COMPARATIVE ANALYSIS OF CLASSIFICATION MODELS FOR DIAGNOSIS TYPE 2 DIABETES

by

Daniah Almadni

Bachelor of Science in Information Technology

King Abdulaziz University

Saudi Arabia, 2011

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Science

in the program of

Computer Science

Toronto, Ontario, Canada, 2016

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Daniah Almadni

COMPARATIVE ANALYSIS OF DATA MINING ALGORITHIMS FOR DIAGNOSIS TYPE 2 DIABETES

Daniah Almadni

Master of Science, Computer Science, 2016

Ryerson University

ABSTRACT

Diabetes mellitus type 2 has become one of the major causes of premature diseases and death in many countries. It accounts for the majority of diabetes cases around the world. Thus, we need to develop a system that diagnoses type 2 diabetes. In this thesis, a fuzzy expert system is proposed using the Mamdani fuzzy inference system to diagnose type 2 diabetes effectively. In order to evaluate the performance of our system, a comparative study has been initiated, and will contrast the proposed system with data mining algorithms, namely J48 Decision tree, multilayer perceptron, support vector machine, and Naïve Bayes. The developed fuzzy expert system and the data mining algorithms are validated with real data from the UCI machine learning datasets. Moreover, the performance of the fuzzy expert system is evaluated by comparing it to related work that used the Mamdani inference system to diagnose the incidence of type 2 diabetes.

ACKNOWLEDGEMENTS

I would like to express my special sincere thanks to my supervisor, Dr. Abdolreza Abhari for his unlimited support, guidance, mentorship, and encouragement. I would like also to express my special appreciation to the committee members, Dr. Alex Ferworn, Dr. Isaac Woungang, and Dr. Alireza Sadeghian for the time and effort they spent to review my thesis and for their valuable comments and feedback.

My gratitude is extended to Dr. Osama Muthaffar fellow at Sickkids Hospital and Dr. Ahmad Alnahari fellow in the Department of Endocrinology and Diabetes at McMaster University for their help in developing the proposed system.

I am very grateful to my husband and my parents, for their support, encouragement, and patience. Without their assistance, completion of this thesis would not have been possible.

Finally, I thank all the members of our research group in the distributed system and multimedia processing (DSMP http://dsmp.ryerson.ca/) lab for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
BP	Blood Pressure
BMI	Body Mass Index
COG	Centre Of Gravity
DM	Diabetes Mellitus type 2
DT	Decision Tree
DPF	Diabetes Pedigree Function
FIS	Fuzzy Inference System
FL	Fuzzy Logic
LOM	Largest value Of the Maximum
MOM	Mean value Of the Maximum
MF	Membership Function
MLP	Multilayer Perceptron
NP	Number of Pregnancy
PIDD	Pima Indian Diabetes Dataset
QP	Quadratic Programming
SVM	Support Vector Machine
TSFT	Triceps Skin Fold Thickness

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes is a chronic disease that occurs when the pancreas cannot produce enough insulin or when the body does not use the insulin effectively. It is a major cause of heart attacks, kidney failure, blindness, lower limb amputation and strokes. In 2014, 422 million people were diagnosed with diabetes compared to 108 million people in 1980. Moreover, in 2014, diabetes global prevalence was estimated to be 9% among adults over the age of 18. It has been reported that an estimated 1.5 million deaths were directly caused by diabetes and that high blood glucose was the direct cause of 2.2 million deaths in 2012 [1]. The World Health Organization estimates that diabetes will be the 7th leading cause of death in 2030 [2]. In addition to this, more than 80% of diabetes-related deaths occur in low and middle-income countries [1].

1.2 Type 2 Diabetes

There are three types of diabetes, namely type 1 diabetes, type 2 diabetes, and gestational diabetes. Type 2 diabetes (also called diabetes mellitus type 2) is the most common form of diabetes since it accounts for 90% of diabetes cases. It is a long term metabolic disorder that is characterised by high blood glucose and insulin resistance. In addition, it results from the body's ineffective use of insulin [3]. There are two main causes of type 2 diabetes, namely an increase in body weight and a lack of physical activity [3, 4]. Rates of this type of diabetes have increased considerably since 1960 in conjunction with increasing rates of obesity [5]. The number of type 2 diabetic patients increased from approximately 30 million in 1985 to around 368 million in 2013 [6, 7]. Until recently, type 2 diabetes was seen only in adults, but is now becoming increasingly common in young people [3].

1.3 Problem Statement

Since diabetes mellitus type 2 has become one of the major causes of premature diseases such as heart disease and kidney disease leading to death in many countries [1], it is important that an expert system be implemented and used in the diagnosis of this condition. Moreover, this system should be accessible and usable for non-specialists, i.e. nurses and the members of the general public.

Physicians diagnose diabetes mellitus type 2 by examining the symptoms exhibited by patients, and then deciding whether a person is diabetic or non-diabetic. In addition, physicians can form an opinion about the severity and stage of the patient's illness. However, in cases where experienced physicians are hesitant, computer-aided disease diagnosis systems can be employed to help the physicians diagnose diabetes mellitus. Indeed, these systems have high success rates.

Despite the fact that an expert's decision is the most important factor in diagnosis, expert systems provide substantial help as they reduce errors resulting from fatigue as well as the time needed for diagnosis. In order to produce a safe and high quality medical systems, it is important that expert systems be used in health care. In general, health information applications help us to reduce human error and to support patient care systems.

Although several systems have been proposed to diagnose diabetes mellitus type 2, the accuracy of different data mining and machine learning techniques is not very high. Researchers have tried to increase the prediction accuracy of the developed systems, but these attempts have failed in most cases. The developed systems encountered certain issues; for instance, some were only focussed on a much younger age group [8], some used only the physiological factors to diagnose this disease [9,10], some tested their systems based on a small number of instances [9,11], and some did not show the prediction accuracy of the proposed systems [10]. Therefore, it is

important to develop a diabetes diagnosis system that is capable of improving accuracy by taking all factors into account.

1.4 Objectives and Proposed Methodology

There are two main objectives of this study. The first of these pertains to developing a fuzzy expert system to efficiently diagnose the incidence of diabetes mellitus. To this end, a fuzzy expert system is built using the Mamdani fuzzy inference system in Matlab. Implementing this system involves four main steps which are fuzzification, rules evaluation, outputs aggregation, and defuzzification. Details regarding these steps are provided in Chapter 3, Section 3.2. The second objective is to investigate and evaluate the performance of the fuzzy expert system. For this purpose, two comparative studies are done. First, the proposed system is compared with regression method and various common classification methods, namely J48, multilayer perceptron (MLP), support vector machine (SVM), and Naïve Bayes. Second, our system is compared with related work that used Mamdani fuzzy inference system to diagnose the diabetes mellitus type 2. The purpose of these studies is to evaluate the performance of the proposed fuzzy expert system. The Pima Indian Diabetes Dataset [53], which includes 768 records and 9 attributes, is used in both studies. Before using the dataset, the physicians are consulted and the regression analysis is applied to the dataset. In light of this consultation and regression analysis, the decision is made to use all of the attributes of the dataset. The dataset is pre-processed using two different methods which are the multiple imputation method and listwise deletion method to handle the values missing in the dataset. Lastly, different evaluation metrics are calculated and the results are compared.

1.5 Contributions

The main contributions of this study are listed as follows:

- We designed and implemented a fuzzy expert system by using a novel fuzzy rules and membership functions that is able to diagnose the incidence of diabetes mellitus type 2. Then, we evaluated our developed fuzzy expert system by comparing its performance with the performance of other well-known data mining algorithms, namely J48, MLP, logistic regression, SVM, and Naïve Bayes.
- We compared our fuzzy expert system with the fuzzy expert system presented in [8] to evaluate the performance of our system. One of the differences between the fuzzy system developed in this thesis and the one in a similar work [8] is that we used all the attributes of the original dataset after consulting the physicians. We employed significant factors which are commonly used to diagnose diabetes mellitus type 2, including physiological and clinical factors. Considering both physiological and clinical factors for automatic diagnosis of diabetes mellitus type 2 is a novel approach that produces a significant improvement in the accuracy of the results.

1.6 Thesis Organization

The remainder of this thesis is organised as follows. Chapter 2 provides introductory information related to this thesis, including an examination of a fuzzy logic, the inference system, and an overview about certain data mining algorithms; there is also discussion regarding associated studies. Chapter 3 offers a detailed description of the proposed methodology of our fuzzy expert system. Following this, Chapter 4 presents and discusses the results of this study; there is also a description of the implementation of the data mining models in Weka and the metrics used for evaluating the classifiers. In addition, this chapter presents a comparative study

between our fuzzy expert system and a fuzzy expert system presented in [8]. Finally, Chapter 5 draws conclusions regarding the research findings and provides guidelines for future directions.

CHAPTER 2

PRELIMINARIES AND RELATED WORK

This chapter describes preliminary information relevant to the topic at hand, including fuzzy Logic, fuzzy set, fuzzy inference system, and some studies related to the use of the fuzzy inference system in medical diagnosis. In addition to this, there are reviews of certain data mining algorithms that will be used for comparison purposes in the results chapter (Chapter 4), and some studies related to the use of these data mining algorithms in diabetes diagnosis.

2.1 Fuzzy Logic

According to Lukasiewicz [12], "Fuzzy logic (FL) is multi-valued logic". He used a number of terms such as “old”, “hot”, and “tall” to study the mathematical representation of fuzziness. Following this, the terms “fuzzy sets” and “fuzzy logic” were introduced by Professor Lotfi A. Zadeh in1965. According to Zadeh, "Fuzzy logic is an addition of the classic logic" [13, 14]. Classical binary logic (Boolean logic) operates with only two values, which are 0 (false), and 1 (true). On the other hand, in fuzzy logic, the range of logical values is extended to all real numbers in the interval between 0 and 1. Fuzzy logic deals with degrees of membership and degrees of truth value. It employs the spectrum of colours rather than relying solely on black and white. This means that things can be partly false and partly true at the same time [15]. Figure 2.1 illustrates how fuzzy logic adds a range of truth values to classic logic.

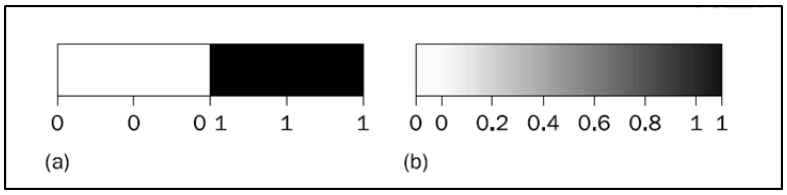

Figure 2.1: Range of Logical Values in Classical and Fuzzy Logic: (a) Boolean Logic; (b) Multivalued Logic [15]

2.2 Membership Function

A membership function (MF) is a distribution that maps each element in the universe of discourse (input space) to a membership value between 0 and 1. Fuzzy sets have several types of membership function, such as trapezoidal membership function, triangular membership function, gaussian membership function, and sigmoid membership function etc. The type of membership function is chosen based on the concept that is being represented, and the context of its use [16]. The triangular and trapezoidal membership functions are used in this study (see Figure 2.2 and Figure 2.3).

The triangular curve is a vector function (x). It is determined by three scalar parameters, which are a, b, and c, as given by:

$$f(x,a,b,c) = \begin{cases} 0, & if \quad x \le a \\ \dfrac{x-a}{b-a}, & if \quad a \le x \le b \\ \dfrac{c-x}{c-b}, & if \quad b \le x \le c \\ 0, & if \quad x \ge c \end{cases} \tag{2.1}$$

Parameters a and c define the "bases" of the triangle shape and parameter b defines the "peak". Figure 2.2 illustrates the triangular membership function.

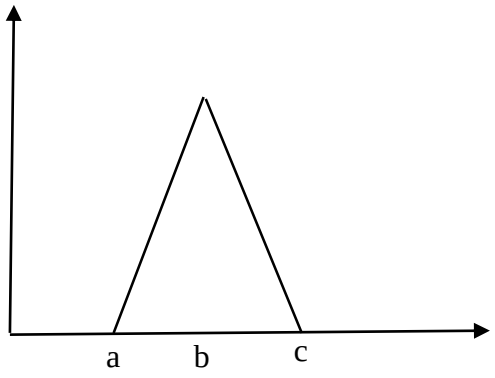

Figure 2.2: Triangular membership function

The trapezoidal curve is a vector function (x), and is determined by four scalar parameters, namely a, b, c, and d, as given by:

$$f(x,a,b,c,d) = \begin{cases} 0, & x \le a \\ (x-a)/(b-a), & a \le x \le b \\ (d-x)/(d-c), & b \le x \le c \\ 1, & c \le x \le d \\ 0, & x \ge d \end{cases} \qquad (2.2)$$

Parameters a and d define the "bases" of the trapezoid shape while parameters b and c define the "shoulders". The trapezoidal membership function is shown in Figure 2.3.

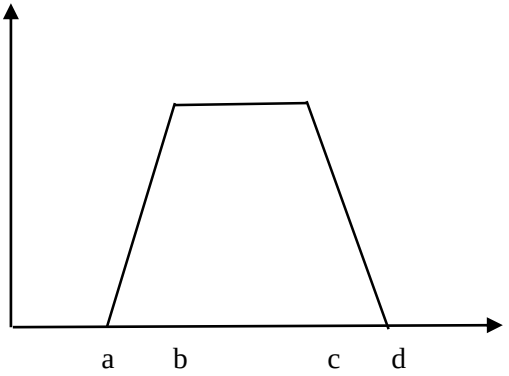

Figure 2.3: Trapezoidal Membership Function

The triangular and trapezoidal membership functions can be combined together, as illustrated in Figure 2.4.

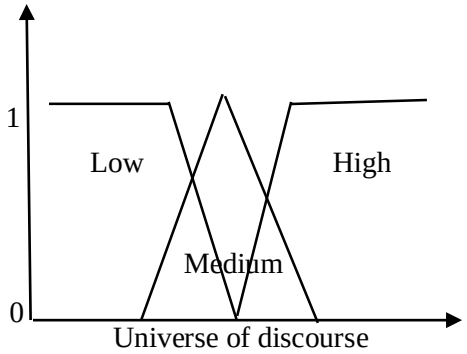

Figure 2.4: A Basic Fuzzy Set of Triangular and Trapezoidal Membership Functions

2.3 Crisp and Fuzzy Sets

Let X be the universe of discourse and its element be denoted as x. In the classical logic, crisp set A of X is defined by a function called characteristic function of A:

$f_A(x): X \longrightarrow \{0, 1\}$

$$f_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

The crisp set maps universe X to a set of two elements, which are 1 and 0. For any element x of universe X, characteristic function f_A (x) equals 1 if x is an element of set A, and equals 0 if x is not an element in the set A [13, 17, 18, 19].

In fuzzy logic, fuzzy set A of universe X is defined by function μ_A, called the membership function of set A:

$\mu_A(x) : X \longrightarrow [0,1]$

$$\begin{cases} 1 & if\ x\ is\ totally\ in\ A \\ 0 & if\ x\ is\ not\ in\ A \\ 0 < \mu\_A < 1 & if\ x\ is\ partialy\ in\ A \end{cases}$$

For any element x of universe X, membership function $\mu_A$is equal to the degree to which x is an element of set A. This degree is a value between 0 and 1, and represents the degree of membership of element x in the set A [13, 17, 18, 19].

2.4 Operations of Fuzzy Sets

Union, intersection, and complement are the three main operations of fuzzy sets. Assume that A and B are two fuzzy sets and x is an element in the universe of disclosure (X). Both fuzzy sets A and B are defined by their membership functions, which are μ_A and μ_B.The three basic operations of fuzzy sets are defined as [13, 17, 18, 19]:

a) Intersection:

$$\mu_{A\cap B}(x) = \mu_A(x) \cap \mu_B(x) = \min(\mu_A(x), \mu_B(x))$$

b) Union:

$$\mu_{A\cup B}(x) = \mu_A(x) \cup \mu_B(x) = \max(\mu_A(x), \mu_B(x))$$

c) Complement:

$$\mu_{\neg A}(x) = 1 - \mu_A(x)$$

The graphical representations of those three main operations are shown in Figure 2.5

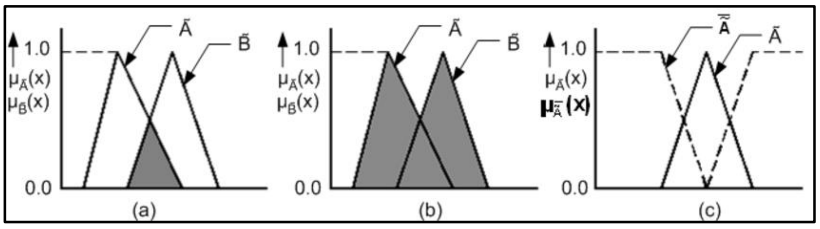

Figure 2.5: Fuzzy Sets Operations [20]

2.5 Linguistic Variables and Linguistic Values

The roots of fuzzy set theory can be traced back to the idea of linguistic variables. A linguistic variable is a fuzzy variable. The values of linguistic variables derive from artificial language or natural language, such as words or sentences. For instance, the sentence 'Age is old' implies that the linguistic variable "Age" accepts the linguistic value "old". In fuzzy expert systems, linguistic variables are used in fuzzy rules. For example, if Age is old then incidence of diabetes is also high. The range of possible values of a linguistic variable represents the universe of discourse of that variable. For instance, the universe of discourse of the linguistic variable Age may have a range between 0 and 100, and might involve fuzzy subsets such as very young,

young, middle aged, old, and very old. Moreover, every fuzzy subset represents a linguistic value of the corresponding linguistic variable [21].

2.5.1 Fuzzy IF-THEN Rules

One of the most important parts of the fuzzy inference system is the rule-based knowledge. The rules of the fuzzy inference system can be created after defining the membership functions using the linguistic variables and linguistic values. The rules of the fuzzy inference system map the inputs to the outputs. The fuzzy rules can be broken down into two parts, namely antecedent(s) or premise(s) and consequence(s) or conclusion(s). An antecedent might have one or more (AND) or (OR) operators [15, 18, 22]. The form of the fuzzy rule can be defined as:

Rule 1: IF a isX1 OR b is Y1 THEN c is Z1.

Rule 2: IF a isX2 AND b is Y2 THEN c is Z2.

Rule 3: IF a isX3 THEN c is Z3.

where a, b, and c are the linguistic variables and X, Y and Z are the linguistic values.

2.6 Fuzzy Inference System

A fuzzy inference system (FIS) uses fuzzy set theory to map inputs to outputs. There are two types of fuzzy inference system, namely Mamdani [23] and the Sugeno [24]. These two types have been successfully used in a variety of applications such as decision analysis, data classification, and expert systems.

In 1975, the Mamdani method was proposed by Professor Ebrahim Mamdani at the University of London. The Mamdani inference approach is used in the present thesis. This method is the most commonly used fuzzy methodology due to the following reasons:

- Mamdani has the ability to acquire human knowledge in a way that is both intuitive and human-like [25].
- Mamdani has expressive power, is easy to formalize as well as being intuitive. Due to the interpretable nature of the rules, it is widely used in the decision support application [26].
- Mamdani fuzzy inference system is more interpretative than Sugeno fuzzy inference system [27].
- Mamdani is more transparent than Sugeno when it comes to representing human knowledge. The Mamdani method is usually used in modelling human expert knowledge [28].

The Mamdani fuzzy inference process involves four main steps, namely fuzzification, rule evaluation, aggregation, and defuzzification [15]. These steps are represented in Figure 2.6 and described below:

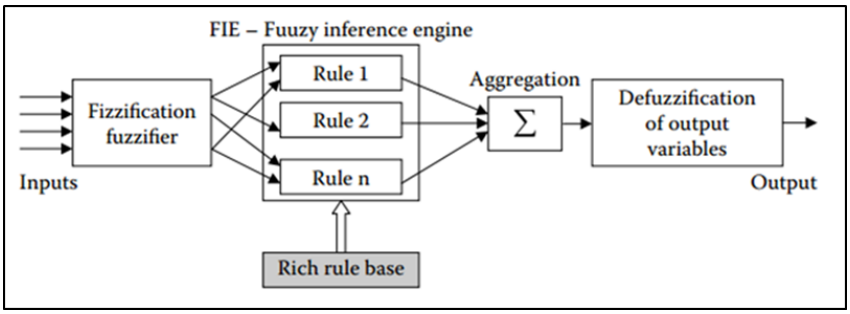

Figure 4.6: Mamdani FIS Process [15]

Step 1: Fuzzification

In this step, the crisp input values are taken and mapped to the degree of membership functions for each fuzzy set [15, 29]. Figure 2.7 shows an example of the fuzzification step.

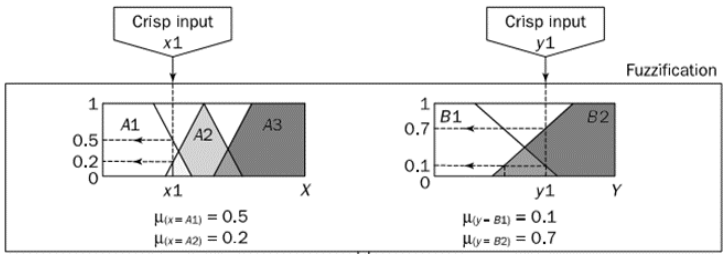

Figure 2.7: Example of Fuzzification [15]

As we can see in Figure 2.7, the crisp input x_1 corresponds to the membership functions A1, A2, and A3 to the degrees of 0.5, 0.2, and 0, respectively.

Step 2: Rule Evaluation

In the second step, the fuzzified inputs are applied to the antecedents of the fuzzy rules. If a fuzzy rule has more than one antecedent, the fuzzy operator (AND or OR) is used to get a single result of the antecedent evaluation. This result (the truth value) is then applied to the consequent membership function. This means that the fuzzy rules are mapped from input(s) to output using the membership functions. It is important to note that the number of rules depends on the number of membership functions (linguistic variables) [15]. An example of the rules evaluation is illustrated in Figure 2.8.

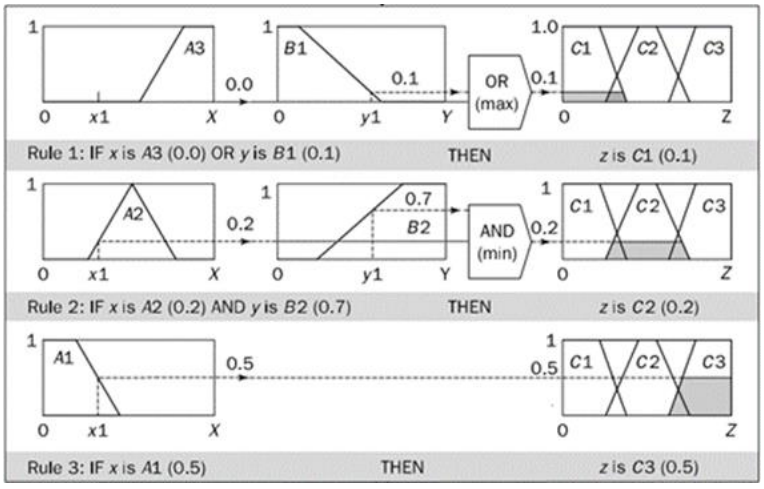

Figure 2.8: Example of the Rules Evaluation [15]

Step 3: Aggregate output(s)

In this step, the outputs of all rules are unified into one fuzzy set. Therefore, the inputs of the aggregation step form a list of scaled or clipped consequent membership functions; the output of the aggregation process is a single fuzzy set. OR operator is used to aggregate the output fuzzy sets in Mamdani [15]. An example of the aggregation of the outputs is shown in Figure 2.9 below.

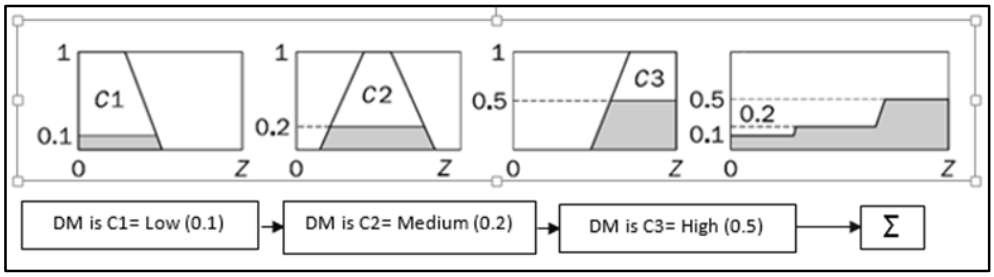

Figure 2.9: Example of Aggregation of the outputs [15]

Step 4: Defuzzification

This step is the last phase of the Mamdani fuzzy inference process. The final result of the fuzzy inference system must be a single number. As such, these defuzzification methods are used to generate the single number from the single fuzzy set that is obtained from the aggregation step. The input of the defuzzification process is the aggregated fuzzy set, while the output is a crisp number. There are various defuzzification techniques, such as bisector of area (Bisector), centre of gravity (COG) (also known as centroid of area), mean value of maximum (MOM), largest (absolute) value of maximum (LOM), and finally smallest (absolute) value of maximum (SOM) [15, 29]. The most common defuzzification method is the centroid approach. This method finds

the point that indicates the centre of gravity of the aggregated fuzzy set [30, 31].The centre of gravity is calculated using the following formula:

$$COG = \int \mu_A(x) * xdx \; / \int \mu_A(x) * dx \qquad (2.3)$$

where $\int$ denotes an algebraic integration and $\mu_A(x)$ is the value of membership function of set A. An example of the centroid defuzzification method is shown below in Figure 2.10.

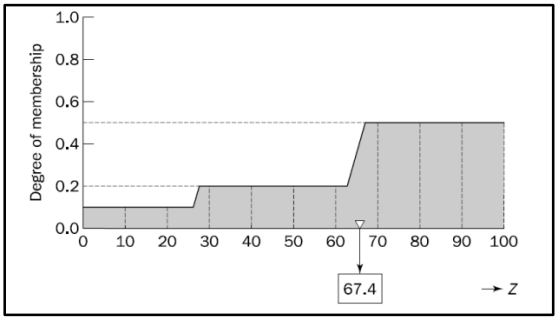

Figure 2.10: Example of Defuzzification [15]

2.6.1 Works Related to Fuzzy Inference System in Medical Diagnostic Systems

There are several developed applications that use the fuzzy inference system to diagnose different types of diseases. This section will review the literature pertaining to some of these applications.

In [32], Almadni and Abhari proposed a comparative analysis of classification models in the diagnosis of type 2 diabetes. In this research, the fuzzy expert system was developed using the Mamdani inference system to diagnose type 2 diabetes. Following this, the system in question was compared with two classification models, namely logistic regression and support vector machine. The aim was to evaluate the performance of the developed fuzzy system. The Pima Indian Diabetes Data Set, which is provided in Chapter 4, Section 4.1, was used to test the system. The results showed that the fuzzy expert system performed best compared with the other

models. However, the proposed fuzzy expert system in this thesis is more accurate than the fuzzy expert system presented in [35]. Indeed, the membership functions and the fuzzy rules of the developed system in this thesis are better than the other [35], and this increased the accuracy of the system significantly.

In [8], the fuzzy expert system for the diagnosis of diabetes using a fuzzy determination mechanism was implemented by Kalpana and Kumar. Their system diagnosed youths (from 25 to 30 years of age). The Mamdani fuzzy inference method was applied and the Pima Indian Diabetes Dataset (PIDD) was also used. The PIDD consisted of 9 attributes and 768 records. Some of the instances that relate to young patients were used. Moreover, six of the nine attributes of the original dataset were used to build this system. Details regarding the PIDD will be provided in Chapter 4, Section 4.1.

In [33], Adeli and Neshat proposed an expert system for heart disease diagnosis using fuzzy logic. The Mamdani inference technique was used to build this system. In addition, the Heart Disease Data Set of the V.A. Medical Centre, Long Beach, and Cleveland Clinic Foundation database were used to implement the system. This data set included 13 attributes and 303 instances. However, this study used 11 out of 13 attributes of the original data set. Input attributes included age, sex, chest pain type, cholesterol level, resting electrocardiography, blood sugar, blood pressure, maximum heart rate, old peak, exercise, and thallium scan. The output relates to the presence of heart disease in the patient. There were five fuzzy sets of the output that indicate the exact stage of the heart disease development process: healthy, mild, moderate, severe, and very severe.

In [34], Parvin and Abhari proposed a fuzzy database for heart disease diagnosis. They improved on the previous work of Adeli et al. by including all the attributes of the original dataset and

increasing the number of rules. However, their system was implemented in the form of a fuzzy database management system, which diagnoses the severity of the patient's heart disease.

In [35], Hamidzadeh, Javadzadeh and Najafzadeh proposed a fuzzy rule based diagnostic system for detecting the lung cancer disease. The system has nine input attributes and one output, with the input attributes including chest pain, bone pain, smoking, weight loss, persistent cough, coughing up blood, hoarseness of voice, age, and shortness of breath. The output attribute is based on advances of the tumour and the spreading of the tumour. The output has four fuzzy sets, namely no cancer, stage 1, stage 2, and stage 3.The system was tested based on the data obtained from 62 patients. The Mamadani inference engine was used to map the input attributes to stage the cancer.

In [36], Neshat et al. built a fuzzy expert system for diagnosis of liver disorders. The Mamdani inference method was applied in this study. The liver disorders dataset was used, as this is part of the UCI database. This dataset has 345 samples and 6 indicators, which are aminotransferase (sgpt alamine), gammagt, aspartate aminotransferase (sgot), means corpuscular volume (MCV), alkaline phosphates (alkphos), and number of half-pint equivalents of alcoholic beverages drunk per day. Indeed, these indicators were used to split the data into two sets, namely healthy liver and unhealthy liver.

In [37], Kadhim, Alam, and Kuar designed and implemented a fuzzy expert system for back pain diagnosis. The input attributes of this fuzzy expert system are body mass index (BMI), gender, age, and the clinical observation symptoms. This system was tested using clinical data that belonged to 20 patients with different back pain diseases. Visual Prolog programming language was used to implement this Mamdani fuzzy expert system.

In [38], Muthukaruppan et al. proposed a hybrid particle swarm optimisation-based fuzzy expert system. The purpose of implementing this system was to diagnose coronary artery disease. Since the Heart disease dataset of the V.A. Medical Center, Long Beach, and the Cleveland Clinic Foundation database include several input attributes, the decision tree (DT) algorithm was applied to the input attributes so as to unravel them and contribute to the diagnosis. The output of the decision tree algorithm was converted into membership functions and the fuzzy rule base in order to build the fuzzy system.

In [10], Thirugnanam et al. proposed a novel approach for the diagnosis of diabetes mellitus. This approach consisted of two stages to predict whether or not a person has diabetes. During the first stage, fuzzy logic and neural network were applied to the training data as an individual approach and the results were stored in a database. In the second stage, a rule-based algorithm was applied to obtain the final results. A survey was then carried out in order to collect the dataset. The input attributes of the dataset were age, gender, family history, smoking, quantity of vegetable and fruit intake, having high blood sugar, taking medicines for blood pressure, physical activity, body mass index, waist to hips ratio, frequency of urination, hunger, thirst level, poor wound healing, itching over the entire body, gestational diabetes, and frequent intake of non-vegetarian food.

2.7 An Overview of Certain Classification and Statistical Methods

2.7.1 J48 Decision Tree

The J48 decision tree classifier is a predictive machine-learning algorithm. It assigns a target value for a new sample based on different attribute values of the dataset. This classifier structures a tree and makes the model more understandable. Thus, it works well in solving many classification problems. In addition, it does not make any prior assumptions about the data, and

it can process both nominal and numerical data. However, the class (output attribute) must be nominal. The decision tree building process is unstable since a slight change in the data may lead to a quite different decision tree [39]. There are two main steps when it comes to inducing a tree, namely the growing phase and the pruning phase. During the first phase, a tree is grown to a sufficiently large size. In most cases, this tree has a large number of redundant nodes. In the second phase, the tree is pruned by removing the redundant nodes [40].

The Pseudo-Code for the J48 growing Algorithm

Algorithm Grow (root, T)
  Input: R root node, T set of training cases
  Output: A trained decision tree, appended to root
if T is pure then
  Assign class of pure cases to root;
  return;
end
if no split yield minimum number of cases split off then
  Assign class of node as the majority class of T
  return;
end
Find an optimal split that divides T into subsets T_n (n=1:N);
foreach n= 1 to N do
  Create a node $l_n$under root;
  Grow (l_n,T_n);
end

The gain ratio should be calculated to obtain the optimal split while growing the tree.

$$GainRatio(X) = \frac{Gain(X)}{SplitInfo(X)}$$

The gain in information made by the split is simply the difference between the amount of information needed to classify a case before and after making the split.

$$Gain(X) = Info(T) - Info_X(T)$$

The gain ratio divides the gain by the evaluated split information.

$$SplitInfo(X) = -\sum_{i=1}^{n+1} \frac{T_i}{T} \cdot \log_2 \frac{T_i}{T}$$

The split information is the weighted average calculation of the information using the proportion of cases which are passed to each child.

The Pseudo-Code for the J48 Pruning Algorithm

```
input: node with an attached subtree
output: pruned tree
leaf-error = estimated leaf error;
if a node is leaf then
   return leaf-error;
else
      subtree-error = ∑ Prunet_i; //t_i ∈ children(node)
      branch-error = error if replaced with the most frequented branch;
  if leaf-error is less than subtree-error and branch-error
      make this node a leaf node;
      error= leaf-error;
  else if branch-error is less than leaf-error and subtree-error
       replace this node with the most frequented branch;
       error = branch-error;
    else
       error = subtree-error;
   return error;
```

end

This algorithm works from the bottom of the tree upward and removes or replaces branches to minimise the predicted error.

2.7.2 Multilayer Perceptrons

In the present study, the standard back propagation-based multilayer perceptron (MLP) architecture of ANN was used. This architecture is most commonly used for ANN in medical research.

Multilayer perceptrons (MLP) are feed-forward neural networks that include multilayer nodes with at least one hidden layer. Each node is a neuron that has a non-linear activation function which defines its output given a set of inputs. A back-propagation learning approach is used by MLP to train the network to find the weight that maps an input to an output. A neural network can solve either classification or regression problems based on the activation function [41].

The simplest Neural Network consists of only one neuron (called a perceptron). If the product of input value and weight is greater than 0, the output of the perceptron will be 1, otherwise the output is 0. The perceptron is trained to learn how to modify the weights [42]. The pseudo- code of training the perceptron is represented as the following:

The Pseudo- Code for Training the Perceptron

while there is input-output to classify do

Select one pair of an input and an output (a_n, b_n)

Compute the output of perceptron $c = f(a_1 * w_1 + a_2 * w_2 + \cdots + a_n * w_n)$

if c $\leq$ 0 and b_n > 0 then

$$w_{i+1} = w_i + a_n$$

else if c > 0 and $b_n \leq 0$ then

$$w_{i+1} = w_i - a_n$$

else

No adjustment

end if

end while

If the output data cannot be split into two classes, then it is not possible to use the perceptron as a classifier. This problem could be solved by adding layers of neuron to the network.

2.7.3 Logistic Regression

Logistic regression is a common type of a generalised linear model. "Logistic regression models the probability of some event occurring as a linear function of a set of predictor variables". Instead of predicting the value of the dependent variable, the logistic regression approach estimates the probability p that the dependent attribute will have a given value. For example, rather than predicting whether a patient is a diabetic or non-diabetic, the logistic regression method tries to estimate the probability of a patient being diabetic. The actual state of the dependent attribute is determined by looking at the output (estimated probability). If the output is greater than 0.50 then the prediction is closer to YES (Diabetic), otherwise the estimated probability is closer to NO (Non-diabetic). Thus, the probability p is called the success probability in logistic regression [43].

Logistic regression tries to find the relation between independent and dependent variables. It can be used as a predicting model because the method estimates how the output changes with the change in input variables. If the model fits dependent variables to independent ones with minimum error, it will be considered a good model.

If there are two or more independent variables, multiple linear regression can estimate the value of the output. Multiple linear regression was used in the present study. The goal of linear regression is to find β coefficients, so that the predicted value is as close to the actual value as it can be. The following is the equation of the multiple regression model:

$$\varepsilon\, Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots$$

where βi describes the rate of change of output when Xi is increased by one unit, assuming that all other variables are held constant [44].

2.7.4 Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) solves the Quadratic Programming (QP) optimisation problem which emerges during the training of support vector machines. It does so by breaking down the QP problem into smaller optimisation sub-problems. SMO is faster than the other SVM algorithms and has better scaling than any other SVM algorithms, since it uses the smallest possible QP problem. It consists of two parts, namely an analytical solution to a QP problem of the two Lagrange multipliers, and a set of heuristics designed to efficiently choose which multipliers to optimise [45, 46]. The algorithm of SMO is provided in [45].

SVM is a supervised learning approach. SVM maps the training data into another space higher than the original space and divides the instances belonging to different categories by separating these instances linearly and non-linearly. SVM tries to keep separation boundary between two different categories (classes) as wide as possible. The perpendicular bisector of the shortest line connecting the two classes is called hyperplane. The hyperplane which is the farthest from both classes. The training instances closest to the hyperplane are called support vectors. The support vectors are very important, because they determine the hyperplane, while the other instances

might be forgotten. After drawing the hyperplane, the test instances are mapped into the same training space. A class value is determined for each test instance by SVM model [47]. Figure 2.10 shows SVM for linearly separable data.

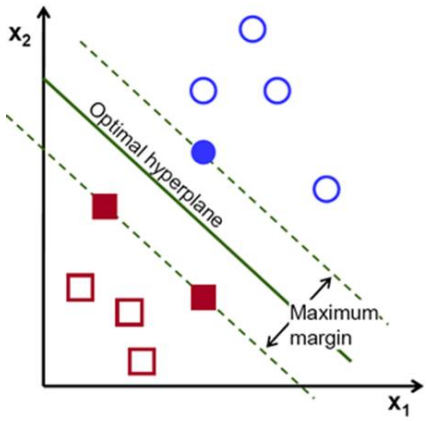

Figure 2.11: SVM for linearly separable

As we can see, the circles and squares are data points belonging to two different classes. In this study, there are two classes which are Diabetic and Non-Diabetic. For each data point (x,y), x is the condition of the patient, y is either 1 or -1 denoting the class to which point x belongs. The classes can be fully separated by the optimal hyperplane. The separation boundary is the optimal hyperplane that leaves the maximum margin from the classes (Diabetic and Non-Diabetic). The margin is the distance between the hyperplane and the closest data point (called support vectors) to the hyperplane. To avoid misclassification, SVM tries to maximise the margin. The classification is done based on the hyperplane function:

$$\boldsymbol{w}^T \boldsymbol{v} + b = 0$$

where *w* is a weight vector (w1,w2) and *b* is the bias. It will be computed by SVM in the training process.

A binary SVM classifies data point $\boldsymbol{v_i}$ if it is above the separation boundary as class 1 ($\boldsymbol{w}^T \boldsymbol{v} + b > 0$) and data point $\boldsymbol{v_i}$ as class -1 if it is below the separation boundary ($\boldsymbol{w}^T \boldsymbol{v} + b < 0$) [48].

2.7.5 Naïve Bayes

One of the most efficient and effective inductive learning algorithms for machine learning and data mining is Naïve Bayes. It depends on the Bayes rule of conditional probability. It assumes that the parameters of a class are independent from each other. In other words, it assumes that the effect of an attribute value of a given class is independent of the values of the other attributes. Therefore, every parameter makes an independent contribution to the prediction of the final result. This assumption can sometimes negatively impact the accuracy of the model. Naïve Bayes can only work with nominal classes. It is easy to implement and, in some cases, outperforms many other complex algorithms. It is a powerful probabilistic representation that is robust in the face of noise and can handle null values. Moreover, practical dependencies exist among variables in this classifier. The Naïve Bayes formula is shown below [49].

$$P\,(E \mid H) \;=\; P(E1 \mid H) \;*\; P(E2 \mid H) \;*\; P(E3 \mid H) \;* \ldots * P(En \mid H) \;*\; P(H) \,/\, P(E)$$

where H is a hypothesis (an output attribute) and E is the evidence (set of input attributes).

2.7.6 Works Related to Certain Classification and Statistical Methods

Sarwar and Sharma [50] carried out a comparative analysis of machine learning techniques in relation to the prognosis of type 2 diabetes. They applied various data mining algorithms in order to diagnose diabetes and analyse the efficiency of these algorithms in predicting the results. They used a dataset containing 500 randomly selected people whose ages ranged from 5 years to 78 years. The dataset consisted of 10 physiological parameters, namely age, family history, weight, height, sex, fatigue, drinking, smoking, thirst, and frequency of urination. Naïve

Bayes, Artificial Neural Networks (ANN), and the decision tree were used to analyse the dataset. These models were implemented using Matlab. Moreover, the standard back propagation-based multilayer perceptron (MLP) architecture of ANN was used to build this model.

Chen and Tan [9] published a paper entitled "prediction of Type-2 Diabetes based on several element levels in blood and chemometrics". They evaluated the levels of eight elements including lithium, zinc, chromium, copper, iron, manganese, nickel and vanadium in the whole blood of type-2 diabetes mellitus patients. Following this, they compared the patients with age-matched healthy controls in order to investigate the feasibility of combining them with an ensemble model for diagnosing diabetes. The dataset included 158 samples, among which 105 were collected from healthy adults while the remaining 53 were taken from diabetic patients. In addition, the collected data set was equally divided into two parts, namely training data and testing data. Chromium and iron were picked out of the eight elements since they were the most important. These two elements were also used for modelling. Fisher Linear Discriminate Analysis (FLDA), Support Vector Machine (SVM), and Decision Tree (DT) were used to build the member models.

Kumari and Singh [11] implemented a neural network system to diagnose diabetes mellitus. A standard questionnaire was administered to 100 hundred patients in order to collect the data. The dataset included thirteen physiological attributes, including age, gender, weight, height, weight loss, thirst, hunger, appetite, nausea, fatigue, vomiting, blurred vision, and bladder, skin, and vaginal infections. In addition, this dataset was divided into two parts. 80% was for the training part and 20% was for the testing part. The neural network was designed and tested by using Matlab software. There were 28 nodes in total, 13 of which were input nodes, while 13 were hidden nodes, and 1 was an output node. An initial weight was assigned to each input. An

output was obtained in binary; zero value means the person is not suffering from diabetes mellitus and a value of one reveals that the person has diabetes mellitus.

Meng et al. [51] conducted a study entitled "comparison of three data mining models for predicting diabetes or pre-diabetes by risk factors". The authors compared the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or pre-diabetes based on 12 risk factors. This study analysed two types of participants; 735 patients were suffering from diabetes while 752 normal controls were also recruited. A survey was carried out to obtain information on demographic characteristics (such as age, gender, marital status, and level of education), anthropometric measurements (such as height and weight), family diabetes history, and lifestyle risk factors (such as cigarette smoking, alcohol intake, tea and coffee consumption, work stress, physical activity, and sleep duration). Following this, the three models were built using 12 input attributes and 1 output variable from the survey information. SPSS software (version 14.1) was used to construct these three models. Statistical analyses were performed using SPSS (version 13.0). The three models were evaluated based on confusion matrix, accuracy, sensitivity and specificity.

Cedeno and Andina [52] applied the Artificial Metaplasticity on Multilayer Perceptron (AMMLP) as a data mining method to diagnose type 2 diabetes. The Pima Indian Diabetes Dataset was used to validate the developed model. Following this, the results of the AMMLP model were compared with the decision tree (DT), and Bayesian classifier. Accuracy, sensitivity, specificity and confusion matrix were used to examine the performances of the algorithms.

CHAPTER 3

METHODOLOGY

This chapter will examine the methodology used to develop the fuzzy expert system, which includes the Mamdani fuzzy inference system. This system is implemented in Matlab. The process of developing the Mamdani fuzzy inference system involves four steps. First, the fuzzy sets, along with their membership functions, are generated for each attribute. Second, the fuzzy rules are defined and evaluated based on the membership functions. Following this, the outputs are aggregated in order to achieve a single fuzzy set. Lastly, the centroid defuzzification method is used to obtain the final result (crisp number) of the system. An example is provided to help in understanding the process of the system.

3.1 Comparison Framework Used in the Thesis

A framework of the comparison consists of different components as shown in Figure 3.2 described as follows:

- **Fuzzy Expert System:** This is the main component of the comparison framework. The design and the implementation of the proposed fuzzy expert system that used the Mamdani fuzzy inference system is described in the next section (Section 3.2).

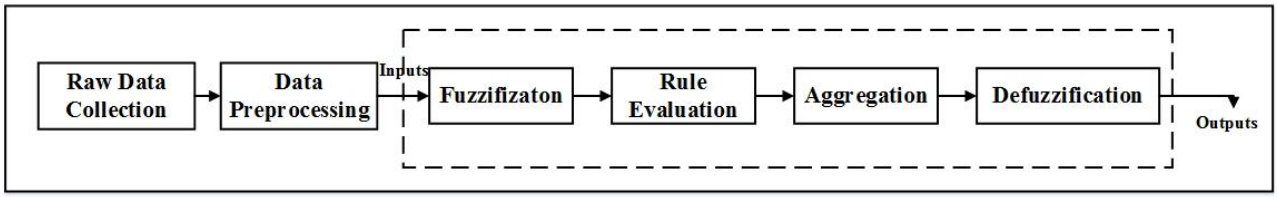

Figure 3.1: Fuzzy Expert System

- **Implementing Data Mining Algorithms:** Several data mining algorithms, which are J48, MLP, logistic regression, SVM, and Naive Bayes, are implemented in this thesis.

An overview of these algorithms is given in Chapter 2, Section 2.8 and details about the implementation of the data mining algorithms are provided in Chapter 4, Section 4.3.

- **Comparing the Fuzzy expert System with the Data Mining Algorithms**: The performance of the data mining algorithms is compared with the performance of the fuzzy expert system proposed in this thesis. The purpose of this comparison is to evaluate the performance of our proposed system. Chapter 4, Section 4.4 gives details about this comparison.
- **Implementing the Fuzzy Expert System Proposed in the Related Work:** In order to evaluate the performance of our system, a fuzzy expert system presented in [8] is implemented and provided in Chapter 4, Section 4.5.
- **Comparing the Fuzzy Expert System with the Related Work:** After replicating the related work, a comparative analysis is conducted between it and the developed fuzzy system in this thesis. Details about this comparative analysis are given in Chapter 4, Section 4.5.

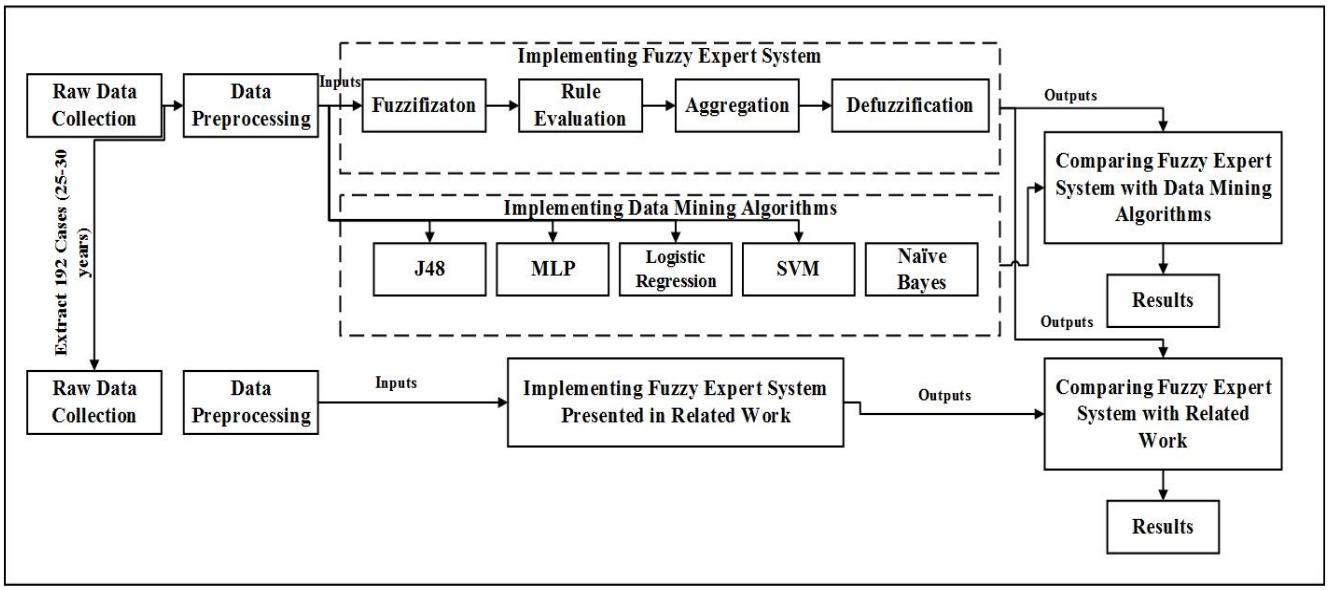

Figure 3.2: Comparison Framework

3.2 Fuzzy Inference System

The following sections will explain the Mamdani fuzzy inference system step by step.

3.2.1 Fuzzification

This is the first step of the Mamdani fuzzy inference system. In the fuzzification step, the fuzzy sets for the input attributes and the output, along with the membership functions are defined. The attributes of the Pima Indian Diabetes Dataset are used. Details about this dataset are provided in Chapter 4, Section 4.1.

Fuzzy Sets for the Input attributes and for the Output of Diabetes Mellitus:

- Age: {Young, Middle Aged, Old, Very Old}.
- Glucose: {Low, Normal, Medium, High, Very High}.
- Insulin: {Low, Medium, High}.
- Body Mass Index: {Normal, Medium, High}.
- Number of Pregnancies: {Absent, Average, High}.
- Triceps Skin Fold Thickness: {Normal, Medium, High}.
- Diabetes Pedigree Function :{ Low, Medium, High, Very High}.
- Diastolic Blood Pressure: {Low, Medium, High, Very High}.
- Output: {Low, Medium, High}.

Table 3.1: Ranges of the Output of Fuzzy Expert System

Output	Range	Fuzzy set
Result	< 0.5	Low
	0.4 – 0.6	Medium
	0.5 – 1	High

Table 3.1 presents the ranges of each fuzzy set of the output. This research considered low as “Non-Diabetic”, while medium and high are treated as “Diabetic”.

Once all of the attributes and their fuzzy sets were defined, the range values were prepared for all fuzzy sets of each attribute based on the data collected from the physicians. Following this, the formulae were constructed by using the ranges to generate the membership functions. Triangular and trapezoidal membership function formulas were used in this study. These two types were selected over other types of membership functions, since their structures are simple. Kummar et al. used only the triangular membership function. However, combining the triangular and trapezoidal membership functions gave us better results.

The membership functions of each input attribute are calculated as the following:

1. The membership function of age is calculated as follows:

$$\mu_{Young}(Age) = \begin{cases} 1, & if \quad Age \leq 20 \\ \frac{35 - Age}{14}, & if \quad 21 \leq Age \leq 35 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Middle}(Age) = \begin{cases} 1, & if \quad Age = 35 \\ \frac{Age - 25}{10}, & if \quad 25 \leq Age \leq 35 \\ \frac{45 - Age}{10}, & if \quad 35 \leq Age \leq 45 \end{cases}$$

$$\mu_{Old}(Age) = \begin{cases} 1, & if \quad Age = 45 \\ \frac{Age - 35}{10}, & if \quad 35 \leq Age \leq 45 \\ \frac{55 - Age}{10}, & if \quad 45 \leq Age \leq 55 \end{cases}$$

$$\mu_{Very\ Old}(Age) = \begin{cases} 1, & if \quad Age \geq 55 \\ \dfrac{Age - 45}{10}, & if \quad 45 \leq Age \leq 55 \\ 0, & Otherwise \end{cases}$$

2. The membership function of glucose concentration in blood is calculated as the following:

$$\mu_{Low}(Glucose) = \begin{cases} 1, & if \quad Glucose \leq 70 \\ \dfrac{94 - Glucose}{24}, & if \quad 70 \leq Glucose \leq 94 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Normal}(Glucose) = \begin{cases} 1, & if \quad Glucose = 105 \\ \dfrac{Glucose - 70}{35}, & if \quad 70 \leq Glucose \leq 105 \\ \dfrac{140 - Glucose}{35}, & if \quad 105 \leq Glucose \leq 140 \end{cases}$$

$$\mu_{Medium}(Glucose) = \begin{cases} 1, & if \quad Glucose = 140 \\ \dfrac{Glucose - 105}{35}, & if \quad 105 \leq Glucose \leq 140 \\ \dfrac{175 - Glucose}{35}, & if \quad 140 \leq Glucose \leq 175 \end{cases}$$

$$\mu_{High}(Glucose) = \begin{cases} 1, & if \quad Glucose = 175 \\ \dfrac{Glucose - 140}{35}, & if \quad 140 \leq Glucose \leq 175 \\ \dfrac{210 - Glucose}{35}, & if \quad 175 \leq Glucose \leq 210 \end{cases}$$

$$\mu_{Very\ High}(Glucose) = \begin{cases} 1, & if \quad Glucose \geq 199 \\ \dfrac{Glucose - 175}{24}, & if \quad 175 \leq Glucose \leq 199 \\ 0, & Otherwise \end{cases}$$

3. The membership function of serum insulin is calculated as follows:

$$\mu_{Low}(Insulin) = \begin{cases} 1, & if \quad Insulin \leq 15 \\ \dfrac{89 - Insulin}{74}, & if \quad 15 \leq Insulin \leq 89 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Medium}(Insulin) = \begin{cases} 1, & if \quad Insulin = 89 \\ \dfrac{Insulin - 15}{74}, & if \quad 15 \le Insulin \le 89 \\ \dfrac{194 - Insulin}{105}, & if \quad 89 \le Insulin \le 194 \end{cases}$$

$$\mu_{High}(Insulin) = \begin{cases} 1, & if \quad Insulin \ge 194 \\ \dfrac{Insulin - 89}{105}, & if \quad 89 \le Insulin \le 194 \\ 0, & Otherwise \end{cases}$$

4. The membership function of body mass index (BMI) is calculated as follows:

$$\mu_{Normal}(BMI) = \begin{cases} 1, & if \quad BMI \le 25 \\ \dfrac{35 - BMI}{10}, & if \quad 25 \le BMI \le 35 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Medium}(BMI) = \begin{cases} 1, & if \quad BMI = 35 \\ \dfrac{BMI - 25}{10}, & if \quad 25 \le BMI \le 35 \\ \dfrac{45 - BMI}{10}, & if \quad 35 \le BMI \le 45 \end{cases}$$

$$\mu_{High}(BMI) = \begin{cases} 1, & if \quad BMI \ge 55 \\ \dfrac{BMI - 35}{10}, & if \quad 35 \le BMI \le 45 \\ 0, & Otherwise \end{cases}$$

5. The membership function of number of pregnancies (NP) is calculated as follows:

$$\mu_{Absent}(NP) = \begin{cases} 1, & if \quad NP \le 0.5 \\ \dfrac{4 - NP}{3.5}, & if \quad 0.5 \le NP \le 4 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Average}(NP) = \begin{cases} 1, & if \quad NP = 4 \\ \dfrac{NP - 0.5}{3.5}, & if \quad 0.5 \le NP \le 4 \\ \dfrac{8 - NP}{4}, & if \quad 4 \le NP \le 8 \end{cases}$$

$$\mu_{High}(NP) = \begin{cases} 1, & if \quad NP \geq 8 \\ \frac{NP-4}{4}, & if \quad 4 \leq NP \leq 8 \\ 0, & Otherwise \end{cases}$$

6. The membership function of triceps skin fold thickness is calculated as follows:

$$\mu_{Normal}(TSFT) = \begin{cases} 1, & if \quad TSFT \leq 5 \\ \frac{20-TSFT}{15}, & if \quad 5 \leq TSFT \leq 20 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Medium}(TSFT) = \begin{cases} 1, & if \quad TSFT = 20 \\ \frac{TSFT-5}{15}, & if \quad 5 \leq TSFT \leq 20 \\ \frac{35-TSFT}{15}, & if \quad 20 \leq TSFT \leq 35 \end{cases}$$

$$\mu_{High}(TSFT) = \begin{cases} 1, & if \quad TSFT \geq 35 \\ \frac{TSFT-20}{15}, & if \quad 20 \leq TSFT \leq 35 \\ 0, & Otherwise \end{cases}$$

7. The membership function of diabetes pedigree function is calculated as follows:

$$\mu_{Low}(DPF) = \begin{cases} 1, & if \quad DPF \leq 0.1 \\ \frac{0.4-DPF}{0.3}, & if \quad 0.1 \leq DPF \leq 0.4 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Medium}(DPF) = \begin{cases} 1, & if \quad DPF = 0.4 \\ \frac{DPF-0.2}{0.2}, & if \quad 0.2 \leq DPF \leq 0.4 \\ \frac{0.6-DPF}{0.2}, & if \quad 0.4 \leq DPF \leq 0.6 \end{cases}$$

$$\mu_{High}(DPF) = \begin{cases} 1, & if \quad DPF = 0.6 \\ \frac{DPF-0.2}{0.2}, & if \quad 0.4 \leq DPF \leq 0.6 \\ \frac{0.6-DPF}{0.3}, & if \quad 0.6 \leq DPF \leq 0.9 \end{cases}$$

$$\mu_{Very\ High}(DPF) = \begin{cases} 1, & if \quad DPF \geq 0.9 \\ \dfrac{DPF - 0.8}{0.1}, & if \quad 0.8 \leq DPF \leq 0.9 \\ 0, & Otherwise \end{cases}$$

8. The membership function of diastolic blood pressure (BP) is calculated as follows:

$$\mu_{Low}(BP) = \begin{cases} 1, & if \quad DPF \leq 111 \\ \dfrac{142 - BP}{31}, & if \quad 111 \leq DPF \leq 142 \\ 0, & Otherwise \end{cases}$$

$$\mu_{Medium}(BP) = \begin{cases} 1, & if \quad BP = 139 \\ \dfrac{BP - 127}{12}, & if \quad 127 \leq BP \leq 139 \\ \dfrac{153 - BP}{14}, & if \quad 139 \leq BP \leq 153 \end{cases}$$

$$\mu_{High}(BP) = \begin{cases} 1, & if \quad BP = 157 \\ \dfrac{BP - 142}{15}, & if \quad 142 \leq BP \leq 157 \\ \dfrac{157 - BP}{15}, & if \quad 157 \leq BP \leq 172 \end{cases}$$

$$\mu_{Very\ High}(BP) = \begin{cases} 1, & if \quad BP \geq 172 \\ \dfrac{BP - 157}{15}, & if \quad 157 \leq DPF \leq 172 \\ 0, & Otherwise \end{cases}$$

After defining the fuzzy sets and their membership functions, the crisp inputs were transformed into the degree to which they belonged to each of the appropriate fuzzy sets. For example, here we look at a patient with the following attributes: Age= 46, Glucose= 155, Insulin= 495, Body Mass Index=36, Number of times pregnant=8, Triceps Skin Thickness= 26, Diabetes Pedigree Function = 0.543, Diastolic Blood Pressure= 82. The fuzzy sets and membership functions of the input attributes of this patient are shown in the following table.

Table 3.2: Example of the Fuzzification Process as it pertains to the attributes of the given patient

Input attributes	Fuzzy sets	Membership functions
Age = 46	Young	0
	Middle age	0
	Old	0.9
	Very Old	0.1
Glucose = 155	Low	0
	Normal	0
	Medium	0.57
	High	0.43
	Very High	0
Insulin= 495	Low	0
	Medium	0
	High	1
BMI= 35	Normal	0
	Medium	1
	High	0
NP= 8	Absent	0
	Average	0
	High	1
TSFT = 26	Normal	0
	Medium	0.6

	High	0.4
DPF = 0.543	Low	0
	Medium	0.285
	High	0.715
	Very High	0
BP = 82	Low	1
	Medium	0
	High	0
	Very High	0

As we can see in Table 3.2, the input column contains the crisp input of each attribute, while the fuzzy set column contains the fuzzy sets of each attribute, and the membership function column presents the degree (value) of membership functions generated for the fuzzy set of each attribute. For example, once the crisp input BMI=34 is obtained, it is fuzzified against the appropriate linguistic fuzzy sets. The crisp input Glucose level corresponds to the membership functions Low, Normal, Medium, High, and Very High to the degrees of 0, 0, 0.57, 0.43, and 0, respectively. In addition, the crisp input TSFT maps the membership functions Normal, Medium, and High to the degrees of 0, 0.6, and 0.4, respectively. In this manner, all of the crisp inputs are fuzzified over all the membership functions used by fuzzy rules.

3.2.2 Rule Evaluation

In this step, the fuzzy rules were generated and evaluated. In this system, some of the rules have a single antecedent and a single consequent, whereas other rules have multiple antecedents and

a single consequent. The number of fuzzy rules in the system is 28. These rules are defined by physicians (domain experts) and presented in the following table.

Table 3.3: Fuzzy Rules of Fuzzy Expert System

1	If (Glucose is Low) Then (DM is Low)
2	If (Glucose is Very High) Then (DM is High)
3	If (Glucose is High) Then (DM is High)
4	If (Glucose is Medium) Then (DM is Medium)
5	If (Glucose is Medium) & (BMI is High) & (TSFT is High) Then (DM is High)
6	If (Glucose is Medium) & (BMI is Medium) & (DPF is High) Then (DM is High)
7	If (Glucose is Medium) & (BMI is Medium) & (DPF is Very High) Then (DM is High)
8	If (Glucose is Medium) & (INS is Medium) & (BMI is Low) & (Age is Young) & (NP is Absent) Then (DM is Low)
9	If (Glucose is Medium) & (INS is Medium) & (BMI is Low) & (Age is Young) & (NP is Average) Then (DM is Low)
10	If (Glucose is Medium) & (INS is Medium) & (BMI is Medium) & (Age is Middle) & (NP is Absent) Then (DM is Low)
11	If (Glucose is Medium) & (INS is Medium) & (BMI is Medium) & (Age is Middle) & (NP is Average) Then (DM is Low)
12	If (Glucose is Medium) & (INS is High) & (BMI is Medium) & (TSFT is High) Then (DM is High)
13	If (Glucose is Medium) & (INS is High) & (BMI is High) & (TSFT is High) Then (DM is High)
14	If (Glucose is Medium)& (BMI is Medium) & (TSFT is High) & (NP is High) Then (DM is High)
15	If (Glucose is Medium)&(BMI is Medium)&(TSFT is High)&(NP is Average) Then (DM is High)
16	If (Glucose is Medium) & (Age is Very Old) & (BP is High) Then (DM is High)

17	If (Glucose is Medium) & (Age is Very Old) & (BP is Very High) Then (DM is High)
20	If (Glucose is Normal) & (BMI is Normal) Then (DM is Low)
21	If (Glucose is Normal) & (INS is Low) Then (DM is Low)
22	If (Glucose is Normal) & (BMI is Medium) & (Age is Young) Then (DM is Low)
23	If (Glucose is Normal)&(INS is High)&(BMI is Medium)&(TSFT is High) Then (DM is Medium)
24	If (Glucose is Normal) & (INS is High) & (BMI is High) & (TSFT is High) Then (DM is Medium)
25	If (Glucose is Normal)&(INS is Medium)&(BMI is Medium)&(TSFT is High)Then(DM is Medium)
26	If (Glucose is Normal) &(INS is High) & (BMI is Medium) & (TSFT is High) Then (DM is Medium)
27	If (Glucose is Normal)&(INS is Medium)&(BMI is Medium)&(TSFT is High)Then(DM is Medium)
28	If (Glucose is Normal) & (BMI is Medium) & (Age is Young) Then (DM is Low)

After generating the fuzzy rules, these rules had to be evaluated. As such, the fuzzified inputs were taken, and applied to the rule antecedents. Following this, the result of the antecedent evaluation (the truth value) was applied to the rule consequent membership function by using the clipping method or the scaling method. Clipping is the most commonly used method since it can be calculated without difficulty and generates an aggregated output surface that can be defuzzified easily. With this method, the consequent membership function is cut to the truth level of antecedent. This means that this method slices the top of the membership function. In terms of the scaling method, the consequent membership function is adjusted by multiplying all its membership functions by the antecedent truth value. The clipping method was used in this study. Figure 3.2 presents an example of the rule evaluation process applied to the fuzzy expert system using information from the same patient mentioned at the beginning of this section.

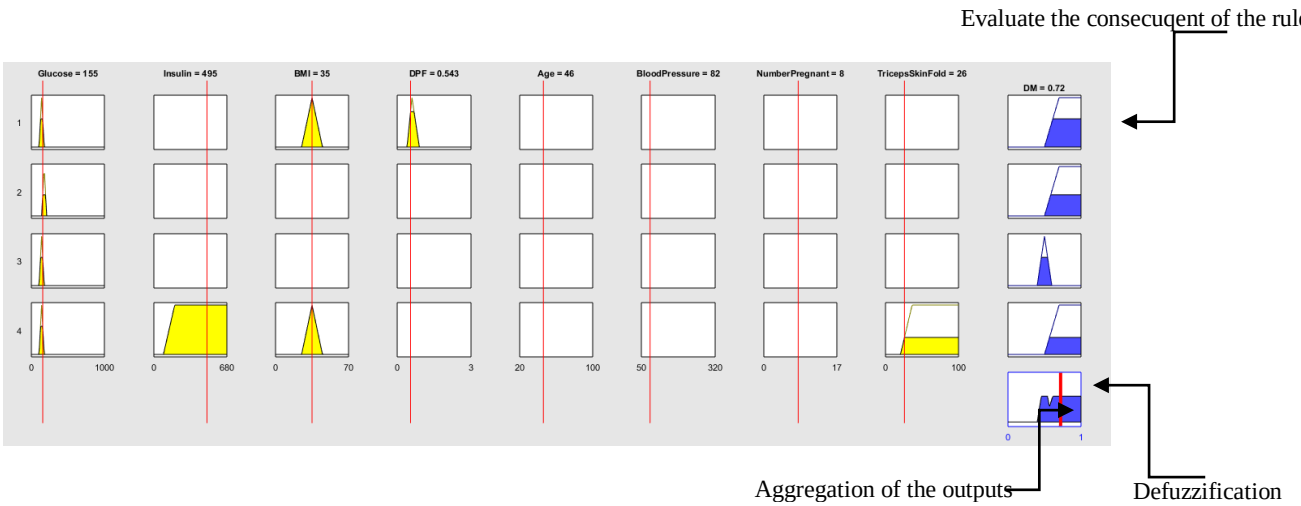

Figure 3.3: Example of the Evaluation Process of Fuzzy Expert System

In Figure 3.2, each row represents a rule and each column represents an attribute. The final result column (DM) shows the fired rules for each input. The fired rules are selected for a particular input from a set of rules. The following is the list of the fired rules for the same patient analysed in the previous section:

1. If (Glucose is Medium(0.57)) &(BMI is Medium (1)) &(DPF is High (0.715))then (DM is High (0.57))
2. If (Glucose is High (0.43)) then (DM is High (0.43))
3. If (Glucose is Medium (0.57)) then (DM is Medium (0.57))
4. If (Glucose is Medium (0.57)) & (INS is High (1)) & (BMI is Medium (1)) & (TSFT is High (0.4)) then (DM is High (0.4))

In this system, some of the fuzzy rules have several antecedents; the fuzzy operator (AND) was used to obtain a single number that represents the result of the antecedent evaluation. This number was then applied to the rule consequent. This fuzzy expert system used the classical fuzzy union operation shown in Section 2.4.1 to evaluate the junction of the rule antecedents. For instance, the AND operator was used for the first rule in the above list to obtain the final result, which was 0.57. In order to evaluate the first rule, the membership functions in Table 3.2

were applied to the rule antecedents. Following this, the result of the antecedent evaluation (0.57) was applied to the rule consequent. All of the rules were evaluated in this manner.

3.2.3 Aggregation of Rules

After defining and evaluating the rules, the clipped consequent membership functions were aggregated to obtain a single fuzzy set output. Max operation was used to aggregate the outputs of the fuzzy expert system. For example, we applied Max operation to the example of the patient discussed in Section 3.2.1. We took the maximum of the medium fuzzy sets (0.57) and the maximum of the high fuzzy sets (0.57). Figure 3.3 illustrates the single fuzzy set of the fuzzy expert system, which is a combination of these two fuzzy sets.

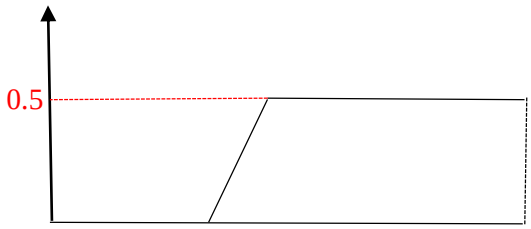

Figure 3.4: The Result of the Aggregation Step for the Fuzzy Expert System

3.2.4 Defuzzification

This step is the last step in the Mamdani fuzzy inference system. It is used to obtain a single crisp number from the single aggregated fuzzy set. The centroid defuzzification method was used in this study. The equation of this method is provided in Section 2.6. The centre point of the aggregated fuzzy set(s) is found by using the equation of the defuzzification method. The final crisp number of the fuzzy expert system was 0.72, thus indicating that the patient in question has diabetes.

Since our fuzzy expert system is built based on human knowledge (i.e. domain experts), changing the membership functions and the fuzzy rules will change the performance of the fuzzy system.

CHAPTER 4

RESULTS

The previous chapter described how the fuzzy expert system from the Mamdani fuzzy inference system was developed using the Matlab software. In this chapter, several common data mining algorithms, namely J48, multilayer perceptron (MLP), logistic regression, support vector machine (SVM), and Naïve Bayes are implemented using the Weka software. In addition, these algorithms and our proposed fuzzy expert system are applied to the Pima Indian Diabetes Dataset (PIDD) in order to measure the performance of our system and compare it with well-known machine learning and statistical algorithms, Moreover, a comparative analysis is done between our fuzzy expert system and a fuzzy expert system proposed in related work to evaluate the performance of our proposed system. Some of the instances of the PIDD are used to test our system and the related work.

4.1 Dataset

In order to compare and validate the findings, the system is tested on the most commonly used Pima Indian diabetes dataset [53], which belongs to the National Institute of Diabetes and Digestive and Kidney Diseases. It is part of the UCI machine learning dataset available to researchers. This dataset contains 768 instances and 9 attributes. The input attributes are age, glucose concentration in blood 2 hours after having breakfast (Glucose), serum insulin in blood 2 hours after having breakfast (Insulin), body mass index (BMI), number of pregnancies (NP), triceps skin fold thickness (TSFT), diabetes pedigree function (DPF), and diastolic blood pressure (BP). The output of the system is either 0 or 1. 0 is interpreted as "no diabetes mellitus" and 1 is interpreted as "diabetes mellitus".

Explanation of the indicators of Type 2 Diabetes Mellitus:

1. **Age:** Age is an indicator for diabetes. Age has four fuzzy sets young, middle age, old, and very old.
2. **Glucose**: Glucose is the main source of energy found in the blood [1]. Glucose has five fuzzy sets low, Normal, medium, high, and very high.
3. **Insulin**: Insulin is the hormone excreted by the pancreas to help mobilize glucose from the blood into the cells to be used for energy. If the cells do not respond well to insulin then glucose is not able to enter the cells [1]. As a result, the cells fail to get the fuel they need, and glucose increment in the blood stream. Insulin has three fuzzy sets low, medium and high.
4. **Body Mass Index (BMI)**: BMI is considered as an assessment of evaluating the weight of the body in relation to the height of a person. This attribute contains of three fuzzy sets in the developed system are Normal, medium and high.
5. **Diabetes Pedigree Function (DPF)**: DPF is the statistical classification of individuals with diabetes in the family pedigree. DPF includes four fuzzy sets low, medium, high, and very high.
6. **Number of pregnancy**: It is categorized as absent, medium, and high. If a person is male then number of pregnancy will be absent.
7. **Triceps skin fold thickness**: Is a value used to measure body fat.
8. **Diastolic blood pressure**: Blood pressure is another contributing factor. It increases the heart's workload, and causes heart thickness and stiffness. This thickness and stiffness decrease normal functionality of the heart. There are three types of blood pressure which are systolic, diastolic, and average. Abnormal systolic blood pressure is commonly

associated with diabetes disease and heart disease [54]. In this thesis, systolic blood pressure is used. This attribute has four fuzzy sets low, medium, high and very high.

4.2 Data Pre-processing

This step is one of the most important phases in the data mining process. It prepares and transforms the initial dataset. Raw data is generally incomplete, inconsistent, and noisy. Analysing data that has such problems can produce misleading results. Thus, some data pre-processing methods can be applied to raw data before running an analysis. Data pre-processing methods involve replacing missing values, normalisation, data discretisation, data transformation, data integration, feature extraction, etc. In this study, two comparative studies were conducted using the Pima Indian Diabetes Dataset, which has missing values for some of the attributes. In the first comparative study, the fuzzy expert system was compared with classification models and regression model. In this study, all attributes and instances of the original dataset were used. In the second study, the fuzzy expert system was compared with the fuzzy expert system presented in [8]. In this study, 192 cases from the lower age range (25 to 30 years old) were extracted from the original dataset (PIDD). Figure 4.1 and Figure 4.2 give summaries of the missing values in the original dataset and the dataset including 192 instances.

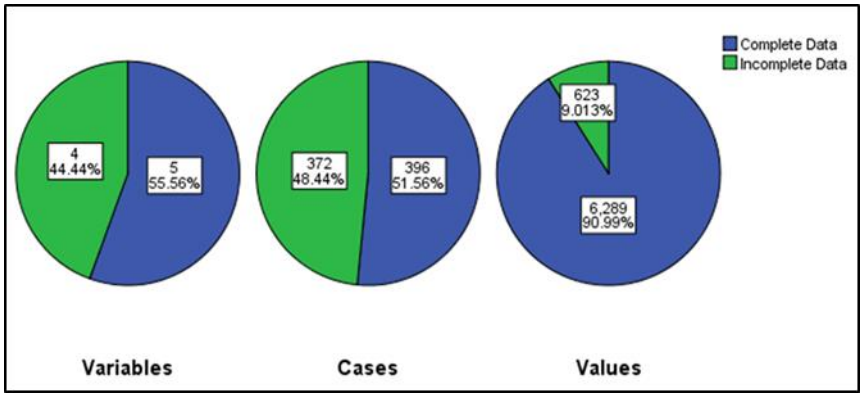

Figure 4.1: Summary of Missing Values in the Original Dataset

In the above figure, the first pie chart displays the number and the percentage of missing variables. It demonstrates that four of the nine variables (attributes), namely insulin, body mass index, triceps skin fold thickness, and diastolic blood pressure, have missing values. The second pie chart shows the number of cases (instances) that are missing some values. The number of cases that have at least one missing value is 372, while 396 cases are complete. The last pie chart illustrates that 9% of all values are missing, whereas approximately 91% of the values are present.

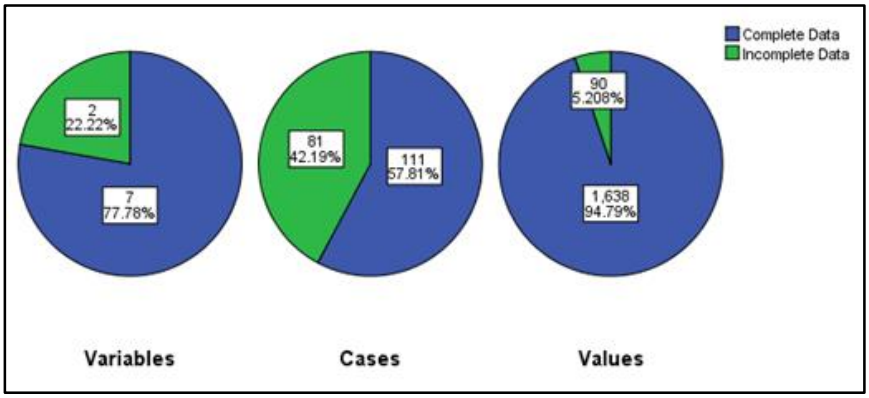

Figure 4.2: Summary of Missing Values in the Dataset including 192 instances

As we can see in the above figure, the first pie chart shows that two of the nine variables, which are insulin and body mass index, have missing values. The second pie chart displays that the number of cases that have at least one missing value is 81, while 111 cases are complete. The third pie chart presents the percentage of the missing values in the dataset which is 5%, while around 95% of the values are present.

In this thesis, two different methods are used to handle the missing values in the original dataset and the dataset that includes 192 instances which are:

- **Multiple Imputation Method**

It is important to select a method to that is capable of replacing these missing values with plausible values. In this study, the multiple imputation technique [55] was selected based on the percentage and pattern of the missing values. Multiple imputation is an approach that replaces each deficient or missing value with more than one acceptable values representing a distribution of possibilities. It looks at the pattern of the available data, and based on probability judgment, attempts to find the best matches, replacing the missing values with imputed values. Replacement is performed repeatedly in order to find the perfect fit. IBM SPSS Statistics version 22 was used to perform the multiple imputation process. The missing values of the original dataset were replaced using the multiple imputation method, with the exception of four records. These four records were deleted from the dataset because of a lack of sufficient data. Also, all missing values of the dataset that includes 192 instances were replaced using the multiple imputation method. Table 4.1 and 4.2 give summaries of cases in the original dataset and the dataset including 192 instances after the application of the multiple imputation method. These datasets are named dataset 1 and dataset 2.

Table 4.1: Summary of the Cases in the Original Dataset after the application of the Multiple Imputation Method (Dataset1)

Class	Number of cases in each class	Total number of cases
Diabetic	269	764
Non-diabetic	495	

Table 4.2: Summary of the Cases in the Dataset including 192 Cases after the application of the Multiple Imputation Method (Dataset 2)

Class	Number of cases in each class	Total number of cases
Diabetic	56	192
Non-diabetic	136	

- **Listwise Deletion**

In this method, an entire case (instance) is excluded from the dataset if any single value is missing [56]. Based on the analysis of the original dataset that is represented in Figure 4.1, 372 cases have at least one missing value. These cases were deleted using IBM SPSS Statistics version 22. Also, the dataset that contains 192 has 81 incomplete cases, so these cases were removed from the dataset. Summaries of the instances in both datasets after the application of the listwise method are provided in Table 4.3 and Table 4.4. The datasets are named dataset 3 and dataset 4.

Table 4.3: Summary of the Cases in the Original Dataset after the application of the Listwise Method (Dataset3)

Class	Number of cases in each class	Total number of cases
Diabetic	130	396
Non-diabetic	266	

Table 4.4: Summary of the Cases in the Dataset including 192 Cases after the application of the Listwise Method (Dataset4)

Class	Number of cases in each class	Total number of cases
Diabetic	32	111
Non-diabetic	79	

4.3 Implementing the Data mining Algorithms

Weka was used to build J48, MLP, logistic regression, SVM, and Naïve Bayes. There are two test methods which are percentage split and cross validation. Both methods were used to test each model. A simple way to use one dataset for both training and estimation the performance of an algorithm on unseen data is to split the dataset. This method splits the dataset into a training dataset and a test dataset. In this study, a supervised (resample) filter [57] was applied to the instances using Weka. This filter produces a random subsample of a dataset using sampling with replacement. It also maintains the class distribution in the subsample. In order to use this filter, the dataset must have a nominal class attribute. The number of instances in the generated training and testing datasets can be specified. As such, we selected 70% of the cases for training and used the remaining 30% as the testing dataset. After dividing the dataset into training and testing, the algorithm was run on the training dataset and a model was implemented and assessed on the testing dataset, following which a classification accuracy was obtained.

In k-fold cross validation, k is the number of splits to make in the dataset. k=10 is selected in this study. This splits the dataset into 10 parts, and the algorithm is run 10 times. Every time the algorithm is run, it is trained on 90% of the data and tested on 10% of the data; for each run of the algorithm, a change is made in terms of which 10% of the data the algorithm is tested on.

This means that each data instance is used as a training instance exactly nine times and as a test instance one time. The accuracy is not a standard deviation and mean, but instead an exact accuracy score of how many correct predictions are made.

4.4 Comparing the Fuzzy Expert System with the Data Mining algorithms

Several performance metrics were used to evaluate the performance of our fuzzy expert system and the other data mining models for the incidence of diabetes, which are confusion matrix, accuracy, specificity, sensitivity, precision, and F-Measure. The first evaluation metric calculated was the accuracy, which is the fraction of true results (both true positives and true negatives) among the total number of cases examined. Following this, specificity, sensitivity, and precision were calculated. When it comes to medical diagnosis, "specificity (also called true negative rate) refers to the test's ability to correctly detect patients who do not have diabetes" [58], whereas "sensitivity (also called recall, or true positive rate) relates to the test's ability to correctly detect patients who do have diabetes" [58]. In other words, sensitivity is the proportion of correct positive classifications (TP) from cases that are actually positive. On the other hand, precision is the proportion of correct positive classifications (TP) from cases that are predicted to be positive. Finally, the F-Measure (also called F-Score) was computed. This metric gives the harmonic mean of precision and sensitivity. It is important to note that the model with highest accuracy, specificity, sensitivity, precision, and F-measure is the best predictive model [58].The equations of the performance metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (4.1)$$

$$Specificity = \frac{TN}{FP + TN} \times 100 \quad (4.2)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (4.3)$$

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (4.4)$$

$$F - measure = \frac{Precision \times Recall}{Precision + Recall} \times 100 \quad (4.5)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives and false negatives, respectively.

Suppose there is a study evaluating a new test that screens people for a diabetes mellitus disease. A person taking the test either has diabetes mellitus or does not have diabetes mellitus. The result of the test can be positive (classifying the person as having diabetes mellitus) or negative (classifying the person as not having diabetes mellitus). The result of the test for each person may or may not match the person's actual status. In that setting:

- True positive: Diabetic people correctly identified as diabetic
- False positive: Non-diabetic people incorrectly identified as diabetic
- True negative: Non-diabetic people correctly identified as non-diabetic
- False negative: Diabetic people incorrectly identified as non-diabetic

In this section, several experiments were done to compare the fuzzy expert system with the data mining algorithms. These experiments are presented as the following:

- **Experiment 1**

  In the first experiment, we applied MLP, logistic regression, SVM, and Naïve Bayes to the pre-processed dataset that we applied the multiple imputation method to (dataset 1)

and used 10-fold cross validation. Also, we applied the fuzzy expert system to dataset 1. The results of this experiment are presented in Table 4.5, Table 4.6, Table 4.7, Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, and Figure 4.7.

Table 4.5: Confusion Matrix of the Classifiers Using 10-Fold Cross Validation

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	241	28
	Non-diabetic	29	466
J48	Diabetic	161	108
	Non-diabetic	82	413
MLP	Diabetic	176	93
	Non-diabetic	78	417
Logistic regression	Diabetic	154	115
	Non-diabetic	63	432
SVM	Diabetic	143	126
	Non-diabetic	53	442
Naïve Bayes	Diabetic	163	106
	Non-diabetic	90	405

Table 4.6: Prediction Accuracy of the Classifiers Using 10-Fold Cross Validation

Classifier	Accuracy
Fuzzy Expert System	92.5%
J48	75%
MLP	77.6%
Logistic Regression	76.7%
SVM	76.5%
Naïve Bayes	74%

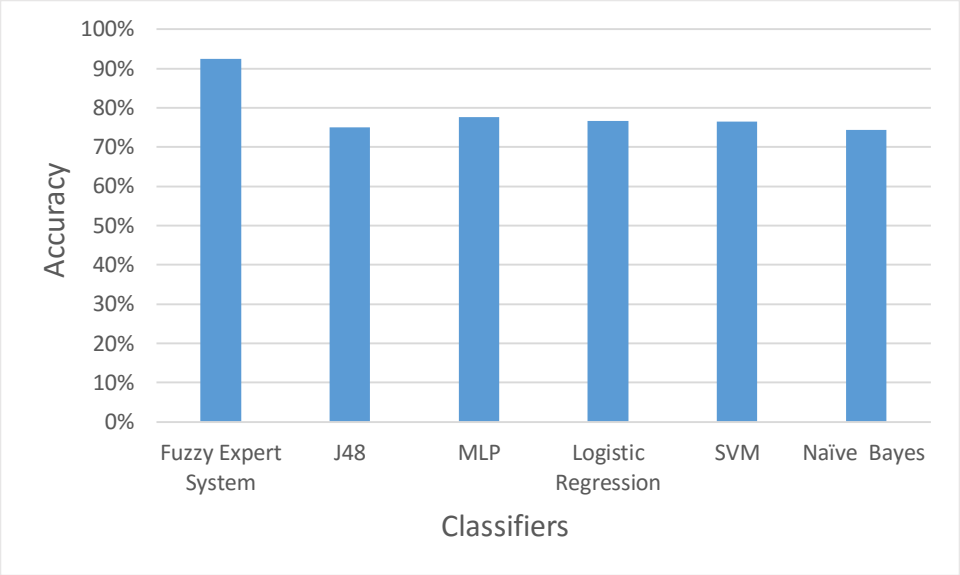

Figure 4.3: Accuracy of Each Classifier Using 10 Cross Validation

As we can clearly see, the fuzzy expert system has the highest accuracy with the lowest number of false positives and negatives. This figure also illustrates that logistic regression and SVM classifiers perform at nearly the same rate of accuracy. Lastly, Naïve Bayes has the lowest prediction accuracy with the highest sum of false positives and negatives cases.

Table 4.7: Results of the Classifiers (10-fold cross validation)

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	94%	89.6%	89%	89.5%
J48	83%	59.8%	66%	62.8%
MLP	84%	65%	69%	67%
Logistic Regression	87%	57%	71%	63%
SVM	89%	53%	72%	61%
Naïve Bayes	81.8%	60.6%	64%	62%

Figure 4.4: Specificity of Each Classifier Using 10-Fold Cross Validation

As we can see in the above figure, the fuzzy expert system has the highest number of true negative cases followed by SVM. On the other hand, Naïve Bayes has the lowest true negative rate.

Figure 4.5: Sensitivity of Each Classifier Using 10-Fold Cross Validation

Figure 4.5 shows that the fuzzy expert system has the highest number of true positive cases. MLP has a higher true positive rate compared to logistic regression, SVM, J48, and Naïve

Bayes. However, SVM has the highest number of false negative cases and the lowest number of true positive cases.

Figure 4.6: Precision of Each Classifier Using 10-Fold Cross Validation

The bar graph above illustrates that the fuzzy expert system has the highest precision value i.e. the lowest number of false positive errors committed by this classifier. By comparison, Naïve Bayes has the lowest precision value, with the large number of false positive cases compared to the other classifiers.

Figure 4.7: F-measure of Each Classifier Using 10-Fold Cross Validation

As we can clearly see, that the fuzzy expert system has the highest F-measure value, which ensures that both precision and recall are reasonably high. Also, we can see that J48 and Naïve Bayes have the same F-measure values. The F-measure value for SVM is slightly lower than these.

- **Experiment 2**

  After we pre-processed the dataset using the multiple imputation method (dataset 1), we split it into training dataset and testing dataset. Then, we applied the fuzzy expert system, MLP, logistic regression, SVM, and Naïve Bayes to the training and testing datasets. The results of this experiment are shown in Table 4.8, Table 4.9, Table 4.10, Table 4.12, Table 4.13, Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11, and Figure 4.12.

Table 4.8: Confusion Matrix for Each Classifier of Training Dataset

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	166	23
	Non-diabetic	18	328
J48	Diabetic	140	49
	Non-diabetic	24	322
MLP	Diabetic	129	60
	Non-diabetic	32	314
SVM	Diabetic	102	87
	Non-diabetic	34	312
Logistic Regression	Diabetic	107	82
	Non-diabetic	38	308
Naïve Bayes	Diabetic	114	75
	Non-diabetic	53	293

Table 4.9: Prediction Accuracy for Each Classifier of Training Dataset

Classifier	Accuracy
Fuzzy Expert System	92%
J48	86%
MLP	83%
Logistic Regression	78%
SVM	77%
Naïve Bayes	76%

Table 4.10: Confusion Matrix for Each Classifier of Testing Dataset

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	75	5
	Non-diabetic	11	138
J48	Diabetic	44	37
	Non-diabetic	26	123
MLP	Diabetic	47	33
	Non-diabetic	19	130
SVM	Diabetic	42	38
	Non-diabetic	17	132
Logistic Regression	Diabetic	49	31
	Non-diabetic	23	126
Naïve Bayes	Diabetic	51	29
	Non-diabetic	21	128

Table 4.11: Prediction Accuracy for Each Classifier of Testing Dataset

Classifier	Accuracy
Fuzzy Expert System	93%
J48	73%
MLP	77%
Logistic Regression	76%
SVM	76%
Naïve Bayes	78%

Figure 4.8: Accuracy of Each Classifier based on the Testing Dataset

As we can see, the fuzzy expert system is the best classifier among its rival classifiers. The above figure shows that the performance of logistic regression and SVM classifiers is almost identical, with an equal sum of true positives and negatives. However, the accuracy of MLP is slightly higher than these. Finally, the least accurate classifier is J48.

Table 4.12: Results of the Classifiers using Training Dataset

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	95%	88%	90%	89%
J48	93%	74%	85%	79%
MLP	91%	68%	80%	73.7%
Logistic Regression	89%	57%	71%	64%
SVM	90%	54%	75%	63%
Naïve Bayes	85%	60%	68%	67%

Table 4.13: Results of the Classifiers Using Testing Dataset

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	93%	94%	87.2%	91%
J48	83%	54%	62%	58%
MLP	87%	59%	71%	64%
Logistic Regression	85%	61%	68%	65%
SVM	89%	53%	71%	60%
Naïve Bayes	86%	64%	71%	67%

Figure 4.9: Specificity of Each Classifier based on Testing Dataset

Figure 4.9 shows that the fuzzy expert system has the highest true negative rate. MLP has almost the same number of true negative cases as SVM. However, J48 has the highest number of false positives relative to the other classifiers.

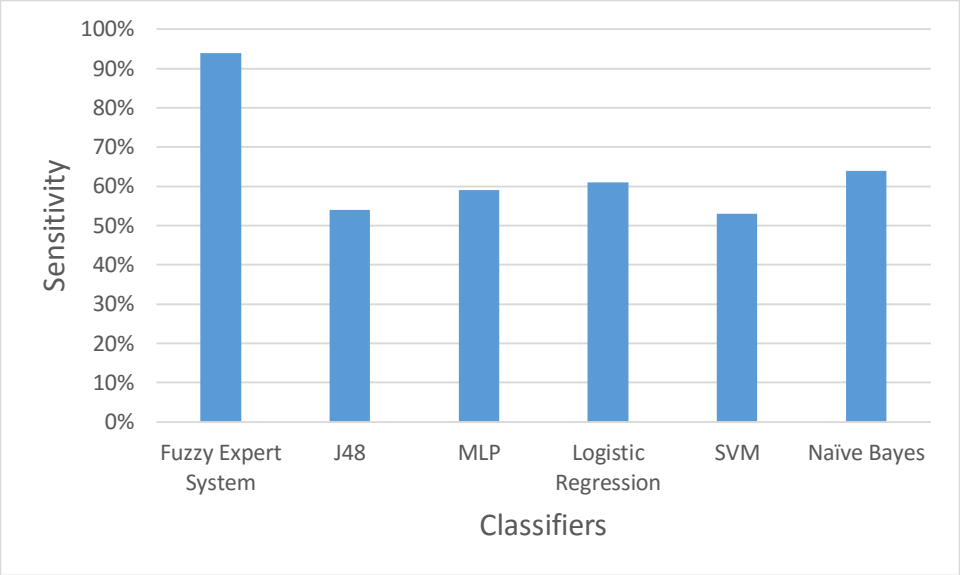

Figure 4.10: Sensitivity of Each Classifier based on the Testing Dataset

The above figure illustrates that the fuzzy expert system has the highest true positive rate. On the other hand, SVM performs very poorly with a very low true positive rate, i.e. very high number of positive cases misclassified as (Non-diabetic) negative.

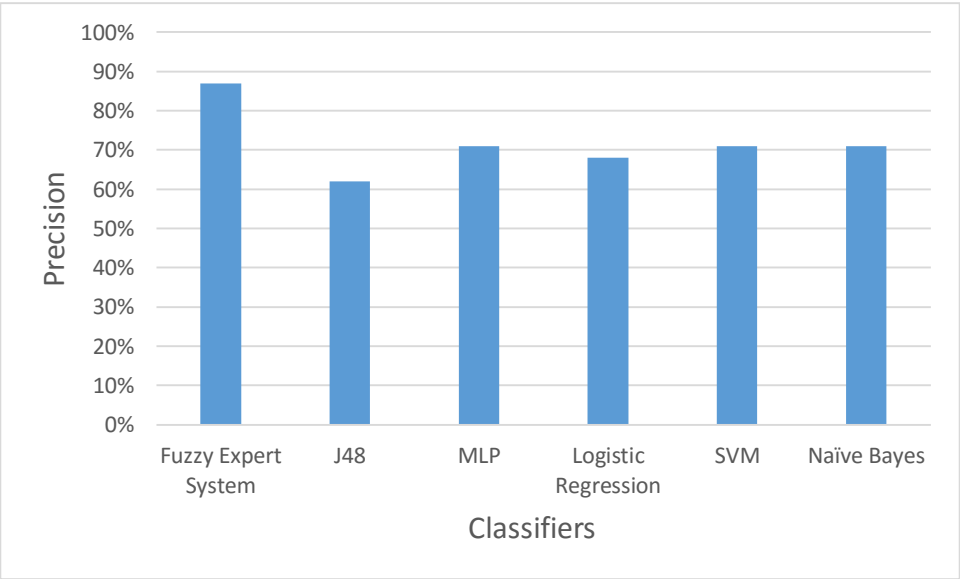

Figure 4.11: Precision of Each Classifier based on the Testing Dataset

It is clear from the bar graph above that the fuzzy expert system has the highest precision value However, J48 is the least precise classifier. It has the highest number of false positives relative to the other classifiers mentioned above in the testing part.

Figure 4.12: F-measure of Each Classifier based on the Testing Dataset

The results show that the fuzzy expert system has the highest value for both precision and recall. Thus, it has the highest value for F-measure in both the training and testing datasets. J48 is considered the second best classifier for the training dataset, whereas Naïve Bayes comes in second rank in the testing dataset.

- **Experiment 3**

  In this experiment, we used the dataset that we applied the listwise method to (dataset 3). After that, we applied fuzzy logic to the pre-processed dataset. In addition, we applied MLP, logistic regression, SVM, and Naïve Bayes to dataset 3 and used 10-fold cross

validation. Table 4.14, Table 4.15, Table 4.16, Figure 4.13, Figure 4.14, Figure 4.15, Figure 4.16, and Figure 4.17 .

Table 4.14: Confusion Matrix of the Classifiers Using 10-Fold Cross Validation

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	120	10
	Non-diabetic	14	252
J48	Diabetic	94	37
	Non-diabetic	52	214
MLP	Diabetic	86	45
	Non-diabetic	46	220
Logistic regression	Diabetic	75	56
	Non-diabetic	30	236
SVM	Diabetic	72	59
	Non-diabetic	28	238
Naïve Bayes	Diabetic	82	49
	Non-diabetic	48	218

Table 4.15: Prediction Accuracy of the Classifiers Using 10-Fold Cross Validation

Classifier	Accuracy
Fuzzy Expert System	94%
J48	77.5%
MLP	77%
Logistic Regression	78%
SVM	78%
Naïve Bayes	75.5%

Figure 4.13: Accuracy of Each Classifier Using 10 Cross Validation

Figure 4.13 shows that the prediction accuracy of the fuzzy expert system is higher than the other classifiers. Logistic regression and SVM have a very similar prediction accuracy. Also, J48 and MLP perform almost the same. However, Naïve Bayes has the lowest prediction accuracy compared to the other classifiers.

Table 4.16: Results of the Classifiers (10-fold cross validation)

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	94.7%	92%	89.7%	91%
J48	80%	71.8%	64%	68%
MLP	82.6%	65.6%	65%	65%
Logistic Regression	88.7%	57%	71%	63.6%
SVM	89%	55%	72%	62%
Naïve Bayes	82%	62.6%	63%	62.8%

Figure 4.14: Specificity of Each Classifier Using 10-Fold Cross Validation

The above figure shows that the fuzzy expert system has the highest specificity value followed closely by SVM and logistic regression. However, J48 has the lowest specificity value compared to the other classifiers.

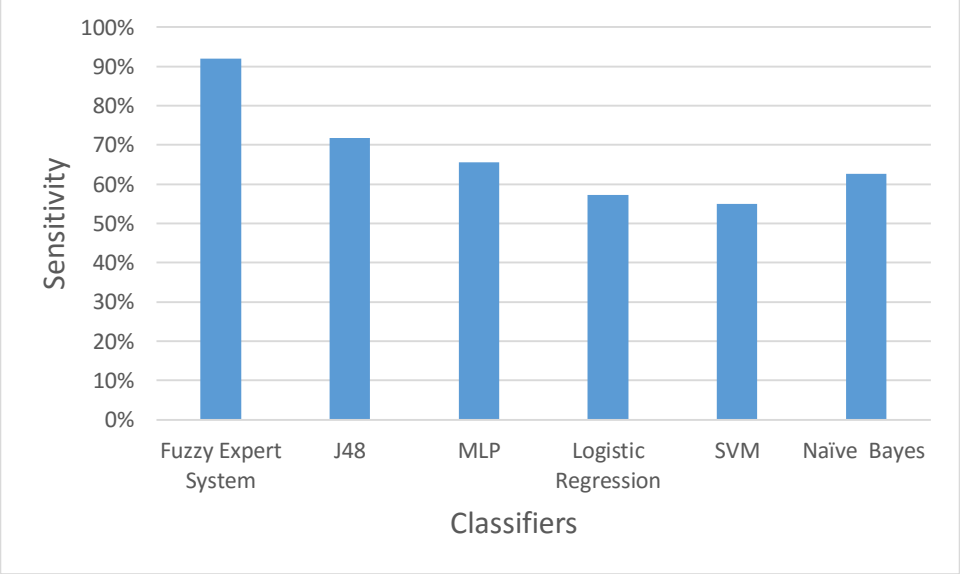

Figure 4.15: Sensitivity of Each Classifier Using 10-Fold Cross Validation

Figure 4.15 illustrates that the fuzzy expert system is the best classifier among its rival classifiers. On the other hand, SVM has the lowest sensitivity value compared to the others.

Figure 4.16: Precision of Each Classifier Using 10-Fold Cross Validation

As we can see, the fuzzy expert system has the best precision value. SVM and logistic regression are considered the second best classifiers. MLP and J48 rank third. However, Naïve Bayes is the least precise classifier.

Figure 4.17: F-measure of Each Classifier Using 10-Fold Cross Validation

The above bar chart reveals that the F-measure of the fuzzy expert system is higher than its rival classifiers. The F-measure values of the other classifiers are similar.

- **Experiment 4**

In this experiment, we divided the dataset that we applied the listwise method to (dataset 3) into a training part and a testing part. Then, we applied fuzzy logic and data mining algorithms to dataset 3. The results are shown in Table 4.17, Table 4.18, Table 4.19, Table 4.20, Table 4.12, Table 4.22, Figure 4.18, Figure 4.19, Figure 4.20, Figure 4.21, and Figure 4.22.

Table 4.17: Confusion Matrix for Each Classifiers of Training Dataset

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	84	6
	Non-diabetic	9	178
J48	Diabetic	70	20
	Non-diabetic	17	170
MLP	Diabetic	77	13
	Non-diabetic	15	172
Logistic regression	Diabetic	56	34
	Non-diabetic	34	153
SVM	Diabetic	44	46
	Non-diabetic	23	164
Naïve Bayes	Diabetic	56	34
	Non-diabetic	34	153

Table 4.18: Prediction Accuracy for Each Classifier of Training Dataset

Classifier	Accuracy
Fuzzy Expert System	94%
J48	86.6%
MLP	89.8%
Logistic Regression	73%
SVM	75%
Naïve Bayes	75.5%

Table 4.19: Confusion Matrix for Each Classifier of Testing Dataset

Classifier	Desired Results	Prediction	
		Diabetic	Non-diabetic
Fuzzy Expert System	Diabetic	36	4
	Non-diabetic	5	74
MLP	Diabetic	31	9
	Non-diabetic	11	68
Logistic regression	Diabetic	17	23
	Non-diabetic	11	68
SVM	Diabetic	19	21
	Non-diabetic	9	70
J48	Diabetic	29	11
	Non-diabetic	6	73
Naïve Bayes	Diabetic	23	17
	Non-diabetic	14	65

Table 4.20: Prediction Accuracy for Each Classifier of Testing Dataset

Classifier	Accuracy
Fuzzy Expert System	92%
J48	85.6%
MLP	83%
Logistic Regression	71.4%
SVM	74.8%
Naïve Bayes	74%

Figure 4.18: Accuracy of Each Classifier based on the Testing Dataset

The above figure shows that the fuzzy expert system has the highest accuracy followed by MLP and J48, respectively. SVM and Naïve Bayes have similar prediction accuracy percentages. Logistic regression has a slightly lower accuracy than SVM and Naïve Bayes.

Table 4.21: Results of the Classifiers using Training Dataset

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	95%	93%	90%	91.5%
J48	91%	77.8%	80.5%	79%
MLP	92%	85.6%	83.7%	84.6%
Logistic Regression	81.8%	47.8%	60.6%	53.4%
SVM	87.7%	49%	65.7%	63%
Naïve Bayes	81.8%	62%	62%	62%

Table 4.22: Results of the Classifiers using Testing Dataset

Classifier	specificity	sensitivity	precision	F-measure
Fuzzy Expert System	93.7%	90%	87%	88.5%
J48	92%	72.5%	83%	77.4%
MLP	86%	77.5%	77.8%	75.6%
Logistic Regression	86%	42.5%	71.4%	64%
SVM	88.6%	47.5%	68%	56%
Naïve Bayes	82.3%	57.5%	62%	59.7%

Figure 4.19: Specificity of Each Classifier based on the Testing Datasets

Figure 4.19 illustrates that the fuzzy expert system and J48 have very similar true negative rates. However, Naïve Bayes has the highest number of false positives relative to the other classifiers.

Figure 4.20: Sensitivity of Each Classifier based on the Testing Dataset

It is clear from the figure above that the fuzzy expert system has the highest true positive rate followed by MLP and J48, respectively. On the other hand, logistic regression has the lowest true positive rate in both the training and testing datasets.

Figure 4.21: Precision of Each Classifier based on the Testing Dataset

As we can see, the fuzzy expert system has the highest precision value followed by MLP and J48. On the other hand, Naïve Bayes has the lowest precision value. This means that it has the highest number of false positives relative to the other classifiers mentioned above.

Figure 4.22: F-measure of Each Classifier based on the Testing Datasets

As we can clearly see, that the fuzzy expert system has the highest F-measure value followed by MLP and J48. However, SVM has the lowest F-measure value compared to the other classifiers.

The results of the four experiments are listed below:

Using 10 cross validation:

- In general, the results of the classifiers using dataset 1 are similar to the results of the classifiers using dataset 3.
- The accuracy of all the classifiers using dataset 3 is slightly higher than the accuracy of the all the classifiers using dataset 1.
- Using dataset 1 and dataset 3, the specificity, sensitivity, precision, and F-measure of all the classifiers are similar.

Using the percentage split:

- The evaluation metrics of the fuzzy expert system using dataset 3 are slightly higher than the evaluation metrics of the fuzzy expert system using dataset 1.
- Using dataset 1 and dataset 3, the specificity, sensitivity, precision, and F-measure of the fuzzy expert system are very similar.
- The accuracy and sensitivity of MLP and J48 using dataset 3 is higher than the accuracy and sensitivity of MLP and J48 using dataset 1.

- The accuracy and sensitivity of logistic regression, SVM, and Naïve Bayes using dataset 3 is slightly lower than the accuracy and sensitivity of logistic regression, SVM, and Naïve Bayes using dataset 1.

4.5 Implementing and Comparing the Fuzzy Expert System with Related Work

In this section, our aim is to evaluate the performance of the proposed fuzzy expert system by comparing it with related work that used the Mamdani fuzzy inference system to diagnose the incidence of type 2 diabetes [8]. For this purpose, we implemented the system that is proposed by Kalpana and Kumar [8] in Matlab. Their system diagnosed patients of a very young age. The specified age range is very limited, spanning from 25 to 30 years old. On the other hand, our system diagnosed a more inclusive range of ages, from 18 to 100 years-of-age. Kalpana and Kumar applied their system to the PIDD and we used the same dataset. However, they used six of the nine attributes of the original dataset, while we used all the attributes of the dataset. The attributes that they did not use, namely number of pregnancies, triceps skin fold thickness, and diastolic blood pressure, are very important. Since all 768 patients in the dataset were females, the attribute related to the number of pregnancies should be considered to diagnose diabetes. Obesity is a known risk factor for diabetes disease, coronary heart disease, and many other diseases. Body mass index (BMI) and skin fold thickness have been commonly used as indexes of obesity in epidemiological research [59]. Because of this, triceps skin fold thickness should be considered. High blood pressure is more common in diabetic people, especially in the elderly. Around 8 out of 10 patients with type 2 diabetes develop high blood pressure at some stage [54]. Since diabetic people are more at risk of developing high blood pressure, diastolic blood pressure should be taken into consideration.

In this section, we conducted two experiments using cases from the lower age range (from 25 to 30 years old). Kalpana and Kumar did not apply any data pre-process method to the dataset before using it. Since the dataset contains missing values, it is important to pre-process the data by replacing the missing values or removing the instances that contain missing values. Both methods are used to deal with missing values. As a result, we obtained two datasets (named dataset 2 and dataset 4). In the first experiment, we used the multiple imputation method to replace the missing values in the dataset. After that, we applied our system and the proposed system presented in [8] to the pre-processed dataset (named dataset 2). In the second experiment, we applied the listwise deletion approach to the dataset. Then, we applied both fuzzy expert systems to dataset 4. The results of these experiments are presented in Table 4.23, Table 4.24, Table 4.25, Table 4.26, Figure 4.23, and Figure 4.24.

After applying both systems on the datasets that consist of 192 cases from the lower age range, we used five performance metrics, which are confusion matrix, accuracy, specificity, sensitivity, precision, and F-measure, to evaluate the systems. The equations of these metrics are provided in Section 4.3.

Table 4.23: Confusion Matrix for Each Fuzzy Expert System

Classifier	Desired Results	Prediction	
		Diabetic	Non-Diabetic
Fuzzy Expert System	Diabetic	46	10
	Non-Diabetic	8	128
Fuzzy Expert System [8]	Diabetic	22	34
	Non-Diabetic	25	111

Table 4.24: Results of Each Fuzzy Expert System

Classifier	Accuracy	Specificity	Sensitivity	Precision	F-measure
Fuzzy Expert System	90.6%	94%	82%	85%	83.5%
Fuzzy Expert System [8]	69%	81.6%	39%	46.8%	59.5%

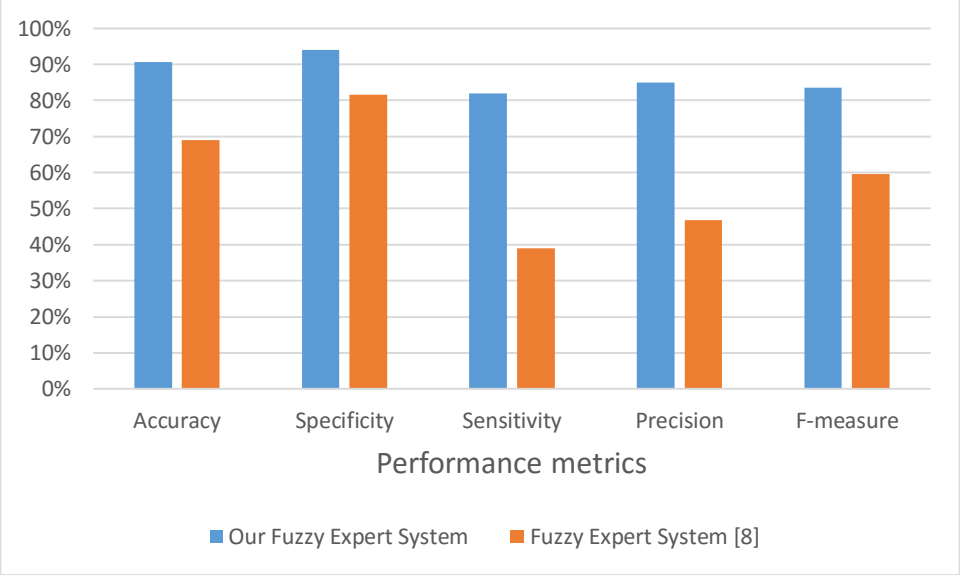

Figure 4.23: The Performance metrics of the Fuzzy Expert Systems Using Dataset 2

Table 4.25: Confusion Matrix for Each Fuzzy Expert System

Classifier	Desired Results	Prediction	
		Diabetic	Non-Diabetic
Fuzzy Expert System	Diabetic	27	5
	Non-Diabetic	4	75
Fuzzy Expert System [8]	Diabetic	14	18
	Non-Diabetic	17	62

Table 4.26: Results of Each Fuzzy Expert System

Classifier	Accuracy	Specificity	Sensitivity	Precision	F-measure
Fuzzy Expert System	92%	95%	84.4%	87%	85.7%
Fuzzy Expert System [8]	68.5%	78.5%	43.8%	45%	44.5%

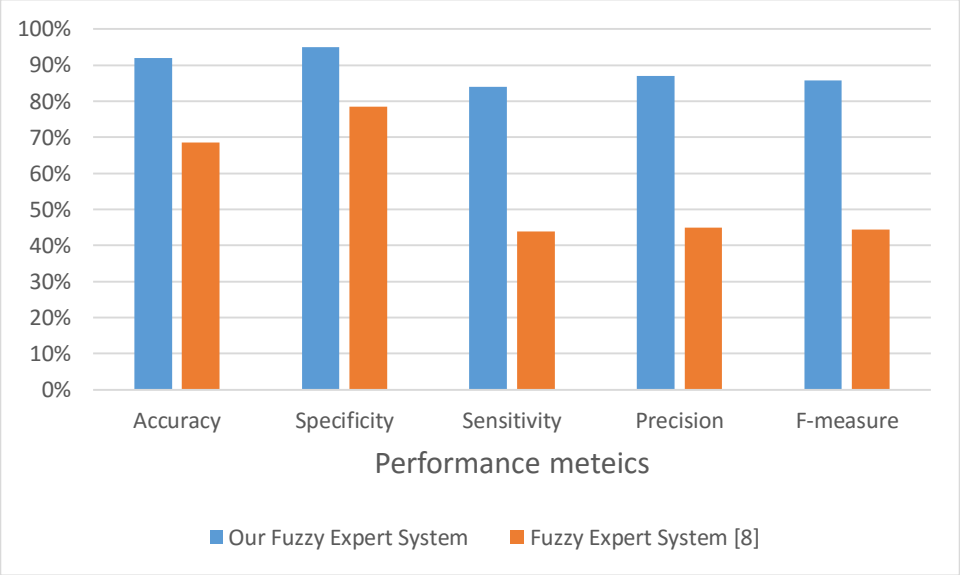

Figure 4.24: The Performance metrics of the Fuzzy Expert Systems Using Dataset 4

Figure 4.23 and Figure 4.24 show that our fuzzy expert system performs more accurately than the fuzzy expert system proposed in [8]. However, the fuzzy expert system proposed in [8] has a very low true positive rate. This means that it has a large number of false negative cases.

In general, the performance of the fuzzy expert system using dataset 4 is slightly higher than the performance of the fuzzy expert system using dataset 2. On the other hand, the performance of the fuzzy expert system [8] using dataset 4 is slightly lower than the performance of the fuzzy expert system [8] using dataset 2.

Table 4.27: Sample of the Dataset with Predicted Results from two Fuzzy Expert Systems

Cases	Glucose	INS	BMI	DPF	Age	BP	NP	TSFT	Actual Results	Our Fuzzy System	Fuzzy System [8]
1	180	78	34	.271	26	64	3	25	1	0.79	0.31
2	170	225	34.5	.356	30	64	3	37	1	0.78	0.32
3	139	160	31.6	.361	25	80	5	35	1	0.58	0.3
4	100	90	32.9	.867	28	66	2	20	1	0.22	0.3
5	151	120	35.5	.692	28	62	6	31	0	0.7	0.4

The above table presents a sample of the dataset with actual results from the dataset and the predicted results by our fuzzy expert system and the other fuzzy expert system presented in [8], which are shown in the last two columns. It is important to notice that our fuzzy expert system used all of the 8 input attributes, while the other fuzzy expert system [8] used only the first five input attributes to predict the outcome. Also, our system assumes that the person has diabetes if the result is more than 0.5. However, the other system assumes that the person is diabetic if the outcome is greater than 0.35. As we can see from Table 4.26, our system diagnoses the first three cases correctly, whereas the other system fails to do so. However, neither system could diagnose the last two cases correctly.

CHAPTER 5

CONCLUSIONAND FUTURE WORK

5.1 Conclusion

In this research, we developed a fuzzy expert system in Matlab to diagnose diabetes mellitus type 2. The proposed methodology included the four steps of the Mamdani fuzzy inference system, namely fuzzification, rule evaluation, aggregation of the outputs, and difuzzification. We also implemented a logistic regression model and various popular machine learning models in Weka for comparison purposes. We validated our proposed system and the other models using real data from UCI machine learning datasets (called Pima Indian Diabetes Dataset). Before we applied them to the dataset, we performed a data pre-processing step. In this step, we used two different methods, multiple imputation and listwise deletion, to handle the missing values. As a result of this step, we got two datasets (named dataset 1 and dataset 3). Following this, we applied the developed fuzzy expert system, logistic regression, and machine learning techniques to the datasets to predict the outcome. Based on these findings, we calculated six performance metrics which are confusion matrix, accuracy, specificity, sensitivity, precision, and F-measure. When we used the dataset 1 and 10 cross validation method, we found that the fuzzy expert system performed the best among of all the classifiers. The fuzzy expert system achieved a prediction accuracy of 92.5%, with a specificity of 94%, a sensitivity of 90%, a precision of 89%, and an F-measure of 90%. The fuzzy expert achieved a prediction accuracy of 94%, with a specificity of 95%, a sensitivity of 92%, a precision of 90%, and an F-measure of 91%, using dataset 3 and 10 cross validation. Also, we found that the performance of the fuzzy expert system is better than its rival classifiers using dataset 1 and percentage split method. The fuzzy expert gave a prediction accuracy of 93%, with a specificity of 95%, a sensitivity of 88%, a precision of 90%, and an F-measure of 89%. In addition, the fuzzy expert performed with a prediction

accuracy of 92%, with a specificity of 94%, a sensitivity of 90%, a precision of 87%, and an F-measure of 89% when we used dataset 3 and percentage split method. In this study, we also implemented a fuzzy expert system presented in related work that used the Mamdani fuzzy inference system. Then, we compared our proposed fuzzy expert system with the other fuzzy expert system presented in the related work to evaluate the performance of our system. We used some of the instances of the Pima Indian Diabetes Dataset since the system presented in the related work can diagnose a certain age group (from 25 to 30 years old). We applied the multiple imputation method and the listwise method to the dataset before using it. We obtained two datasets (named dataset 2 and dataset 4). Finally, we compared the performance of our system with the performance of the system presented in the related work. We found that the performance of our system was better than the performance of the system presented in the related work using both datasets. Our system achieved a prediction accuracy of 92%, with specificity of 95%, sensitivity of 84%, precision of 87%, and F-measure of 86% while the fuzzy system [8] gave an accuracy of 69%, specificity of 79%, sensitivity of 44%, precision of 45%, and F-measure of 45%, using dataset 4.

In conclusion, the proposed fuzzy system was successfully implemented to diagnose and predict the incidence of type 2 diabetes, to overcome many issues existing in related works, to help specialists, and to reduce human error when diagnosing this disease.

5.2 Future Work

With regard to a future research direction, it would be very interesting to develop an interface for the fuzzy expert system in order to enhance its usability. In this study, the Mamdani fuzzy inference system was used to develop the fuzzy expert system. Another possible research avenue could be the use of the Sugeno fuzzy inference system, which is the other type of fuzzy inference

system. Also, another possible research direction could be the use of the adaptive neuro fuzzy inference system (ANFIS) to build a system which diagnoses the incidence of type 2 diabetes. In terms of membership functions, this study used two types of membership function, namely the triangular membership function and the trapezoidal membership function. Future research could study other types of membership functions, a number of which were discussed in Chapter 2. While designing the fuzzy expert system, this study defuzzified the result of the aggregation step by using the centroid defuzzification method. It is certainly worth conducting future research by using the other types of defuzzification approaches provided in Chapter 2, Section (2.8). Indeed, this would make it possible to observe the differences between the results when using different defuzzification methods. Also, the Pima Indian Diabetes dataset was used in this study; adding another dataset to this dataset that we used would make the system more accurate.

APPENDIX A

PROPERTIES OF FUZZY SETS

A.1 Properties of Fuzzy Sets

There is a similarity between the properties of fuzzy sets and the properties of classical sets. Classical sets represent a special type of fuzzy set, in which membership values are a subset of the interval [0, 1]. The common rules of the classical set theory are applied to fuzzy set theory as follows [13, 17, 18, 19]:

Here, A, B, C are three random fuzzy sets.

Commutatively:

$$A \cup B = B \cup A$$

$$B \cap A = A \cap B$$

Associativity:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

Distributivity:

$$A \cup (B \cup C) = (A \cup B) \cap (A \cup C)$$

Idempotency:

$$A \cup A = A$$

$$A \cap A = A$$

The result of union between the same set and intersection between the same set is the same set.

Identity:

$$A \cup \emptyset = A$$

$$A \cup X = A$$

where

∅indicates the null set and X indicates the universal set.

$$A \cap \emptyset = \emptyset$$

$$A \cap X = X$$

Transitivity:

$$A \subseteq B \subseteq C = A \subseteq C$$

Reflexivity of Complementation:

$$(A^c)^c = A$$

De Morgan's Laws:

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

APPENDIX B

DETAILS ABOUT THE IMPLEMENTION OF DATA MINING ALGORITHMS

B.1 Coefficients of Logistic Regression

When we studied the relation between the independents variables (Inputs) and the dependent variable (Outputs), we found the following results:

Table B.1: Coefficients of Logistic Regression

Attribute	Coefficient
Glucose	0.62
Insulin	0.37
Age	0.33
BMI	0.48
TSFT	0.36
NP	0.28
DPF	0.45
BP	-0.01

B.2 Details about the Experiments

Experiment 1

- **J48**

===Classifier model (full training set)=== J48 pruned tree

Glucose <= 127

|BMI <= 26.4: NoDiabetes(1.0/123.0)

|BMI > 26.4

| |Age <= 28: NoDiabetes(22.0/184.0)

| |Age > 28

```
|   |   |Glucose <= 99: NoDiabetes(10.0/55.0)
|   |   |Glucose > 99
|   |   |   |DPF <= 0.2: NoDiabetes(4.0/21.0)
|   |   |   |DPF > 0.2: Diabetes(42.0/98.0)
Glucose > 127
|   |BMI <= 29.9
|   |   |Glucose <= 145
|   |   |   |Insulin <= 132
|   |   |   |   |Glucose <= 140
|   |   |   |   |   |NP <= 1: NoDiabetes(2.0)
|   |   |   |   |   |NP > 1
|   |   |   |   |   |   |Insulin <= 69: NoDiabetes(2.0)
|   |   |   |   |   |   |Insulin > 69: Diabetes(1.0/7.0)
|   |   |   |   |Glucose > 140: NoDiabetes(4.0)
|   |   |   |Insulin > 132: NoDiabetes(26.0)
|   |   |Glucose > 145
|   |   |   |Age <= 25: NoDiabetes(4.0)
|   |   |   |Age > 25
|   |   |   |   |Age <= 61
|   |   |   |   |   |BMI <= 27.1: Diabetes(1.0/12.0)
|   |   |   |   |   |BMI > 27.1
|   |   |   |   |   |   |BP <= 82
|   |   |   |   |   |   |   |DPF <= 0.396: Diabetes(1.0/8.0)
|   |   |   |   |   |   |   |DPF > 0.396: NoDiabetes(3.0)
|   |   |   |   |   |   |BP > 82: NoDiabetes(4.0)
|   |   |   |   |Age > 61: NoDiabetes(4.0)
|   |BMI > 29.9
|   |   |BP <= 61: Diabetes(19.0)
|   |   |BP > 61
|   |   |   |Glucose <= 157
```

| | | |Age <= 30
| | | | |BMI <= 41.8
| | | | | |BP <= 72
| | | | | | |BP <= 65: NoDiabetes(4.0)
| | | | | | |BP > 65
| | | | | | | |DPF <= 0.318: NoDiabetes(1.0/5.0)
| | | | | | | |DPF > 0.318: Diabetes(6.0)
| | | | | |BP > 72: NoDiabetes(1.0/17.0)
| | | | |BMI > 41.8
| | | | | |Glucose <= 142: Diabetes(6.0)
| | | | | |Glucose > 142
| | | | | | |DPF <= 0.371: Diabetes(2.0)
| | | | | | |DPF > 0.371: NoDiabetes(3.0)
| | | |Age > 30: Diabetes(18.0/61.0)
| | |Glucose > 157: Diabetes(12.0/84.0)

Number of Leaves: 26

Size of the tree: 51

- **MLP**

===Classifier model (full training set)===

Sigmoid Node 0

Inputs	Weights
Threshold	1.585009497601915
Node 2	2.912820821216212
Node 3	2.2621900543600173
Node 4	7.975075392675472
Node 5	6.605210513690843
Node 6	3.6432743710523185

Sigmoid Node 1

Inputs	Weights
Threshold	1.5850095101818296
Node 2	2.912820841056781
Node 3	2.2621900647916418
Node 4	7.975075490055289
Node 5	6.6052106412147324
Node 6	3.643274391739627

Sigmoid Node 2

Inputs	Weights
Threshold	8.189155309463654
Attribute Glucose	4.452654126294416
Attribute Insulin	7.998378174388446
Attribute BMI	6.3563873445884
Attribute DPF	8.56736296188997
Attribute Age	3.850471157332341
Attribute BP	1.0844486869966237
Attribute NP	4.8692848451069555
Attribute TSFT	4.201639780229084

Sigmoid Node 3

Inputs	Weights
Threshold	6.319924560476144
Attribute Glucose	10.026495597704018
Attribute Insulin	12.949733979172775
Attribute BMI	0.4187608402316052
Attribute DPF	5.042003574316904
Attribute Age	7.178133110503125
Attribute BP	3.7219263946286363

Attribute NP	13.557242575032163
Attribute TSFT	2.11723652415405

Sigmoid Node 4

Inputs	Weights
Threshold	9.455002333423169
Attribute Glucose	10.483216555100979
Attribute Insulin	9.611162394073595
Attribute BMI	13.442197828639316
Attribute DPF	1.7911707075383776
Attribute Age	8.482012575951366
Attribute BP	2.9361459394089664
Attribute NP	2.0468613672781495
Attribute TSFT	3.0266447876854587

Sigmoid Node 5

Inputs	Weights
Threshold	0.43030402195076467
Attribute Glucose	9.784516785884115
Attribute Insulin	6.355259357834153
Attribute BMI	3.1755865802138663
Attribute DPF	6.0064660668045855
Attribute Age	5.9645775802142555
Attribute BP	4.618731905512688
Attribute NP	7.777981191472427
Attribute TSFT	2.1620793796659536

Sigmoid Node 6

Inputs	Weights
Threshold	11.91615536716494
Attribute Glucose	13.865844281942197
Attribute Insulin	5.102144748331865
Attribute BMI	3.031938246698168

Attribute DPF 0.7588104242692454

Attribute Age 14.9950630806784

Attribute BP 7.355385982165217

Attribute NP 2.1704588604446142

Attribute TSFT 4.894656728936191

Class Diabetes

Input

Node 0

Class NoDiabetes

Input

Node 1

- **SVM:**

Number of kernel evaluations: 14858

Experiment 2

- **J48**

===Classifier model (full training set)=== J48 pruned tree

Glucose <= 128

|BMI <= 26.4: NoDiabetes(1.0/85.0)

|BMI > 26.4

| |NP <= 5

| | |Age <= 34: NoDiabetes(23.0/156.0)

| | |Age > 34

| | | |BP <= 89

| | | | |NP <= 2

| | | | | |TSFT <= 43: Diabetes(1.0/8.0)

| | | | | |TSFT > 43: NoDiabetes(2.0)

| | | | |NP > 2

| | | | | |NP <= 3: NoDiabetes(4.0)

| | | | | |NP > 3

| | | | | | |NP <= 4: Diabetes(1.0/4.0)

```
|  |  |  |  |  |  |NP > 4
|  |  |  |  |  |  |  |INS <= 250: NoDiabetes(1.0/8.0)
|  |  |  |  |  |  |  |INS > 250: Diabetes(3.0)
|  |  |  |BP > 89: NoDiabetes(6.0)
|  |NP > 5
|  |  |Glucose <= 103
|  |  |  |TSFT <= 26: NoDiabetes(9.0)
|  |  |  |TSFT > 26
|  |  |  |  |DPF <= 0.711
|  |  |  |  |  |BMI <= 31.3: Diabetes(1.0/4.0)
|  |  |  |  |  |BMI > 31.3
|  |  |  |  |  |  |BMI <= 38.9: NoDiabetes(14.0)
|  |  |  |  |  |  |BMI > 38.9
|  |  |  |  |  |  |  |BMI <= 39.6: Diabetes(2.0)
|  |  |  |  |  |  |  |BMI > 39.6: NoDiabetes(2.0)
|  |  |  |  |DPF > 0.711: Diabetes(3.0)
|  |  |Glucose > 103: Diabetes(13.0/33.0)
Glucose > 128
|BMI <= 29.9
|  |INS <= 146
|  |  |DPF <= 0.551
|  |  |  |NP <= 1: NoDiabetes(1.0/4.0)
|  |  |  |NP > 1: Diabetes(1.0/12.0)
|  |  |DPF > 0.551: NoDiabetes(1.0/7.0)
|  |INS > 146: NoDiabetes(3.0/28.0)
|BMI > 29.9
|  |Glucose <= 165
|  |  |Age <= 42
|  |  |  |BP <= 61: Diabetes(8.0)
|  |  |  |BP > 61
```

| | | | |BMI <= 47.9: NoDiabetes(19.0/46.0)

| | | | |BMI > 47.9: Diabetes(5.0)

| | |Age > 42: Diabetes(3.0/29.0)

| |Glucose > 165: Diabetes(4.0/53.0)

Number of Leaves: 25

Size of the tree: 49

- **MLP**

===Classifier model (full training set)===

Sigmoid Node 0

Inputs	Weights
Threshold	4.523360028415952
Node 2	9.701341077919976
Node 3	4.967386995156418
Node 4	2.382198240031957
Node 5	7.912077832269816
Node 6	4.102066300512981

Sigmoid Node 1

Inputs	Weights
Threshold	4.523360021893742
Node 2	9.7013410413022
Node 3	4.9673869760416105
Node 4	2.3821982377339115
Node 5	7.912077806197532
Node 6	4.102066295165723

Sigmoid Node 2

Inputs	Weights
Threshold	10.047100301440492
Attribute Glucose	3.677800993550474
Attribute Insulin	3.2596005256068246

Attribute BMI	18.45021528322118
Attribute DPF	5.391886291737243
Attribute Age	2.7980391888163347
Attribute BP	0.9510455162377901
Attribute NP	3.444928380109859
Attribute TSFT	0.8421588851766383

Sigmoid Node 3

Inputs	Weights
Threshold	1.4251449367691842
Attribute Glucose	3.0071731479588784
Attribute Insulin	1.9809711192898232
Attribute BMI	2.710932656045036
Attribute DPF	13.342406316388475
Attribute Age	0.7758548805001481
Attribute BP	0.9759354658766474
Attribute NP	7.39630050009308
Attribute TSFT	1.180720692032757

Sigmoid Node 4

Inputs	Weights
Threshold	3.792386958611866
Attribute Glucose	10.40884223836602
Attribute Insulin	5.3393202310109675
Attribute BMI	8.011418399766695
Attribute DPF	10.951244119436744
Attribute Age	4.099697054487383
Attribute BP	6.0336167577592015
Attribute NP	2.4085443936904007
Attribute TSFT	0.39255345350495685

Sigmoid Node 5

Inputs	Weights

Threshold 12.873155459056886

Attribute Glucose 7.111092051213641

Attribute Insulin 3.591104107996935

Attribute BMI 11.62185732519619

Attribute DPF 0.5322044731146883

Attribute Age 6.867862491619192

Attribute BP 1.931726566358325

Attribute NP 3.5611240038632745

Attribute TSFT 6.413153837512411

Sigmoid Node 6

Inputs Weights

Threshold 11.467470465622704

Attribute Glucose 9.573075676229664

Attribute Insulin 2.701603730859461

Attribute BMI 7.929503384074205

Attribute DPF 9.582364484633198

Attribute Age 3.6427131141366944

Attribute BP 0.24953414350666095

Attribute NP 4.165381243220216

Attribute TSFT 2.259227314729409

Class Diabetes

Input

Node 0

Class NoDiabetes

Input

Node 1

- **SVM**

Number of kernel evaluations: 12192

Experiment 3

- **J48**

===Classifier model (full training set)=== J48 pruned tree

Glucose <= 127

|NP <= 7: NoDiabetes(28.0/226.0)

|NP > 7

| |Insulin <= 110: NoDiabetes(8.0)

| |Insulin > 110

| | |DPF <= 0.347

| | | |Insulin <= 176: Diabetes(2.0)

| | | |Insulin > 176: NoDiabetes(2.0)

| | |DPF > 0.347: Diabetes(6.0)

Glucose > 127

|Glucose <= 165

| |Age <= 23: NoDiabetes(1.0/19.0)

| |Age > 23: Diabetes(34.0/87.0)

|Glucose > 165: Diabetes(5.0/46.0)

Number of Leaves: 8

Size of the tree: 15

- **MLP**

===Classifier model (full training set)===

Sigmoid Node 0

Inputs	Weights
Threshold	2.9150564764518037
Node 2	3.7310232225777633
Node 3	3.274901150556053
Node 4	10.951732631504356
Node 5	8.674615814993862

Node 6	6.504756925946784

Sigmoid Node 1

Inputs	Weights
Threshold	2.9150514540301753
Node 2	3.731019891091492
Node 3	3.2748959732510023
Node 4	10.951700986085664
Node 5	8.674589832429263
Node 6	6.5047392298106566

Sigmoid Node 2

Inputs	Weights
Threshold	10.052655913610304
Attribute Glucose	7.814895417974961
Attribute Insulin	5.024354092306454
Attribute BMI	0.3055046634917537
Attribute DPF	4.985164296972291
Attribute Age	11.979183644865454
Attribute BP	3.306006288016739
Attribute NP	1.1377109496546205
Attribute TSFT	5.0544789838040085

Sigmoid Node 3

Inputs	Weights
Threshold	0.39837753764439365
Attribute Glucose	7.827172555617314
Attribute Insulin	7.558978470129309
Attribute BMI	3.695005598209774
Attribute DPF	2.8362309150988736
Attribute Age	8.491625727312812
Attribute BP	0.6351224835187856

Attribute NP	12.418774848778758
Attribute TSFT	2.885211510337168

Sigmoid Node 4

Inputs	Weights
Threshold	10.44878023681133
Attribute Glucose	13.461810265469062
Attribute Insulin	9.905046665401185
Attribute BMI	9.064478659777988
Attribute DPF	5.559886665820133
Attribute Age	11.319393410268951
Attribute BP	1.1056794464856352
Attribute NP	1.66092015802544
Attribute TSFT	4.51965107342122

Sigmoid Node 5

Inputs	Weights
Threshold	1.2034149092633148
Attribute Glucose	3.2609787518554674
Attribute Insulin	4.587799675167808
Attribute BMI	3.013039102986074
Attribute DPF	8.861904032845038
Attribute Age	1.8535792543499765
Attribute BP	7.331619963086653
Attribute NP	8.79338713429296
Attribute TSFT	3.9549267597492044

Sigmoid Node 6

Inputs	Weights
Threshold	2.1173817013984753
Attribute Glucose	7.911413720619726
Attribute Insulin	5.911314857036954
Attribute BMI	4.718088074345779

Attribute DPF 4.78724934780257

Attribute Age 2.31832946784664

Attribute BP 5.008313127673822

Attribute NP 2.8625940652791604

Attribute TSFT 4.247729242739029

Class Diabetes

Input

Node 0

Class NoDiabetes

Input

Node 1

- **SVM**

Number of kernel evaluations: 5612

Experiment 4

- **J48**

===Classifier model (full training set)=== J48 pruned tree

BMI <= 25.4: NoDiabetes(44.0)

BMI > 25.4

|Glucose <= 157

| |Age <= 22: NoDiabetes(1.0/38.0)

| |Age > 22

| | |DPF <= 0.719

| | | |BP <= 88

| | | | |NP <= 3

| | | | | |BMI <= 45.3

| | | | | | |TSFT <= 24

| | | | | | | |Glucose <= 127: NoDiabetes(16.0)

| | | | | | | |Glucose > 127: Diabetes(1.0/4.0)

| | | | | | |TSFT > 24

| | | | | | | |TSFT <= 32: Diabetes(1.0/13.0)

| | | | | | | |TSFT > 32
| | | | | | | | |Insulin <= 277: NoDiabetes(2.0/32.0)
| | | | | | | | |Insulin > 277: Diabetes(2.0)
| | | | | |BMI > 45.3: Diabetes(4.0)
| | | | |NP > 3
| | | | | |TSFT <= 42
| | | | | | |NP <= 10: NoDiabetes(30.0)
| | | | | | |NP > 10
| | | | | | | |Insulin <= 115: NoDiabetes(4.0)
| | | | | | | |Insulin > 115: Diabetes(2.0)
| | | | | |TSFT > 42
| | | | | | |Age <= 41: NoDiabetes(2.0)
| | | | | | |Age > 41: Diabetes(3.0)
| | | |BP > 88: Diabetes(6.0)
| | |DPF > 0.719
| | | |TSFT <= 34: Diabetes(15.0)
| | | |TSFT > 34
| | | | |BP <= 56: Diabetes(4.0)
| | | | |BP > 56
| | | | | |DPF <= 0.785: Diabetes(2.0)
| | | | | |DPF > 0.785: NoDiabetes(9.0)
|Glucose > 157
| |TSFT <= 33
| | |Age <= 48
| | | |Insulin <= 342: Diabetes(1.0/14.0)
| | | |Insulin > 342
| | | | |DPF <= 0.851: NoDiabetes(4.0)
| | | | |DPF > 0.851: Diabetes(2.0)
| | |Age > 48: NoDiabetes(5.0)
| |TSFT > 33: Diabetes(22.0)

Number of Leaves: 23

Size of the tree: 45

- **MLP**

===Classifier model (full training set)===

Sigmoid Node 0

Inputs	Weights
Threshold	9.821155201776252
Node 2	9.41483227773479
Node 3	8.23408956807997
Node 4	10.229656526531283
Node 5	7.372285712329026
Node 6	10.354954121226488

Sigmoid Node 1

Inputs	Weights
Threshold	9.82138324160123
Node 2	9.41504081981572
Node 3	8.234253227133305
Node 4	10.22986614623246
Node 5	7.3724685563616426
Node 6	10.355203628142979

Sigmoid Node 2

Inputs	Weights
Threshold	8.313410378418778
Attribute Glucose	6.681921390929404
Attribute Insulin	0.47728748547738753
Attribute BMI	6.1002714112938605
Attribute DPF	0.10494330084929024
Attribute Age	16.49665678596859
Attribute BP	13.254598770442904

Attribute NP	1.4281306859007996
Attribute TSFT	2.228977235833143

Sigmoid Node 3

Inputs	Weights
Threshold	2.730086011780161
Attribute Glucose	2.088266555280628
Attribute Insulin	5.182536565636508
Attribute BMI	4.754943324455467
Attribute DPF	0.766759478401586
Attribute Age	9.043074130684214
Attribute BP	0.4873266080764324
Attribute NP	1.6459561405091951
Attribute TSFT	8.877056510297715

Sigmoid Node 4

Inputs	Weights
Threshold	17.866694333114967
Attribute Glucose	2.4170358119879296
Attribute Insulin	2.242592006206089
Attribute BMI	11.00403065096046
Attribute DPF	6.6062250199094485
Attribute Age	6.42844657425157
Attribute BP	6.669398102637736
Attribute NP	0.3249045438543616
Attribute TSFT	2.4308490566721637

Sigmoid Node 5

Inputs	Weights
Threshold	5.386489858989088
Attribute Glucose	3.9085225963031363
Attribute Insulin	8.225679864817499
Attribute BMI	10.351855020747049

Attribute DPF 5.192873940486711

Attribute Age 3.2113268717251504

Attribute BP 9.854665509703583

Attribute NP 0.9013124038597133

Attribute TSFT 6.808968048976017

Sigmoid Node 6

Inputs Weights

Threshold 4.927082092103205

Attribute Glucose 4.908323772428128

Attribute Insulin 1.4092644513911896

Attribute BMI 11.485577810031005

Attribute DPF 1.8915329154756675

Attribute Age 3.0516015149496423

Attribute BP 0.1817780318346392

Attribute NP 0.7288067282590448

Attribute TSFT 8.835601421538446

Class Diabetes

Input

Node 0

Class NoDiabetes

Input

Node 1

- **SVM**

Number of kernel evaluations: 4648

REFERENCES

[1] World Health Organization, "Global report on diabetes," *World Health Organization 2016*, pp. 1-88, 2016.

[2] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLOS Med*, vol. 3, no. 11, pp. 442-462, 2006.

[3] World Health Organization, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus," *World Health Organization*, pp. 1-66, 1999.

[4] National Institute of Diabetes and Digestive and Kidney Diseases, "Causes of Diabetes," *National Institute of Diabetes and Digestive and Kidney Diseases*, pp. 1-16, 2014.

[5] M. Truglio-Londrigan and S. B. Lewenson, *Public Health Nursing*, 2nd ed. Jones & Bartlett Publishers, 2012.

[6] S. Smyth and A. Heron, "Diabetes and obesity: the twin epidemics," *Nature Medicine*, vol. 12, no. 1, pp. 75–80, 2006.

[7] T. Vos, R. Barber, B. Bell, S. Biryukov, A. Bertozzi-Villa, I. Bolliger, and L. Duan, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries" *The Lancet*, vol. 386, no. 9995, pp. 743–800, 2015.

[8] M. Kalpana and A. Kumar, "Fuzzy Expert System for Diagnosis of Diabetes Using Fuzzy Determination Mechanism," *International Journal of Advanced Research in Computer Science*, vol. 3, no. 1, pp. 244–250, 2012.

[9] H. Chen and C. Tan, "Prediction of Type-2 Diabetes Based on Several Element Levels in Blood and Chemometrics," *Biological Trace Element Research*, vol. 147, no. 1–3, pp. 67–74, 2011.

[10] M. Thirugnanam, P. Kumar, S. V. Srivatsan, and C. R. Nerlesh, "Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach," in *International Conference on Modeling Optimization and Computing*, 2012, vol. 38, pp. 1709–1718.

[11] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," in *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, 2013, pp. 373–375.

[12] J. Lukasiewicz, "Philosophical remarks on many-valued systems of propositional logic," in *Polish Logic 1920-1939*, Oxford University Press, 1967, pp. 153–179.

[13] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[14] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex and Decision Process," in *IEEE Transactions on Systems, Man, and Cybernetics*, 1973, vol. 1, pp. 28–44.

[15] M. Negnevitsky, *Artificial intelligence: a guide to intelligent systems*. 2nd ed. Addison-Wesley, 2001.

[16] G. José, *Handbook of Research on Fuzzy Information Processing in Databases*. IGI Global, 2008.

[17] L. A. Zadeh, "Fuzzy Sets and Systems," *International Journal of General Systems*, vol. 17, pp. 129–138, 1990.

[18] G. Chen and T. Tat Pham, *Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems*. CRC Press, 2001.

[19] R. A. Aliev, "Fuzzy Sets and Fuzzy Logic," in *Fundamentals of the Fuzzy Logic-Based Generalized Theory of Decisions*, Springer Berlin Heidelberg, 2013, pp. 1–64.

[20] A. K. Nandi, "GA-Fuzzy Approaches: Application to Modeling of Manufacturing Process," *Statistical and Computational Techniques in Manufacturing*, pp. 145–185, 2013.

[21] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[22] N. H. Phuong and V. Kreinovich, "Fuzzy logic and its applications in medicine," *International Journal of Medical Informatics*, vol. 62, no. 2, pp. 165–173, 2001.

[23] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975.

[24] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, pp. 116–132, 1985.

[25] M. Setnes, R. Babuska, U. Kaymak, and H. R. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cypernetics)*, vol. 28, no. 3, pp. 376–386, 1998.

[26] A. Hamam and N. D. Georganas, "A comparison of Mamdani and Sugeno fuzzy inference systems for evaluating the quality of experience of Hapto-Audio-Visual applications," in *IEEE International Workshop on Haptic Audio visual Environments and Games (HAVE), HAVE 2008*, 2008, pp. 87–92.

[27] W. L. Tung and C. Quek, "A mamdani-takagi-sugeno based linguistic neural-fuzzy inference system for improved interpretability-accuracy representation," in *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2009*, 2009, pp. 367–372.

[28] A. Sancho-Royo and J. L. Verdegay, "Methods for the Construction of Membership Functions," *International Journal of Intelligent Systems*, vol. 14, no. 12, pp. 1213–1230, 1999.

[29] W. Siler and J. Buckley, *Fuzzy expert systems and fuzzy reasoning*. Wiley-Interscience, 2005.

[30] S. Sanyal, S. Iyengar, A. A. Roy, N. N. Karnik, N. M. Mengale, S. B. Menon, and W. G. Feng, "Defuzzification Method for a Faster and More Accurate Control," *arXiv [Computer Science]*, 2010.

[31] S. Naaz, A. Alam, and R. Biswas, "Effect of different defuzzification methods in a fuzzy based load balancing application," *International Journal of Computer Science*, vol. 8, no. 5, pp. 261–267, 2011.

[32] D. Almadni and A. Abhari, "Comparative analysis of classification models in diagnosis of type 2 diabetes," *in the proceedings of Modeling and Simulation in Medicine (MSM) Symposium, Spring Simulation Multi-Conference*, SpringSim 2016, Pasadena, CA, USA, 2016, pp. 772–776.

[33] A. Adeli and M. Neshat, "A Fuzzy Expert System for Heart Disease Diagnosis," in *the proceedings of International Multi Conference of Engineers and Computer Scientist*, Hong Kong, 2010.

[34] R. Parvin and A. Abhari, “Fuzzy database for heart disease diagnosis,” in Proceedings of Medical Processes Modeling and Simulation (MPMS) of the 2012 Autumn Simulation Multi-Conference (SCS/AutumnSim’12), 2012.

[35] J. Hamidzadeh, R. Javadzadeh, and A. Najafzadeh, “Fuzzy Rule Based Diagnostic System For Detecting The Lung Cancer Disease,” *Journal of Renewable Natural Resources Bhutan*, vol. 3, no. 1, pp. 147–157, 2015.

[36] M. Neshat, M. Yaghobi, M. B. Naghibi, and A. Esmaelzadeh, “Fuzzy Expert System Design for Diagnosis of Liver Disorders,” in *International Symposium on Knowledge Acquisition and Modeling, 2008. KAM ’08*, 2008, pp. 252–256.

[37] M. Kadhim, M. Alam, and H. Kaur, “Design and Implementation of Fuzzy Expert System for Back pain Diagnosis,” *International Journal of Innovative Technology &Creative Engineering*, vol. 1, no. 9, pp. 16–22, 2011.

[38] S. Muthukaruppan and M. J. Er, “A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease,” *Expert Systems With Applications*, vol. 39, no. 14, pp. 11657–11665, 2012.

[39] I. Witten and M. Hall, *Data Mining : Practical Machine Learning Tools and Techniques*, 3rd ed. Amsterdam, London: Morgan Kaufmann, 2011.

[40] J. Beck, M. Garcia, M. Zhong, M. Georgiopoulos, and G. Anagnostopoulos, “A Backward Adjusting Strategy and Optimization of the C4.5 Parameters to Improve C4.5’s Performance - Semantic Scholar,” in *Proceedings of the Twenty-First International FLAIRS Conference*, Coconut Grove, 2008, pp. 35–40.

[41] R. Gutierrez, *L18: Multi-Layer Perceptrons.* CSCE 666 Pattern Analysis, 2013.

[42] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review,* vol. 65, no. 6, pp. 386–408, 1958.

[43] M. Kantardzic, *Data mining: Concepts, models, methods, and algorithms*. USA: Wiley-Interscience, 2003.

[44] J. Rawlings, S. Pantula, and D. Dickey, *Applied regression analysis: a research tool*, 2nd ed. Springer Science & Business Media, 1998.

[45] J. C. Platt, “Fast training of support vector machines using sequential minimal optimization,” in *Advances of kernel methods*, Cambridge, MA, USA, 1999, pp. 185–208.

[46] G. Flake and S. Lawrence, “Efficient SVM Regression Training with SMO,” *Machine Learning*, vol. 46, no. 1–3, pp. 271–290, 2002.

[47] V. N. Vapnik, *The nature of statistical learning theory*, 2nd ed. New York: Springer, 2000.

[48] D. Boswell, “Introduction to Support Vector Machines,” University of California, 2002.

[49] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. Amsterdam, London: Elsevier, 2005.

[50] A. Sarwar and V. Sharma, “Comparative analysis of machine learning techniques in prognosis of type II diabetes,” *AI & SOCIETY*, vol. 29, no. 1, pp. 123–129, 2014.

[51] X. Meng, Y. Huang, D. Rao, Q. Zhang, and Q. Liu, “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,” *Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–99, 2013.

[52] A. Marcano-Cedeño and D. Andina, “Data mining for the diagnosis of type 2 diabetes,” in *World Automation Congress (WAC)*, 2012, pp. 1–6.

[53] “UCI Machine Learning Repository: Pima Indians Diabetes Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Accessed: 15-May-2016].

[54] “Diabetes and High Blood Pressure,” *Patient*. [Online]. Available: http://patient.info/pdf/4642.pdf. [Accessed: 26-May-2016].

[55] D. Rubin, “Multiple imputation for nonresponse in surveys,” *John Wiley & Sons*, vol. 81, p. 253, Jun. 2004.

[56] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Wieley-Interscience, 2002.

[57] “Resample.” [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/filters/supervised/instance/Resample.html. [Accessed: 06-Jun-2016].

[58] W. S. W. Ahmad, W. M. W. Zaki, and M. F. A. Fauzi, “Lung segmentation on standard and mobile chest radiographs using oriented Gaussian derivatives filter,” *BioMedical Engineering OnLine*, vol. 14, no. 1, p. 20, 2015.

[59] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, “Global prevalence of diabetes estimates for the year 2000 and projections for 2030,” *Diabetes care*, vol. 27, no. 5, pp. 1047–1053, 2004.