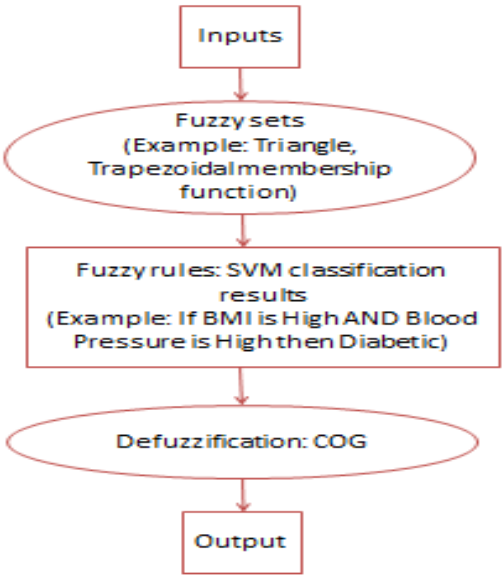

Figure 4.1: SVM-Fuzzy expert system architecture

4.3 Fuzzy sets

As shown in the Figure 4.1, the inputs given to the system need to be fuzzified. In the proposed model, the triangular and trapezoidal MFs are used in the fuzzy sets.

A triangular MF (Zadeh, L.A., 1965) is specified by three parameters {a, b, c} as follows:

Triangle (x; a, b, c) = 0, $x \leq a$; (x-a / b-a), $a \leq x \leq b$; (c-x / c-b), $b \leq x \leq c$; 0, $c \leq x$

A trapezoidal MF (Zadeh, L.A., 1965) is specified by four parameters {a, b, c, d} as follows:

Trapezoidal (x; a, b, c, d) = 0, $x \leq a$; (x-a / b-a), $a \leq x \leq b$; 1, $b \leq x \leq c$; (d-x / d-c), $c \leq x \leq d$; 0, $d \leq x$

Below figures (Figure: 4.2, 4.3, 4.4, 4.5, 4.6) show fuzzy sets of inputs and output classification.

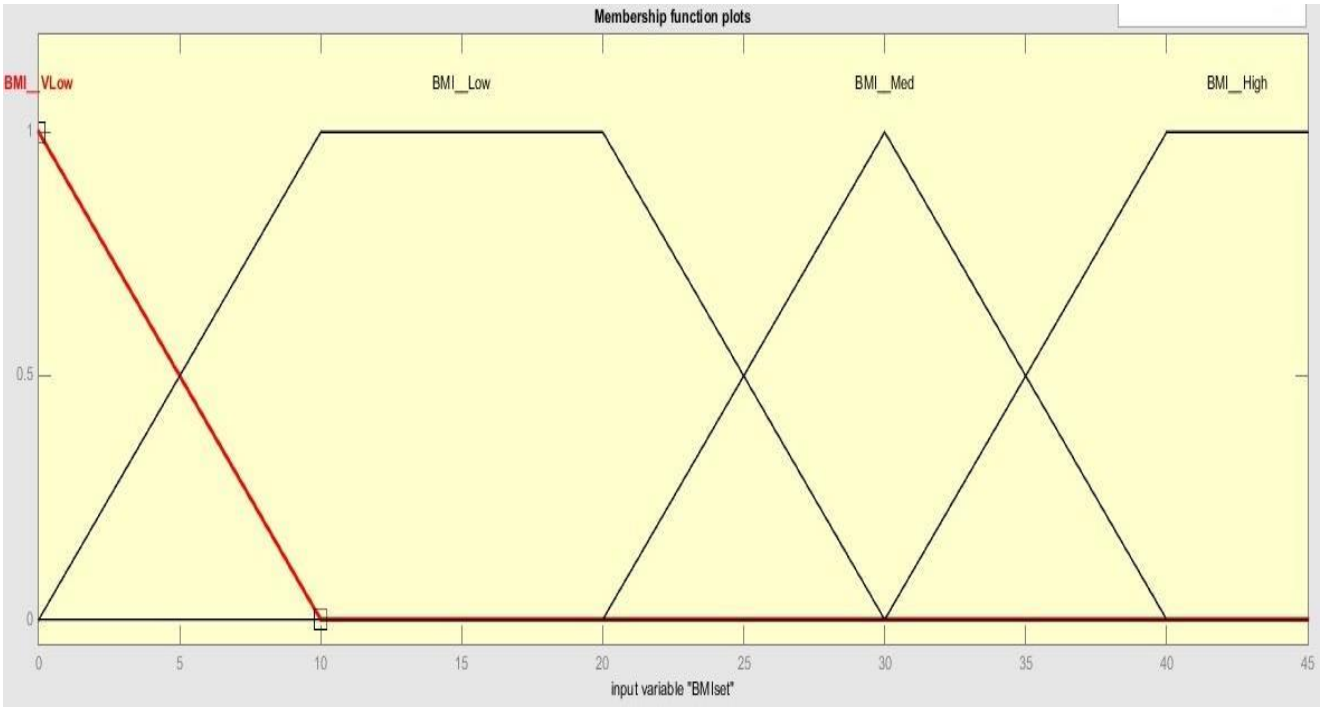

Figure 4.2: Fuzzy sets for the input BMI

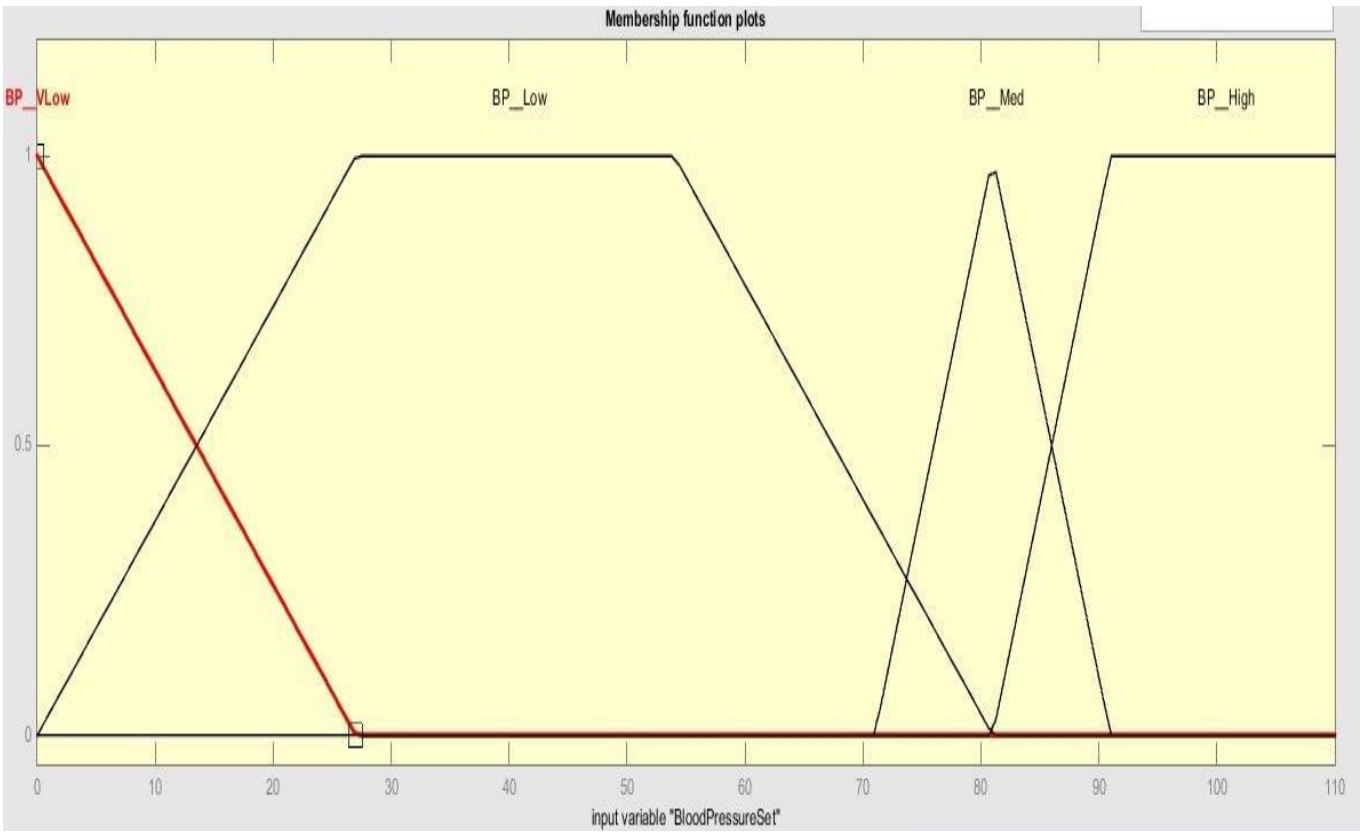

Figure 4.3: Fuzzy sets for the input blood pressure

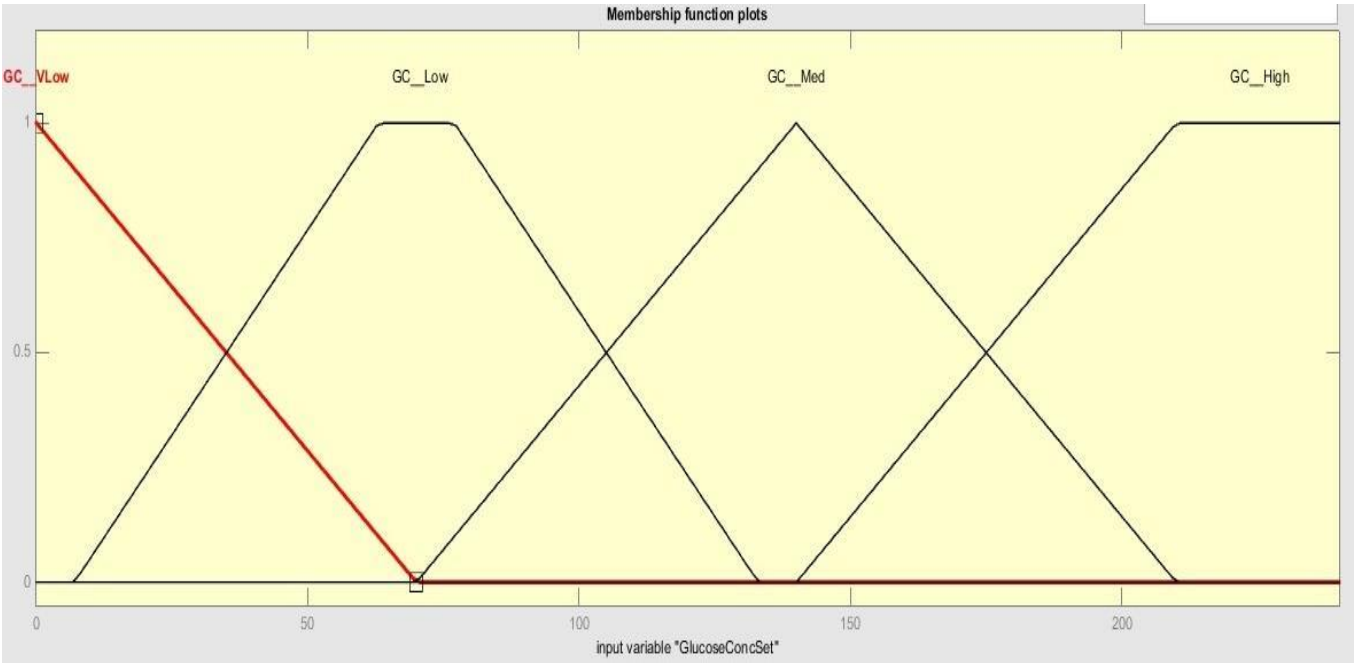

Figure 4.4: Fuzzy sets for the input glucose concentration

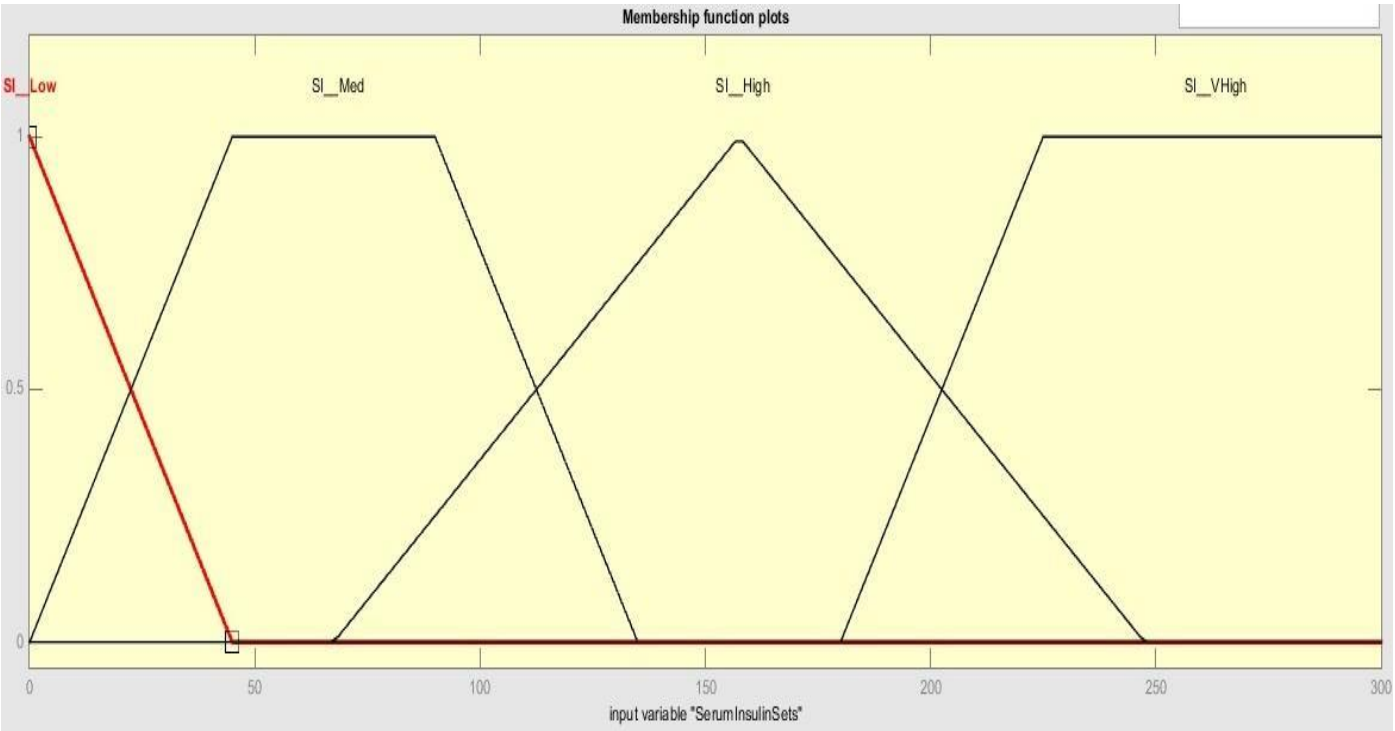

Figure 4.5: Fuzzy sets for the input serum insulin

Figures 4.2, 4.3, 4.4, 4.5 show the fuzzy sets using triangular and trapezoidal MFs for the inputs BMI, blood pressure, glucose level and insulin. Fuzzy sets created for the input BMI

are BMI-Very Low, BMI-Low, BMI-Medium, BMI-High. The sets BMI-Very Low and BMI-Medium are created using triangular MF. The sets BMI-Low and BMI-High are created using trapezoidal MF. Fuzzy sets created for the input blood pressure are BP-Very Low, BP-Low, BP -Medium, BP-High. The sets BP-Very Low and BP-Medium are created using triangular MF. The sets BP-Low and BP-High are created using trapezoidal MF. Fuzzy sets created for the input glucose concentration are GL-Very Low, GL-Low, GL-Medium, GL-High. The sets GL-Very Low, GL-Low and GL-Medium are created using triangular MF. The set GL-High is created using trapezoidal MF. Fuzzy sets created for the input serum insulin are SI-Very Low, SI-Low, SI-Medium, SI-High. The set SI-Very Low is created using triangular MF. The sets SI-Low, SI-Medium, SI-High are created using trapezoidal MF.

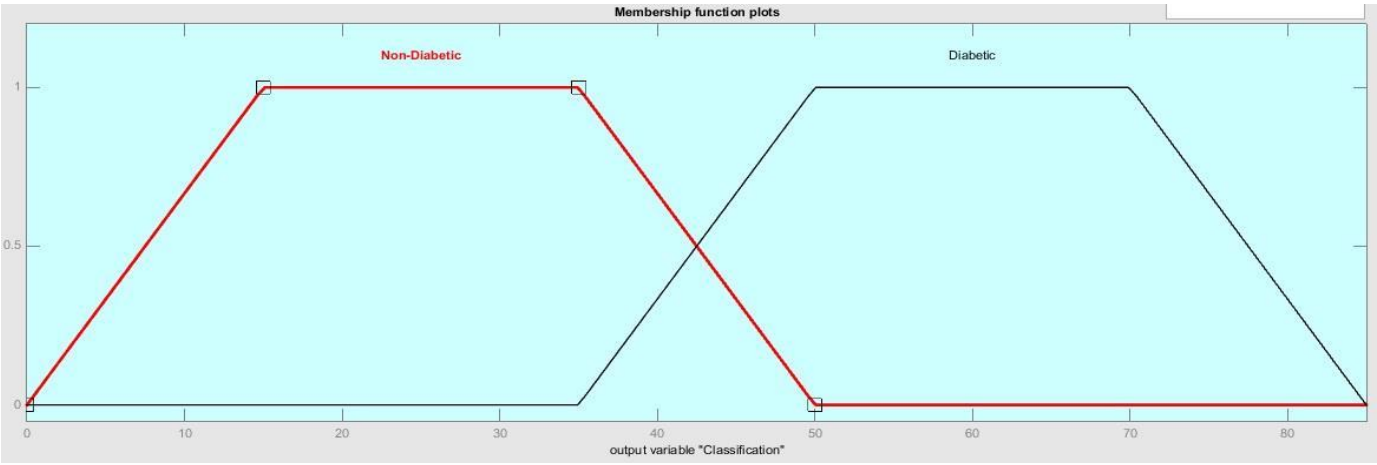

Figure 4.6 Classification fuzzy set

Figure 4.6 shows the fuzzy sets using trapezoidal MF for the output classification. Fuzzy sets created for the output are diabetic and non-diabetic. Below Table 4.1 shows fuzzy sets range of input and output field.

Input and Output field	Range	Fuzzy sets
I/P: 2-hour serum insulin	(-45) – 45 0 - 130 95 - 230 180 - 350	Very Low Low Medium High
I/P: BMI	(-10) – 10 0 - 30 20 - 40 30 – 60	Very Low Low Medium High

I/P: Plasma glucose conc.	(-70) – 70 0 - 130 70 - 210 130 - 330	Very Low Low Medium High
I/P: Diastolic blood pressure	(-25) – 25 0 - 60 70 - 90 80 - 120	Very Low Low Medium High
O/P: Classification	0 – 50 35 - 85	Non-Diabetic Diabetic

Table 4.1: Fuzzy sets and their ranges

Table 4.1 shows ranges for all the created fuzzy sets. The ranges for these sets are chosen based on medical research papers (Diagnosis of Diabetes, 2014; Diabetes and High Blood Pressure, 2014; MATH3220, 2010).

4.4 Fuzzy Rules in SVM-Fuzzy

SVM was trained using three different groups of dataset as described earlier. SVM was trained to group inputs (serum insulin, BMI), (serum insulin, glucose conc.), (serum insulin blood pressure), (BMI, glucose conc.), (BMI, blood pressure) and (glucose conc. blood pressure). Below figures (Figures: 4.7, 4.8, 4.9) shows the SVM training of three groups of Pima dataset using MATLAB. Figure 4.7 below shows the SVM training using smaller dataset to group the inputs under the output class. Figure 4.8 shows the SVM training using bigger dataset to group the inputs under the output class. Figure 4.9 shows the SVM training using random dataset to group the inputs under the output class.

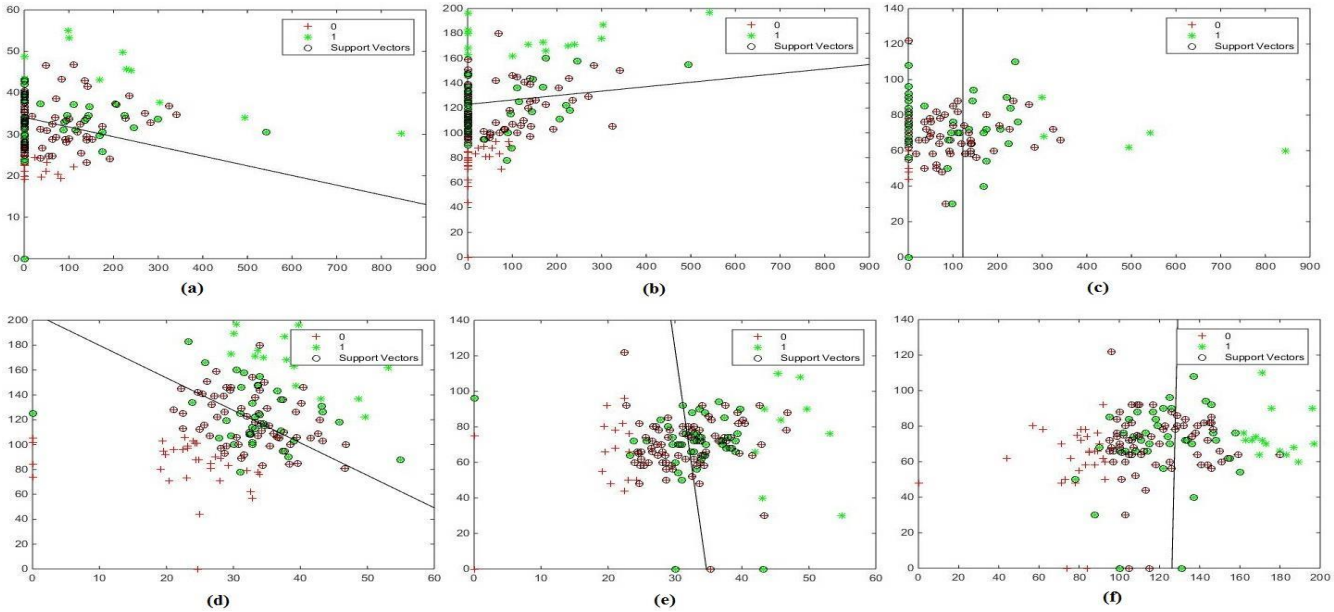

Figure 4.7: Training (a) serum insulin, BMI (b) serum insulin, glucose conc. (c) serum insulin, blood pressure (d) BMI, glucose conc. (e) BMI, blood pressure (f) glucose conc., blood pressure, from smaller dataset to group under the output class

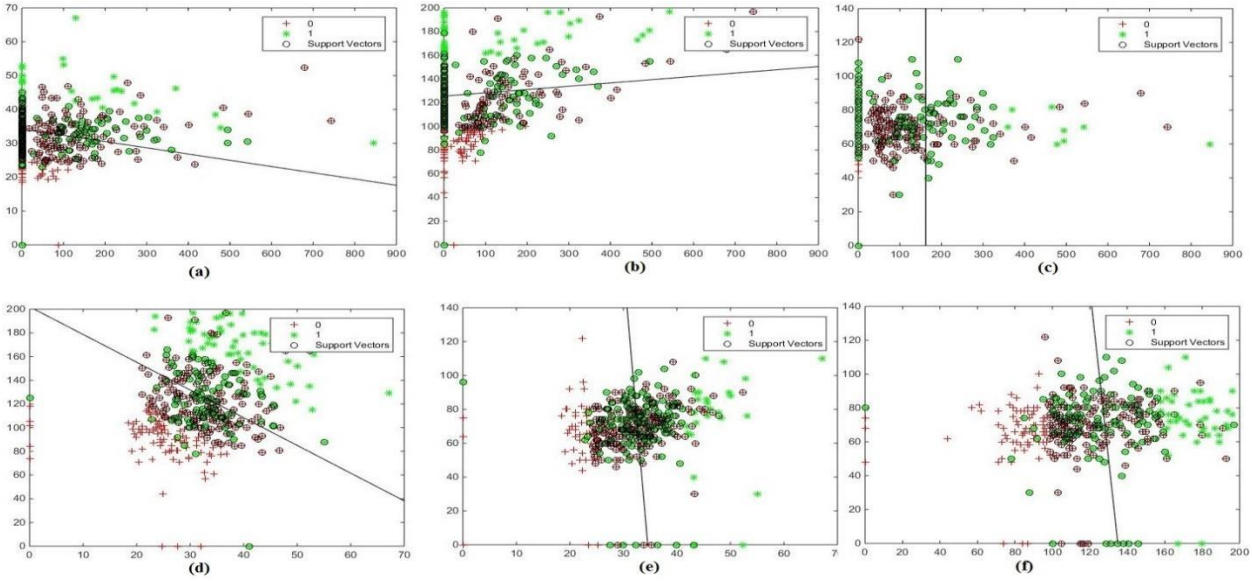

Figure 4.8: Training (a) serum insulin, BMI (b) serum insulin, glucose conc. (c) serum insulin, blood pressure (d) BMI, glucose conc. (e) BMI, blood pressure (f) glucose conc., blood pressure, from bigger dataset to group under the output class

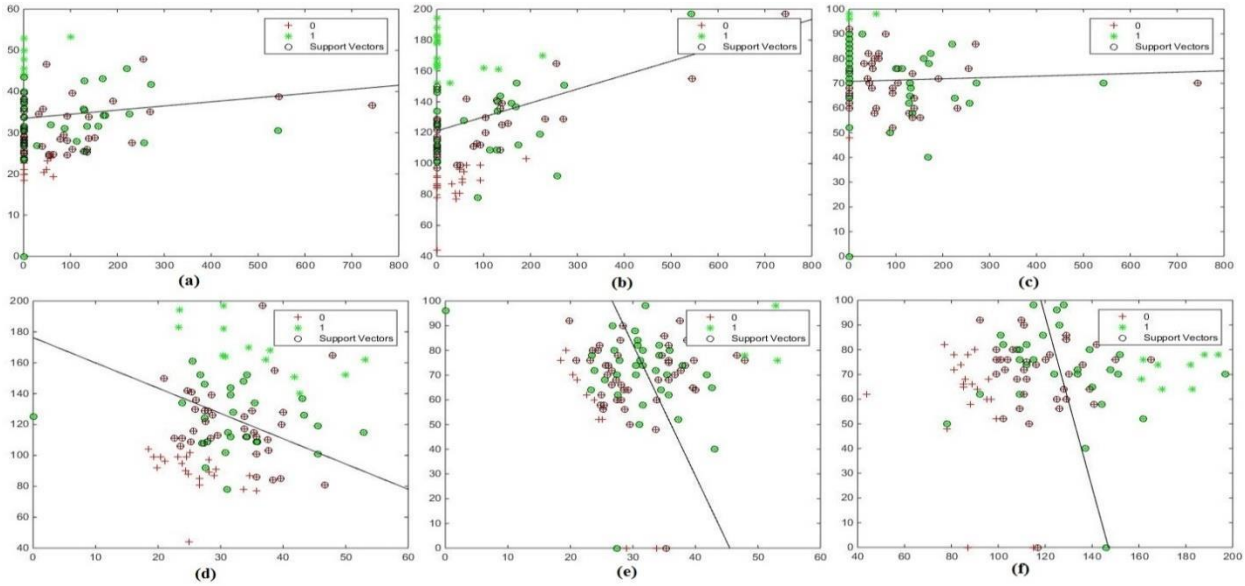

Figure 4.9: Training (a) serum insulin, BMI (b) serum insulin, glucose conc. (c) serum insulin, blood pressure (d) BMI, glucose conc. (e) BMI, blood pressure (f) glucose conc., blood pressure, from random dataset to group under the output class

The median value of range for all input fuzzy sets are calculated and given as inputs to the SVM classifier. Based on these training (Figures: 4.7, 4.8, 4.9) of different sized group of datasets, the SVM performs classification. First SVM classifications were done to the inputs based on training from smaller dataset. Then SVM classifications were done to the inputs based on training from bigger dataset. Then SVM classifications were done to the inputs based on training from random dataset. The SVM classification results were saved and designed as rules in the fuzzy rule base. As the SVM do classification for binary inputs, each fuzzy rule created contains binary inputs and their inference.

For the eighteen input fuzzy sets, totally 96 rules have been used from SVM classification for the fuzzy expert system. An example of a single rule is

SVM is used to classify the inputs glucose level and blood pressure using median value of their fuzzy sets range.

Median value of glucose level-High (mf) = 144 {mf: membership function}

Median value of blood pressure-High (mf) = 114

SVM-classification (144, 114) = diabetic

Rule: If glucose level-High AND blood pressure-High then diabetic.

Below figures (Figures: 4.10, 4.11, 4.12) show the rule viewer from the MATLAB for the fuzzy system designed from SVM classification trained using smaller dataset.

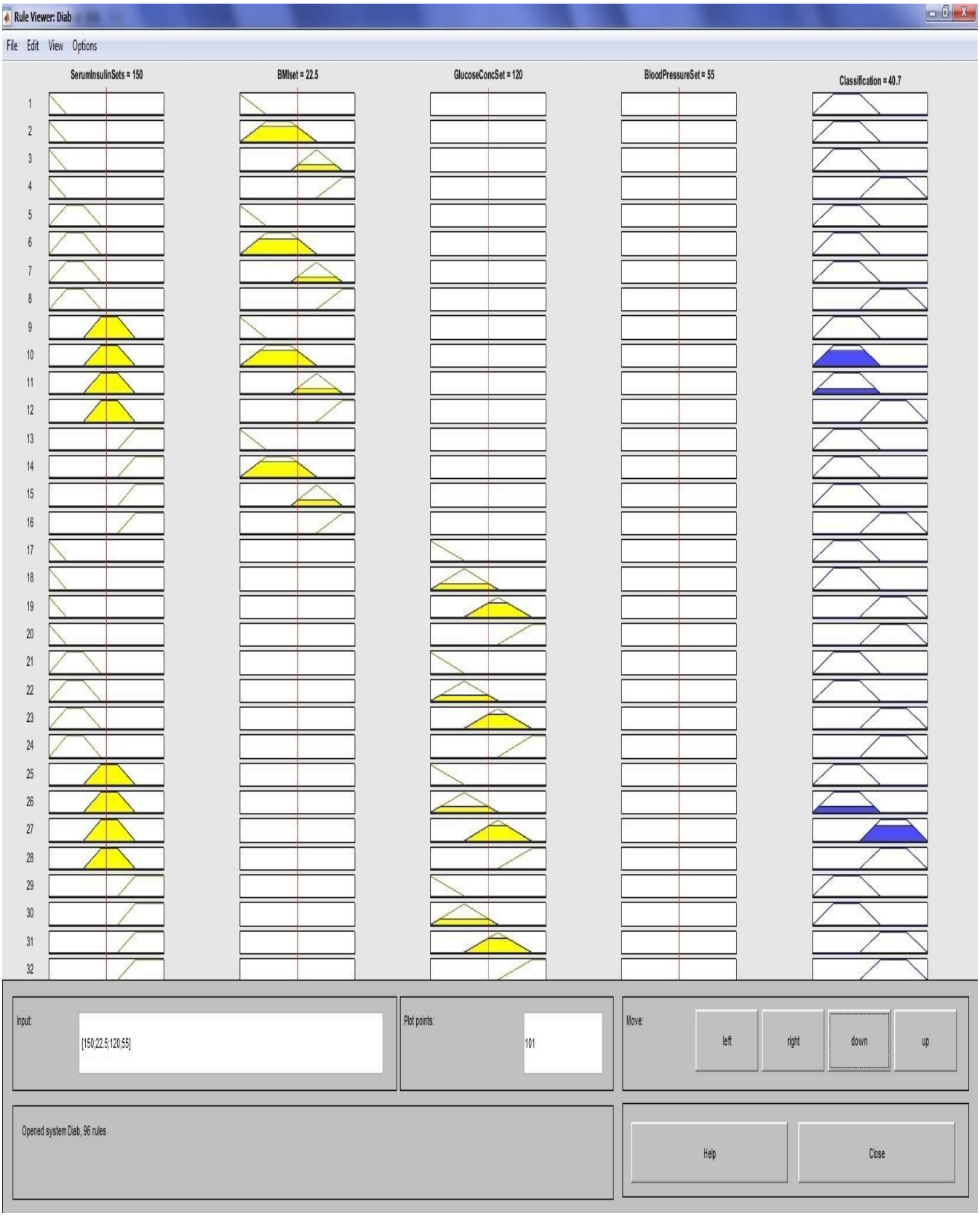

Figure 4.10: Rule viewer (rules 1-32)

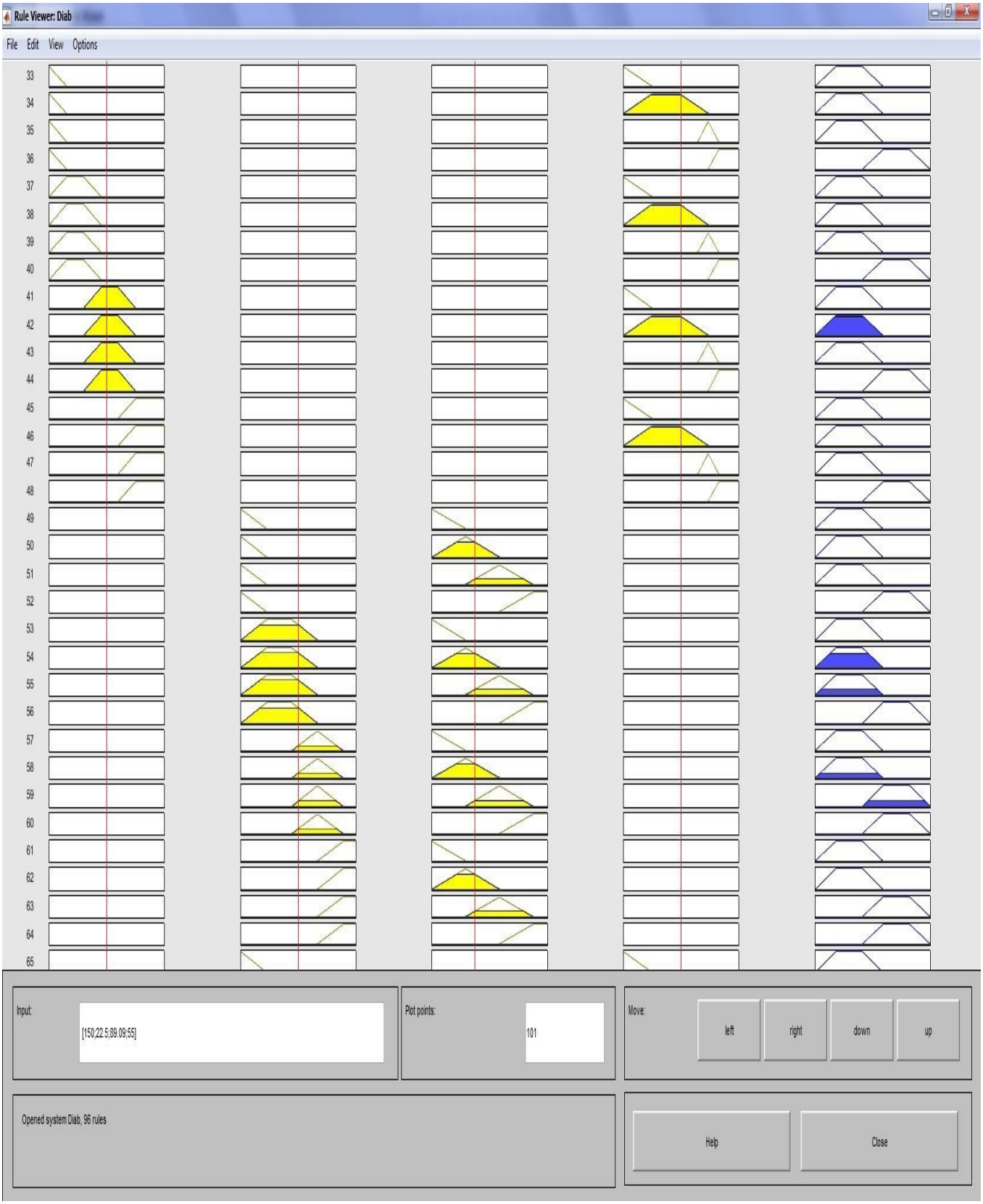

Figure 4.11: Rule viewer (rules 33-66)

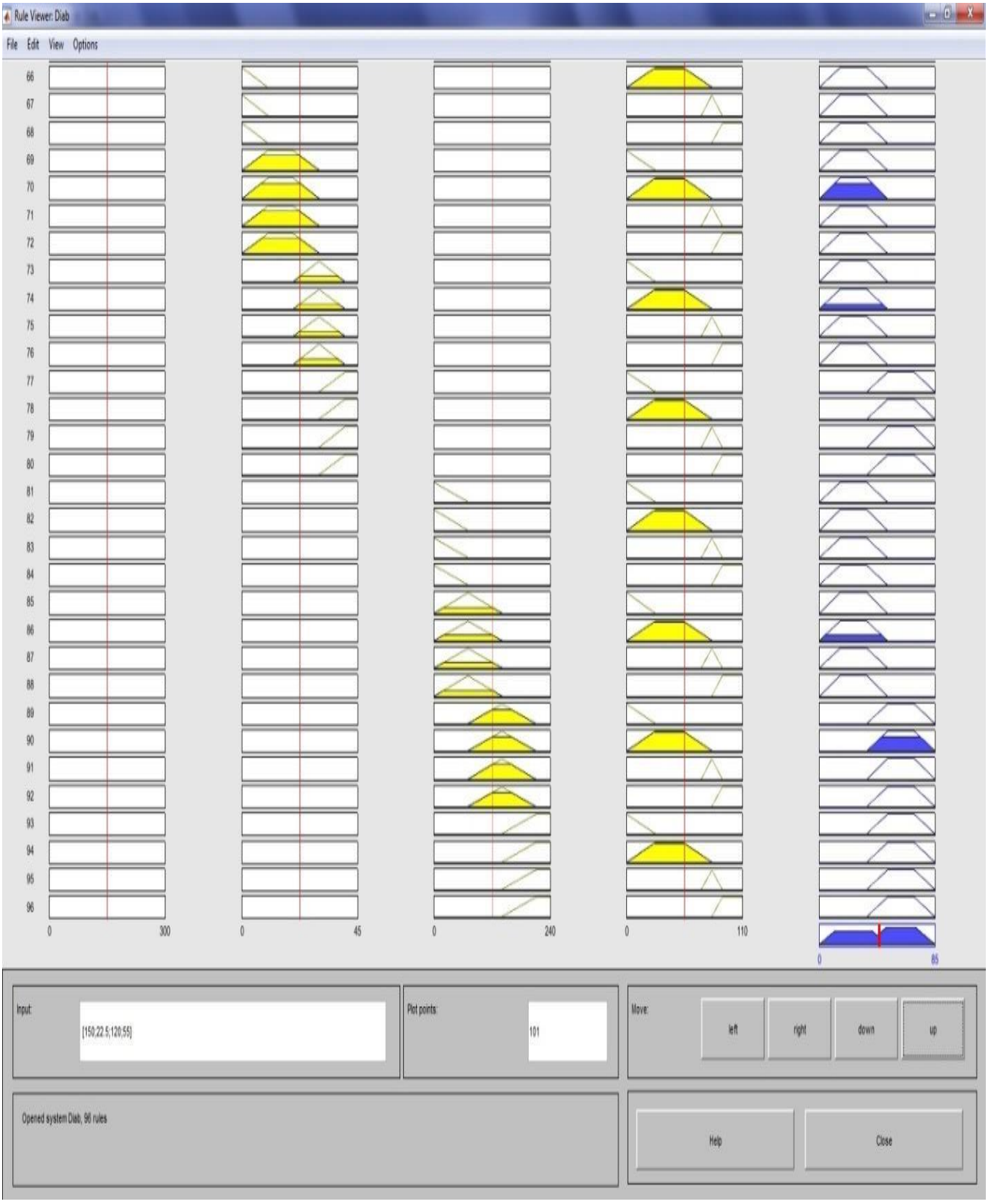

Figure 4.12: Rule viewer (rules 67-96)

Figures 4.10, 4.11, 4.12 show the total rules which are created using SVM classification results. The first, second, third, fourth column in the figure are inputs serum insulin, BMI, glucose conc., blood pressure respectively. The fifth column is the output classification.

4.5 Defuzzification in SVM-Fuzzy

The previous section shows how the fuzzy rules are designed. The fuzzified outputs that are reasoned from the fuzzy rules have to be defuzzified into crisp output values using COG method (Engelbrecht, Andries, 2007). Figure 4.13 illustrates the COG method. The centroid technique appears to provide consistent results. This is a well-balanced method sensitive to the height and width of the total fuzzy region as well as to sparse singletons. This method is also known as center of gravity or center of area defuzzification. This technique was developed by Sugeno in 1985. This is the most commonly used technique and is very accurate. The centroid defuzzification technique (Sugeno, 1985) can be expressed as

$$x^* = \frac{\int \mu_i(x)\, x\, dx}{\int \mu_i(x)\, dx}$$

where x^* is the defuzzified output, $\mu_i(x)$ is the aggregated MF and x is the output variable.

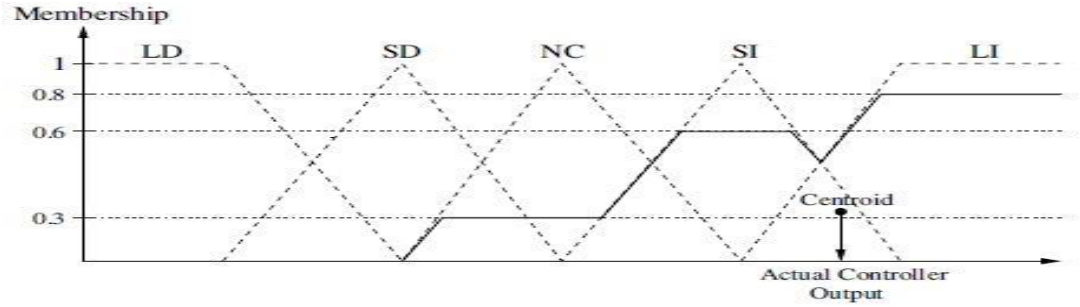

Figure 4.13: Center of Gravity method (Engelbrecht, Andries, 2007)

The fuzzified outputs are defuzzified into crisp output values using COG. Figure 4.13 illustrates the COG method. In this approach, each MF is clipped at the corresponding rule firing strengths. The centroid of the composite area is calculated and the horizontal coordinate is used as the output of the controller (Engelbrecht, Andries, 2007).

4.6 Summary

This chapter has proposed and described the architecture of SVM-Fuzzy. It explains how the fuzzy sets are calculated, how the SVM is used in designing the fuzzy rule base, and how the output fuzzy sets are defuzzified into crisp output values. The fuzzy sets are calculated using triangular and trapezoidal MF. The SVM is trained using different sized datasets individually and the results are put in the fuzzy rules for reasoning. The defuzzification using COG method is explained. The next chapter describes the performance of the SVM-Fuzzy model in type 2 diabetes risk classification.

Chapter 5

Performance Analysis of SVM-Fuzzy Model

5.1 Introduction

The previous chapter describes fuzzy sets, fuzzy rules, how the SVM results are used in rules and defuzzification in the SVM-Fuzzy. This chapter presents the outcomes from the SVM-Fuzzy system on Pima diabetes dataset (Frank, 2010) and describes the performance of SVM-Fuzzy.

5.2 Result Analysis for SVM-Fuzzy system

Fuzzy logic performs an approximate reasoning which gives the output in terms of level between the given ranges. Fuzzy logic handles data which are difficult to classify. In the SVM-Fuzzy, the patterns are extracted from the dataset using SVM for the fuzzy system. Three different groups of Pima diabetes dataset (Frank, 2010) were individually used to train SVM. The testing dataset which contains 50 people records different from the other three groups of datasets was used to test the proposed model. According to Section 4.4, the proposed system resulted from training of smaller, bigger and random datasets showed similar classification. The range for the output fuzzy set was chosen randomly and was divided into two half for diabetic and non-diabetic classification. Then the output fuzzy sets were put in the fuzzy rules based on SVM classification results. The range of output set is 0 to 85 where the non-diabetic fuzzy set is 0-50 and the diabetic fuzzy set is 35-85. To confirm whether the output classification is correct, the outputs were compared with the Pima dataset output class. For example, if the output of a person data is within diabetic fuzzy set range, then the Pima output class must be diabetic for the same data. The outputs are in term of points between 0 and 85. The output which falls between 0 and 50 is taken as diabetic and the output which falls between 35 and 85 is taken as non-diabetic to compare them with Pima output class for correctness.

Below Table 5.1 shows the SVM-Fuzzy model results for the inputs of 50 people records from the testing dataset (see Appendix B) which was used to test the propose model where the SVM is trained using bigger dataset and the classification results were used in the fuzzy rule base. SVM training by using smaller and random datasets showed similar results. Table 5.1 also lists the Pima output class for each input to compare the output class with the SVM-Fuzzy model results.

Serum Insulin	BMI	Plasma glucose concentration	Diastolic blood pressure	Pima Output Class	SVM-Fuzzy Output
95	21.2	82	64	0	40.7796
0	28.9	134	70	1	44.6246
210	39.9	91	68	0	59.9969
0	19.6	119	0	0	43.9993
105	37.8	100	54	0	43.6074
0	33.6	175	62	1	36.9963
0	26.7	135	54	0	46.3955
71	30.2	86	68	0	33.7209
237	37.6	148	84	1	47.8716
60	25.9	134	74	0	41.6833
56	20.8	120	72	0	44.0142
0	21.8	71	62	0	49.2462
49	35.3	74	70	0	36.929
0	27.6	88	78	0	42.7221
0	24.0	115	98	0	37.7515
105	21.8	124	56	0	43.5456
36	27.8	74	52	0	40.6676
100	36.8	97	64	0	40.0847
0	30.0	120	0	1	39.6341
140	46.1	154	78	0	37.6809
0	41.3	144	82	0	47.0871
0	33.2	137	70	0	32.9472
0	38.8	119	66	0	43.2784
0	29.9	136	90	0	27.8404
0	28.9	114	64	0	40.563
0	27.3	137	84	0	32.6874
191	33.7	105	80	1	54.0371
110	23.8	114	76	0	44.6848
75	25.9	126	74	0	42.5646
0	28.0	132	86	0	33.3593
328	35.5	158	70	1	43.2469
0	35.2	123	88	0	25.0036
49	27.8	85	58	0	26.7707
125	38.2	84	82	0	44.0607
0	44.2	145	0	1	45.9066
250	42.3	135	68	1	44.0136
480	40.7	139	62	0	41.414
265	46.5	173	78	0	42.8372
0	25.6	99	72	0	43.4232
0	26.1	194	80	0	40.8433

66	36.8	83	65	0	43.1278
0	33.5	89	90	0	40.8024
0	32.8	99	68	0	42.0927
122	28.9	125	70	1	32.3433
0	0.0	80	0	0	27.1482
0	26.6	166	74	0	38.7786
0	26.0	110	68	0	40.3506
76	30.1	81	72	0	41.5912
145	25.1	195	70	1	43.3073
193	29.3	154	74	0	39.9799

Table 5.1: SVM-Fuzzy Results

5.3 Classification accuracy

The classification accuracy is calculated using true positive (TP), false positive (FP) (Sapna Jain E et al., 2010) and true negative (TN), false negative (FN) (Jiawei et al., 2011) as below.

Classification accuracy = (TP+TN) / (TP+TN+FP+FN) where

TP: These are the positive tuples that were correctly labeled by the classifier. If the outcome from a prediction is p and the actual value is also p. then it is called TP.

TN: These are the negative tuples that were correctly labeled by the classifier.

FP: These are the negative tuples that were incorrectly labeled as positive. However if the actual value is in then it is said to be FP.

FN: These are the positive tuples that were mislabeled as negative.

The proposed system design showed classification accuracy of 96%. Out of 50 input data, 48 showed correct classification. As the fuzzy logic does approximate reasoning, some input data which falls between the points 35 and 50 could be taken as both non-diabetic and low risk diabetic. This kind of approximate reasoning is an example of handling uncertain data. As the fuzzy system output is in terms of points between the given ranges, it is taken as the level of risks for the individual case.

5.4 Summary

This chapter has described the performance of SVM-Fuzzy in type 2 diabetes risk classification. Instead of giving a binary classification, the results from SVM-Fuzzy had

showed the level of risks in terms of points between 0 and 85 as shown in the Table 5.1. The output points which fall between 35 and 50 could be taken either as non-diabetic or low risk diabetic. This is an example of reasoning by fuzzy logic which helps to handle uncertainty in the input data. It has been shown that similar results are obtained from smaller and bigger sized datasets in training the SVM. When the SVM is trained using smaller and bigger sized datasets with the using the SVM results in the fuzzy rule base, the fuzzy system showed similar output which is listed in Table 5.1. Both the smaller sized dataset and bigger sized dataset showed similar result when trained using SVM. The smaller sized data set contains first 100 people records of the bigger sized data set that contains 400 people records. In other words, the smaller sized data set contains 25% proportion of the bigger sized data set. So, in the case of Pima Indian dataset, it was discovered that a relatively small proportion of the full data set (about 25%) is sufficient to train the machine learning algorithm.

Chapter 6

Conclusion and Further Work

In recent work reviewed, most of the risk classification problems for type 2 diabetes are addressed using machine learning techniques. The reviewed papers had drawbacks of handling uncertain input data, to use sample datasets in fuzzy expert system, to classify the degree of risks from input data. The fuzzy logic by nature is able to handle uncertain data. The proposed system have used sample datasets in fuzzy expert system through SVM and classifying the degree of risks with high accuracy.

The drawbacks of previous research work are due to lack of capability in dealing with type 2 diabetes input data that is uncertain and incomplete. The fuzzy technique is able to handle uncertain input data by approximate reasoning. But the fuzzy logic should also need to make use of the dataset of sample cases in risk classification. This formed the motivation for this thesis.

A computational model SVM-Fuzzy which combines machine learning algorithm with fuzzy reasoning was designed, implemented and validated to address the above inadequacies and gaps. The proposed model gave better classification and promising outcomes. Although neuro-fuzzy system (see Section 2.9) has been applied for diabetes risk classification before, the SVM-Fuzzy model described in this thesis which use machine learning algorithm to design the fuzzy rule base gives better classification accuracy than the neuro-fuzzy system developed earlier.

The range of output set is 0 to 85 degree where the non-diabetic fuzzy set is 0-50 and the diabetic fuzzy set is 35-85. The range for the output fuzzy set was chosen randomly and was divided into two half for diabetic and non-diabetic classification. The output is in terms of points between 0 and 85, showing the level of risks in type 2 diabetes patients. To confirm whether the output classification is correct, the outputs were compared with the Pima dataset output class and the classification was confirmed.

The SVM-Fuzzy shows 96 % accuracy in type 2 diabetes risk classification. The fuzzy reasoning performs better when using sample cases extracting through the SVM machine learning technique and in case of a large dataset, 25% of the total dataset is sufficient to train the machine learning techniques as shown with this system results. For some data, the solution may be unpredictable and uncertain, here fuzzy reasoning helps by classifying

them according to level of risk in the output range. Focusing on five research questions listed in Section 1.4, the first three research questions have been successfully answered from the SVM-Fuzzy results as below.

The outcomes are summarized below:

1) *Can the level of risk be modelled as a classification problem?*
   The level of risks is modelled through the SVM-Fuzzy model which shows the type 2 diabetes risks classification output in terms of points between 0 and 85. As the fuzzy system reasons the input data based on the fuzzy sets and fuzzy rules, the output points from 0 to 85 were chosen randomly and divided into two half as diabetic and non-diabetic, to compare the system results with the Pima dataset output class. The reviewed models used for diabetic classification show output as either diabetic or non-diabetic. The SVM-Fuzzy is able to show improved output classification, showing the level of risks in diabetic or whether the patient is pre-diabetic or absolute non-diabetic. The point between 0 and 85 indicates the level of risk in people data where the point 0 indicates pure non-diabetic and the point 85 indicate highest risk in type 2 diabetes.

2) *Can an improved algorithm be developed to provide optimal classification?*

   The SVM-Fuzzy algorithm is developed and evaluated which show better classification accuracy of 96% in type 2 diabetes risk classification when compared to previous hybrid fuzzy model like neuro-fuzzy model described in Section 2.8. Although few reviewed models shows better classification accuracy than the SVM-Fuzzy model, they have drawbacks in their classification. The SVM-Fuzzy algorithm has overcome their drawbacks shown through the results which is able to show level of risks of type 2 diabetes from input data and has the ability to deal with uncertain data through fuzzy reasoning as shown with the SVM-Fuzzy results such as the output points which falls between 35 and 50 could be taken either as low risk diabetic or non-diabetic.

3) *Can mining a smaller subset from the data set would result in same outcomes as bigger data sets?*

   The results of the system were compared using the output from the mining of the first 100 people data and from the first 400 people data of the whole Pima dataset. The experimental results showed same outcomes. This gives a conclusion that 25% of the dataset would be sufficient for data mining in the diabetes dataset studied.

The below remaining two research questions will be investigated in the future as

extension to the current research.

4) *Can the risk classification results be used for developing an intelligent planning system for providing guidelines?*

5) *Can the risk classification algorithm be improved using multi-agent learning?*

Contribution of this research:

- A successful hybrid computational approach to risk classification combining SVM and fuzzy reasoning design. The approach was designed, implemented and successfully evaluated on a benchmark dataset.
- The proposed design showed improved accuracy in risk classification compared to other hybrid fuzzy models.
- The level of risks is classified as points between 0 and 85 which gives the levels in between diabetic and non-diabetic.
- The experiments from the model showed that a relatively small subset of dataset was sufficient to train the machine learning algorithm. The full dataset is very large and would be inefficient. The 25% of the full dataset produced the same outcomes equal to the results when mining from the whole dataset but more efficiently.

Future work will investigate incremental learning for SVM-Fuzzy. An intelligent planning system would be investigated as an extension to SVM-Fuzzy for providing lifestyle guidelines to type 2 diabetes patients according to their risk level. SVM-Fuzzy will be used to derive the risks level and used in a planning system (that will have exercise, lifestyle and diet planning components). The SVM-Fuzzy algorithm will be further enhanced using a multi-agent learning system. It will be further applied and evaluated against other classification problems in complex applications like extracting patterns in structure of proteins in bioinformatics.

Appendix A

Publications

Thirumalaiappan Ramanathan, T., Sharma, D., 2015. An SVM-Fuzzy Expert System Design for Diabetes Risk Classification. International Journal of Computer Science and Information Technologies (IJCSIT), vol. 6 (3), pp. 2221-2226

Thirumalaiappan Ramanathan, T., Sharma, D., 2015. Reasoning Using A Case Based Fuzzy Technique. International Journal of Computational Intelligence Research ISSN 0973-1873, vol. 11 (1), pp. 69-77.

Appendix B

Pima Indian Diabetes Dataset

There are totally 768 people data available in the Pima Indian diabetes dataset from (Frank, 2010). Below shows the sample people data of the dataset.

Attribute information for each record given in the data as a row:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. BMI (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

As described earlier in the Section 3.6, the attributes: number of times pregnant, triceps skin fold thickness, diabetes pedigree function, age are not used as the inputs in this thesis.

Sample Pima Indian Diabetes Dataset

No. of times pregnant	Plasma glucose concentration	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	BMI (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Output Class variable (0 or 1)
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1

1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1

References

Ahlawat, N., Gautam, A., Sharma, N., 2014. International Research Publications House. Use of Logic Gates to Make Edge Avoider Robot. International Journal of Information & Computation Technology, p. 630 vol. 4(6).

Alpaydin, E., 2010. Introduction to Machine Learning. MIT Press. p. 9.

ANSI/PMI 99-001-2008. A Guide to the Project Management Body of Knowledge (4th Edition). PMI. Newton Square. PA.

A Practical Guide to Risk Assessment, 2008. Available from <http://www.pwc.com/en_us/us/issues/enterprise-risk-management/assets/risk_assessment_guide.pdf>. [6 September 2014].

Auria, L., Moro, R. A., 2008. Support vector machines (SVM) as a technique for solvency analysis. DIW Berlin.

Barakat, M.N.H., Bradley, A.P., 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE Transactions on 14.4, pp. 1114-1120.

Bebis, G., 2010. Support Vector Machines (SVM). Available from <http://www.cse.unr.edu/~bebis/MathMethods/SVM/lecture.pdf>. [12 August 2014].

Ben-Gal, I., Ruggeri, F., Faltin, F., Kenett, R., 2007. Bayesian Networks. Encyclopaedia of Statistics in Quality and Reliability.

Blondin, J., 2009. Particle swarm optimization: A tutorial. Available from <http://www.cs.armstrong.edu/saad/csci8100/pso_tutorial.pdf>. [15 August 2014].

Business Dictionary, 2000. Available from <http://www.businessdictionary.com/definition/risk-classification.html>. [6 September 2014].

Dalakleidi, K.V., et al, 2013. A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients. Bioinformatics and Bioengineering (BIBE), IEEE 13th International Conference on. IEEE.

Damodaran, A., 2007. What is Risk? : Lecture notes. Available from <http://people.stern.nyu.edu/adamodar/pdfiles/valrisk/ch1.pdf >. [6 September 2014].

Delashmit, W. H., Manry, M. T., 2005. Recent developments in multilayer perceptron neural networks. In Proceedings of the seventh Annual Memphis Area Engineering and Science Conference, MAESC.

Dey, R., Bajpai, V., Gandhi, G., Dey, B., 2008. Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus. In Industrial and Information Systems. ICIIS. IEEE Region 10 and the Third international Conference on (pp. 1-4). IEEE.

Diabetes and High Blood Pressure, 2014. Available from <http://www.patient.co.uk/health/diabetes-and-high-blood-pressure>. [24 December 2014].

Diagnosis of Diabetes and Prediabetes, 2014. Available from <http://diabetes.niddk.nih.gov/dm/pubs/diagnosis>. [24 December 2014].

Dowla, F. U., Rogers, L. L., 1995. Solving problems in environmental engineering and geosciences with artificial neural networks. MIT Press.

Dulay, N., 2001. Genetic Algorithms. Available from <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html>. [15 August 2014].

Elio, R., Hoover, J., Nikolaidis, I., Salavatipour, M., Stewart, L., Wong, K., 2011. About Computing Science Research Methodology.

Engelbrecht, A. P., 2007. Computational intelligence: an introduction. John Wiley & Sons.

Fischhoff, B., Kadvany, J., 2011. Risk: A very short introduction. Oxford University Press.

Frank, A., Asuncion, A., 2010. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fuzzy Logic. Stanford Encyclopedia of Philosophy. Stanford University. 2006-07-23. Retrieved 2008-09-30.

Health Direct Australia, 2013. Available from <http://www.healthdirect.gov.au/diabetes-types>. [8 August 2014].

Gou, J., Xiong, T., Kuang, Y., 2011. A novel weighted voting for k-nearest neighbor rule. Journal of Computers, 6(5), 833-840.

Government Must Act Now to Stop Diabetes, 2012. Available from <http://www.diabetesaustralia.com.au/Documents/DA/Media Releases/Diabetes

National Election Agenda 2013-2015 media release.pdf>. [8 August 2014].

Han, J., Kamber, M., Pei, J., 2011. Data Mining Concepts and Techniques Third edition.

Jain, S., Aalam, M.A., Doja, M.N., 2010. K-Means Clustering Using WEKA Interface, Proceeding of the 4th National Conference; India Com-2010 Computing For Nation Development, February 25-26.

Joint Publication 2-0, Joint Intelligence. Defense Technical Information Center (DTIC). Department of Defense. 22 June 2007. pp. GL–11. Retrieved February 22, 2013.

Karatsiolis, S., Schizas, C.N., 2012. Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset, Proceeding of the 2012 IEEE 12t International conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November.

Khashei, M., Effekhari, S., Parvizian, J., 2012. Diagnosing diabetes type ii using a soft intelligent binary classification model. Review of Bioinformatics and Biometrics (RBB) 1: 9-23.

Kumari, M., et al, 2014 Prediction of Diabetes Using Bayesian Network (IJCSIT), pp. 5174-5178 vol. 5 (4).

Machine Learning (Theory), 2005.
Available from <http://hunch.net/?p=65>.
[13 August 2014].

Mamdani, E.H., Assilian, S., 1975. An experiment in linguistic synthesis with a fuzzy controller, International Journal of Man-Machine Studies, 7(1), 1-13.

Marinakis, Y., Dounias, G., Jantzen, J., 2009. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbour classification. Computers in Biology and Medicine, 39(1), 69-78.

MATH 3220 Research Report By Jennifer Lamb, 2010.
Available from <http://mercury.webster.edu/aleshunas/Support Materials/C4.5/Lamb -FINALDIABETESPPT.pdf>. [24 December 2014].

Mendel, J.M., 2001. Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall PTR.

Morik, K., Brockhausen, P., Joachims, T., 1998. Combining statistical learning with knowledge-based approach-A case study in intensive care monitoring, in Proc. Eur. Conf.

Mach. Learn., pp. 268–277.

Narashiman, B., Malathi, A., 2014. Fuzzy Logic System For Risk –Level Classification of Diabetes Nephropathy, Green Computing Communication and Electrical Engineering (ICGCCEE), International Conference on 6-8 March.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 2014. Causes of Diabetes. Available from <http://www.niddk.nih.gov/health-information/health-topics/Diabetes/causes-diabetes/Pages/index.aspx#type2>. [2 October 2014].

Nauck, D., Kruse, R., 1997. What are neuro-fuzzy classifiers? In Proc.Seventh International Fuzzy Systems Association World Congress IFSA. Vol. 97(4) pp. 228-233.

Navlakha, S., Bar-Joseph, Z., 2013. Algorithm in Nature. Genetic Algorithms. Available from <http://www.cs.cmu.edu/~02317/slides/lec_9.pdf>. [15 August 2014].
Negnevitsky, M., 2011. Artificial Intelligence: A Guide to Intelligent Systems Third Edition. Pearson.

Novák, V., Perfilieva, I., Močkoř, J., 1999. Mathematical principles of fuzzy logic Dodrecht: Kluwer Academic.

Peterson, L. E., 2009. K-nearest neighbor. Scholarpedia, 4(2), 1883.

Rumsfeld, D., 2011. Known and unknown: a memoir. Penguin.

Safaric, R., 2011. Intelligent Control Techniques in Mechatronics – Genetic algorithm. Available from <http://www.ro.feri.uni-mb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic algorithm/_18.html>. [15 August 2014].

Sammany, M., Zaghloul, K. S. S., 2006. Support Vector Machine Vs an Optimized Neural Network for Diagnosing Plant Diseases. ICENCO.

Selami, S., Firat, H., Adem, K., Huseyin, O., Turgut, Y., Inan, G., 2004. A Neurofuzzy Classification System for the Effects of Diabetes Mellitus on Ophthalmic Artery, Journal of Medical Systems, April, vol. 28(2).

Soliman, O.S., Aboelhamd, E., 2014. Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine." arXiv preprint arXiv:1405.0549.

Stergiou, C., Siganos, D., 2003. Neural Networks. Available from

<http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>. [12 August 2014].

Suykens, J. A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

Tu, J. V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology, 49(11), 1225-1231.

Uusitupa, M., 2002. Lifestylematter in prevention of type 2 diabetes, Diabetes Care, pp. 1650– 1651 vol. 25(9).

Wang, W., Cao, J., Lu, H., Wang, J., 2013. A Default Discrimination Method for Manufacturing Companies by Improved PSO-based LS-SVM. International Journal of Hybrid Information Technology, 6(2), 95-106.

Watts, M. J., 2013. Multi-Layer Perceptrons. Available from <http://mike.watts.net.nz/Teaching/IIS/Lecture10.pdf>. [13 August 2014].

Yang, X.S., Cui, Z.H., Xiao, R., Gandomi, A., Karamanoglu, M., 2013.Swarm Intelligence and Bio-Inspired Computation: Theory and Applications, Elsevier.

Zadeh, L.A.,1965. Fuzzy sets. Information and Control 8 (3): 338 353. doi:10.1016/s0019-9958(65)90241-x.

Zadeh, L.A., 1994. Fuzzy Logic, Neural Networks, and Soft Computing, Communication of the ACM, March, pp. 77-84 vol. 37(3).

Zytkow, M., Jan Rauch, 1999. Principles of data mining and knowledge discovery. Springer.