STATISTICAL MODELLING AND MACHINE LEARNING FOR THE EPIDEMIOLOGY OF DIABETES IN SAUDI ARABIA

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

by

Entissar Almutairi

Department of Electronic and Electrical Engineering, College of Engineering, Design and Physical Sciences, Brunel University London

May 2022

Abstract

Mathematical modelling and machine learning algorithms have been successfully applied to the healthcare domain and epidemiological chronic disease including diabetes mellitus, which is classified as an epidemic due to its high rates of prevalence around the world. Machine learning and statistical techniques are useful for the processes of description, prediction, and evaluation of various diseases, including diabetes. These techniques can be efficient tool in modelling diabetes and the most related risk factors. Although, Machine learning methods have been utilised in different aspects of diabetes research, but most of them were based on diagnosing or detecting the disease, and little research attention has explored the adoption of machine learning methods to study the trends in the prevalence of diabetes and forecast its future in specific populations. Thus, this thesis attempts to apply various machine learning and combination methods for studying diabetes and make future predictions.

This thesis has investigated the application of machine learning and statistical techniques for developing prediction models for diabetes and the relevant risk factors (smoking, obesity, and physical inactivity) in the Kingdom of Saudi Arabia that can be used to support health policy planning and diabetes controlling. Regression, classification, and time series modelling approaches were used for diabetes modelling. Several models were developed namely, Multiple Linear Regression, Adaptive Neuro-Fuzzy Interference System ANFIS, Artificial Neural Network ANN, Support Vector Regression, Bayesian Linear Regression, Support Vector Machine, K-Nearest Neighbour KNN, Linear Discriminant, Neural Network Pattern Recognition, and Neural Network Time Series NARX-NN models. These models integrate historical data on diabetes, smoking, obesity, and inactivity prevalence to achieve its aim for examining the trends in prevalence of diabetes mellitus in the Kingdom of Saudi Arabia, to predict the future level of the disease. A combination of regression models is performed to improve the prediction accuracy using combination methods (Average, Weighted Average, Majority Voting, Weighted Majority, Minimum, and Maximum, and a new combination method consensus model).

Several statistical evaluation metrics were applied to evaluate the performance of regression and time series models: mean squared error, root mean squared error, mean absolute percentage error, and the coefficient of determination R-squared. Classification models' accuracy performance was evaluated. Results from the regression and combined models were validated by comparison with some observed data from existing studies by the World Health Organization, International Diabetes Federation, and Family Health Survey from the Saudi General Authority for Statistics, revealing that improved accuracy was achieved with ANFIS model and the combined weighted average model in comparison to previous studies.

The experimental results demonstrate the effectiveness of regression and combined models compared to classification and time series models. The ANFIS and WAVR models were found to be suitable for diabetes prediction due to their flexibility and high accuracy.

Publications Based on This Research

Almutairi, E., Abbod, M. and Itagaki, T., 2020. “Mathematical modelling of diabetes mellitus and associated risk factors in Saudi Arabia”. *Int. J. Simul. Sci. Technol, 21, pp.1-7.*

Almutairi, E., Abbod, M.‘Forecasting the Prevalence of Diabetes Mellitus in Saudi Arabia using Machine learning classification techniques’ *Sensors (Submitted 8 Jul. 2022, Under 1st review)*

Almutairi, E., Abbod, M. ‘Application and comparison of NARX neural network and ANN models for diabetes prevalence forecasting in Saudi Arabia’. *(In progress).*

Declaration

I declare that the research in this thesis is the author's work and submitted for the first time to the Post Graduate Research Office at Brunel University London. The study was originated, composed, and reviewed by the mentioned author in the Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, UK. All information derived from other works has been referenced and acknowledged.

Entissar Almutairi

May 2022

London, UK

Acknowledgements

First and above all, I am grateful and indebted to Allah for giving me the strength, determination, and the patience to complete this work. Deepest gratitude goes to my supervisor, Dr Maysam Abbod, for his supervision, continuous support, and guidance during my studies. I am particularly grateful for his understanding, patience, and support when I experienced difficulties and challenges.

I would like to express my deepest gratitude to my family, especially my mother, I dedicate this thesis to her. I am grateful for all her prayers, encouragement, love, and support, she has provided me over the years. This acknowledgement would not be complete without stating my sincere heartfelt gratitude to my husband for his endless patience, understanding and support during my study. I am also greatly indebted to my children, for their love, caring, patience and support during my study journey and wish all the best for them in their future. I would like also to extend my thanks to my brothers and sisters for their continuous love, support, and encouragement. Finally, I wish to thank all my friends for their collaboration and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADA	American Diabetes Association
ANFIS	Adaptive Neuro-fuzzy Inference System
ANN	Artificial Neural Network
ARCH	Autoregressive conditional heteroscedastic
ARIMA	Autoregressive integrated moving average
AVR	Average
BLM	Bayesian linear regression model
BMI	Body mass index
CSVM	Cubic Support vector machine
DM	Diabetes mellitus
DT	Decision trees
EMR	Eastern Mediterranean Region
GCC	Gulf Cooperation Council
HHS	Hyperglycaemic hyperosmolar state
IDF	International Diabetes Federation
KNN	K-Nearest Neighbour
KSA	Kingdom of Saudi Arabia
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSVM	Linear support vector machine
MAJ	Majority voting
MAPE	Mean absolute percentage error
MLP	Multi-layer Perceptron
MLR	Multiple linear regression

MOH	Ministry of Health
MSE	Mean square error
NARX	Nonlinear Autoregressive Exogenous
NB	Naïve Bayes
NCDs	Non-communicable diseases
NPR	Neural Net Pattern Recognition
QALYs	Quality adjusted life years
QSVM	Quadratic Support Vector Machine
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root mean square error
SVM	Support Vector Machine
SVR	Support Vector Regression
T1DM	Type 1 Diabetes
T2DM	Type 2 Diabetes
WAVR	Weighted Average
WHO	World Health Organisation
WMAJ	Weighted Majority

Chapter 1
Introduction

1.1. Background

Diabetes mellitus (DM) was first recognised as a disease about 3000 years ago, when ancient Egyptians and Indians mentioned some of its clinical features. “Diabetes” is a Greek word which means “siphon”, denoting excessive urination, and “mellitus”, referring to “honey-sweet” (denoting the sweet taste of urine from diabetic people). The first confirmation of excess sugar in urine and blood was reported in Britain in 1776 [1]. Over the years the clinical understanding of the aetiology of DM has increased, and it is currently defined as “a group of metabolic diseases characterised by hyperglycaemia resulting from defects in insulin secretion, insulin action or both”. DM-related disturbances in carbohydrate, fat, and protein metabolism are linked to chronic hyperglycaemia, and can result in long-term damage, dysfunction, and failure of various organs, especially the heart, eyes, kidneys, blood vessels, and nerves [2].

There are three types of DM classified according to aetiology and clinical picture: type 1 diabetes (T1DM), type 2 diabetes (T2DM), and gestational diabetes. T1DM is usually a result of absolute insulin deficiency, due to the destruction of β cells in the pancreas, mostly due to a cellular-mediated autoimmune process. T2DM is caused by insulin resistance and relative insulin deficiency. Gestational diabetes is recognised or first starts during pregnancy, which is characterised by glucose intolerance of varying degrees of severity. There are other specific rare types of diabetes which can be caused by drugs, surgery, malnutrition, infections, specific genetic syndromes, and other illnesses [3].

T2DM is the most common variant, accounting for 90% of diabetic diagnoses. People with T2DM are usually diagnosed after the age of 40, but younger adults or even children can be affected by this type. The symptoms of this type may not appear on the affected person for many years, and many patients are incidentally diagnosed due to seeking treatment for related or other complications. Unlike T1DM, patients with T2DM are not dependent on insulin therapy, but they may need insulin to control their hyperglycaemia if this is not reached with diet alone, or with oral hypoglycaemic agents [4].

T2DM is multifactorial, and its aetiology is complex. Different risk factors affect the incidence of the disease, but they are not all causative factors *per se*. These associated risk factors might be genetic, demographic (such as age), or factors related to the behaviour of the person, such as diet, smoking, obesity, and physical inactivity. Behavioural risk factors can be amended or changed, thus they are often called “modifiable” risk factors [5].

T2DM is considered as one of the most widespread non-communicable diseases (NCDs) worldwide, with continually increasing prevalence. According to the International Diabetes Federation (IDF), there were more than 460 million people with diabetes in 2019, and it is expected that this figure will increase to 578 million in 2030, and 700 million in 2045. In the Kingdom of Saudi Arabia (KSA), the case of our study, there are currently an estimated 4 million diabetic patients according to the IDF [6].

There are different severe complications linked to diabetes disease, which in turn lead to negative effects on health and productivity. The increased risk of developing cardiovascular disease is 2 to 4 times greater in people with diabetes, especially heart disease and stroke. Moreover, T2DM is commonly associated with progression to end stage renal disease, requiring either dialysis or kidney transplantation to avoid mortality. The risk of lower limb amputation is also up to 25 times greater in people with T2DM than among non-diabetics. In addition, people with T2DM can develop blindness because of retinal damage [7]. In 2019 it was estimated that around 4.2 million deaths occurred in adults aged 20–79 years caused by diabetes and its complications, costing at least USD 760 billion in healthcare expenditures. The latter is projected to reach USD 825 billion by 2030, and USD 845 billion by 2045, representing relative increases of 8.6% and 11.2% on the current level, respectively.

In the last few decades, a variety of studies have sought to predict the incidence of diabetes and its global prevalence, using diverse data and methods of analysis [8][9][10]. Future estimates of the burden of diabetes are very important for health policy planning and identifying the necessary costs of controlling diabetes [11][12]. Recently, machine learning algorithms have been widely used in public health for predicting or diagnosing epidemiological chronic diseases, such as DM. There are many published diabetes modelling studies using different machine learning techniques, including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), and Decision trees (DT) [13][14][15][16][17][18][19]. Most of these models have been related to diagnosing and detecting diabetes at early stage or modelling the disease progression and complications, but little has been done to adopt any of the machine learning methods in attempt to study the trends in the prevalence of diabetes and forecast its future burden using risk factors in specific populations such as in KSA, thus the current research seeks to address this research gap.

1.2. Motivations

Among all countries around the world, low- and middle-income countries are facing the highest burden of diabetes disease due to various reasons, including urbanisation and the improvement in socioeconomic aspects in these countries, which in turn lead to different

lifestyles for their populations, such as the consumption of Western dietary pattern and the sedentary lifestyle, as well as increasing life expectancy [20]. The Eastern Mediterranean Region (EMR) has been classified by the World Health Organization (WHO) as a growing hot spot for T2DM, as the estimations and predictions of its burden go beyond those of other regions around the world. Within the EMR, the Middle East and North Africa (MENA) has been ranked by the IDF as having the highest age-adjusted diabetes prevalence (12.2%) in 2019, and an increase by 38.8% by 2030. In fact, there is growing alarm about this health problem and the impacts it poses for these countries and their health care systems, as they lack readiness for the widespread implications of treating increasing numbers of diabetic patients with associated complications. More awareness of the current high burden caused by diabetes and its complications among those making and planning health decisions in these countries, particularly at the policy level (which tends to be concentrated in governmental institutions rather than at the health system level), in order to plan for inevitably increased demands in the coming decades [6].

KSA is among the richest and largest EMR countries. National social and economic conditions have consistently improved over recent decades, based on the national economy's status as a major oil exporter. However, the Kingdom also has some of the highest levels of diabetes; according to the IDF, the country had the highest estimated number of children and adolescents with T1DM in 2019, and it has some of the highest prevalence of known risk factors, such as obesity. Although more efforts have been taken by the Ministry of Health in Saudi (MOH) regarding the prevention procedures of diabetes, it seems that these procedures need more improvements and more stringent execution. The absence and shortage of national health surveillance systems for monitoring risk factors and morbidity in KSA, which leads to difficulty in understanding past levels of diabetes and prediction of its future burden, which in turn makes this task seriously urgent.

Some published studies have reported useful estimations and predictions of diabetes prevalence rate in KSA [21][22][23]. However, the estimations and predictions of these studies were based only on the changes on the demographic factors (aging and urbanisation) and did not involve some key risk factors that linked to the disease, such as obesity. Thus, the actual trends and future predictions of diabetes prevalence in KSA were most possibly underrated by these studies. Furthermore, one international study provided estimations of diabetes prevalence in the country using body mass index (BMI) data, but its estimations for the year 2008 make it outdated, and it lacked future predictions [24].

Therefore, this thesis uses mathematical modelling by machine learning methods to make indications about the past trends and future predictions in diabetes prevalence and the related risk factors in KSA.

1.3. Aim and Objectives

This thesis aims to examine the application of machine learning and statistical methods for developing prediction models for DM and the relevant risk factors (smoking, obesity, and physical inactivity) in the KSA that can be used to support health policy planning and diabetes controlling.

To achieve this aim, the objectives of this thesis are to:

1. Identify the available local data which are required for training and testing of the models to be developed. The required data include prevalence of diabetes, smoking, obesity, and inactivity, also the demographics data (age and gender) in KSA.
2. Develop different machine learning models and statistical techniques (Adaptive Neuro-fuzzy Inference System, ANN, Support vector regression, Multiple linear regression, and Bayesian linear regression) for diabetes prediction.
3. Improve the performance of the individual regression models by using six traditional ensemble methods (average, weighted average, majority voting, weighted majority, minimum, and maximum) and introduce a new combination method (consensus model). In addition, to present the best combination method, the prediction of the combined model is validated by comparing with observed results from other existing studies.
4. Develop different classification models SVM, KNN, Linear Discriminant, and Neural Pattern Recognition for diabetes prevalence classification.
5. Develop three nonlinear autoregressive models for each independent variable (smoking, obesity, and inactivity) using neural network time series to be utilised to predict the prevalence rate of diabetes using (one-step ahead prediction).
6. Evaluate the performance of all individual regression models, the combined models, and the time series model by using some statistical metrics such as mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE) and R-squared, this is to help select a suitable predictive model. Additionally, the accuracy performance of the classification models was evaluated.

1.4. Contributions to Knowledge

This thesis contributes mathematical modelling using machine learning techniques for studying and predicting trends in the prevalence of diabetes. The novelty here is exploring various approaches of modelling and thus developing different algorithms and combination methods that provides the most accurate predictions, as well as the particular application of these technological solutions to the case of diabetes in KSA.

The key contribution of this research is the use of statistical and machine learning techniques in studying the trends in the prevalence of diabetes and its related risk factors. This thesis also proposes different combination methods to enhance the accuracy of the predictive model. In addition, a new combination method based on consensus theory is introduced. To further validate the outcomes of the predictive model, the observed results are compared with some existing studies. Moreover, this study illustrates the relationships between diabetes prevalence rate and the prevalence of the three behavioural risk factors. Consequently, the novel findings of this research are contextualised relative to existing literature.

Albeit machine learning methods have been utilised in other aspects of diabetes research, most of them were based on diagnosing or detecting the disease, and little research attention has explored the adoption of machine learning methods to study the trends in the prevalence of diabetes and forecast its future in specific populations. Thus, this thesis attempts to apply various machine learning and combination methods for studying diabetes and make future predictions.

Findings obtained from this thesis will contribute to diabetes control programmes in KSA. This is by providing indications to health policy makers and providers about the behavioural factors in diabetes, to help reduce the risk of diabetes complications and increase individuals' healthy life expectancy.

1.5. Research Methodology

This research adopts modelling strategies using machine learning techniques to study diabetes prevalence rate along with the related behavioural risk factors in KSA. It utilises three different approaches of modelling: regression, classification, and time series modelling. The proposed models integrate historical data on diabetes, smoking, obesity, and inactivity prevalence to achieve its aim for examining the trends in prevalence of DM in KSA, and to make predictions and estimations for the expected future level of the disease. These datasets were collected from the published national surveys in KSA. Data for the prevalence of diabetes, smoking, obesity, and inactivity were obtained from Saudi Health Interview Survey, which is provided by the Saudi Ministry of Health, along with other published national surveys.

The pre-processing methods were applied on the datasets namely data imputation and data categorization for classification modelling.

Following the pre-processing, the completed datasets of smoking, obesity, and inactivity only was divided into two parts: training data (from 1999-2013), and testing data (from 2014-2025), which were used for building and evaluating the models, respectively. Only the training dataset of diabetes (morbidity) data was required for building the purposed models. Next, several statistical evaluation metrices were applied to evaluate the performance of regression and time series models: MSE, RMSE, MAPE, and the coefficient of determination R-squared. The performance of the classification models was evaluated based on the accuracy of models. Lastly, the obtained results from the regression models and from the combined model were validated by comparing to some observed data from other existing studies by the WHO, IDF, and Family Health Survey from the Saudi General Authority for Statistics. Improved accuracy was achieved with ANFIS model and WAVR in comparison to these studies.

1.6. Thesis Outline

The thesis comprises eight chapters, the remainder of which are structured as follows:

Chapter 2 provides a general background on diabetes, its complications, and related risk factors. Furthermore, the mathematical models and machine learning techniques used for diabetes studies are comprehensively reviewed.

Chapter 3 presents the theoretical background of modelling along with the methodological framework of the development process of the experimental design of diabetes modelling and presents and discusses related issues.

Chapter 4 applies and develops the individual regression models used in this thesis (Multiple Linear Regression, Bayesian Linear Regression, Support Vector Regression, Artificial Neural Networks, and Adaptive Neuro -Fuzzy Inference Model). Their results are discussed, analysed and then their performance is evaluated by different statistical metrics.

Chapter 5 introduces six traditional ensemble methods (average, weighted average, majority voting, weighted majority, minimum, and maximum) and a new combination method (consensus model) for the use to combine the predictions for each individual regression model used in Chapter 4. The experimental results obtained from each combination method are then discussed and analysed, and the results of the best combined model are compared with the results of the individual regression models and with some observed results from other existing studies.

Chapter 6 presents the developments in classification modelling performance, including linear discriminant, support vector machines, K-nearest neighbours, and neural net pattern recognition, to investigate their ability to classify diabetes prevalence rates and the predicted trends of the disease according to the associated risk factors (smoking, obesity, and inactivity). The obtained results from each classification method are discussed and analysed, and their accuracy performance is evaluated.

Chapter 7 explores the implementation of time series modelling approach by developing three nonlinear autoregressive models for each independent variable (smoking, obesity, and inactivity) using neural network time series. Then providing a comparison between the time series model and ANN regression model used in Chapter 4.

Chapter 8 is the final chapter which provides a summary of the thesis, highlighting its main conclusions, and suggesting future research directions arising from the findings of this thesis.

Chapter 2
Literature Review of Mathematical and Machine Learning Models in Diabetes

2.1. Introduction

Mathematical and machine learning modelling has been successfully applied to the healthcare domain and epidemiological chronic diseases, including DM, which is classified as an epidemic due to its high rates of prevalence around the world. Mathematical modelling has become a useful method in the processes of description, prediction, and evaluation of various diseases, including diabetes. This chapter gives an overview of diabetes in terms of the world prevalence rates, associated risk factors, main types, and the complications related to the disease. It then reviews literature on the mathematical and machine learning models that contribute to the study of diabetes for clinical and economic purposes in most important aspects, including the dynamics of glucose and insulin, the progression of diabetes and its complications, the costs of diabetes-associated healthcare, and the cost-effectiveness of interventions for treating or preventing diabetes.

2.2. Epidemiological Background of Diabetes and Worldwide Prevalence

2.2.1. Overview

Over recent centuries there has been a significant change in mortality rates worldwide, with dramatically decreasing mortality and extended life expectancy due to developments in health interventions that address the most common communicable diseases such as cholera, typhoid, smallpox, and others. Although some developing countries still face these diseases, lifestyle-related chronic diseases have become the most dangerous threats to health nowadays, including cancer, cardiovascular disease, diabetes, and heart disease [25]. This shift in the pattern of diseases and their causes is called epidemiological transition, which was originally formulated by Omran in 1971 [26]. Rapidly and continuously changing lifestyles and urbanisation contributed to this transition, which led to an increase in the risk factors of non-communicable diseases. Consequently, the economic burden of these diseases is increasing around the world, particularly in developing countries [27]. Diabetes is a clear example of the phenomenon of epidemiological transition [28].

Diabetes is the main health problem addressed by this research. DM is a significant disorder of the metabolism which leads to chronic hyperglycaemia, and causes abnormalities in the

metabolism of carbohydrates, fats and proteins as a result of a deficient production of insulin, or a resistance to the insulin produced, or both [29].Patients with T1DM need insulin injections to survive, while T2DM, which represents most cases, is a defect in the secretion and function of insulin, meaning some diabetics of this type need insulin but most do not (with appropriate diet and physical activity), as they continue to produce insulin [4][30].

Diabetes is a serious health problem that is growing significantly around the world because of increasing population density, urbanisation, an aging population, and a high prevalence of obesity and lack of exercise [22]. In the last few decades, a variety of studies predicted the incidence of diabetes and its global prevalence using diverse data and methods of analysis. Future estimates of the burden of diabetes are very important for health policy planning, identifying the necessary costs of controlling diabetes [8]. King et al. [21] estimated diabetes prevalence in a 1998 study by the number of diabetics aged 20 years and over for every country in the world in three time points: 1995, 2000, and 2025. Other variables were calculated such as gender proportion, urban-rural proportion and age groups of the population who suffer diabetes. The study took data from the WHO's global database, collected from 75 societies representing 32 countries. A set of five-year age-and sex-specific estimations of the prevalence of diabetes was selected from rural and urban parts of various countries. Estimates were made according to two criteria: the sample of population had to be valid and unbiased, and diabetes had to be diagnosed using the recommended WHO diagnostics methods. In order to estimate the number of diabetes cases in every country in the world, data gathered from the WHO was linked to demographic estimates and projections released by the United Nations.

In some countries no estimates were available, so prevalence estimates from other countries with similar demographic and socioeconomic features were used. The study assumed that, besides ethnicity, other factors contribute to diabetes trends, such as population size, sex, age structure, and urbanisation level. All data sources were analysed using logistic regression modelling. The global prevalence of diabetes in 1995 was estimated to be 4.0%, predicted to increase to 5.4% by the year 2025. This was higher in developed than developing countries. The number of adults with diabetes in the world was predicted to rise from 135 million in 1995 to 300 million in 2025. The number of diabetic adults was projected to rise from 135 million in 1995 to 300 million in 2025. The age group 45-64 years had the majority of diabetic people in developing countries, while most people with diabetes were over 65 in developed countries.

Following King et al.'s study of 1998, Wild et al. [22] developed an updated report in 2004, adding new data and various techniques to estimate age-specific diabetes prevalence. This study estimated the prevalence of diabetes, and the number of diabetics in all age groups, for

the years 2000 and 2030. For this study, diabetes prevalence data according to age and sex were collected from a restricted range of countries and extrapolated to all 191 states represented by the WHO. For people aged 20 and over the data were obtained using population-based studies, using WHO criteria for diagnosing diabetes. The prevalence of T1DM in people aged under 20 in specific countries was predicted from the reported data using the same methods used by the IDF, while T2DM people in the same age group were excluded from the estimate, because there were no data available. A set of criteria for age- and sex-specific estimates of the prevalence of diabetes was extrapolated for other countries, including similarities in ethnicity, socioeconomic factors, and geographical proximity.

In order to generate smoothed, age-specific estimates, DISMOD II software was used, which is a mathematical model for analysing estimations of disease with regard to occurrences, prevalence, and mortality rates. The age- and sex-specific prevalence from the study, remission rates (assumed to be zero), and relative risk estimations of mortality of people with diabetes were used as inputs into the model. Estimates of the prevalence, occurrence, and mortality, consistent with each other, were provided as outputs of the model. These estimates were applied to United Nations population estimates for individual countries for 2000 and 2030. As in previous studies, urban and rural areas of developed countries were assumed to have similar trends of diabetes prevalence, however urbanisation is known to increase risk factors associated with diabetes in developing countries, such as changes in dietary habits, obesity, lack of physical activity, and stress, these factors account for the differing diabetes trend among urban and rural populations. It was estimated that the global prevalence of diabetes for all age-groups was 2.8% in 2000, projected to rise to 4.4% in 2030; a total of 171 million diabetic people in 2000 was predicted to increase to 366 million by 2030.

A 2010 study by Shaw et al. [23] aimed to predict the number of diabetes cases globally for 2010 and 2030. Studies were collected from the 91 countries in which they were published between January 1989 and March 2009. A total of 133 studies that used a population-based method to evaluate the prevalence of diabetes, applying the diagnostic measures of the WHO or the American Diabetes Association (ADA) were selected. Age- and sex-specific diabetes prevalence in people aged 20-79 was calculated using logistic regression modelling. These calculations were applied to the estimates of the national populations to estimate the number of diabetic people for all 216 countries for 2010 and 2030. It was estimated that the global prevalence of diabetes within the 20-79 age group was 285 million adults in 2010, projected to rise to 439 million by 2030.

A further study that addresses the global prevalence of diabetes is the 2014 study of Guariguata et al. [8] using age-specific studies of the prevalence of diabetes and applying the

Analytic Hierarchy method to choose studies systematically, making estimates for 219 countries and territories. Of the 744 sources of data available, 174 were selected to represent 130 states. For the remaining countries where no data existed, the estimates were based on countries with similar characteristics of ethnicity and socioeconomic life. In order to generate smoothed, age-specific prevalence estimates for people aged 20-79, logistic regression modelling was used, and the estimates were applied to the population of every country for 2013 and 2035. The number of people with diabetes in 2013 was estimated to be 382 million, predicted to increase to 592 million by 2035.

The *2016 Global Report on Diabetes* gives WHO's most recent available figures and expectations [10]. This study aimed to estimate global trends in diabetes and the number of diabetic adults. The likelihood of reaching the worldwide diabetes objective was also estimated. The study evaluated the trends in diabetes prevalence over the period 1980 to 2014 for adults aged 18 and over, in 200 countries and territories, divided into 21 areas according to geography and level of income. The data sources used in this study were national statistics that relied on population studies, and which included a test for one or more diabetes biological variables (FPG, 2hOGTT, or HbA1c (the study used any one of the following criteria to define diabetes: fasting plasma glucose (FPG) of 7.0 mmol/L or more, diabetes diagnosis history, or the use of oral hypoglycaemic medications or insulin). After the data were analysed, the prevalence of diabetes from sources which used other measures to define diabetes were modified to correspond to the existing definition, then those data were modelled statistically to get estimations of diabetes trends for every country and year. The results show that the prevalence of diabetes doubled globally from 4.7% to 8.5% within the period 1980 to 2014, which means the number of diabetic people was estimated to increase from 108 million in 1980 to 422 million in 2014.

Following the WHO study, an official and trusted study by the *IDF Diabetes Atlas 2017* (a global reference report) reported global diabetes estimates [31]. This report took diabetes data from seven areas, including 131 countries, to evaluate the prevalence of diabetes around the world. These data were collected from multiple sources, including health surveys, formal reports from health ministries, and unofficial communications. A total of 221 data sources were chosen. For countries where no sources were available, data were obtained from other countries with common ethnic and demographic features. This study considered various age groups, including children and adolescents aged 0-19 years, people aged 20-79, and a more expansive adult age group (18-99). A linear regression model was used to estimate diabetes prevalence by age and gender. Changes in sex, age, and variances in urban and rural areas were used to provide estimates and projections for every country. According to this study, based on data available in 2017, there will an estimated 425 million people with diabetes by

2045. Table 2.1 presents a summary of diabetes prevalence around the world according to the last published statistics by IDF in 2019.

IDF regions	Number of diabetics by years			Percentage of increase
	2019	2030	2045	
North America & Caribbean	48 million	56 million	63 million	33%
South & Central America	32 million	40 million	49 million	55%
Africa	19 million	29 million	47 million	143%
Europe	59 million	66 million	68 million	15%
South-East Asia	88 million	115 million	153 million	74%
Middle East & North Africa	55 million	76 million	108 million	96%
Western Pacific	163 million	197 million	212 million	31%
World	463 million	578 million	700 million	51%

Table 2.1: Prevalence of diabetes around the world for 2019,2030, and 2045. Source: IDF Atlas 9th Edition

2.2.2. Type 1 and Type 2 Diabetes

There are three types of DM classified according to aetiology and clinical picture: type 1 diabetes (T1DM), type 2 diabetes (T2DM), and gestational diabetes. There are other specific rare types of diabetes which can be caused by drugs, surgery, malnutrition, infections, specific genetic syndromes, and other illnesses. In this section only the two main types of diabetes will be described. Figure 2.1 shows diabetes mellitus types.

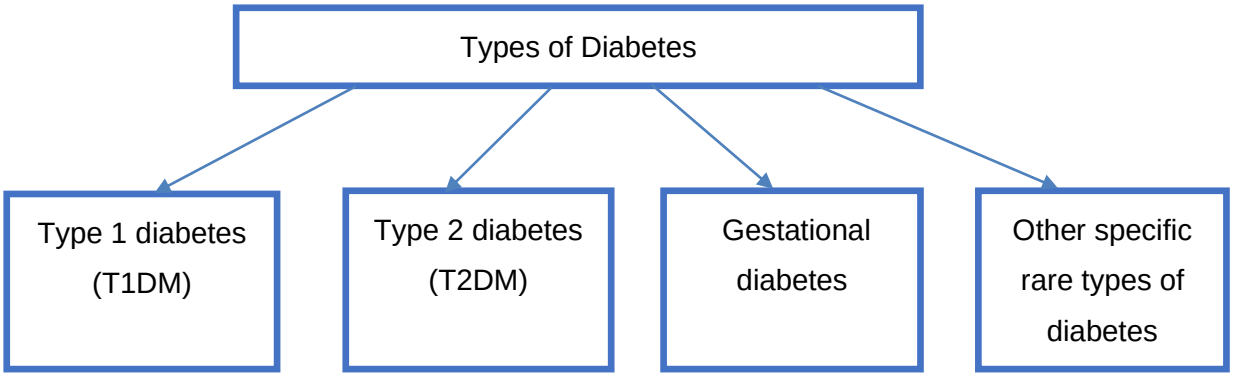

Figure 2.1: Diabetes Mellitus Types

T1DM is described as insulin deficiency as a result of an auto-immune disorder leading to the destruction of β-cells in the pancreas which is responsible for producing insulin. Genetic predisposition in some people, plus the impact of environmental factors which are not yet accurately defined, can lead to the development of T1DM. The destruction of β-cells may take place for months or years before it is sufficiently advanced for the disease to be diagnosed [32]. The speed of β-cells damage differs, but it is faster in young patients, such as children and adolescents, and slower in older individuals. This type of diabetes makes up about 5-10% of cases [33]. T1DM can be simply diagnosed in young people based on an increase of glucose levels, while it is more difficult to diagnose in older people. Patients with this type of diabetes require insulin treatment throughout their lives to survive [34].

T2DM can be classified as a result of resistance to insulin in most cases. The process of developing T2DM is more gradual than T1DM, starting at the stage of pre-diabetes, in which individuals may have either impaired glucose tolerance (IGT) or impaired fasting glucose (IFG), or both. The progression of this type of diabetes usually takes several years to be diagnosed, or before a person notices the first complications of the disease. People with this type of diabetes do produce enough insulin; the problem is that the normal level of insulin cannot be recognised by their body cells, because of insulin resistance. The genetic pre-disposition of individuals and the impact of lifestyle factors contribute to increased risk of developing T2DM, which is more common with increasing age. T2DM can be treated by changes in lifestyle, including adopting a healthy diet (particularly fewer refined sugars) and increasing physical activity, which improves insulin sensitivity, or pancreatic activity can be facilitated using medication. Compared to T1DM, patients with T2DM are not completely dependent on insulin treatment, but insulin may be required to manage glycaemia, either regularly or during exacerbated conditions (i.e., blood sugar peaks) [35].

2.2.3. Diabetes-Related Complications

People with T1DM or T2DM are susceptible to several complications that affect their health and quality of life. These complications can be long-term or short-term, according to the situation. Health problems associated with diabetes may lead to hospital care, permanent disability or even death. The most serious acute or short-term complications are hypoglycaemia, hyperglycaemic hyperosmolar state (HHS), and diabetic ketoacidosis (DKA). Long-term or chronic complications, which can take years to occur, include retinopathy (an eye disease), nephropathy (kidney disease), neuropathy (peripheral nerve disease), cardiovascular disease (heart disease), and stroke [36].

Acute metabolic changes result in short-term complications, primarily brought on by inappropriate insulin treatment, excessive food, delayed or forgotten meals, intensive

(unaccustomed) physical exercise, or other causes. People with T1DM are the most affected by DKA and hypoglycaemia, whereas HHS without ketoacidosis more often accompanies T2DM [37].

Hypoglycaemia is clinically described as blood glucose lower than 70mg/dl.[38] If insulin is taken excessively by a patient, without identifying the required dosage, it can lead to hypoglycaemia. Patients may become susceptible to hypoglycaemia by changing their dietary habits, eating times, or the content or amount of food [39]. To diagnose hypoglycaemia, a patient needs to satisfy Whipple's criteria: the main signs and symptoms associated with hypoglycaemia must exist, accompanied by a lower level of blood glucose, and the symptoms disappear after glucose management [40]. Hypoglycaemia can be identified by symptoms such as sweating, heart palpitations, tingling, discomfort, and extreme hunger. Common symptoms of hypoglycaemia include trouble focusing, tiredness, problems talking, feeling unbalanced, and extreme irritation. Preventing and dealing with hypoglycaemia can be a difficult task, requiring the restoration of a normal level of glucose in the blood. The condition hypoglycaemia can come on very quickly, most commonly in a mild form which is often treated immediately by raising the blood glucose level to about 55-70mg/dl. This can be achieved by having glucose drugs or sugary drinks. However, omitting or delaying treatment may lead to severe hypoglycaemia which can result in coma and neural defects, particularly at night, when hypoglycaemia can be very dangerous. Over 50% of severe hypoglycaemia cases occur during the night. Thus, measuring glucose regularly, educating patients and insulin therapy help to reduce the risk of hypoglycaemia [37].

DKA is a serious problem commonly associated with T1DM. The clinical definition of DKA is an absolute deficiency of insulin with hyperglycaemia, acidosis and hyperketonaemia [41]. A lower level of efficient insulin along with increasing levels of counter-regulatory hormones, which includes glucagon, cortisol, growth hormone, and catecholamines, results in DKA. The causes of DKA contribute to metabolic changes in proteins, fats and carbohydrates. DKA leads to symptoms such as extreme thirst, polyuria, losing weight, weakness, sleepiness, and coma in the later stages. Prevention of DKA play an important role in managing diabetes treatments properly. Which means that diabetic patients are required to control their blood glucose effectively, taking the correct dosage of insulin. They have to be aware of the potentially lethal complications [36][37].

Hyperglycaemic hyperosmolar state (HHS) is the new term for the condition previously known as hyperglycaemic hyperosmolar non-ketotic coma. The new term much better represents the clinical aspects of the condition, since most people affected with HHS are not in a coma, and have ketones that can be detected [42]. HHS is an intensive increase in blood glucose levels

(>600mg/dl), along with higher serum osmolality (>320 mosm/kg), without major ketosis or acidosis. Ketones, in small quantities, might be found in blood and urine [40]. HHS is less common than diabetic ketoacidosis DKA, but it does represent 10-20% of mortalities, which is dramatically higher than the mortality rate of DKA, which only accounts for 1%. Hyperglycaemic hyperosmolar state is distinguished by insulin deficiency and is different from DKA only in the intensity of metabolic acidosis, dehydration, and ketosis. Even though HHS appears more frequently in elderly people who suffer from T2DM, records show that young patients and people with T1DM can also be affected [41]. HHS can be avoided through suitable diabetes education, sufficient treatment, and recurrent personal observation of blood glucose levels.

Some factors may accelerate the incidence of HHS, such as dryness and medication (e.g., thiazides, sympathomimetic agents, and corticosteroids). It is recommended that these medications are used more carefully especially by weak patients, such as older people. Staying in hospital is a requirement for patients with HHS, which may be extended according to their situation. However, if their condition is not complex, it can be treated by similar methods as DKA [40].

Chronic or long-term complications are not actually associated with metabolic changes of diabetes *per se*, or are less directly linked to these changes [43]. Most of the effects of chronic diabetes are caused by the progression of cells, and the most serious of these problems are microvascular disease (retinopathy, neuropathy and nephropathy) and macrovascular disease (atherosclerosis). Microvascular problems are associated with the length and intensity of hyperglycaemia in T1DM and T2DM [44]. The risk of developing microvascular disease is highest in patients with poor diabetes management, and lowest in those with better control.[29] While microvascular disease affecting the eyes, kidneys, and nervous system may develop in patients with T1DM or T2DM, the risk of developing macrovascular disease such as cardiovascular disease is higher in people with T2DM [45].

Microvascular disease related to diabetes affecting the retinas, kidneys, glomerulus, and peripheral nerves can be described by common pathophysiological characteristics. Evidence provided by medical tests and animal experiments shows that severe hyperglycaemia is the essential cause of every form of microvascular disease. The duration of hyperglycaemia, along with its degree of strength, are highly related with the severity of and likelihood of developing microvascular complications in diabetic people [46].

Diabetic neuropathy can be characterised as the appearance of symptoms of peripheral neural disorder, whether patients have indicators of the disorder or not, after excluding any other reasons. Like other microvascular problems, the risk of diabetic neuropathy is related to the

extent and degree of hyperglycaemia. Most treatment methods for diabetic neuropathic problems deal with the symptoms. Amendments or changes to risk factors such as glycaemia, hyperlipidaemia, high blood pressure, smoking, and weight problems are significant protective dimensions against diabetic neuropathy.

Diabetic retinopathy is one of the ocular problems that might affect individuals with diabetes and is a major contributor to serious sight disabilities in adults. Research shows that diabetic retinopathy is linked to the duration of diabetes. It appears that 60% of individuals who have had diabetes for more than 15 years are affected by retinopathy, and it may affect 90% of those who have had diabetes for 25 years. A visual disease may be accompanied by several chronic diseases such as cardiovascular disease, renal failure or high blood pressure, which can result in serious morbidity and mortality. Chronic hyperglycaemia can cause microvascular problems in the eyes, specifically in the retina, causing damage to the small blood vessels, leading to diabetic retinopathy and consequent loss of sight. Microvascular changes result in weak capillary walls and cause aneurysms, that may rupture, leading to blood loss. Procedures to manage diabetic retinopathy involve glycaemic regulation with blood pressure administration to minimise the risk of disease progression. Laser photocoagulation is a significant treatment method, particularly for maculopathy and proliferative retinopathy [47].

Worldwide, diabetes is a major contributing factor to final-stage kidney problems. The interaction between metabolic and hemodynamic components leads to diabetic nephropathy, that stimulates the shared pathways, resulting in kidney failure. The renin-angiotensin system (RAS) significantly contributes to the pathophysiology of nephropathy. The first sign of nephropathy is microalbuminuria, that develops into albuminuria followed by kidney failure. The main risk factors involved in the recurrence, intensity and development of diabetic nephropathy include hyperglycaemia, high blood pressure, a long period of diabetes, excessive protein, smoking and age at which the disease started. The pathological transitions that occur in the kidneys of sufferers from diabetic nephropathy involve a rise in the thickness of the glomerular basement membrane, producing microaneurysm, and mesangial nodule production. Some mechanisms that lead to diabetic retinopathy contribute to the pathogenesis of diabetic nephropathy. The treatment of diabetic nephropathy begins with accurate glycaemic management. Angiotensin transforming enzyme inhibitors (ACEIs) help reduce the risk of progressive nephropathy and cardiovascular incidents in diabetics. Also, ACEIs and angiotensin receptor blockers (ARBs), along with their attributes, might have preventive impacts on the kidneys.

Macrovascular problems result from deteriorated protein in arterial walls, leading to thickening of these walls. This affects the major blood vessels in the circulatory system and increases the chances of stroke (cerebrovascular disease), cardiovascular disease, and peripheral vascular disease. Because of these complications, diabetic patients may be susceptible to gangrene, ulceration and extremity amputation. Macrovascular disease is the main reason for death among diabetic individuals. Blood pressure and fat management can be useful to glycaemic control and avoiding death due to macrovascular problems, but following a multifactorial treatment technique strongly minimises the three risk factors, and is considered essential to most effective procedures for proper protection from cardiovascular problems [44][48].

Long-term complications can be prevented through additional approaches besides glycaemic management, such as health education strategies that greatly enhance the consequences for diabetics. Patients with diabetes are advised to have frequent eye tests (annually), to facilitate early detection of retinopathy or other eye diseases (in addition to being good ophthalmological practice, particularly for older people). Foot infections can be minimised through educating patients about foot care and smoking cessation. Medical care and surgical treatment could decrease morbidity. Diabetics are at greater risk of high blood pressure and hypercholesterolemia. Awareness of proper protection and care procedures regarding the risk factors may minimise the prevalence of cardiovascular problems. People with diabetic nephropathy can reduce kidney dysfunction by extreme management of hypertension and microalbuminuria [49].

2.2.4. Risk Factors of Diabetes

Despite the complicated pathogenesis associated with diabetes, a range of factors that maximise the risk of developing the disease are specified. The risk factors for T1DM are family history, ethnic background (with a greater risk for whites than any other ethnic group), and childhood viral infections [50]. It appears that T1DM requires many years to develop. The chances of the child of a father with T1DM having the disease is 1 in 17, while the risk of a child born to a mother under the age of 25 years with T1DM is 1 in 25. In the case where the mother gave birth after age 25, the risk reduces to 1 in 100. If the mother had diabetes earlier than age 11, the child's risk is doubled. For parents who both have T1DM, the child's risk ranges from 1 in 10 to 1 in 4 [51].

There are several risk factors for T2DM, some of which are changeable, while others are not. Risk factors for T2DM that are non-changeable include age, ethnic background, genetic predisposition, history with gestational diabetes, and reduced bodyweight at birth. When T2DM is diagnosed in a father or mother before the age of 50, the potential risk of the child

becoming diabetic is 1 in 7. If the diagnosis of either of the parents happens after age 50, the risk becomes 1 in 13. In cases where both parents suffer from T2DM, the average child's risk is around 1 in 2 [51]. The occurrence and prevalence of diabetes increases as people get older. The Centre for Disease Control and Prevention announced that during 2005, diabetes prevalence in individuals over 20 in the US reached 20.6 million (9.6%), and diabetes was most prevalent in the group aged over 60, accounting for 10.3 million (20.9%). Diabetes in African-Americans tends to be higher than among Whites. Estimates of Native Americans diagnosed with diabetes range from 5% to 50%, in various tribes with various populations.

A slight difference can be found in diabetes prevalence by sex. Genetic predisposition plays a significant part, however non-genetic factors, particularly lifestyle factors, including eating habits and physical exercise, seem to be the leading cause. Modifiable risk factors related to lifestyle include increased body weight, lack of exercise, deficient healthy food, high blood pressure, drinking alcohol, and smoking. Obesity is the main risk factor contributing to the development of diabetes; the proportion fat distributed in the body, particularly a raised rate of fat at the waist or hips, particularly increases the risk of diabetes. A variety of studies report that lower levels of physical exercise increase the risk of diabetes. The latest report of 10 potential cohort studies examining medium intensity physical exercise and diabetes suggests that individuals who maintain medium intensity exercise have a 30% reduced chance of diabetes compared to others who are less active. Some elements of food such as fats and refined carbohydrates, along with total calories, have been associated with diabetes. It has been demonstrated that smoking is an independent factor increasing the risk of diabetes. Psychosocial factors such as excessive stress, depression, lack of social support and weak mental health are also connected with a higher risk of developing diabetes [50].

2.3. Mathematical Modelling for Diabetes

2.3.1. Models for Glucose-insulin Dynamics

Mathematical modelling of physiological operations is an essential technique for understanding these operations and improving technology to deal with them. This is specifically applied in the area of diabetes study and glycaemic management. A wide variety of mathematical models presented in the literature focus on the dynamics of glucose and insulin. Models range from simple linear models relating only to glucose and insulin to complicated nonlinear models that relate to chemical changes in the pancreas regarding beta cells. Although the simple linear models provide analytical assessment using essential management algorithms, they do not study the dynamic actions of the actual process. The variety of medical therapy choices for diabetes are increasing beyond the standard traditional solution of insulin therapy. To deal with the varied outcomes of these therapies, mathematical

models are required to contain the dynamics that are directly influenced by the therapies. A variety of existing models, along with their significant attributes, are reviewed in this section.

One of the leaders in this area was Bolie, who in 1961, used mathematical models to explain the connection between insulin and glucose usage. Because of the restricted possibility at that time for simulating complicated models, it was naturally rather simple, consisting of two compartments, one for glucose and the other for insulin. A set of two ordinary differential equations were included that constituted a second order system. The simulation using this model was done by an analogue computer. In fact, the linearity of the model was a major disadvantage, as the level of both glucose and insulin may be negative, which, from a physical viewpoint, is not possible. Improving this model, Bolie added an important element: the system of glucose and insulin is significantly dampened, and for that reason comes quickly back to a constant state [52]. The system of the two ordinary differential equations (Bolie model) can be given as:

$$\frac{dg}{dt} = -\alpha . g - \beta . i + G_P \qquad (2.1)$$

$$\frac{di}{dt} = -k . i - k_{sec} . g + I(t) \qquad (2.2)$$

where g and i are the blood glucose and insulin concentrations, respectively, and α, β, k, and k_{sec} are the first order rate constants. G_P and $I(t)$ are the rate of endogenous glucose production and the rate at which exogenous insulin enters the general circulation, respectively.

In 1964, the model was improved by Ackerman et al., who proposed a model of linear differential equations of the metabolic processes for glucose in the oral glucose tolerance test (OGTT) [53]. Modelling of the dynamics of glucose and insulin started with the minimal model presented by Bergman et al. in the early 1980s [54][55]. This was designed to quantify the responsiveness of the pancreas and the sensitivity of insulin in the intravenous glucose-tolerance test (IVGTT) in non-diabetic people [56].

The model had the simplest explanation of the most basic physiological factors. It was the earliest mathematical modelling using computers to evaluate insulin sensitivity. This model contains variables and three differential equations that explain the plasma quantity of glucose and insulin and the concentration of insulin in a remote compartment, which accounts for the dynamics of subcutaneous insulin infusion or the mechanics of gut glucose intake from carbohydrate food. Despite the particular use of the Bergman minimal model for analysing data during the intravenous glucose-tolerance test (IVGTT) in non-diabetic people, its natural simplicity triggered a great number of scientists to implement and enhance the model in

subsequent diabetes modelling studies. As of 2002, over 500 studies based on the minimal model had been published [57]. Additional information and facts related to this model and associated models are available in the literature [58].

A number of researchers mentioned that despite its value in physiological studies, the minimal model's minimal range of constants entails some problems. Firstly, the model comprises a couple of sections, the first using two equations, and the second adding a third. In the second section, the concentration of plasma glucose is considered a recognised strengthening function. This means that to fit the parameters of the model it must make two actions, applying the captured insulin concentration as an input to obtain the variables in the two first equations, then using the captured glucose as an input to derive the variables in the third equation. Secondly, some of the mathematical outcomes generated by the model are obviously not logical. Thirdly, the artificial non-visible factor is inserted to account for the delay in the motion of insulin [59].

Another exhaustive model regarding the regulation of blood glucose which examines the mechanism of glucagon alongside insulin, and their relation, was presented by Cobelli et al. involving three subsystems [60]:

1. The glucose subsystem explained by just one compartment model of division and metabolic process involving net hepatic glucose equilibrium, such as the variance between producing and absorbing liver glucose, kidney secretion of glucose, insulin-dependent glucose usage through muscles, and insulin-independent glucose usage through the neurological system.
2. The insulin subsystem characterised by a five-compartment model involving insulin storage in the pancreas, plasma insulin, liver and portal plasma insulin, and insulin in the interstitial substance.
3. The glucagon subsystem characterised by just one compartment model, involving plasma glucagon and glucagon in the interstitial substance.

The authors admitted the obstacles to verifying this complicated model against trial data, but they indicated that several parts of their particular model depend on outcomes from both whole body and personal organ studies [60]. A variety of simulations of their model in various cases illustrated sensible outcomes for every case. The main shortcoming of the model was that it is complicated in comparison to other options. In fact, due to its complexity, the model has not been as broadly quoted as the Bergman minimal model. Using five compartments for distributing insulin is exclusive to this model among all the models studied, and this attribute was never transferred to subsequent models by Claudio Cobelli. A recently available example of this is a model in which a range of independent sub-models are grouped together [61]. One

of these sub-models represents the glucose absorption and absorption operations in the gastrointestinal system. The model resulted from an extensive test, including more than 200 healthy people and 14 with T2DM, taking a meal with three types of traceable glucose. The purpose of developing the model is to work as a simulator tool to evaluate cure programmes for diabetes patients. For this reason, a MATLAB version of the model, referred to as the glucose insulin model (GIM) [62], has been developed for use by analysts.

A further physiological model introduced by Sorensen in 1985 [63] is considered to be an extremely comprehensive whole-body glucose metabolism model applying a great number of synchronised differential equations, based on earlier research by Guyton et al. Mass balances of glucose within a range of compartments (brain, lungs, heart, liver, gut, kidney, and periphery) are involved in this model. In each compartment sources and sinks of glucose mass are included. Considering the glucose balance of the liver as an example, the equation shows the complexity associated with the total model and the interaction between the several body parts. Glucose and insulin balance equations for the heart and lungs, brain, liver, gut, kidney and periphery are modelled, along with a mass balance for plasma glucagon. There are 16 differential equations in total. This is an illustration of a bottom-up modelling method, which considers all physiological actions that contribute, small or large. This is the opposite of the minimal methods which seek the key contributing factors to get agreement with the consequences by model fitting, taking into account individual variation. The bottom-up method is likely to be even less ideal for coping with patient variation. Disadvantages of this model are its complexity, the difficulty of evaluating patient-specific variables [64] and the lack of dynamics of subcutaneous insulin absorption. Such a model was produced to explain the interactions of glucose and insulin in non-diabetic subjects, therefore it is not able to predict a realistic hyperglycaemic manifestation of T1DM [65].

A model introduced by Hovorka et al. in 2002 provides an effective compromise in between simplicity and precision. It is a nonlinear model describing the dynamics of glucose and insulin using a range of differential equations [66][67]. The inputs of this model include the level of subcutaneously infused insulin and the quantity and time of a meal. The outputs are a temporary description of plasma glucose and concentrations of insulin. The model consists of three subsystems which represents subcutaneous and plasma insulin, insulin motion, and plasma glucose. There are two compartments in the glucose subsystem: a compartment for plasma and a compartment that is non-accessible. There are also two compartments representing the subcutaneous insulin absorption. The action of the insulin subsystem takes into consideration the physiological impacts of insulin on transporting glucose, its elimination, and its endogenous creation. These insulin procedures express themselves at a constant rate dependent on time, related to every process of metabolism. The constants of the model are

quantities that are hard or impossible to determine, even though the parameters of the model were previously determinable because of their physiological importance. The nonlinearity in the model does not come from just the insulin motion, but also from the impacts of physiological saturation. For example, the excretion of renal glucose is zero, which is less than a particular threshold (160mg/dl); insulin-independent peripheral glucose absorption is constant, higher and relative to glucose concentration, less than a further threshold (80mg/dL). The model includes glucose gut uptake dynamics. Hovorka et al. later modified their model, rechecking the uptake kinetics related to the transport of subcutaneous insulin.[68]

As an extension of Bergman and Cobelli's models, Dalla Man et al. [61] proposed a model relating to the whole-body dynamics of insulin and glucose, aiming to record several physiological actions after a meal. The principal cause of the Dalla Man model being established was the accessibility of abnormally instructive details regarding the concentrations of insulin and glucose in plasma. It is composed of a network of glucose and insulin compartments that relate to the management of glucose in insulin excretion, the action of insulin on glucose usage, and endogenous production. The system of the model can be given as:

$$\dot{G}_P(t) = EG_P(t) + R_a(t) - U_{ii}(t) - E(t) - K_1.G_P(t) + K_2.G_t(t) \qquad G_P(0) = G_P b \qquad (2.3)$$

$$\dot{G}_t(t) = U_{id}(t) + K_1\,.G_P(t) - K_2.G_t(t) \qquad G_t(0) = G_{tb} \qquad (2.4)$$

$$G(t) = \frac{G_P}{V_G} \qquad G(0) = Gb \qquad (2.5)$$

where G_P and G_t (mg/kg) are glucose masses in plasma and rapidly equilibrating tissues, and in slowly equilibrating tissues, respectively, G (mg/dl) plasma glucose concentration; suffix b indicates basal state; EG_P is the endogenous glucose production (mg/kg/min); R_a is the glucose rate of appearance in plasma (mg/kg/min); E is renal excretion (mg/kg/min); $U_{ii}$and U_{id} are the insulin-independent and -dependent glucose uses respectively (mg/kg/min); $V_G$is the distribution volume of glucose (dl/kg); K_1andK_2 are the rate factors.

In the study, 204 individuals were included, and measurements were performed for 420 minutes during and after a meal. Similar measurements were conducted on 14 individuals with T2DM. The normal data made up one parameter set and the data from diabetic patients made up the other parameter set. To measure the glucose circulation through the gastrointestinal tract, the meal was tagged using radioactivity. In order to evaluate extra moves in the body, two further tagged tracers were infused intravenously, and estimates were created using the complicated tracer-to-trace rate clamp method mentioned by Basu et al. [69]. Several later improvements were made to the Dalla Man model [70] allowing it to simulate metabolic

situations of T1DM. The β-cells subsystem that is not active is substituted in the primary model with the subcutaneous insulin subsystem to model the dynamics of exterior insulin infusion in people with T1DM [62].

2.3.2. Models for Diabetes Progression

The majority of models described assume a steady status of the disease, which is not realistic for severe chronic progressive diseases such as diabetes [71]. Furthermore, improved, newer antidiabetic factors are concentrated in medication which can modify the progress of diabetes. For this reason, the progression of the disease must be included in the model, with the purpose of understanding and studying the long-term impact of antidiabetic agents at various levels of progression. Models of disease progression, particularly for diabetes, which combine long-term population studies with antidiabetic agents have been introduced and evaluated [72]. Some efforts have been made to mathematically model the progress of T2DM. It is critical to present a realistic expression of the long-term physiological adaptation to enhancing insulin resistance for efficient clinical studies and assessing diabetes protection and treatments. It is difficult to produce a reliable model of diabetes progression, because the long-time duration of the disease makes trial validation of modelling the hypotheses difficult. With this perspective, it is significant that the basic suppositions of model equations accurately represent proven physiology, while mathematical formulae of models provide only physically possible solutions.

Damage to ß-cell mass happens in both types of diabetes. Chronic hyperglycaemia can result in an increase of ß-cell mass, while excessive hyperglycaemia can cause a decrease in ß-cell mass. To assess ß-cell response and the operations working to decrease in ß-cell mass, Topp et al. [73] introduced a model focusing on ß-cell mass as a dynamic factor along with insulin and glucose concentration. The ß-cell mass model is among only a few attempts to model the glucose-insulin system over the long term. It was developed to assist in the prediction of diabetes aetiology over weeks and months as opposed to hours, and to describe self-regulation of the glucose-insulin system using β-cell mass. However, it is not verified for realistic data. The variables of the model were obtained from physiological data and the literature about previous models. This is a three-compartment mathematical model of the glucose-insulin system, including β-cell mass, insulin, and glucose. The analytical aspect focuses on long-term factors instead of the short-term system, which means the changes happen over days instead of minutes or hours. Thus, the concentrations of glucose and insulin in this model can be regarded as essentially levels of glucose and insulin. Three differential equations represent glucose, insulin and β-cell mass. The change rate of glucose in the model is described by a differential equation containing terms that represent the occurrence rate of glucose, minus the effectiveness of glucose and the sensitivity of insulin.

The change rate of insulin is described by a differential equation containing terms that represent insulin secretion, which depends on glucose and β-cell mass, in a Hill function, and a clearance function. The change rate of β-cell mass is described by a differential equation which contains terms that represents the natural rate of death of ß-cells, the growth rate of ß-cells, which depends on the level of glucose, and a reduction in ß-cells which also depends on glucose. In the last equation the combination of variables produces a system that conforms to changing concentrations of glucose. As an example, when the glucose level is less than 100mg/dl, there is a decline in ß-cell mass that consequently reduces insulin then returns glucose into an acceptable level of 100mg/dl. When the glucose level is between 100 and 250mg/dl, the ß-cell mass rises, meaning insulin maximises and glucose levels decline close to 100mg/dl. When the level of glucose is less than 250mg/dl, the system adapts to the glucose level. When the level of glucose is more than 250mg/dl, glucose poisoning can decrease ß-cell mass, which reduces levels of insulin and causes ever-rising levels of glucose in a runaway fashion. An extension of this model by Ribbing et al. is a semi mechanistic pharmacokinetic-pharmacodynamic model, demonstrating the dynamics of fasting plasma glucose, insulin and ß-cell mass, and the impact of antidiabetic therapies in a heterogeneous population.

A model of diabetes progression over the long-term developed by De Gaetano et al. [74] summarises the pathophysiological operation of T2DM. This model explains the simultaneous development of ß-cell mass, pancreatic ß-cell duplication reserve, dominant glycaemia and dominant insulinemia, based on variables which represent glucose absorption in tissue dependent on insulin, hepatic net glucose outcomes, ß-cell insulin excretion capability, and insulin removal by plasma. In recent times, the effectiveness of this model was assessed by Hardy et al. [75] who applied the outcomes of the diabetes prevention program study. A mechanistic model based on population was produced by Nie et al.[76] designed for muscle pyruvate dehydrogenase kinase-4 mRNA changing over time, quantifying disease progression and the impact of diet and plasma variables on pyruvate dehydrogenase kinase-4 mRNA. While the implementation of epidemiological models of non-communicable diseases is rather uncommon, a few age-designed models and population-based models have previously been presented.[71].

2.3.3. Models for Complications

People with T1DM and T2DM have a greater risk of developing microvascular and macrovascular complications, the main causes of nephropathy and retinopathy, that result in the aetiology for neuropathy and the resection of ulcerated limbs. They may also cause coronary heart disease and stroke. Mathematical models to understand and predict the various problems and complications associated with diabetes have been developed. The

complications include nephropathy, retinopathy, foot ulcers and cardiovascular disease. However, because of the opacity and complexity of the complications of diabetes, mathematical models in this particular area are uncommon.

Among the mathematical models available is the Cardiff Diabetes Model, which is a discrete event model operating stochastic simulations designed in Visual Basic script, part of Microsoft Excel. Its structure is dependent on the Eastman model of T2DM, first introduced in 1997. It is designed to assess the effect of common recent treatments in about 10,000 people recently diagnosed with T2DM. In this model non-diabetic and T2DM people are included. It applies the UK Prospective Diabetes Study (UKPDS) risk engine method along with the Framingham risk equation. Typical outcomes from the simulation are the occurrence of microvascular (retinopathy, neuropathy, nephropathy) and macrovascular complications (congestive coronary disease, unexpected death, and cerebrovascular disease), described as deficit neural with signs and symptoms lasting longer than one month. The UKPDS outcomes model was designed by the University of Oxford based on survival equations predicted by applying data from the UKPDS on 3,642 patients. A detailed description of the model was presented in an article published in *Diabetologia* [77].

Briefly, the UKPDS outcomes model is a probabilistic discrete-time model which uses an integrated set of parametric relative risk models to calculate the total risk for earliest incidence of 7 main complications related to diabetes, heart failure, ischemic coronary disease, stroke, kidney failure, blindness, amputation and myocardial infarction (fatal and not fatal). The estimations of this model are dependent on properties of the patient such as age and gender, as well as risk factors which differ over time including level of A1C and systolic blood pressure. The simulated patients begin with a pre-specified health position and may suffer one or other chronic complications or even die in any of the yearly periods as the simulation continues [77].

Another model that considers the complications of diabetes is the EAGLE model, which applies Monte Carlo simulation and risk equations. The latest version of this model includes the impact of several variables [78]. The fundamental composition of this model is a Markov process, with annual time periods and first and second order estimations. Probabilities of transition are based on the situation of the simulated patient, along with relevant computations which are internally defined. Several input variables are involved including demographics (such as age, gender and duration of diabetes), physiological features (such as A1C and systolic blood pressure), pre-existing complications, and lifestyle factors (such as smoking). However, the primary determining factor in the progression of diabetes and the development of its related complications is A1C, which is simulated over time with respect to predetermined targeted A1C.

Up to 20 effects, including hypoglycaemia, macular edema, retinopathy, neuropathy, end stage renal disease, stroke, and diabetic foot syndrome, are predicted in accordance with data provided by clinical and epidemiological studies including the UKPDS, the Wiscons Epidemiological Study of Diabetic Retinopathy (WESDR) and the Diabetes Control and Complications Trial (DCCT). In this model, a cohort of up to 50,000 virtual patients is used for every simulation, using distribution suppositions to identify the properties of patients. Occurrences are designated to patients individually over time. Outcomes are the mean rates of every cohort simulation and the mean rates of several versions of similar simulations, as demanded [78].

The Archimedes model is a mathematical illustration of the pathophysiology, physiology, symptoms, indications, attitudes, tests, therapies, strategies, sources, and consequences of T1DM and T2DM, along with other diseases such as hypertension, congestive heart failure and stroke [79]. It applies an object-oriented simulation method with differential equations to reconstruct standard details equivalent to an in-patient chart, clinical textbooks, medical training regulations, and clinical studies. A range of disease-related subjects are covered by this model, which takes into consideration biological features, care operations, patient and healthcare professional behaviours, strategies, costs, and resources. The Archimedes model provides biological parameters over time, which means that any incident can happen at any time and, in contrast to the Markov model, it does not have any specific cases. This model has an effective role in modelling diseases and complications together, allowing it to cope with pathology, symptoms, a variety of therapies, and treatments with several impacts [80].

The Archimedes model includes over 100 biological parameters, signs, tests, therapies and consequences of the complications associated with diabetes, as well as their administration. As an example, coronary artery disease is modelled by two main attributes known as slow occlusion and fast occlusion, compatible with the development of atherosclerotic plaque in heart blood vessels, the acute closure of a coronary artery as a result of plaque, and the progression of an occlusive thrombus. Any of these types of closure may occur at any time in any of the four coronary arterial blood vessels, with the effect that a quantity of the distal myocardium is impacted, myocardial contractility, cardiovascular outcome, and so on. The equations which identify the duration of these attributes are obtained from datasets of population-based studies, for example the Framingham study, including several parameters such as age, gender, high-density lipoprotein, smoking, cholesterol level, overall cholesterol, diastolic blood pressure, hypertrophy associated with the heart and FPG [79].

The Archimedes model deals with these parameters differently from other similar regression models. The most significant aspect is that the equations in Archimedes do not compute the

potential risk of a result, for instance a myocardial infarction, but model the occlusion of specific coronary arteries. An additional significant distinction is that they involve FPG as an ongoing factor and combine the level of FPG and the period over which the FPG has been increased. Additional parameters required for these types of equations are computed in other areas of the model. The model involves equations for the string of biological incidents which happen along with an infarction, such as myocardial problems, decline in myocardial contractility, and decline in cardiac outcome, along with the therapy for those incidents. Strokes are managed in the same way [79].

The key attribute regarding nephropathy is the gradual decrease of glomerular performance. The quantity and sort of protein released in the urine and the subsequent symptoms of diabetic nephropathy are functions of the value of this attribute, which is, in turn, an event for the individual's FPG, blood pressure level, and a factor called glycaemic load which provides not just the level of FPG but also the period of time for which the FPG continues to be increased to various levels. Retinopathy identifies the path of this complication. Medical symptoms such as retinal haemorrhages, difficult and soft exudates and micro aneurysms are features of the retinopathy attribute. Thus, these symptoms determine the stages employed in evaluating the progression of retinopathy. Consequently, retinopathy develops constantly with the stages physicians identify based on the symptoms of the disease, and the stage the patient is in can be tracked by the model at any time. Just like nephropathy, the progress of the retinopathy attribute is caused by a person's FPG, glycaemic load, and blood pressure [79].

In the present version of the model, the primary medical symptom of diabetic neuropathy is a lack of sensation. The incidence and progress of this complication are modelled by determining the principal attribute, known as neuropathy, caused by a person's FPG, blood pressure and glycaemic load. In the model, complications including foot ulcers and diabetic foot are effects of the neuropathy attribute. A publication has been introduced which aims to validate the Archimedes model as regards diabetes and its complications, with clinical studies giving information about 74 validation exercises, including 18 trials and studies [81].

2.3.4. Economic Models in Diabetes

The health economics aspects of diabetes are among the most important topics addressed by the diabetes models available, which use data and provide outcomes (e.g., the costs of diabetes healthcare, the cost-effectiveness of medical treatments, medication and medical or other interventions). Modelling studies of diabetes have presented several types of valuable information for policy makers and healthcare providers. Several models have been designed to predict the burden of diabetes in the future and the influence of interventions, such as the amendment of risk factors, therapies and other things, on future disease trends [82][83]. Some

models estimate the likely incidence of serious complications associated with diabetes, such as myocardial infarction, ischaemic heart disease, blindness, stroke, renal failure etc. over a lifetime. They assess the results in terms of health economics, including quality adjusted life expectancy; one example is the UKPDS [77]. Other sorts of models have been created to evaluate the economic influence of diabetes in terms of direct medical care costs related to major diabetic complications [84].

A wide range of randomised managed trials relating to diabetes clinical results and the effect of interventions have been used as a major source of data for physicians and decision makers engaged in diabetes care. However, even though these trails remain an important source of data, frequently they do not present results for the long-term such as >5-10 years. The outcomes obtained from clinical trials apply only to the population recruited and the protocol employed [80][85]. As an alternative, physicians and policy makers are forced to depend on their own autonomous decisions, and there are vast discrepancies in the behaviours of physicians in practice. It is very difficult for the human mind to cope with complexity, variation and uncertainties regarding health and diseases [85]. There is a significant increase in the worldwide incidence rates of diabetes and its risk factors, which means that countries need to be aware of the future trends in levels of the disease and the resulting economic burden.

Accordingly, the framework of the decision analytical modelling, as a useful tool to offer details of economic evaluation, is widely supported by popular diabetes organisations including the ADA [85]. Decision analytical modelling considers the predicted costs and outcomes of decisions using data from a variety of sources and mathematical methods, normally using computer software. This type of modelling aims to present effective evidence for decision makers to help them make the best decision (e.g., would it be beneficial for an alternative medication to be adopted?) [86]. Recently, there has been increasing use of computer simulation modelling technology to examine diabetes, including the clinical, epidemiological, and economic aspects of the disease [80].

A great number of studies have been published in the literature which describe various diabetes models and their relevant outcomes [77][83][84][87][88]. In these models, diverse methods are applied, various data sources are involved, and several result measures are recorded. Regarding economic evaluation in healthcare and planning of services, mathematical models can represent patients' experiences as individuals or cohorts, assessed in accordance with incidents or discrete increases of time [89]. Practically, the most popular mathematical models which represent decision-analytic modelling are decision-trees and state transition models (including the Markov model and first-order Monte Carlo microsimulation). There are other modelling methods such as agent-based models, discrete event simulations

and dynamic models [89][90]. These models are briefly outlined below, along with their structures, advantages, and limitations.

2.3.4.1. Decision Tree Models

These are a simple type of decision analytical modelling for economic evaluation [86][89]. Optional alternatives are demonstrated using a sequence of paths that represent decisions and potential events over time [86][91]. Every optional event has branches which represent the potential situation and its regarded chances. These chances might rely not only on various procedures but additionally on properties of the patient, such as sub-groups with various risk factor backgrounds. At the end of the tree every pathway has a result, for instance signs or symptoms, medical status, costs, survival, quality adjusted life years (QALYs), or death [86][91]. For every option the expected value of the medical result can be computed as a measured average of all potential results, using the pathway chances as weights. Decision trees work effectively for evaluating events which have specific recurrences and a specific steady time [91]. The benefit of decision trees is that they are easy to formulate and realise [89].

Decision trees are preferred because of their simplicity and transparency, and they could be an effective method of making alternatives clear [86]. Nevertheless, a shortcoming of decision-tree models is that the chance of every potential event remains steady, meaning it does not easily cope with the time dependent elements of the economic evaluation of diseases. In fact, with chronic diseases, the chances of potential events change along with age, health condition and time. Thus, decision trees are usually not suitable in cases where an issue is complicated, for example when there is a risk of events continuing or when the number of events is significantly increased and then the branches might become uncontrollable. Therefore, decision-tree models are usually not used to model chronic diseases such as diabetes [89][92].

2.3.4.2. State-transition Models

These are based on events of interest distributed into health states outlined in accordance with population properties including age, disease level and treatment. Clinical background, age, and treatment are contained in the model by being integrated within the description of the health states or within the determination of the transition possibility. Transitions develop from one health state to another at specified periods, normally one year, depending on the transition possibilities [92]. The main supposition is that all people must exist in one of a specific range of states during a single period. State transition models are commonly represented as cohort models, even though they follow individual subjects [89]. State-transition modelling is a manageable, easy-to-use, transparent technique for decision analytic

modelling which includes both Markov model cohort simulation and individual-based first-order Monte Carlo microsimulation [93]. With the Markov model, the ratio of patients in each health state each year is specific, and the transition possibilities are based on the present state. By simulation, the number of patients in the population moving to each state at each interval of time can be calculated [92].

In contrast to decision trees, which present a series of events as a large number of possible paths, Markov models give more direct and flexible results, including continuing results over time. Patients stay in one of the specific ranges of health states at any point in time and develop transitions among these health states over discrete periods or cycles. The chance of remaining in a state or transferring to another in each cycle is identified by a collection of determined transition possibilities [86]. For chronic diseases such as coronary heart disease or diabetes, the model variables such as rate of progression, quality-of-life measures or costs can change over time, so the time of an event or the rate of progression plays a significant role. Events might also reoccur. According to the conditions, Markov models are commonly recommended to assess interventions and compare them for estimates of life expectancy, QALYs, and predicted costs.[91]

With Monte Carlo models, every potential event is simulated for every individual in the cohort and the final statistics are calculated by gathering the events within the simulated time period for the modelled population. As an illustration, to simulate a 5% probability of having diabetes in a specified year, the computer produces a random number between 1 and 100, and if that number is 5 or lower, the computer system adds the simulated individual as having diabetes in that year. This provides one potential experimental observation, but the simulation recurs several times, and as the number of operations gets bigger, the average values approximate the values that might be calculated by using a Markov model. For diabetes, both Markov and Monte Carlo models have been used to describe the progress of the disease and assess the cost-effectiveness of treatment options [92].

The main advantage of cohort state transition (Markov) models is that they are easy to practically formulate, handle, connect and evaluate with user-friendly computer software as long as the range of states is not too big [93]. State transition models are helpful when the risk of an event continues over time, if the timing of events is essential, or if events can happen several times [89]. A major restriction of Markov models is that the transition possibilities rely only on the present health state, independent of any historical background. This assumption (the Markovian assumption) can be very restricting in clinical situations when existing features become powerful determinants of what occurs in the following. This restriction can be handled by presenting impermanent states, which subjects enter for just one period, or a range of

impermanent states which must be entered within a fixed cycle. A Markov model can deal with memory by developing states which have history, however this significantly expands the number of states, leading to massive models which are hard to handle [86][93].

With individual level state transition models there is no restriction of the Markovian assumption, since they simulate only one individual over time. Such microsimulations are assessed using first-order Monte Carlo simulation, where an individual experiencing a specific transition relies on a random number [93]. While cohort models are assessed as individual cohorts developing within the states together, not differentiating one person from another, excluding state descriptions, individual-level state transition models maintain tracking of every individual's historical background. This significantly reduces the number of states. However, major disadvantages are that they require intensive calculation, usually demanding the simulation of millions of individuals to achieve steady values of the results, making them even more difficult to handle [93].

Markov models, alone or in conjunction with decision trees, are often the most popular models applied to economic evaluation in healthcare and diseases, although several other methods exist [86].

2.3.4.3. Discrete Event Simulations (DES)

These model the development of patients within healthcare operations or programmes, measuring their attributes and the consequences over unlimited periods. Discrete event simulations can be described as individual-level models in which events rely on prior incidents experienced by patients through the model as well as their attributes. In these simulations there is no restriction on using the same periods or the Markovian assumption [86]. With DES, the time is modelled by computing if the following event arises after the incidence of every event. Staff members or hospital services such as beds or other resources, can be clearly modelled by applying shift patterns [89]. DES is a manageable modelling framework which simulates the changes in a system over time using a series of discrete individual events [94]. The possibilities of specific events happening vary over time and are influenced by connections among individuals, populations, and the situation after every event. A collection of system principles and probabilities control the behaviour of the population or individuals in the model, and these might be modified depending on the intervention modelled.

As every event arises, the costs and resources depending on the event, the outcome, and the duration of the event are calculated. The method of DES is widely implemented for assessing screening applications in public health economic modelling of non-communicable diseases. Along with other model frameworks which simulate connections among the population which are modelled with the environment, DES is well prepared for dealing with connections in

complicated systems [89][94]. Nevertheless, because of the huge quantity of potential parameters in DES and the requirement to simulate lots of individuals, models might be slow and time consuming in performing the calculations specially when evaluating uncertainty. Large amounts of data are needed for every disease consequence, and DES often demands experience and time to develop [89].

2.3.4.4. Agent-based Modelling (ABS)

This is an individual-level method, that simulates agents, for instance patients, through potentially adaptable behaviour based on their environment [89]. Similar to DES, ABS enables the possibility of events happening through the system which are being modelled to alter as a result of interactions between individuals (agents), and between agents and the environment, over time. Thus, ABS is able to cope with interventions which have several components and interactions in complicated systems, not health related. Agents in the model can modify their behaviour in accordance with interactions with each other and the environment [94]. ABS is commonly used to model infectious diseases.

However, ABS is complicated in comparison to DES, and significantly more data is needed in order to present a heterogeneous population. For every event and state of the disease, costs and morbidity are calculated to obtain estimates of the cost effectiveness of interventions which affect the behaviour of agents (some other social costs and consequences can also be calculated in the same way). By the combination of all agent-level outcomes, the population-level outcomes can be obtained. The Archimedes model is an illustration of ABS applied to public health for economic modelling of non-communicable diseases such as diabetes [95][96]. This valuable model has been developed to simulate a variety of interventions, modelling a wide collection of clinical results with modified physiological risk factors. Hence, it is an excellent illustration of simulating interventions that include several components. It enables interactions between parameters among individuals, as well as with the system in general [59].

2.3.4.5. Dynamic Models

These rely on differential equations and are often employed for modelling transmission impacts of infectious diseases and estimating costs using sets of mathematical equations usually applied in cost simulations. As decision models seek to inform decisions, they must evaluate and examine various optional methods. Thus, decision models are normally suitable for complete economic evaluations, described as comparison analysis of different programmes of action where both costs and outcomes are assessed [91]. The interaction between populations and their environments is permitted in system dynamics models. While the model operates, the possibilities of events occurring in the model (the system) alter with

feedback [94]. Because a rising number of variables being included affects the system (demanding an increased quantity of data), this model can be created to be significantly complicated.

As a result, system dynamics models are becoming better suited to simulating interactions in complex non-health sector systems and calculating the impacts of interventions that include several components, better than other model frameworks. In these models, costs can be attached to the stage of disease or any other variable in the model, and a comparison of costs and health results with and without the intervention can be made. System dynamics models can normally be manifested graphically, which helps stakeholders communicate with the model. This form of model is effectively implemented for infectious diseases where the complex features of an infectious disease need to be identified [89][86]. It is also used for the risk factors of non-infectious diseases [94]. A disadvantage of the system dynamics models is that the dynamic variables involved in the model; in other words, the level of change of variables over time is deterministic, even though they can model parametric uncertainty [94].

2.4. Regression and Classification Machine Learning Models

Machine learning algorithms have been widely used in public health for predicting or diagnosing epidemiological chronic diseases, such as DM. There are many published diabetes modelling studies which applying different machine learning techniques including SVM, ANN, KNN, and DT models or hybrid techniques. These models have been used for different purposes such as diagnosing or detecting diabetes at early stage, and for modelling the disease progression and complications. A summary of all studies discussed in this section are provided in Table 2.2. A comparative study conducted by Faruque et al. [97] using different machine learning models, including LSVM, decision tree classification, Naïve Bayes, and KNN used the evaluation metrics of accuracy, recall, and precision to compare the performance of the classification models on predicting diabetes. The results obtained from this study indicated that the best performance was achieved by LSVM model, when applied on the dataset collected from Bangladesh Medical Centre for Diabetes Classification.

Another study by Ratna Patil et al. [98] aimed to evaluate the performance of classification algorithms on the prediction of diabetes. In this study, the PIMA Indian data repository were used, which included a total of 768 samples. This data was divided into training and testing sets with 70% for training (n = 583 samples), and 30% for testing (n = 230 samples). This study examined the implementation of eight machine learning models namely: Logistic Regression, KNN, SVM, Gradient Boost, Decision tree, MLP, Random Forest, and Gaussian Naïve Bayes. The results showed that the highest accuracy was achieved by the Logistic

Regression model, with 79.54% and RMSE of 0.4652; the lowest accuracy was given by the Multilayer Perception (MLP), with 64.07% and RMSE of 0.5994. The authors suggested improving the obtained results by using outlier detection before classification. Additionally, Bukhari et al. [99] constructed an ANN model using different numbers of neurons, from 5 to 50, in hidden network layers. This study aimed to predict female diabetes utilising the PIMA Indians diabetes dataset. Eight features were considered to train the ANN model, and the results showed accuracy of 93% when using the validation set.

A further study performed by Hasan et al. [100] for diabetes diagnosis and prediction by using six machine learning models: AdaBoost, k-NN, Decision trees, XG Boost, Naïve Bayes, and random forest. The PIMA Indian Diabetes dataset was used to train the models, containing a total number of 768 female patients, with 268 diabetic patients (positive) and 500 non-diabetic patients (negative). In this study eight different attributes were involved: glucose, insulin, pregnant, pressure, triceps, BMI, pedigree, pressure, and age. The authors mentioned that the pre-processing steps of the data is contributed to achieve a good result, the procedures involved in preparing the data are feature selection, data standardization, outlier rejection, substitution with the mean for missing values, and k-fold cross-validation (fivefold in this study). An ensemble method has been also implemented which used to enhance the performance using multiple classifiers. In ensemble approaches, the combination of the outcomes obtained by different models can enhance the accuracy of the prediction. AdaBoost and XGBoost were the best models used together. For evaluating the performance, area under the curve (AUC) was chosen as an evaluation metric. This study was capable to achieve an AUC score of 0.95 which considered the best compared to other studies.

An additional study by Abdulhadi et al. [101] developed a variety of machine learning models for the purpose of predicting the presence of diabetes in females using the PIDD dataset. They addressed the problem of missing values using the mean substitution technique, and all attributes were rescaled using a standardization method. The constructed models are linear discriminant analysis (LDA), LR, SVM (linear and polynomial), and RF. Based on the results of this study, the highest accuracy score was achieved by the RF model, with 82%.

In another research, Oleiwi et al. [102] proposed a classification model aimed to an early detection of diabetes using machine learning algorithms. This study has been designed to use significant features and deliver results which are close to the clinical outcomes. Three classification models have been trained, namely a Multi-layer Perceptron (MLP), a Radial Basis Function Network (RBF), and a Random Forest (RF), mainly to obtain the best classifier model for predicting diabetes. Their findings showed that the RBF model outperformed other models, with accuracy of 98.80%.

Further to the studies that predict or diagnosed diabetes, some existing studies have addressed the use of machine learning techniques to construct predictive models for diabetes complications. An example is the model developed by Kantawong et al. [103] to predict some complications related to diabetes, particularly hyperlipidaemia, coronary heart disease, kidney disease, and eye disease. A dataset of 455 records was used in this study. Selection and cleaning process were carried out on the dataset which reducing the number of records used to build the model. The number of features and the final number of records which used to train the model were not mentioned by the authors. An iterative decision tree (ID3) algorithm was chosen to construct the model. For evaluating the performance of the proposed model, 10-fold cross validation method was used, giving an accuracy of 92.35%. It should be noted that the high score of accuracy obtained by this study is not sufficient to evaluate the performance of the model, specifically when training unbalanced data. The main reason for this is that when the model trains the data a minority class can be ignored, and all the predictions classified as the majority class and the good accuracy scores still achieved.

Dagliati et al. [104] developed different classification models including LR, NB, SVMs, and random forest to predict the onset of retinopathy, neuropathy, and nephropathy in T2DM patients. The authors used different time scenarios for making the predictions: 3, 5, and 7 years from the first visit to the Hospital for diabetes treatment. The dataset used to train the proposed models was collected by Istituto Clinico Scientifico Maugeri (ICSM), Hospital of Pavia, Italy, for longer than 10 years. These data involve a total number of 943 records including the features of gender, age, BMI, time from diagnosis, hypertension, glycated haemoglobin (HbA1c), and smoking habit. The problem of unbalance and missing data was managed by applying missForest approach, while the problem of unbalanced class was overcome by oversampling the minority class. The obtained results of this study show that the highest accuracy score was achieved by LR with 77.7%.

A further study performed by Islam et al. [105] developed HbA1c regression models to predict the average amount of glucose accumulated in the blood over the last 2–3 months. Predicting the levels of HbA1c in advance helps determine direct relationships with diabetes and avoid the future risk of complications. In this study the dataset was collected from the Diabetes Research in Children Network (DirecNet) trials on a total of 170 patients having T1DM. Furthermore, various methods for feature extraction and selection were used to prepare the dataset. The findings obtained by this study show that the best performance was achieved by the constructed model which involved two ensemble methods RF and extreme gradient boosting (XGB) with low mean absolute error (MAE) of 3.39 mmol/mol and a high score of coefficients of determination (R-squared) of 0.81.

Sr. No.	Researcher Name and year	Methods/Techniques	Results and Findings
1	Faruque et al., 2019 [97]	SVM, C4.5 Decision Tree, Naïve Bayes (NB), and KNN	Researchers found that the C4.5 decision tree model outperformed the other classifiers with an accuracy of 73%
2	Ratna Patil et al., 2018 [98]	Logistic Regression (LR), KNN, SVM, Gradient Boost, Decision tree, MLP, Random Forest, and Gaussian Naïve Bayes	The highest accuracy was achieved by the Logistic Regression model, with 79.54% and RMSE of 0.4652
3	Bukhari et al., 2021 [99]	Artificial Neural Network (ANN)	The ANN model showed an accuracy of 93%
4	Hasan et al., 2020 [100]	AdaBoost, k-NN, Decision trees, XG Boost, Naïve Bayes, and random forest	An ensemble method has been implemented indicating that AdaBoost and XG Boost were the best models used together. This study achieved a score of area under the curve (AUC) of 0.95 which considered the best compared to other studies
5	Abdulhadi et al., 2021 [101]	Linear discriminant analysis (LDA), LR, SVM (linear and polynomial), and Random Forest (RF)	The highest accuracy score was achieved by the RF model, with 82%
6	Oleiwi et al., 2020 [102]	Multi-layer Perceptron (MLP), a Radial Basis Function Network (RBF), and a Random Forest (RF)	Their findings showed that the RBF model outperformed other models, with accuracy of 98.80%
7	Kantawong et al., 2020 [103]	An iterative decision tree (ID3) algorithm	10-fold cross validation method was used, giving an accuracy of 92.35%
8	Dagliati et al., 2018 [104]	LR, NB, SVM and RF	The obtained results show that the highest accuracy score was achieved by LR with 77.7%

9	Islam et al. 2020 [105]	HbA1c regression models, involving two ensemble methods RF and extreme gradient boosting (XGB)	Best performance was achieved by the constructed model with low mean absolute error (MAE) of 3.39 mmol/mol and a high score of coefficients of determination (R-squared) of 0.81.

Table 2.2: Summary of machine learning techniques in diabetes research

2.5. Time Series Forecasting Models

During the past few decades, a great deal of attention has been devoted to time series analysis which have become a popular research topic in different fields such as finance, energy, electricity, and medicine etc. Time series forecasting is an important task in time series analysis, and it considered as a powerful tool for describing a complex system using observed data. Autoregressive integrated moving average (ARIMA) models have been the most popular and widely used in the time series forecasting domain. These models gained their popularity due to its statistical characteristics as well as the well-known Box–Jenkins approach in the model development procedures.

In recent years, there are several studies which applying time series to analyse and model clinical datasets. In the following some of the time series applications for forecasting the number of patients in different diseases will be presented. However, there are a limited number of studies that using time series method for diabetes incidence research. A summary of all studies discussed in this section are provided in Table 2.3.

A study by Earnest et al. [106] used autoregressive integrated moving average (ARIMA) models to estimate the number of beds occupied during a severe acute respiratory syndrome (SARS) outbreak in Tan Tock Seng Hospital Singapore. Their results showed that the ARIMA (1, 0, 3) model was efficiently able to predict and describe the number of beds occupied throughout the SARS outbreak. For evaluating the performance of their model, they measured the MAPE for the training and validation set which were 5.7% and 8.6% respectively, where they conclude that it was reasonable for use in the hospital setting. In addition, three-day forecasts of the number of the required beds were also provided by the model. They also found that the total number of admissions and probable cases admitted on the previous day were independent predictive factors of bed occupancy.

Another study by Liu, et al. [107] on the prediction of haemorrhagic fever with renal syndrome (HFRS) in China applied ARIMA for the period of 1978 to 2008. The best model was chosen according to the minimum value of Akaike Information criterion (AIC) and by using Ljung-Box test. To evaluate the validity of their proposed model, they used the MAPE evaluation metric between the observed and fitted HFRS incidence. Lastly, the fitted ARIMA model was used to predict the incidence of the years from 2009 to 2011.

An additional study by Dan et al. [108] forecast the mortality rate of malaria by applying SARIMA model, using a total of 216 data points for predicting malaria mortality rate for the period of 1996 to 2013, using Box–Jenkins methodology to build their ARIMA model. One-year ahead forecast was used. They selected the best model in accordance with the minimum values of AIC and Schwarz Bayesian Information Criterion (SBC). They concluded that ARIMA (1,1,1) × (0,0,1)12 model was the best model for forecasting the malaria mortality rate for the upcoming period of January 2014 to December 2014.

Villani et al. [109] where used the time series modelling to forecast the prehospital EMS demand for emergencies, and found that 41454 prehospital diabetic emergencies were attended within the period from January 2009 until December 2015. They recommended that the SARIMA (0, 1, 0) × (0, 1, 0)12 model was the best fit with a MAPE of 4.2%.

Singye and Unhapipat [110] used a time series to model and predict diabetes patient dataset obtained from Jigme Dorji Wangchuk National Referral Hospital (JDWNRH) of Bhutan by applying Box–Jenkins approach. The collected dataset was for the period from January 2006 to December 2016, and it was divided into two sub-datasets: the training dataset (January 2006 – December 2015) and validation dataset (January 2016 – December 2016). ARIMA was used to model diabetes patients' datasets. They evaluated different models of ARIMA using the Bayesian Information Criterion (BIC) and Ljung-Box Q statistics. They found that ARIMA (0, 1, 1) is the best model to describe and predict the future trends of diabetes incidence rate, which in turn help to properly plan and allocate resources for emergencies.

Sr. No.	Researcher Name and year	Methods/Techniques	Study purpose
1	Earnest et al. [106]	Autoregressive integrated moving average (ARIMA) models	Estimate the number of beds occupied during a severe acute respiratory syndrome (SARS) outbreak in Tan Tock Seng Hospital Singapore.
2	Liu, et al. [107]	ARIMA models	For prediction of haemorrhagic fever

			with renal syndrome (HFRS) in China
3	Dan et al. [108]	Applying SARIMA model using Box–Jenkins methodology to build their ARIMA model.	Forecast the mortality rate of malaria
4	Villani et al. [109]	Using SARIMA time series modelling	Forecast the prehospital EMS demand for emergencies
5	Singye and Unhapipat [110]	Using ARIMA time series modelling	Model and predict diabetes patient dataset obtained from Jigme Dorji Wangchuk National Referral Hospital

Table 2.3: Summary of time series forecasting models

2.6. Summary

This chapter provides a general background of diabetes, its main types, complications, and the related risk factors. The epidemiology of diabetes and its global prevalence were highlighted with a review of different studies that predict the incidence of diabetes and its global prevalence. In addition, this chapter presents the existing mathematical models involved in diabetes studies in terms of several aspects, including dynamics of glucose and insulin, the progression of diabetes and its complications, the cost of healthcare and the cost-effectiveness of interventions for treating or preventing diabetes.

The latter part of the chapter reviewed literature on adopted machine learning techniques in diabetes research. Different modelling approaches were presented such as regression, classification, and time series along with performance evaluation of the models using different statistical measures. Critical issues related to datasets, pre-processing strategies, and data imputation were explained by these studies.

Most of these studies were based on diagnosing and detecting diabetes at early stage or modelling the disease progression and complications; however, not much has been done in adopting any of the machine learning methods for studying the trends in the prevalence of diabetes and forecast its future burden using risk factors in specific populations.

In the next chapter, the theoretical background of modelling along with the methodological framework of the development process of the experimental design of diabetes modelling and the related issues are presented and discussed.

Chapter 3
Theoretical and Experimental Modelling of Diabetes

3.1. Introduction

As indicated in Chapter 1, this research was implemented using modelling strategies to study diabetes prevalence rate in KSA. Three different approaches of modelling were used: regression, classification, and time series modelling. This chapter presents a theoretical background of modelling and describes the methodologies involved in this study. In addition, this chapter briefly provided a description of the experimental datasets used in this thesis. The performance evaluation methods for comparing between models are also discussed.

A general definition of the word “model” has been given as “a simplified representation of a real-world situation used to help answer a specific question” [111]. A mathematical model according to Bender [112] is “…an abstract, simplified, mathematical construct related to a part of reality and created for a particular purpose”. Another definition of the mathematical model has been given by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Task Force on Good Research Practices - Modelling Studies, which defined it as “a logical mathematical framework that permits the integration of facts and values and that links these data to outcomes that are of interest to health-care decision makers” [12]. Briefly, the operational definition of a mathematical model for this study is: a model whose parts are represented by mathematical terminologies, for instance equations, variables, functions, constants, etc. [113].

3.2. Mathematical Modelling in Healthcare

Modelling can be used as a useful tool to support and manage several aspects in healthcare domain when appropriately applied, taking its limitations into consideration. Modelling in healthcare can be seen as a way of combining mathematical techniques and computational skills with expert knowledge (particularly of healthcare professionals) to come up with a convenient strategy for healthcare problems. Therefore, it could be said that the main aim of a model is to guide decision-makers in different areas involved in health and human life [11][111]. For example, models can help in predicting trends of disease prevalence and mortality rates, and assess the impact of alternative health policy scenarios, or be used to conduct a comparison of the effectiveness and cost-effectiveness of various treatments [114][115]. Moreover, models can provide an appropriate evaluation of the future burden of the diseases with regard to the history of the disease incidence and progression and modern developments of treatment [116]. Using models allows the combination of different data from

several local sources, utilising effectiveness data from trials, thus providing a helpful tool for making decisions. Also, policy makers can use models in the processes of assessing and comparing different alternatives of future policies and intervention strategies, which in turn contributes to resource allocation and appropriate planning [12][115].

One of the main useful aspects of modelling is the ability to deal with various types of inputs from several sources, and to detect logical relations between these inputs and desired outputs. Models allow the combination and integration of different types of data from a variety of sources such as prospective studies, expert opinions, assumptions, and trials. However, the diversity of data sources is linked to data of varying quality, which can lead to the uncertainty of parameters used in a model [12][117]. Therefore, models and their results should not be represented as statements of scientific fact, but as tools to aid decision making. Also, the results of the model should never be introduced as unconditional claims of effectiveness or estimates, instead, these outputs should be conditioned by input data and assumptions [12][115]. Moreover, modellers and decision-makers when examining the models' results must be more aware about the different biases that impacting individual opinions and trials. Models should be applied after they are mathematically validated and judged by experts. If the model is opposed to the real position (as perceived by experts), then it should be doubted, and its outcomes should be reviewed carefully. Despite these limitations, the role of modelling techniques is considerable in the field of modern healthcare [11][111].

3.3. Steps of Modelling Process

Although it is not possible to define a step-by-step modelling process to obtain a suitable model for solving a particular problem, it is possible to follow some general basic steps that can guide and help the process of modelling [111]. Importantly, the modelling process should not be a linear process; rather it is cyclical, with potentially overlapping steps of modelling. For example, at any step we might return to an earlier stage to revise and do some modifications, and then continue the process from that point [89][118]. Some guiding steps that one should expect during the modelling process are as follows.

A preliminary step in developing a model is understanding and identifying the research questions and objectives carefully [118][119]. In addition, data under study, intervention comparators, outcomes, and model perspectives should be defined [120]. Data under study involves all statistics and information about the target population, those who will influenced by the interventions [121]. Interventions depend on model objectives, which may pertain to any use of resources that are supposed to effect upon relevant outcomes for the population represented within the model. In healthcare, this might entail screening programmes, medical

techniques, and public awareness campaigns, etc. The outcomes of the model can be specific to a disease such as the incremental cost per QALY gained, or specific to the system being modelled, such as test failures, waiting times, or general costs and cost effectiveness [89].

Once the research problem has been defined, the determination of the model structure is the next important step in modelling process. This step is basically based on different factors, which includes the sort of question posed, the nature of the health event to be modelled, and the type and amount of data available to inform the model parameters [111][120]. It is worth mentioning that the level of complexity of a chronic disease model depends on the amount of data available; if the required data are limited, some assumptions might be necessary with regard to the technique of modelling a chronic diseases and natural history [114][122]. Generally, there are two main components in disease modelling, especially chronic diseases, which are: natural history of the disease, and various interventions that can be combined to answer policy questions [114][123]. It can be very complex to consider the natural history of the chronic diseases, whereas the model design should be as simple as possible [111].

Thus, it is recommended to simplify some of the factors, and to ignore others that do not seem as important if the decision can be made by a more comprehensive reasonable structure [12][118]. When developing a new model, it is good to have the skill of balancing between simplicity and complexity [124]. In addition, it is advisable to consider both the coherent theory of the modelled disease and the available evidence regarding causal linkages between variables when choosing the structure of a model as they should be consistent with this structure. However, this does not mean to prove that all causal linkages, as an example, showing that the effect size is statistically significant ($p<0.05$). Alternatively, this means that there is no conflict between the assumed linkages and the available evidence and they are consistent with most acceptable theories [12].

Data collection is the next stage after selecting the appropriate modelling structure. The flexibility of models allows to deal with different types of data. Moreover, data can give a description and information about the system’s behaviour, and suggestions about the important variables that should be considered within the model [111][118]. The required data for modelling process can be gained from various sources which differ in their quality and amount [12][117]. These sources may include databases, health records, clinical trials, observational studies, etc. expert opinions and assumptions can be also used if there is no reliable data on some variables [12][120]. Describing all data sources with their strengths and limitations is an essential task, and also any assumptions must be clearly addressed, as modelling outcomes are conditional by such data inputs and assumptions [12]. Data collection can be a limiting factor in model development and analysis for most problems in healthcare.

Even though this has been changed due to the development and implementation of computerized patient tracking, data confidentiality issues still cause data collection challenges, but patient data is much more stringently protected than other forms of consumer information (due to ethical obligation among healthcare professionals and systems) [111][125]. Even putting aside, the issue of patient data confidentiality, data collection in healthcare remains a resource-intensive task, which often requires preliminary surveys and population studies. These surveys can be quite expensive and time-consuming, and even on completion the data may be corrupted by survey bias [126][111].

After collecting the relevant modelling data, we need to move to the data processing (cleaning) step, where the modelling data can be prepared. It is often stated that the data cleaning and pre-processing takes more than half of the total project time of any inclusive data analytics study [127]. In fact, the information about model variables is often obtained from data collected for other objectives, and can result in data bias, inaccuracies, or errors. One more possible problem related to healthcare modelling is that the question or the case to be modelled might be too complicated to determine at first. In this condition, the modelling process can be started as a process of exploration, with a conceptual model of the problem as its outcome. It might be unclear at the beginning what data is suitable to describe the system. In these conditions and to improve model's quality, inclusive cleaning of the data is often required. Data cleaning can include data entry, checking data to avoid errors, determining bias within data, removing outliers, and removing duplicated data [111][119].

After the data processing step, we must identify variables and units. In this stage variables need to be distinguished and determined, in terms of whether they are dependent or independent, and random or deterministic [118][121]. For example, an independent variable is the variable on which others depend, and which is not affected by them. In many modelling applications, time is an independent variable. To simplify the model, some variables might be neglected, treated as constants, or be aggregated into one. While setting up the variables, we must also decide their units, such as days as a unit of time. Moreover, the relationships between variables should be established by drawing a diagram of the model if possible. Also, to make the model simple, we might suppose that some of the relationships are simpler than they really are.

For instance, we may make assumption that the relation between two variables is linear instead of a more complicated relation [118]. Another important aspect is determining the types of outcome parameters. This highly based on the nature of the modelled health event. For example, in some predicting models of diabetes, the main modelling outcome would be the forecasted prevalence of diabetes and the number of diabetics. Comparing to models of

other chronic diseases such as ischaemic heart disease, they might use different outcome parameters which consider quality of life or length of survival of patients, for example, life years gained (LYG), QALYs, and deaths prevented or postponed (DPPs) [128][115]. Once the relationships between variables have been established, then equations and functions for these variables can be determined. The usage of proportion and ratio, and rate of change, is essential in building up relations between variables to shape equations [121]. For instance, it may decide that two variables are relative to each other or finding that a known scientific equation or formula fits to the model. Several computational science models consist of differential equations, or equations involving a derivative [118].

When appropriate data becomes available, statistical analysis can be applied to study the system. This analysis can be applied by using some forms of statistical analysis that may include regression analysis, descriptive statistics, risk analysis, or a combination of these [111]. There are many other types of statistical analysis that can also be utilised, such as probability distributions, statistical inference, and hypothesis testing, etc [129]. The obtained results of the data analysis are used to define the most important variables for the problem, to examine a model's validity, and to set an accurate predictive model [130]. This analysis is a necessary task for dealing with uncertainty problems of some model variables, which resulted by for example, subjective estimates, assumptions, and uncertain values [11]. Statistical analysis is often helping a modeller check their basic assumptions concerning the system, so if they were incorrect this will force the modeller to take a step backwards and reshape another conceptual model for the problem. This might happen when a modeller decided that a variable assumed to be insignificant becomes significant or vice versa [111]. The process of the analysis work must represents an important part for all modelling studies [12].

After thoroughly understanding the problem and designing the model, the next stage is the model implementation, whereby the model outcomes will be calculated, and the predictions about the system under study will be produced. The implementation of the model may involve a computer (computer programs and packages), or might be performed using some analytical methods (algebra, calculus, and graphs) [111][118]. Computer simulation is a common way of using computer programs for modelling implementation [131]. When simulating a system, it is easy to address the behaviour of the model with no need to understand all the analytical details of the system. Therefore, simulation is often described as a black box; where input and output are apparent, but the way the output is produced may not be completely understood [132]. Computer simulation is one of the main forms of systems analysis that produces data under the instructions given by a model [133]. Simulation methods have their advantages and disadvantages. One of their most important features is that even highly complicated problems can be captured in a simulated model, with no need for detailed knowledge of the mechanics

of the system and without requiring mathematical expertise from the side of the modeller [111][134]. However, this shortage of transparency can lead to hiding logical errors in the model, often making incorrect outcomes [111][135].

A second way can be used for model implementation is mathematical analysis tools. When the modeller can use equations to describe the system, then numerical or analytical solutions can be adequate for solving the model with no need for simulation [111][118]. This presents several solid advantages compared to simulation. For instance, analytical methods provide accurate reproducible solutions without the demand of (often expensive) software. Moreover, by using the analytical methods it is often gain a deeper knowledge about the workings of a system. However, the obtained analytical outcomes of complex models are often hard or even not possible to achieve [111][136].

A third way that aggregates simulation and analysis is numerical analysis. In this approach, the modeller develops equations using mathematical techniques to perform the model, and then makes these equations as simple as possible [111][137]. The modeller can then use computers to calculate approximate solutions for these equations. This process keeps some of the power of mathematical analysis, and is particularly helpful when analytical method cannot be defined for solving particular problem [111][121]. Solutions obtained from models can gave an exact outcome or can gave a simulation of the event [118].

After a model is implemented, it needs to be tested for validation. This is an important step for a model to be accepted by decision makers and healthcare providers [138]. Valid outcome predictions of new patients should be provided by the prediction model. Basically, the dataset to develop a model is mainly concerned to learn for the future. Thus, validation is an essential part of the process of predictive modelling [119]. According to the ISPOR Task Force, model validation approaches can be classified into three main categories [12]. First, internal validation via internal testing and “debugging”. this type of validation can be carried out using null or extreme input values to examine if they produce the expected output values [114][119]. Between-model validation is the second type of validation, where the model is validated against other models dealing with the same problem (convergent validity). The third type is external validation, where the model outputs are compared with observed data [12][114]. However, models are based on the available evidence at the time they are developed. Therefore, as mentioned earlier, models should never be considered as perfect, unchangeable or statements of scientific facts [12].

A final step to consider is applying the prediction model. Once the final model is tested, tuned, and implemented, it can be applied to examine attributes of the system as illustrated by the model [111]. It is worth to mention that no matter how well a model is tested, tuned, and

implemented, it can only examine the aspects of the system it is prepared to study. When applying the model, we need to be careful that it is applied only in the appropriate ranges for the independent data. For example, the accuracy of the model may be limited for time periods of a few days and when applied to time periods of several years, the accuracy of the model will not be as it is [118]. There are many ways can be used for exploring the attributes of the system. For example, models usually have varying levels of sensitivity for input parameters; consequently, it is very useful to define which parameters have the most significant impacts on the system, which helps define where interventions could be made, and where further data collection efforts could be focused [111].

There is no need to accurately quantify parameters that exert negligible impacts on the system to the extent necessary for parameters that have great impacts on the system. Analysis that concentrates on determining which parameters have the major impact on the system is generally called sensitivity analysis. The process of the analysis work is an iterative, and the main modelling variables should be tested utilising different values or ranges [118][114][128][139]. This analysis is a necessary task for dealing with uncertainty problems of some model variables, related to subjective estimates, assumptions, and uncertain values [11]. This process is an important part of all modelling studies [12]. Two main types of sensitivity analysis can be utilised: one-way or univariate and multi-way or multivariate.

In the first main type of sensitivity analysis, one variable only can be tested at a time. After estimating the base-case scenario, the outcome variable is re-estimated holding all parameters constant aside from the one parameter chosen. The application of this method can be repeated as many variables in a model as required [140]. A common applied subtype of univariate sensitivity analysis is the “threshold analysis”. In this kind of analysis, there will be a changed in the size of one input parameter over a range, which will be followed by determination of the level above or below in which the conclusions change, that is the ‘threshold’ point where none of the alternatives is better than others. The usage of threshold analysis is more common in models with cost effectiveness analyses [140].

In the second main type of sensitivity analysis, multiple variables can be tested simultaneously. This type of sensitivity analysis can be two-way, three-way, or n-way analysis. For instance, varying values of a range for two parameters can be examined at the same time by two-way analysis. The assessed interventions for both parameters should be in common, lastly, assessed the impact of alteration on the outcomes of two mutually exclusive interventions [140]. One more common use of models is defining some type of optimal behaviour. For example, if a health ministry has a budget for a limited number of physicians, they might be interested to know where to deploy those physicians to obtain optimal patient

care. Generally, this type of question can be applied as a model using optimization method. This question can be usually written as “which selection of parameters minimizes the cost such that the desired result occurs?” Solving these types of questions has become a field in itself, and can be carried out by a number of different means [111][121].

3.4. Data Collection

As previously discussed, data collection is a fundamental step in the process of building models. This research requires the use of historical data on diabetes, smoking, obesity, and inactivity prevalence data for the starting year of modelling (1999), and for as many time points as possible thereafter, to achieve the research aim and develop the models. As previously mentioned, concerning data collection, modelling data can be obtained from various sources which differ in their quality and amount. The main sources of data were the published national surveys in KSA. Data for the prevalence of diabetes, smoking, obesity, and inactivity in KSA were obtained from Saudi Health Interview Survey [141], which is provided by the Saudi Ministry of Health, along with other published national surveys [142][143][144][145].

All these population-based studies were implemented at the national level, included all regions in KSA, with good sampling sizes and high rates of responses. Thus, they were more likely represent the population of KSA. These population-based national studies include adults aged 15 years and over. In addition, the diagnostic criteria used as a diabetes detection method was either WHO or ADA criteria. In this research, obesity as a risk factor was defined according to the definition of body mass index (BMI ≥ 30 kg/m^2); for smoking, only data for current smokers was taken; and for inactivity, inactive people were classified as those who did not meet the criteria for the “active” category (30 minutes or more of at least moderate to intensity activity for three or more times per week). The data is arranged according to age and gender (demographic factors). Data were divided into six ten-year age bands (25-34, 35-44, … 75$^+$ years old) for men and women. The data for men and women are hereinafter referred to as the “men data” and “women data” respectively.

3.4.1. Data Pre-Processing and Missing Data Imputation

As previously discussed, enhancing the quality of the models depends mainly on the quality of the data used for modelling, which refers to how suitable the data is to be used in relation to the number of samples, the importance of the features used in the analysis, and the existence of outliers in the dataset. Accordingly, data pre-processing is considered as a critical step in modelling processes [127][119]. As mentioned earlier, the main data inputs for modelling diabetes were the morbidity data (diabetes prevalence), and the prevalence of the three risk factors (smoking, obesity and inactivity). Data collection was undertaken using

published national surveys, which use credible, standardised, and validated measuring tools. However, the results of these studies were presented in different ways. For example, the age variable of the participants under study varied with respect to the whole age range and with respect to the age-group bands. Because of the deficiency and differences in data from KSA, it was necessary to make some sensible assumptions, and to perform a method to impute the missing data in order for the data to be prepared for the modelling processes.

As indicated earlier, it is normal for the most of models to employ variety of data inputs including assumptions from different sources [12]. In order to treat the overlap between the developed model's age groups with those used by studies, it was necessary to make some assumptions. For example, in some studies,[142] it was assumed that the prevalence rate for the age group 25-34 years is the average of the prevalence rates for the study age groups of 14-29 and 30-44 years, and so on; similar assumptions have been applied to data extracted in other studies, needed [144][145].

Another essential step is to solve the issues regarding to missing values, which is considered as an important part in mathematical modelling of data [146]. As there is no fixed standard method to deal with missing values,[147] some researchers tend to disregard these missing values, eliminate all attributes with missing values, or remove any record with missing values [148][149][150]. However, if the percentage of missing data is high, a recovering approach should be employed carefully [147]. Data imputation is the way where to deal with incomplete values problem. This way can be defined as the procedures to calculate an estimated values to replace the missing ones in the database, then generating a complete dataset [151]. Many approaches have been applied to treat this issue whether statistical or machine learning-based methods. An example of statistical technique is Bayesian inference or likelihood-based approach [152]. Another statistical way to solve this issue which have a large background is the use of mean/mode methods or those based on regression. However, these methods may lead to a bias in the estimated data [153]. Machine learning algorithms have been widely used to treat this issue and generate a smoother dataset. The performance of these algorithms varies depending on the system, datasets, and the pattern of missing values. One of the most useful aspect of using a machine learning model is the flexibility and higher order interaction among the missing values in the attribute [154].

Some of the most frequently used machine learning techniques deploy ANN, due to its ability to handle a large number of problems. Many researchers have used ANN to implement the imputation process [155][156][157].

Aydilek et al. [158] applied a hybrid method using support vector regression and genetic algorithm with fuzzy c-means clustering to calculate missing values. In their study they

implemented a comparison between the results of the proposed method with other different techniques. They concluded that better imputation accuracy is achieved by the hybrid approach. Another study by Kuppusamy et al. [154] developed a hybrid prediction model to impute the missing data in two mixed medical databases. In their study the constraint-based hybrid prediction model was designed using WLI fuzzy clustering and the grey fuzzy neural network. They combined the Grey Wolf Optimizer with the ANFIS network model, named the Grey Fuzzy Neural Network. The main aim of the proposed model was to define the optimal parameters for designing the membership function. They evaluated their experimental results using MSE and RMSE metrics. They reported that the performance of the proposed model was better than that of traditional methods, which they attributed to the hybrid behaviour of the clustering and neural network.

In this research, ANFIS method implemented in MATLAB software was used to impute all the missing data between the years from 1999 to 2025 for diabetes, smoking, obesity, and inactivity, and then to generate a smooth, completed dataset. The key advantages of ANFIS in comparison with other training algorithms is that the smoothness property and the more achieved learning capability [154]. ANFIS is a data-driven modelling approach for defining the behaviour of a complicated dynamical system [159]. The systematic aim of an ANFIS model is to generate unknown fuzzy rules from a specific input/output dataset [160].

In this research, to estimate the missing data, an ANFIS structure with two inputs and one output was constructed. For instance, the collected data of diabetes or smoking with their available year were taken as inputs, while the missing data by years that need to be predicted were taken as outputs. In order to train the ANFIS model, two Gaussian membership functions were used for the input variable, and for the output variable the type of membership function was linear. In addition, a hybrid method was applied, whereby the error tolerance was set as 0, and the number of epochs was set as 100. After imputing the missing values in the training set, the full dataset train again by running the same imputation method to predict the missing values in the testing set. This step was applied only for smoking, obesity, and inactivity data, while the expected percentage of diabetes was considered as a target variable when applying the purposed models. Finally, the complete dataset of smoking, obesity, and inactivity was divided into two parts: training data (from 1999-2013) and testing data (from 2014-2025), which are used for building and evaluating the model, respectively. Only the training dataset of diabetes (morbidity) data was required for building the purposed models, thus after imputing the missing data, a full dataset of the prevalence rate of diabetes from 1999 to 2013 was obtained, as shown in Table 3.1. Prevalence rates of smoking, obesity, and inactivity after generating all the missing data from 1999 to 2025 (training and testing data) are shown in Tables 3.2, 3.3, and 3.4.

Men morbidity data (training set)								Women morbidity data (training set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
1999	3.69	7.01	21.06	24.91	28.75	28.75	9.7	3	5.03	22.1	23.2	24.4	24.4	7
2000	4.10	7.15	21.76	26.89	30.38	30.38	9.8	3.21	5.42	22.26	24.50	26.08	26.08	7.1
2001	4.51	7.35	22.45	28.89	32.01	32.01	10.0	3.41	5.80	22.48	25.81	27.77	27.77	7.3
2002	4.93	7.62	23.15	30.88	33.65	33.65	10.2	3.62	6.19	22.77	27.12	29.47	29.47	7.5
2003	5.34	8.04	23.86	32.88	35.30	35.30	10.5	3.82	6.58	23.14	28.43	31.16	31.16	7.8
2004	5.75	8.68	24.56	34.88	36.94	36.94	10.9	4.03	6.97	23.59	29.74	32.86	32.86	8.1
2005	6.17	9.60	25.26	36.89	38.59	38.59	11.3	4.24	7.36	24.12	31.06	34.56	34.56	8.5
2006	6.58	10.74	25.96	38.89	40.23	40.23	11.8	4.45	7.75	24.69	32.37	36.26	36.26	9.0
2007	7.00	11.89	26.67	40.90	41.88	41.88	12.3	4.65	8.14	25.25	33.69	37.96	37.96	9.4
2008	7.41	12.82	27.37	42.91	43.53	43.53	12.7	4.86	8.54	25.78	35.00	39.67	39.67	9.8
2009	7.83	13.46	28.08	44.92	45.19	45.19	13.1	5.07	8.93	26.24	36.32	41.37	41.37	10.2
2010	8.25	13.88	28.78	46.93	46.84	46.84	13.4	5.28	9.32	26.61	37.64	43.08	43.08	10.5
2011	8.66	14.16	29.49	48.94	48.50	48.50	13.6	5.48	9.71	26.91	38.97	44.79	44.79	10.7
2012	9.08	14.36	30.20	50.96	50.16	50.16	13.8	5.69	10.11	27.13	40.29	46.50	46.50	10.9
2013	9.5	14.5	30.9	53	51.8	51.8	13.9	5.9	10.5	27.3	41.6	48.2	48.2	11.00

Table 3.1: Prevalence rate of diabetes (training data) for men and women

Men smoking data (training set)								Women smoking data (training set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
1999	14.2	13.3	9.2	7.5	7.1	7.3	21.1	0.6	0.7	1.1	0.7	0.7	0.5	0.9
2000	16.5	14.5	10.5	7.79	7.18	7.37	21.4	0.7	0.8	1.26	0.7	0.73	0.53	0.98
2001	18.8	16.9	11.8	8.3	7.29	7.47	21.71	0.8	0.9	1.39	0.71	0.77	0.58	1.05
2002	21.1	20.3	13.2	9.06	7.43	7.61	22.08	0.9	1.03	1.55	0.74	0.82	0.65	1.13
2003	23.5	23.8	14.8	10.1	7.62	7.78	22.53	1.03	1.13	1.73	0.79	0.89	0.73	1.23
2004	25.8	26.1	16.7	11.5	7.84	7.99	23.04	1.13	1.23	1.93	0.9	0.98	0.84	1.35
2005	28.1	27.4	19.2	13.1	8.11	8.23	23.7	1.3	1.4	2.2	1.1	1.09	0.96	1.5
2006	28.1	27.6	20.6	15.1	8.39	8.5	24.16	1.3	1.4	2.34	1.42	1.2	1.1	1.6
2007	28.2	27.9	22.4	17.1	8.68	8.76	24.7	1.35	1.45	2.52	1.88	1.31	1.23	1.71
2008	28.6	28.4	23.9	19	8.94	9	25.17	1.37	1.47	2.68	2.43	1.41	1.35	1.81
2009	29.8	29	25.1	20.7	9.17	9.22	25.56	1.37	1.47	2.79	2.98	1.5	1.46	1.89
2010	30.9	29.7	25.9	22.1	9.36	9.39	25.84	1.35	1.45	2.87	3.45	1.57	1.55	1.94
2011	31.3	30.2	26.4	23.2	9.51	9.52	26.04	1.32	1.42	2.91	3.8	1.63	1.61	1.98
2012	31.4	30.5	26.7	24	9.62	9.63	26.17	1.28	1.38	2.93	4.04	1.67	1.66	2
2013	31.4	30.7	26.5	24.7	9.7	9.7	26.2	1.1	1.3	3	4.2	1.7	1.7	2.1
Men smoking data (testing set)								Women smoking data (testing set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
2014	32.3	31.2	27.3	25.1	10.2	10.1	26.44	1.5	1.5	3.4	4.5	1.8	1.7	2.40
2015	32.8	31.6	27.7	25.4	10.4	10.3	26.62	1.5	1.5	3.5	4.7	1.9	1.8	2.50
2016	33.4	32.1	28.2	25.7	10.6	10.5	26.80	1.6	1.6	3.7	5.1	2	1.9	2.70
2017	33.9	32.6	28.6	25.9	10.8	10.7	26.97	1.6	1.6	3.8	5.4	2.1	2	2.80
2018	34.4	33.0	29.0	26.0	11.04	10.9	27.15	1.7	1.7	3.9	5.7	2.2	2.1	2.90
2019	35.0	33.5	29.5	26.2	11.3	11.1	27.33	1.7	1.7	4.1	6.02	2.2	2.2	3.00
2020	35.5	34.0	29.9	26.3	11.5	11.3	27.50	1.8	1.8	4.3	6.3	2.3	2.3	3.10
2021	36.0	34.4	30.4	26.3	11.7	11.5	27.68	1.8	1.8	4.4	6.6	2.4	2.4	3.20
2022	36.6	34.9	30.8	26.4	11.9	11.7	27.86	1.8	1.8	4.5	6.9	2.5	2.5	3.30
2023	37.1	35.4	31.2	26.5	12.1	11.9	28.03	1.9	1.9	4.7	7.3	2.6	2.6	3.50
2024	37.7	35.8	31.7	26.6	12.3	12.1	28.21	1.9	1.9	4.8	7.6	2.7	2.7	3.60
2025	38.2	36.3	32.1	26.6	12.6	12.3	28.39	1.9	1.9	4.9	7.9	2.7	2.7	3.70

Table 3.2: Prevalence rate of smoking (training and testing) data for men and women

Men obesity data (training set)								Women obesity data (training set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
1999	12.5	17.6	17.8	16.4	16.4	16.4	13.1	20.05	31.5	32.4	28.7	28.7	28.7	20.3
2000	12.9	19.5	20.6	18.3	16.7	16.7	15.4	20.4	31.6	33.8	29.9	29.2	29.2	21.6
2001	13.4	21.0	22.8	20.0	17.2	17.2	17.4	21.0	32.2	36.3	32.1	29.8	29.8	24.0
2002	14.1	22.8	24.8	21.6	17.8	17.8	19.3	21.7	35.4	40.0	35.3	30.6	30.6	27.6
2003	15.0	24.8	26.7	23.2	18.6	18.6	21.1	22.7	42.0	44.3	39.1	31.7	31.7	31.8
2004	16.1	27.8	29.1	25.0	19.7	19.7	23.2	23.8	45.2	48.1	42.5	33.0	33.0	35.4
2005	17.3	29.2	29.9	26.2	20.8	20.8	24.5	25.2	45.9	50.7	44.9	34.5	34.5	37.9
2006	18.7	31.4	31.3	27.6	22.1	22.1	26.0	26.6	45.9	52.2	46.7	36.1	36.1	39.5
2007	20.1	33.4	32.5	29.0	23.3	23.3	27.5	28.1	46.0	53.2	48.4	37.8	37.8	40.6
2008	21.4	35.0	33.5	30.3	24.5	24.5	28.9	29.5	46.2	54.1	50.6	39.3	39.3	41.8
2009	22.5	36.2	34.4	31.6	25.5	25.5	30.2	30.6	46.9	55.0	53.7	40.6	40.6	43.1
2010	23.4	37.0	35.1	32.7	26.3	26.3	31.5	31.6	48.5	55.9	57.1	41.7	41.7	44.6
2011	24.1	37.2	35.6	33.9	27.0	27.0	32.6	32.3	49.3	56.8	60.1	42.5	42.5	45.9
2012	24.6	37.3	35.9	35.0	27.4	27.4	33.7	32.9	49.5	57.3	62.0	43.1	43.1	46.7
2013	25	37.4	35.9	35.9	27.8	27.8	34.5	33.3	49.5	57.6	63.1	43.6	43.6	47.2
Men obesity data (testing set)								Women obesity data (testing set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
2014	25.4	37.8	38.1	36.6	28.3	28.3	35.1	33.7	50.2	58.2	63.6	44.2	44.2	48.1
2015	25.8	38.1	38.5	37.5	28.8	28.8	36.0	34.2	50.8	58.8	64.7	44.9	44.9	49.0
2016	26.3	38.4	39.0	38.5	29.3	29.3	36.9	34.7	51.3	59.4	65.7	45.6	45.6	50.0
2017	26.8	38.6	39.4	39.4	29.9	29.9	37.8	35.2	51.9	60.0	66.6	46.3	46.3	50.9
2018	27.3	38.9	39.8	40.4	30.4	30.4	38.7	35.7	52.5	60.7	67.5	47.0	47.0	51.9
2019	27.8	39.2	40.2	41.4	30.9	30.9	39.6	36.2	53.1	61.3	68.4	47.7	47.7	52.8
2020	28.2	39.5	40.7	42.3	31.4	31.4	40.5	36.7	53.6	61.9	69.2	48.4	48.4	53.7
2021	28.7	39.8	41.1	43.3	32.0	32.0	41.4	37.2	54.2	62.5	70.0	49.1	49.1	54.7
2022	29.2	40.1	41.5	44.2	32.5	32.5	42.3	37.7	54.8	63.1	70.9	49.8	49.8	55.6
2023	29.7	40.3	42.0	45.2	33.0	33.0	43.2	38.2	55.4	63.7	71.7	50.5	50.5	56.6
2024	30.2	40.6	42.4	46.2	33.6	33.6	44.1	38.7	55.9	64.4	72.5	51.2	51.2	57.5
2025	30.6	40.9	42.8	47.1	34.1	34.1	45.0	39.3	56.5	65.0	73.3	51.9	51.9	58.4

Table 3.3: Prevalence rate of obesity (training and testing) data for men and women

Men inactivity data (training set)								Women inactivity data (training set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
1999	89.51	91.31	94.61	96.71	97.61	97.61	93.9	97.91	97.91	98.01	98.61	99.58	99.58	98.1
2000	89.02	89.94	93.42	95.38	96.69	96.69	93.21	94.68	93.12	94.57	96.12	99.62	99.62	96.36
2001	87.76	88.24	91.76	93.69	95.38	95.38	91.89	92.63	89.59	92.01	94.61	99.64	99.64	94.83
2002	85.88	86.04	89.51	91.44	93.59	93.59	89.99	90.34	86.42	89.25	92.94	99.58	99.58	93.04
2003	83.26	83.27	86.57	88.57	91.26	91.26	87.41	87.55	83.62	86.34	90.97	99.35	99.35	90.98
2004	79.85	79.95	82.93	85.07	88.38	88.38	84.13	83.97	81.17	83.41	88.52	98.75	98.75	88.68
2005	76.2	75.9	78.7	80.9	85.02	85.02	80.5	77.4	77.6	78.7	84.03	97.46	97.46	85.5
2006	71.17	72.18	74.09	76.75	81.39	81.39	75.92	77.16	77.37	78.22	82.08	95.45	95.45	83.87
2007	66.45	68.27	69.49	72.49	77.76	77.76	71.53	76.69	76.01	76.37	79.79	93.43	93.43	81.70
2008	61.95	64.72	65.21	68.59	74.39	74.39	67.38	75.98	75.02	75.19	79.02	92.13	92.13	79.90
2009	57.95	61.73	61.51	65.26	71.49	71.49	63.74	75.07	74.38	74.69	79.23	91.52	91.52	78.55
2010	54.59	59.38	58.50	62.60	69.14	69.14	60.71	74.17	74.11	74.77	79.83	91.28	91.28	78.00
2011	51.89	57.63	56.17	60.59	67.33	67.33	58.32	73.45	74.20	75.31	80.58	91.22	91.22	77.50
2012	49.79	56.41	54.45	59.14	66.00	66.00	56.48	72.99	74.65	76.17	81.35	91.24	91.24	77.20
2013	48	55.7	53.2	58.2	65	65	55	72.7	76.1	78	82.5	91.3	91.3	77.3
Men inactivity data (testing set)								Women inactivity data (testing set)						
Years	25-34	35-44	45-54	55-64	65-74	75+	Total	25-34	35-44	45-54	55-64	65-74	75+	Total
2014	46.94	54.73	52.08	57.15	64.14	64.14	54.1	72.87	72.75	75.03	80.13	89.98	89.98	76.6
2015	45.98	53.56	50.50	55.80	62.91	62.91	53.3	72.57	70.97	73.39	78.82	89.29	89.29	76.1
2016	45.23	52.40	48.93	54.45	61.68	61.68	52.7	72.26	69.19	71.75	77.5	88.61	88.61	75.7
2017	44.62	51.24	47.36	53.10	60.46	60.46	52.3	71.96	67.42	70.12	76.18	87.92	87.92	75.3
2018	44.11	50.08	45.79	51.75	59.23	59.23	51.9	71.66	65.64	68.48	74.86	87.24	87.24	74.8
2019	43.67	48.92	44.22	50.41	58.00	58.00	51.6	71.36	63.85	66.85	73.55	86.55	86.55	74.4
2020	43.28	47.75	42.65	49.06	56.78	56.78	51.4	71.06	62.08	65.21	72.23	85.87	85.87	73.9
2021	42.92	46.59	41.08	47.71	55.55	55.55	51.2	70.76	60.3	63.57	70.91	85.19	85.19	73.5
2022	42.59	45.43	39.50	46.36	54.32	54.32	50.9	70.45	58.52	61.94	69.6	84.5	84.5	73
2023	42.27	44.27	37.93	45.01	53.10	53.10	50.8	70.15	56.74	60.3	68.28	83.82	83.82	72.6
2024	41.97	43.10	36.36	43.66	51.87	51.87	50.6	69.85	54.96	58.66	66.96	83.13	83.13	72
2025	41.67	41.94	34.79	42.31	50.64	50.64	50.5	69.55	53.19	57.03	65.65	82.45	82.45	71.7

Table 3.4: Prevalence rate of inactivity (training and testing) data for men and women

3.4.2. Data Categorisation

In classification modelling it was required to discretise the continuous variable (output values) into multiple categories. These categories were chosen and assigned numerical classes: low (1), medium low (2), medium (3), high (4), and extremely high (5). MATLAB software was used to implement this transformation of continuous data, and the discretise function was applied

to group the morbidity data (training data from 1999 to 2013, along with the predicted morbidity data from 2014 to 2025).

3.4.3. Data Analysis

Building a good predictive model requires a good choice of input variables. The choice of these inputs should be such that the functions of the model perform accurately between the inputs and outputs. The predictor variables comprised age, gender, smoking, obesity, and inactivity. The predictor behavioural variables smoking, obesity, and inactivity were collected according to age and gender (demographic variables). These data were divided into six ten-year age bands (25-34, 35-44, … 75^{+} years old) for men and women. The morbidity data of diabetes was considered as the response variable. All these data were prepared for the models as shown in Tables 3.1, 3.2, 3.3, and 3.4. Moreover, correlation analysis was carried out to determine the correlation between the variables in the data in terms of statistical significance, Table 3.5 shows that demographic as well as behavioural risk factors significantly contributed to the increased level of diabetes, with a significance level of 0.05; however, smoking, obesity, and physical inactivity were the most significant factors.

Variables	P-value
Gender	0.02
Age	0.01
Smoking	0.000
Obesity	0.001
Inactivity	0.001

Table 3.5: Relationship between diabetes prevalence and the related risk factors with P-value

3.5. Modelling Approaches

This section discusses the developed machine learning modelling approach used in this thesis. Three types of modelling were applied: regression, classification, and time series. Different machine learning models for regression and classification were developed. This section starts with an overview of the concept of machine learning and the main types of learning, with a concentration on the supervised learning, which is the umbrella category of all modelling approaches considered in this research.

3.5.1. Machine Learning

Machine learning (ML) is one of the most important and powerful sub-branches of artificial intelligence (AI), and it is closely related to statistics, thus statisticians and mathematicians commonly call it “statistical learning”. It is concerned with the development of algorithms and

techniques that allow machines to learn and gain intelligence based on the past experience [161][162]. The power and effectiveness of these techniques derived from the ability to identify and understand the input data by extracting patterns and creating models, then making decisions and predictions based on it [163]. Basically, ML algorithms rely on trial and error, which is quite different from the traditional algorithms directed by programming, based for example on if-else decision statements [17]. The training set is the name given to the dataset that is subjected to the learning process. Different types of learning can be applied, depending on the range of the study and the nature of the data in the training set [164]. There are four main categories of ML tasks, namely supervised, unsupervised, active, and reinforcement learning [17]. When the dataset includes one or more target variables (outputs) where the analyst aims to describe their current behaviour and assess future behaviour employing other variables (inputs) from the dataset, this is supervised learning.

The target variable can be called a label, dependent variable, or response variable. If there are no target variables in the dataset, this will be an unsupervised learning type. Contrastingly, when the target variable is partly available (i.e., both labelled and unlabelled data are utilised in the training process), this is known as semi-supervised learning [164]. There is also a third type of learning known as active learning, in which the most informative sample can be chosen as a label to train the model. Reinforcement learning is the fourth type, where the system interacts directly with a dynamic environment. It is fundamentally used in independent systems, because of its independence relation to its environment [163]. However, the most common applied type in real-world applications, especially in disease prediction and diagnosis, is the class of supervised ML. In this study, the development of multiple supervised ML models was proposed deploying different algorithms for the prediction of DM. In the next sections, the supervised learning approaches applied in subsequent chapters of this thesis are briefly discussed.

3.5.2. Supervised Learning Methods

As previously mentioned in the above section, when the target variable(s) are included in the dataset, the type of learning will be a supervised learning. The family of supervised learners have various methods which vary from each other in different perspectives. There are two kinds of learning tasks in supervised learning: regression and classification. The type of prediction will be regression if the target variable takes continuous values, and it will be classification if it takes discrete values [163][165]. The following sections give a general overview of the concept of regression and classification and associated techniques applied in later chapters.

3.5.2.1. Regression Modelling

Regression techniques are the most common methods of supervised machine learning. These methods aim to predict or describe a specific numerical value depend on a set of past data [166]. Montgomery et al. defined regression analysis as a statistical process that attempts to describe and model the relationship between a dependent variable and one or more corresponding value(s) of other variables [167]. These processes of establishing a statistical model demonstrate the mathematical explanation of a variable based on other variables. Thus, the main objective of regression analysis is to find a function to estimate the change of the response variable according to the change in one or more predictors. This regression function is defined by a finite number of unknown parameters, so the regression analysis mainly aimed to estimate these parameters, this will be based on the observed pairs of X and Y, to shape the regression equation that measures the covariate impacts on the data. This procedure is also known as fitting the model to the data. Linear regression is the simplest method, where the mathematical equation of the line is used to model a dataset, this model has a single regressor X which has a straight-line relationship with a response Y:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.1}$$

where the intercept β_0 and the slope β_1 are unknown parameters; and ε is the random error element. These parameters are generally known as regression coefficients. There is a simple and often helpful interpretation of these coefficients.

The slope β_1 is defined as the process when the mean of the distribution of Y is changed according to a unit change in X. If $X = 0$ is included in the range of data on X, then β_0 (the intercept) is calculated as the mean of the distribution of the response Y when $X = 0$. If the $X = 0$ is not included in the range of X , then there is no practical interpretation of β_0 [167].

Regression techniques range between simple, such as linear regression; to complex, such as polynomial regression, decision trees, neural nets, random forest regressions and support vector regression [168]. In this research five methods of regression modelling were applied: multiple linear regression model, Bayesian linear regression model, support vector regression, ANFIS model for regression and ANN regression model. A full description of these models with its implementation and its results is provided in Chapter 4.

3.5.2.2. Classification Modelling

Classification methods are another popular class of supervised machine learning which predict or determine a class value. As mentioned earlier, if the outcomes take discrete values, known as categories, then the process of classification follows [165]. According to Aggarwal,

the classification problem can be expressed as: “Given a set of training data points along with associated training labels, determine the class label for an unlabelled test instance” [169]. Both structured and unstructured datasets can be used for implementing classification models. Dataset samples are classified in accordance with specific label or category and for unseen inputs, the label or category that will be determined to it is predicted by the same technique. A classifier algorithm is “an algorithm that learns from the training set and then assigns new data point to a particular class” [170]. A classification predictive model determines some proper mapping function from training dataset and estimate the class label by the aid of the mapping function for the new inputs data.

The classification task can be in different forms. When the classification problem has two possible outcomes then this will be a binary classification, such as disease diagnosis (“infected” or “not infected”). Multi-label classification is another type of classification task where there are more than two possible outcomes. For example, obesity classification according to BMI (body mass index) into classes such as low-risk, moderate-risk, and high-risk [170]. Some of the most common applied methods in classification problems are SVM methods, probabilistic, rule-based, instance-based, neural networks, and decision tree methods [169]. In this research some classification models were applied using classification learner in MATLAB. An overview of these classifiers with their implementation and results is given in Chapter 6.

3.5.2.3. Time Series Modelling

A time series can be defined as a sequence S of past measured data Y_t of a variable Y observed over equal periods of time t. The first observation available on Y is $t = 1$, and $T = t$ will be the last. The observation period is denoted by the completed set of times $t = 1,2, \dots \dots, T$ [171]. The measurement of the observations is usually taken at equally spaced intervals, such as every minute, hour, or day, etc., thus the order of observation arrival is substantial [172]. A definition of time series analysis is “the endeavour of extracting meaningful summary and statistical information from points arranged in chronological order to diagnose past behaviour as well as to predict future behaviour” [173].

There are various objectives of using time series such as to forecast the future based on the past knowledge, to understand any phenomenon behind the measures, or clearly describe the prominent characteristics of the series. Predicting the future using observed time series is increasingly important in the most vital fields of science and engineering, such as medicine, economics, business, finance, and telecommunication. The horizon size is a significant aspect of the forecasting process. If it is challenge to forecast one-step of a time series, it would be more difficult to perform a multi-step forecasting, this due to further obstacles, such as errors

accumulated, increased uncertainty, and reduced accuracy [171]. Time series analysis has the ability to deal with statistical methods in order to analysis and model an order sequence of observations.

This modelling process leads to stochastic process model for the system that produced the data. In traditional statistical analysis, it is often neglecting the correlation of data in time. For example, in regression analysis the hypothesis that has been made for serial uncorrelated residuals is often disregard in practical applications. However, appropriate techniques and models for time series analysis may often be considered as an easy extended way of linear regression analysis where past observations of the target variable are involved as a descriptive variable in a simple type of linear regression model [174]. Some of the most common classical statistical models for time series include Autoregressive (AR), moving average (MA), ARIMA , and vector autoregression (VAR) models [173][172].

These traditional models have been widely applied for time series forecasting, and they are still used in many tasks, from academic studies to industrial modelling. In the last two decades, machine learning methods have been greatly involved to address these predictive tasks. These methods, also known as data-driven based models or black-box models. In these models only historical data were used to examine the pattern of stochastic dependency between the past and the future [171]. Examples of these models are nearest neighbour regression, support vector machines, and decision trees. Moreover, ANNs have been successfully developed for modelling and forecasting nonlinear time series and it was performing better than the classical statistical methods [175][176]. In this thesis, Neural Networks has been also applied for time series modelling. A description of this strategy, its implementation, and its results is presented in Chapter 7.

3.6. Model Evaluation

This section describes the different performance metrics that were used to evaluate the performance of the proposed models which applied in the later chapters. Evaluating the performance of the model is the last and important step in modelling development [177]. During this step, the developed model is examined through the collected datasets and the performance evaluation metrics will define if the model is optimized and how reliable and robust the outcomes are, then it can be applied to predict new real data [178]. One of the most important tasks in evaluating a model is defining the objectives of the developed model in which its performance will be assessed. Furthermore, it is essential to use more than one performance evaluation measure, this is because that the model that needs to be evaluated may have more than one objective, or it might be evaluated by different performance

measures used by the end user. In addition, using multiple evaluation methods will help to take advantages of all the important features of the model and minimise its limitations [179].

A variety of different metrics have been applied in the literature to evaluate the outcomes of prediction models and rank them appropriately [17][14][146][154]. These metrics have their own advantages and disadvantages in explaining the outcomes of the compared models. In this study, the metrics were chosen in accordance with relevance, explain ability, and popularity in both statistics and machine learning. Consequently, in order to validate the proposed models and reach reliable and accurate outcomes, different measures have been used according to the type of models. For regression models, the prediction accuracy is evaluated based on MSE, RMSE, MAPE, and the coefficient of determination R^2. While for classification models there are several methods for evaluating the model's performance, this study compared classifier performance in terms of accuracy, which is generally the most important consideration in the healthcare field.

3.6.1. Mean Square Error

MSE is the most popular and simple interpreted metric for many types of regression models [151]. It measures how close a regression line is to a set of data points. This can be calculated by taking the distances (errors) from the points to the regression line and then calculate their square values [180]. It is substantial to square them to eliminate any negative indications and it is also helping to allow more weight for considerable differences. It is known as the mean squared error where this stands for its way of calculating the average of a set of errors. The lower the value of the MSE, the closer the fit of the regression line to the data, resulting in better forecasting. The MSE is expressed by the following equation:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \tag{3.2}$$

where *n* the number of data points; Y_i is the actual values; and $\hat{Y}_i$ is the predicted values.

3.6.2. Root Mean Square Error

RMSE is another popular and excellent error metric for numerical predictions. It measures the accuracy of models by taking the square root of MSE between the actual and predicted output [181]. It is sensitive to the outliers as it is scale-dependent, and it is also affected by the larger errors. Lower RMSE values indicate better model performance [151]. RMSE is presented in the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{3.3}$$

where *n* is the number of data points, Y_i is the actual values, and $\hat{Y}_i$ is the predicted values.

3.6.3. Mean Absolute Percent Error

MAPE is another common evaluation metric, because it is simple to calculate and easy to understand. It can be defined as the mean or average of the absolute percentage errors of predictions [182]. It can be calculated by taking the summed average of the absolute percentage errors (the actual values minus the predicted values divided by the actual) and then divided by the number of samples. This measure can be a very good indication of the quality of the evaluation method, and it is easy to understand for a wide range of users. because it calculates the error in terms of percentages [183]. In addition, because using the absolute value, any problem with positive and negative errors will be prevented. The MAPE calculation is given by the following equation:

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i} * 100}{n} \tag{3.4}$$

where *n* the number of data points; Y_i is the actual values; and $\hat{Y}_i$ is the predicted values.

3.6.4. Coefficient of Determination

The coefficient of determination (R^2) is a statistical metric that measure how well the data fits the regression model by indicating the deviation of the predicted values from the regression line. The R^2 value is normally between 0 and 1. A value close to 1 indicates that the model is perfectly fits the data, while a low value or close to 0 implies a poor fit of the model. It is scale-independent, and it is sensitive towards the variance in observations [184]. The coefficient of determination (R^2) is provided by the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}\,)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

where *n* is the number of data points; y_i is the actual values; $\hat{y}$ is the predicted values; and $\bar{y}$ is the mean (average) of the actual values.

3.6.5. Accuracy Rate

There are several methods to evaluate the performance of machine learning classification models. Accuracy rate is one of the most important metrics in performance evaluation of classification models. It is very common measure for evaluating the success of a classifier and is used in many different studies. The accuracy rate of a model can be defined as the proportion of correctly classified predicted cases to the total number of predicted cases. The accuracy rate is defined according to the following equation [17]:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{3.6}$$

3.7. Simulation Software

All the analyses and computations in this thesis were carried out using MATLAB software. This software was chosen because it is a proprietary high level programming language and is considered as one of the most popular programs for scientific, numerical computing. The name MATLAB is a portmanteau of “Matrix” and “Laboratory”. As illustrated by the name, it relies on interactive systems of matrixes. The implementation of all models in MATLAB (version 2018a) will be explained in the following chapters.

3.8. Summary

This chapter overviewed the theoretical background of modelling in terms of definitions, uses in healthcare, and developmental steps. A description of the experimental datasets used in this thesis, with the processes of preparing the data and dealing with missing values, was also presented. The statistical and machine learning methods and procedures adopted while carrying out the research were described. Lastly, various methods of model performance evaluation that are used for comparison were discussed. An explanation of the developed models with their implementation and results is provided in the following chapters.

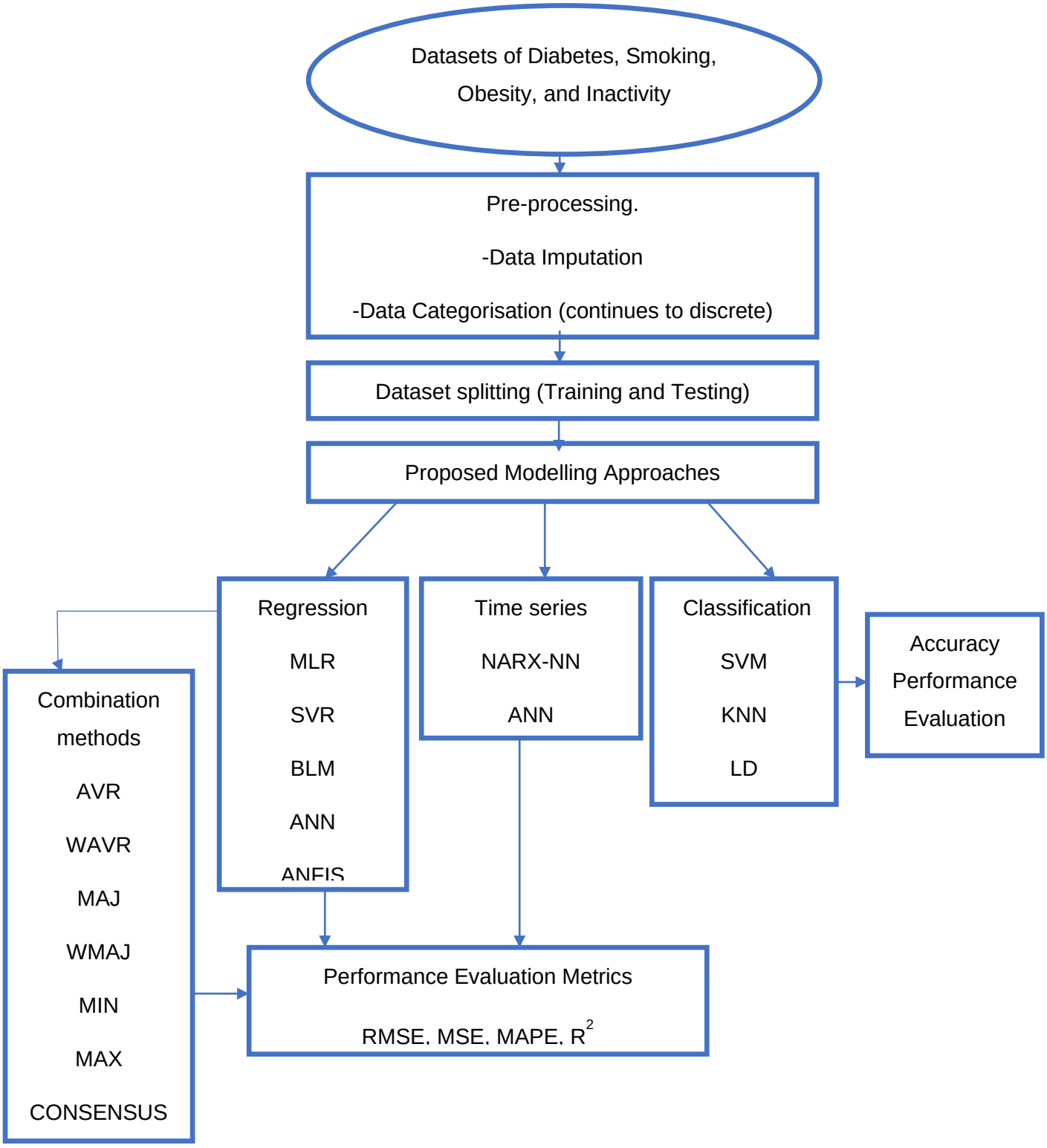

Figure 3.1: Proposed workflow for the research.

Chapter 4
Regression Modelling

4.1. Introduction

As mentioned in Chapter 3, three types of modelling approaches were applied while carrying out this research: regression, classification, and time series. This chapter specifically describe the first type, regression modelling. It briefly describes the five machine learning models developed for regression, including Multiple Linear Regression, Bayesian Linear Regression, Support Vector Regression, Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference Model (Section 4.2), to highlight their operational properties. The implementation of these individual models is illustrated in Section 4.3. Section 4.4 analyses and discusses the results of these models and compares their performance in terms of various evaluation performance metrics. A summary of the chapter is given in Section 4.5.

4.2. Machine Learning Regression Methods

This section gives a brief description of the five machine learning models used for regression modelling of diabetes disease forecasting: Multiple Linear Regression, Bayesian Linear Regression, Support Vector Regression, Artificial Neural Networks and Adaptive Neuro - Fuzzy Inference Model. This is mainly to present the mathematical background of these models and to highlight their individual operational qualities.

4.2.1. Multiple Linear Regression

Multiple linear regression is one of the most common types of linear regression analysis. It is an extended form of simple linear regression, with a relationship between more than two variables [185]. In predictive analysis, this technique describes the relationship between one dependent (response) variable and two or more independent (predictor) variables. The general model of multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (4.1)$$

where Y is the dependent variable; $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_n$ are the coefficients; and $X_1 + X_2 + \cdots + X_n$ are the independent variables.

4.2.2. Adaptive Neuro-Fuzzy Inference System

The ANFIS model is a combined model of fuzzy systems and ANN [186]. The main parts of the FIS are fundamental rules, which contain the choices of fuzzy logic rules “If-Then”, a set

of membership functions, and the fuzzy logic inference procedures from the fundamental rules to obtain the output. In order to map the inputs with the outputs, two common fuzzy inference systems (FIS) can be employed in different applications: Mamdani and Sugeno inference systems.

The fuzzy rules in the two inference models give different results, therefore their actions of defuzzification and combination are also different. However, the Sugeno system is believed to be computationally more efficient than the Mamdani; in the former, the resultant parameter is a linear equation or constant coefficient. Supposing that we have a system including two inputs, x and y, and the output is f, and the based rule has two fuzzy if-then rules, then the description of rules for the linear equation Sugeno FIS can be presented as the rule 1 (R1) and rule 2 (R2):

$$\text{R1: if } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ then } f_1 = p_1 x + q_1 y + r_1 \tag{4.2}$$

$$\text{R2: if } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ then } f_2 = p_2 x + q_2 y + r_2 \tag{4.3}$$

where $A_i$and B_i are the membership functions of each input x and y; and p_i, q_i and $r_i$are the linear parameters in the resulting part of the Sugeno fuzzy inference system.

ANFIS model can be considered successful due to the strength of its results. Moreover, as with other machine learning techniques and as a neural network, ANFIS has a high ability to generalise. On the other hand, there are some limitations of ANFIS model regarding the type, number, and position of membership functions [187].

4.2.3. Artificial Neural Networks

Neural Network and ANN are mathematical models based on the concept of Artificial Intelligence, which simulates the biological neuronal activity of the human brain. This modelling approach is a valuable tool that simulates the functionality of the human brain when dealing with complex relations between the inputs and outputs in any systems [188]. There are many types of ANN architectures, the most common of which is Multi-Layer Perceptron (MLP), which is commonly used for prediction. It comprises three tiers: an input layer, hidden layers, and an output layer. Supposing that the input vector is $\vec{x}$ and the weight vector is $\vec{w}$, and the activation function is a sigmoid function (which is the most commonly used function type), the output is given by:

$$Y = sigmoid(\vec{w}^T.\vec{x}) \quad (4.4)$$

where the sigmoid(x) is

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (4.5)$$

One of the characteristic advantages of Neural Network technique is its ability to deal with noisy, incomplete, or missing data, requiring no previous assumptions. In addition, it has capabilities to deal with complex relations between input and output variables, and consequently to predict the output of new data input. However, overfitting and overtraining are considered as limitations of Neural Networks. Also, regarding the selection of parameters, in Neural Network there is no formal way to select the suitable parameters for the model, which may influence the accuracy of its prediction.

4.2.4. Support Vector Regression

Support Vector Machine (SVM) is a popular method developed by Vapnik [189]. The generalized concepts of SVM have been applied to regression problems such as modelling and prediction, and accordingly called Support Vector Regression (SVR). SVR has been effectively utilized to deal with forecasting issues in many areas as diverse as pharmacology, economics, and power systems analysis. SVR is less popular than SVM, but it has been verified that it is a valuable technique in estimating the real value of a function [190]. One of the most useful features of SVM is that the complexity of its computation does not rely on the dimensional parameters of the input space. Moreover, SVR shows better generalization ability, with high performance and accurate prediction. Fundamentally, SVR is a linear approach with one output, dealing with a space of high dimensional feature established by nonlinear mapping of the N-dimensional input vector into a K- dimensional feature space (K > N) utilising the function ϕ(x). The learning process is moved to the minimization of the error function, which is defined by the so called ε-insensitive loss function $L_\varepsilon(d, y(x))$:

$$L_\varepsilon\big(d, y(x)\big) = \begin{cases} |d - y(x)| - \varepsilon, for\ |d - y(x)| \geq \varepsilon \\ 0\ , for\ |d - y(x)| < \varepsilon \end{cases} \quad (4.6)$$

where ***ε*** is the assumed accuracy; *d* is the destination; *x* is the input vector; and y*(x)* is the actual output under the effect of x.

The actual output of the SVR is defined by:

$$y(x) = \sum_{j=1}^{K} \omega_j \varphi_j(x) + b = w^T \varphi(x) + b \quad (4.7)$$

where $w = [\omega_0, \omega_1, \dots, \omega_K]^T$ is the weight vector; and $\varphi(x) = [\varphi_0(x), \varphi_1(x), \dots, \varphi_K(x)]^T$ is the basis function vector.

4.2.5. Bayesian Linear Regression

Bayesian linear regression is based on a generative method which is different from a discriminant one, which depends on Bayesian inference to build linear regression models [189]. Once the model is specified, the posterior distribution of parameters and forecasts of the model are computed by the method. This statistical analysis enables the method to define the complexity of the model through training, which produces a model with few possibilities to overfit. In contrast to simple linear regression model, the responses in Bayesian Linear Regression are assumed as samples from the probability distribution, for example the normal (Gaussian) distribution, which is:

$$Y \sim N(\beta^T X, \sigma^2) \quad (4.8)$$

The product of the parameters β and the inputs X is the mean of the Gaussian, where the normal deviation is σ. As well as the responses, in Bayesian Models, the parameters are also supposed to be sampled from a distribution. The aim is to define the posterior probability distribution for the parameters of the model with given X inputs and Y outputs, as in Eq. (4.9):

$$P(\beta \backslash Y, X) = \frac{P(Y \backslash \beta, X) * P(\beta \backslash X)}{P(Y \backslash X)} \quad (4.9)$$

The final result obtained from modelling by Bayesian Linear regression is not a single estimate, but rather a distribution range which can be used to produce inferences regarding new observations. This distribution enables determination of uncertainty in the model, which is considered one of the advantages of Bayesian Modelling methods. When the volume of data increases, the uncertainty of the result declines, presenting a better level of certainty in the approximation [191].

4.3. Implementation

To construct a forecast model of diabetes using regression modelling approach, five different regression machine learning models (described in the previous section) were developed and employed as base learners in this chapter. These models describe the development of diabetes disease in a given population in this case, the Saudi population, and they were structured to represent and utilise the available epidemiological data. The forecast model of diabetes was parameterised using nationally representative epidemiological and demographic data. This model integrates data on the population (adults age > 25) and for both genders (men and women), along with the behavioural (modifiable) risk factors (smoking, obesity, and inactivity) dataset, and it was programmed and implemented in MATLAB version R2018a.

This section describes the processes followed in the implementation of the five regression machine learning models involved in this chapter. Each model is trained to forecast the prevalence of diabetes using the training dataset which represents 56% of the data and validated using the testing dataset which represents 44% of the data (the required modelling datasets are described in Chapter 3). As mentioned in Chapter 3 the training data consists of predictor variables (age, gender, smoking, obesity, and inactivity) and the response variable (diabetes morbidity data). The implementation for each model is described below.

4.3.1. Multiple Linear Regression Model (MLR)

To establish this model in MATLAB, a constrained linear least-squares solver “lsqlin” with bounds or linear constraints was used to determine the regression positive coefficients for the MLR model using the training dataset. The optimization toolbox lsqlin function was used as the following: coefficients = *lsqlin*(X,Y,[ ],[ ],[ ],[ ], lb, ub), where X is the independent (predictor) variables (gender, smoking, obesity, inactivity); Y is the dependent (response) variable (the prevalence of diabetes morbidity); and lb and ub are the constraints (equal to zeros and ones, respectively). After calculating the model coefficients, the multiple linear regression model is represented by the following equation:

$$Y = 1 + 2.7 \times 10^{-10} X_1 + 0.2215 X_2 + 0.1738 X_3 + 0.0148 X_4 \qquad (4.10)$$

where Y is the dependent variable (diabetes prevalence); X_1, X_2, X_3, X_4 are the independent variables gender (men=1, women=0), smoking, obesity, and inactivity, respectively.

Figures 4.1 and 4.2 show 3D surfaces for the dependent variable and the independent variables for men and women training data respectively.

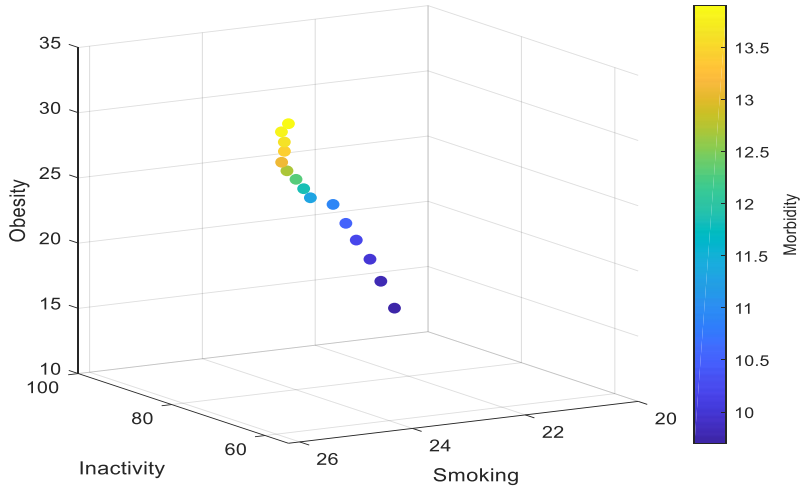

Figure 4.1: 3D surface of the three risk factors and morbidity of diabetes (men training data)

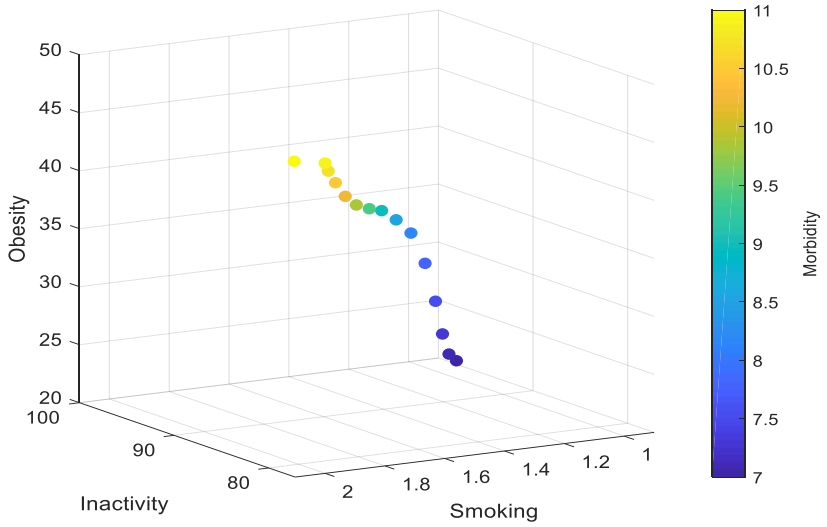

Figure 4.2: 3D surface of the three risk factors and morbidity of diabetes (women training data)

4.3.2. Adaptive Neuro-Fuzzy Inference System Model (ANFIS)

ANFIS was modelled using the MATLAB Neuro-Fuzzy Designer app, determining the number and type of membership functions, and the optimization method. To predict the prevalence of diabetes, the same training dataset that used in the previous model were used to create an ANFIS structure with three inputs (smoking, obesity, and inactivity) and one output (diabetes prevalence) for both men and women. In order to train the ANFIS model, the number of membership functions was selected as 2 for each input; Gaussian membership function was chosen for the type of function; and for the output variable, the type of membership function was linear. In addition, a hybrid method is implemented as optimization algorithm of the training, the error tolerance is set to 0, and the maximum number of epochs considered for

training is set as 300. Figure 4.3 represents a typical ANFIS structure with three inputs, one output, and eight rules.

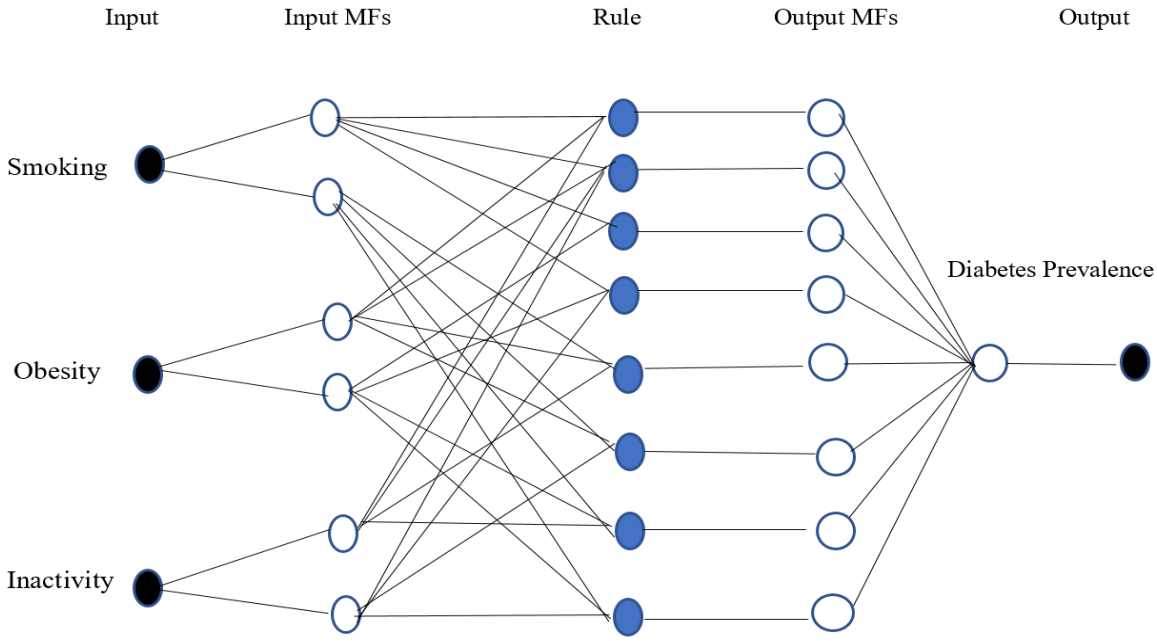

Figure 4.3: ANFIS model architecture with three inputs, one output, and eight rules

4.3.3. Artificial Neural Network (ANN)

To apply this model a neural fitting tool (*nftool*) is used from the neural network toolbox in MATLAB, which is a two-layer feed-forward network with sigmoid hidden neurons and linear output neurons (*fitnet*). In this model, inputs are defined as X and targets as Y, with samples set in rows. The training dataset has been used to create an ANN structure with 3 inputs (smoking, obesity, and inactivity) and one output (diabetes prevalence) for both men and women, and the number of neurons in the fitting network's hidden layer was set to be 10. The training functions are varied and can be selected according to the type and size of a problem. To train the ANN model, Levenberg–Marquardt algorithm was chosen, which is suitable for training small- and medium-sized networks, and it is considered to be an effective and fast training function [192]. The structure of the ANN model has three input variables, with 10 neurons for the hidden layer, and one output variable, as seen in Figure 4.4. The training process of the neural network is allowed to be started by itself sufficiently until it is automatically stopped after six epochs, when it achieves the best validation performance. The training performance graphs of the constructed neural networks are shown in Figure 4.5 and Figure 4.6 for the men and women training datasets, respectively.

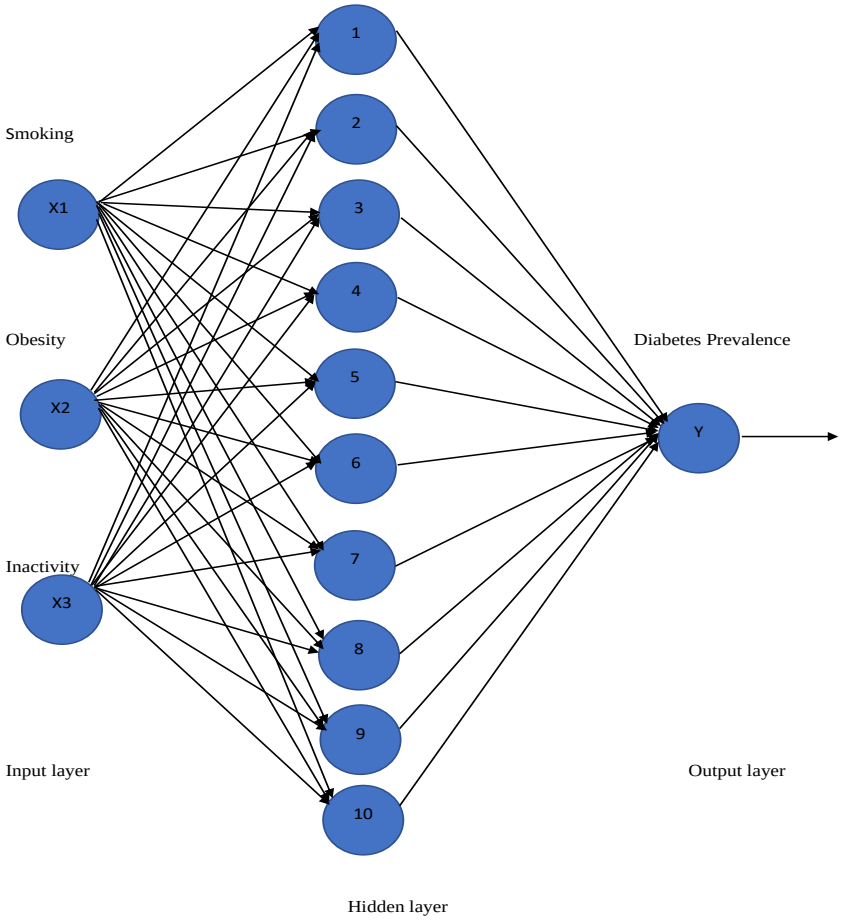

Figure 4.4: ANN architecture

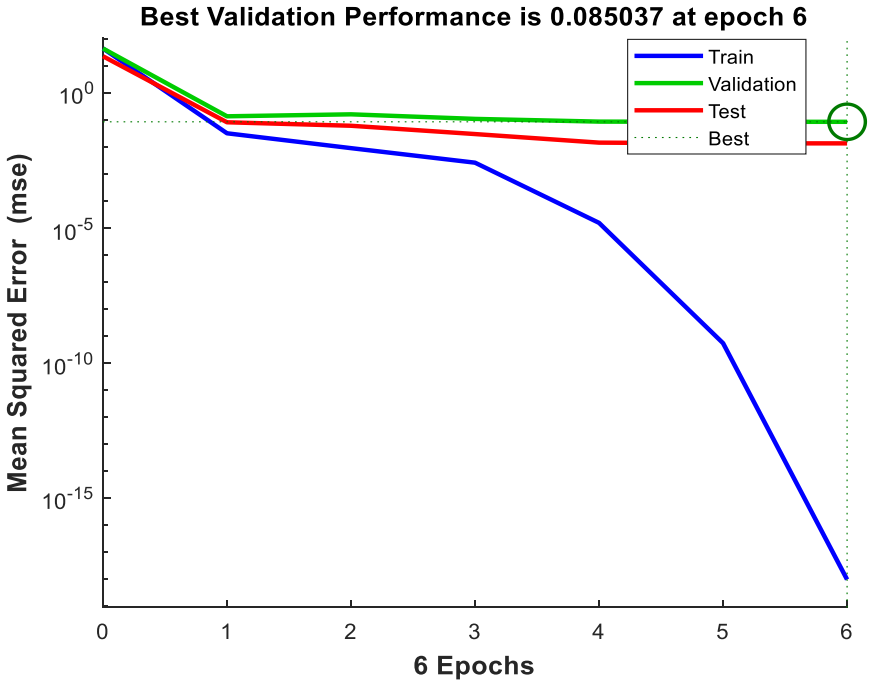

Figure 4.5: ANN training performance (men data)

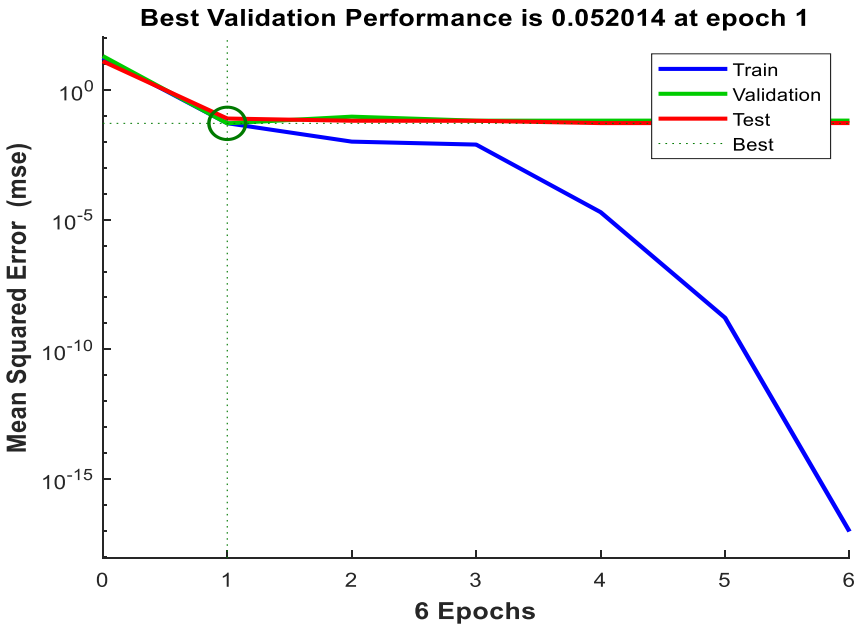

Figure 4.6: ANN training performance (women training data)

4.3.4. Support Vector Regression (SVR)

The SVR regression model was applied using the *fitrsvm* tool in the Statistics and Machine Learning Toolbox [MATLAB, R2018a] (https://mathworks.com/help/stats/fitrsvm.html). As with the above trained models, SVR model was trained using the training data, with the input values (independent variables) in the matrix and the target values (dependent variable) in the vector. SVR aims to find an optimal hyperplane by transforming the original feature space to a high-dimensional one utilising kernel functions. Some of the most popular kernel functions include linear kernel, polynomial function, Gaussian radial basis function (RBF), and hyperbolic tangent [189]. In this research, SVR model was trained with a default linear kernel, automatic hyperparameter tuning, and Sequential Minimal Optimization. The default settings contain the Kernel Scale auto unit, which assigns a proper scale factor using a heuristic procedure based on subsampling with “Standardize” unit, which standardises each variable using mean and standard deviations, then the obtained SVR model can be used to predict diabetes prevalence using the test dataset.

4.3.5. Bayesian Linear Regression Model

To create Bayesian linear regression model, the function (*bayeslm*) was used from the Econometrics Toolbox/ Bayesian Linear Regression Models in MATLAB2018. (https://uk.mathworks.com/help/econ/bayeslm.html). Firstly, *bayeslm* is used to create a prior model object appropriate for predictor selection: p = 3; PriorMdl = *bayeslm* (NumPredictors p)

This creates a diffuse prior model for the linear regression parameters, which is the default model type and identify the number of predictors p.

Then, the estimate function is applied to the prior model object, the predictors X, and the response Y (the training data) as follows: posteriorMdl = *estimate* (priorMdl ,X,Y);

By default, estimate returns a model object that represents the posterior distribution.

Finally, to predict responses of Bayesian linear regression model, forecast function is applied to the model object representing the posterior distribution as follows: forecast (posteriorMdl,x); where x representing the testing dataset.

4.4. Results and Discussion

This section presents the results obtained from each individual regression model described in section 4.3. Analysis of these results are carried out to compare the prediction power of the five models by evaluating their performance accordance with four important statistical evaluation measures, these measures are MSE, RMSE, MAPE, and the coefficient of determination R^2, respectively, given in equations (3.2, 3.3, 3.4, 3.5) (as presented in Chapter 3).

Table 4.1 shows the regression modelling results of the total prevalence of diabetes among men and women respectively, population ≥ 25 years of age, from 1999 to 2013 (training data). The results showed that diabetes prevalence is still increasing over this period for both men and women, however the prevalence rate was higher among men than women. In men, the total population prevalence of diabetes increased from 9.7% in 1999 to 13.9% in 2013, an absolute increase of 4.2 percentage points (pp) which represents an annual increase of 0.3 pp. The prevalence of diabetes among the total women population increased also by 0.3 pp each year from 7% in 1999 to 11% in 2013. Table 4.9 shows the projected estimates by regression modelling techniques for the total prevalence of diabetes in men and women Saudi population aged ≥ 25 years using (test data) during the period from 2014 to 2025 assuming the observed trends from 1999 to 2013 continue. Among men, the total predicted diabetes prevalence rate is estimated to rise during the same period from about 14.2% in 2014 to around 17.6% in 2025. In women, the estimated diabetes prevalence is predicted to increase from around 12.4 % in 2014 to about 17.3% in 2025. Figure 4.7 illustrates the total estimations of the prevalence of diabetes for men and women from 1999 to 2025.

Table 4.2 represents the results of the total prevalence of diabetes using regression modelling techniques for both men and women by including the gender variable (1 for men and 0 for women) in the training data. The projected estimates of the total prevalence of diabetes for men and women at the same time using (the test data) is shown in Table 4.10. The models demonstrate similar results to the modelling results of men and women taken separately.

Tables 4.3 to 4.8 show the results of the prevalence of diabetes according to age groups for men and women. Overall, it can be seen that there was a steady increase in diabetes prevalence among men and women in all age groups, but the prevalence was lower in younger age groups, and it increased with age. The highest prevalence of diabetes from all the age groups was found in the population aged 55-74 years. The highest growth in diabetes prevalence was in men aged 55-64 years old, which increased in prevalence from 24.91% in 1999 to 53% in 2013, an absolute increase of 28.1 pp in 14 years. The highest prevalence of diabetes for women was in the age group 65-74 years' old, which indicated an increase from 24.4% in 1999 to 48.2% in 2013, with an absolute increase of 23.8 pp in 14 years.

The results of the predicted prevalence of diabetes according to age groups for men and women for the period from 2014 to 2025 using the test data is demonstrated by Tables 4.11 to 4.16. In men, those aged 55 to 75^{+} years old showed the highest projected diabetes prevalence reaching over 60% by the year 2025. The lowest prevalence rates would belong to the youngest age group (25-34), predicted to be around 12.3% by 2025. Estimated diabetes prevalence ≥35% by 2025 was observed only in men 45 to 54 years old. Figure 4.8 demonstrates the estimations of the prevalence of diabetes for men by each age group from 1999 to 2025.The results for the projected diabetes prevalence in women population showed that by 2025, diabetes prevalence would reach 50% and higher for women aged between 55 and 75^{+} years old. The highest prevalence rate of diabetes in women by 2025 would be found in the group aged 65-74, which is predicted to be around 70%. Figure 4.9 shows the estimations of the prevalence of diabetes for women by each age group from 1999 to 2025.

The results in Figure 4.10 show that the prevalence of the behavioural risk factors of smoking and obesity also increased during the period from 1999 to 2025: smoking increased from 11% to 16.05%, while obesity sharply increased from 16.7% to 51.7%. The prevalence rate of inactivity is expected to have dropped significantly to 61.1% by 2025, compared to 96% in 1999, however this percentage remains dangerously high. According to [193], "Saudis are not active enough to meet the recommended guidelines for moderate to vigorous PA", which has been attributed to the rapid increase of urbanization, the nature of weather, and cultural characteristics. In addition, the prevalence of risk factors such as obesity, smoking, and physical inactivity varied according to gender. The prevalence of smoking was higher among men than women: 21.1% against 0.9% in 1999, and 28.4% against 3.7% by 2025. Women had a higher prevalence rate of obesity than men: 20.3% against 13.1% in 1999, and 58.4% against 45% by 2025 (Figure 4.10). Moreover, the prevalence rate of physical inactivity was higher among women than men: 98.1% against 93.9% in 1999, and 71.7% against 50.5% by 2025.

After discussing the obtained results of diabetes prevalence prediction by each regression model, their performances were evaluated based on the measures of MSE, RMSE, MAPE, and R-squared between the observed and predicted prevalence of diabetes for men and women, as presented in Table 4.17. Based on the results, it can be observed that the overall performance of all regression models was reasonably good. The values of R^2 were range between 0.92 to 0.99 and the evaluation corresponding results in terms of MSE and RMSE were range between 0 and 0.4 and with MAPE ranged between 0 and 0.03 for all models. However, the best results were provided by ANFIS model with RMSE = 0.04 and R^2 = 0.99 for men training data, and RMSE = 0.02 and R^2 = 0.99 for women training data. Among other models, BLM and MLR were giving a good accuracy in terms of RMSE and R^2 for men and women training datasets respectively.

The comparison graphs for actual vs predicted values for the total diabetes prevalence for men and women by all models are given in Figures 4.11 and 4.12, respectively. These graphs show different performance of the models but, they both show close proximity of ANFIS model predictions with the experimental data signifying the validity and accuracy of this model compared to other models. In addition, figures 4.13 to 4.18 demonstrate comparison graphs for actual vs predicted values for diabetes prevalence for men and women of each age group by all models. These plots also indicated a good performance of ANFIS model on the experimental data of each age group. A poor performance has been given by MLR model across all the experimental data of each age group for both men and women, while an acceptable performance of this model has been showed for the age group of (55-64) for both genders. SVR and BLM models performing weakly on the data of the first three age groups aged from 25 to 54 for both men and women, whereas they present a reasonable performance for the rest of age groups. Not bad performance has been presented by ANN model in the experimental data of each age group with exception of 35-44 age group where some curvature can be seen on the presented chart figure 4.14.

Furthermore, Figures 4.19 and 4.20 show the overall comparison of all regression models on performance metrics, which clearly indicated a reduced error by ANFIS model for both men and women datasets.

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	9.3	9.70	9.70	9.88	9.70	6.8	7.00	7.00	6.70	6.92
2000	9.6	9.80	9.70	10.03	9.81	7.0	7.12	7.26	7.08	7.23
2001	10.0	10.00	9.70	10.20	9.98	7.3	7.28	7.40	7.32	7.08
2002	10.4	10.20	9.70	10.43	10.21	7.5	7.50	7.50	7.55	7.49
2003	10.8	10.50	9.71	10.70	10.51	7.9	7.79	7.70	7.81	7.90
2004	11.2	10.90	9.80	11.04	10.88	8.3	8.14	8.14	8.18	8.23
2005	11.6	11.30	10.65	11.37	11.33	8.7	8.55	8.78	8.86	8.90
2006	12.0	11.80	11.90	11.78	11.81	9.1	8.99	8.99	9.20	9.25
2007	12.3	12.30	12.36	12.18	12.27	9.4	9.43	9.25	9.71	9.47
2008	12.7	12.70	12.69	12.56	12.71	9.8	9.84	9.69	10.13	9.77
2009	13.0	13.10	13.09	12.89	13.10	10.1	10.19	10.19	10.40	10.14
2010	13.2	13.40	13.35	13.17	13.38	10.3	10.48	10.53	10.45	10.13
2011	13.5	13.60	13.70	13.40	13.61	10.5	10.71	10.78	10.50	10.64
2012	13.7	13.80	14.09	13.58	13.80	10.6	10.88	10.88	10.52	10.63
2013	13.8	13.90	14.01	13.73	13.91	11.2	11.00	11.85	10.50	11.21

Table 4.1: Total diabetes prevalence results for men and women using regression models (training data), 1999-2013

	MLR		ANFIS		ANN		SVR		BLM	
Year	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
1999	9.34	6.18	9.70	7.00	9.65	7.08	9.30	6.66	9.66	6.59
2000	9.80	6.40	9.80	7.10	9.80	7.15	9.62	6.89	9.81	6.86
2001	10.19	6.81	10.00	7.30	9.99	7.28	9.93	7.23	10.01	7.18
2002	10.58	7.42	10.20	7.50	10.22	7.52	10.28	7.69	10.24	7.61
2003	10.94	8.14	10.50	7.79	10.49	7.81	10.65	8.24	10.55	8.11
2004	11.37	8.76	10.90	8.10	10.91	8.11	11.10	8.74	10.92	8.59
2005	11.70	9.19	11.30	8.50	11.34	8.61	11.51	9.18	11.22	9.09
2006	12.00	9.46	11.80	9.00	11.84	8.90	11.95	9.43	11.71	9.35
2007	12.31	9.65	12.30	9.40	12.30	9.40	12.39	9.67	12.18	9.65
2008	12.60	9.84	12.70	9.80	12.70	9.85	12.80	9.89	12.60	9.91
2009	12.86	10.08	13.10	10.20	13.09	10.24	13.17	10.10	13.00	10.14
2010	13.09	10.34	13.40	10.48	13.42	10.47	13.47	10.29	13.38	10.29
2011	13.29	10.56	13.60	10.73	13.63	10.72	13.73	10.45	13.68	10.42
2012	13.50	10.71	13.80	10.89	13.75	10.91	13.96	10.56	13.90	10.51
2013	13.61	10.81	13.90	11.00	13.82	10.94	14.11	10.62	14.13	10.50

Table 4.2: Total diabetes prevalence results for both men and women using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	4.4	3.69	3.79	3.92	3.90	3.4	3.10	3.10	3.22	3.04
2000	4.5	3.06	4.10	4.10	4.07	3.5	3.21	3.21	3.29	3.21
2001	4.7	2.82	4.51	4.37	4.40	3.6	3.41	3.28	3.40	3.36
2002	5.0	2.69	5.04	4.70	4.77	3.7	3.62	3.54	3.52	3.53
2003	5.3	2.71	5.34	5.11	5.21	3.9	3.82	3.82	3.70	3.76
2004	5.7	2.91	5.75	5.59	5.73	4.0	4.03	4.03	3.92	4.02
2005	6.1	2.80	6.17	6.10	6.24	4.3	4.24	4.24	4.21	4.44
2006	6.6	3.92	6.50	6.62	6.68	4.5	4.45	4.45	4.46	4.62
2007	7.1	4.68	7.01	7.10	7.03	4.8	4.65	4.65	4.72	4.81
2008	7.5	5.52	7.42	7.58	7.44	5.0	4.86	4.88	4.98	5.01
2009	7.9	6.39	7.84	8.07	7.97	5.2	5.07	5.07	5.19	5.17
2010	8.2	7.23	8.26	8.48	8.46	5.4	5.28	5.28	5.38	5.33
2011	8.5	8.00	8.67	8.78	8.79	5.5	5.48	5.48	5.53	5.44
2012	8.6	8.69	9.08	9.01	9.06	5.6	5.69	5.69	5.65	5.52
2013	8.8	9.50	9.50	9.16	9.06	5.7	5.90	5.90	5.78	5.58

Table 4.3: Diabetes prevalence results for men and women aged 25-34 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	7.0	7.01	7.01	6.56	6.83	6.6	5.03	4.89	4.81	4.20
2000	7.8	7.20	7.20	6.92	7.27	6.6	5.42	5.01	5.22	5.37
2001	8.4	7.30	7.30	7.30	7.46	6.8	5.80	6.20	5.58	5.99
2002	9.1	7.60	7.60	7.78	7.63	7.5	6.19	6.34	6.28	6.42
2003	9.9	8.00	8.01	8.36	7.92	8.9	6.58	6.75	7.44	7.47
2004	11.1	8.70	8.08	9.15	8.77	9.5	6.97	6.98	8.08	7.90
2005	11.7	9.60	9.94	9.97	9.59	9.7	7.36	7.55	8.44	7.71
2006	12.6	10.70	10.74	10.80	10.71	9.7	7.75	7.67	8.46	7.83
2007	13.4	11.90	11.87	11.65	11.81	9.8	8.14	8.19	8.57	7.91
2008	14.0	12.80	12.78	12.42	12.74	9.8	8.54	8.67	8.68	8.18
2009	14.5	13.50	13.38	13.05	13.45	10.0	8.93	8.99	8.84	8.59
2010	14.8	13.90	13.89	13.52	13.93	10.3	9.32	9.42	9.10	9.17
2011	14.9	14.20	14.21	13.86	14.23	10.4	9.71	9.84	9.22	9.59
2012	14.9	14.40	14.39	14.08	14.43	10.5	10.11	10.14	9.22	9.88
2013	15.0	14.50	14.52	14.22	14.54	10.4	10.50	10.39	9.12	10.12

Table 4.4: Diabetes prevalence results for men and women aged 35-44 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	21.8	21.06	21.20	21.45	21.05	21.8	22.10	22.10	21.83	21.96
2000	22.7	21.80	22.00	21.98	21.86	21.9	22.30	22.30	21.92	22.33
2001	23.5	22.50	22.64	22.46	22.50	22.2	22.50	22.57	22.25	22.45
2002	24.4	23.20	23.27	23.04	23.19	23.0	22.80	22.99	22.97	22.98
2003	25.3	23.90	23.90	23.69	23.89	23.7	23.10	23.10	23.68	23.11
2004	26.3	24.60	24.44	24.57	24.87	24.5	23.60	23.60	24.41	23.62
2005	27.8	25.30	25.76	24.91	24.80	24.8	24.10	24.10	24.72	24.25
2006	28.3	26.00	26.01	26.00	25.95	25.2	24.70	24.70	25.13	24.62
2007	29.1	26.70	26.70	26.86	26.72	25.4	25.30	25.30	25.36	25.20
2008	29.7	27.41	27.39	27.71	27.51	25.6	25.80	25.80	25.66	25.86
2009	30.2	28.07	28.08	28.49	28.27	25.9	26.20	26.20	25.93	26.21
2010	30.4	28.82	28.76	29.19	29.00	26.2	26.60	26.58	26.29	26.65
2011	30.5	29.54	29.46	29.76	29.63	26.5	26.90	26.92	26.58	26.74
2012	30.5	30.17	30.15	30.23	30.14	26.7	27.10	27.10	26.83	26.88
2013	30.2	31.10	30.39	30.71	30.78	27.1	27.30	27.25	27.31	27.53

Table 4.5: Diabetes prevalence results for men and women aged 45-54 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	24.9	24.90	24.90	24.58	24.68	23.7	23.20	23.20	23.27	23.77
2000	26.8	26.93	26.88	26.26	26.99	24.2	24.50	24.50	23.98	24.37
2001	28.7	28.88	28.89	27.94	29.05	25.3	25.80	25.80	25.11	25.34
2002	30.6	30.91	30.98	29.79	30.99	26.9	27.10	27.12	26.73	26.74
2003	32.6	32.90	32.90	31.80	32.89	28.8	28.40	28.38	28.66	28.45
2004	35.0	34.90	34.90	34.23	35.21	30.4	29.70	28.76	30.43	30.02
2005	36.9	36.90	36.99	36.44	36.46	31.4	31.10	30.99	31.83	31.33
2006	39.1	38.90	38.89	38.95	38.72	32.3	32.40	32.39	32.81	32.34
2007	41.4	40.90	40.89	41.48	40.84	33.0	33.70	33.68	33.79	33.46
2008	43.5	42.90	42.86	43.84	42.94	34.2	35.00	35.00	34.92	34.72
2009	45.6	44.93	44.89	46.01	45.16	35.9	36.30	36.31	36.45	36.37
2010	47.3	46.80	46.90	47.82	47.05	37.9	37.60	37.42	38.08	38.07
2011	49.0	49.00	49.00	49.46	49.11	39.6	39.00	39.00	39.51	39.47
2012	50.5	51.01	51.00	50.81	50.92	40.7	40.30	40.30	40.39	40.35
2013	51.7	52.96	53.00	51.89	52.71	41.3	41.60	41.59	40.88	40.89

Table 4.6: Diabetes prevalence results for men and women aged55-64 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	29.4	28.75	28.76	30.45	30.16	28.8	24.40	24.65	25.35	24.96
2000	29.7	30.38	29.94	30.99	30.46	29.3	26.10	25.95	26.19	26.17
2001	30.3	32.01	31.77	31.79	31.81	30.0	27.80	27.99	27.21	27.20
2002	30.9	33.65	33.65	32.86	32.73	30.8	29.50	29.64	28.55	29.13
2003	31.8	35.30	35.31	34.25	34.06	32.0	31.20	31.29	30.34	31.15
2004	32.9	36.94	37.00	36.02	36.98	33.3	32.90	33.17	32.30	33.21
2005	34.2	38.59	38.44	38.05	37.96	34.9	34.60	34.63	34.32	35.00
2006	35.6	40.23	40.21	40.24	40.64	36.6	36.30	36.36	36.20	36.15
2007	36.9	41.88	41.95	42.39	42.16	38.3	38.00	39.55	38.28	37.66
2008	38.1	43.53	43.57	44.45	44.65	39.9	39.70	39.82	40.30	39.44
2009	39.3	45.19	45.03	46.20	46.06	41.3	41.40	41.16	42.26	41.51
2010	40.2	46.84	46.90	47.63	47.07	42.5	43.10	43.25	44.05	43.53
2011	40.9	48.50	48.69	48.74	49.16	43.3	44.80	44.73	45.35	45.34
2012	41.3	50.16	50.33	49.51	49.35	43.9	46.50	46.69	46.37	46.38
2013	41.8	51.80	51.56	50.13	50.49	44.4	48.20	48.36	47.25	47.69

Table 4.7: Diabetes prevalence results for men and women aged 65-74 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
1999	30.3	28.75	29.86	30.41	29.99	29.2	24.42	24.41	25.35	25.01
2000	30.9	30.40	30.40	30.95	30.60	29.7	26.09	25.71	26.19	26.07
2001	31.7	32.00	31.71	31.75	31.99	30.4	27.78	27.27	27.20	27.33
2002	32.8	33.70	33.70	32.82	32.57	31.3	29.50	29.50	28.55	29.02
2003	34.2	35.30	35.02	34.21	34.17	32.4	31.21	31.20	30.32	31.13
2004	36.1	36.90	36.90	35.99	36.95	33.8	32.90	32.82	32.28	33.24
2005	38.0	38.60	38.60	38.01	37.71	35.5	34.60	34.59	34.32	35.02
2006	40.3	40.20	40.20	40.20	40.71	37.2	36.30	36.30	36.20	36.17
2007	42.4	41.90	41.95	42.36	42.50	39.0	38.00	38.00	38.29	37.63
2008	44.5	43.50	43.56	44.41	44.46	40.7	39.71	39.71	40.32	39.41
2009	46.2	45.20	45.22	46.16	46.04	42.1	41.36	41.42	42.28	41.52
2010	47.6	46.80	46.83	47.60	47.24	43.3	43.09	42.93	44.06	43.63
2011	48.9	48.50	48.57	48.70	49.01	44.1	44.87	44.84	45.36	45.21
2012	49.6	50.20	50.22	49.47	49.21	44.8	46.57	46.54	46.38	46.48
2013	50.3	51.80	51.85	50.09	50.62	45.3	48.10	48.23	47.25	47.64

Table 4.8: Diabetes prevalence results for men and women aged +74 using regression models (training data), 1999-2013

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	14.1	14.2	14.5	13.8	14.4	12.9	12.2	11.5	11	12.6
2015	14.2	14.4	15.1	13.9	14.7	13.4	12.8	11.7	11	13.1
2016	14.4	14.8	15.8	14.0	15.1	14.6	13.1	12.3	11	14.2
2017	14.6	15.2	16.6	14.1	15.5	15.1	13.8	13.1	11	14.6
2018	14.7	15.5	17.0	14.1	15.9	15.7	14.4	14.1	11	15.1
2019	14.9	16.0	17.7	14.2	16.3	16.8	14.8	15.5	11	16.1
2020	15.1	16.6	18.2	14.2	16.7	16.8	15.7	15.9	11	16.0
2021	15.2	17.1	18.6	14.3	17.1	17.3	16.4	16.4	11	16.5
2022	15.4	17.7	18.8	14.3	17.5	17.9	17.0	16.8	12	16.9
2023	15.5	18.3	18.9	14.4	17.9	19.0	17.4	17.0	12	18.0
2024	15.7	19.0	19.1	14.4	18.3	19.5	18.1	17.1	12	18.4
2025	15.9	19.6	19.2	14.5	18.8	20.1	18.7	17.2	12	18.9

Table 4.9: Total diabetes prevalence results for men and women using regression models (test data), 2014-2025

	MLR		ANFIS		ANN		SVR		BLM	
Year	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
2014	13.75	11.02	13.88	12.65	14.12	11.30	14.24	10.79	14.22	10.55
2015	13.94	11.20	14.07	13.75	14.35	11.56	14.40	10.92	14.31	10.63
2016	14.13	11.41	14.41	14.92	14.61	11.84	14.56	11.08	14.38	10.69
2017	14.32	11.58	14.90	15.86	14.87	12.15	14.70	11.21	14.43	10.78
2018	14.50	11.77	15.53	16.85	15.15	12.51	14.83	11.35	14.51	10.87
2019	14.69	11.96	16.24	17.67	15.45	12.86	14.97	11.49	14.55	10.92
2020	14.89	12.11	17.07	18.40	15.72	13.28	15.10	11.61	14.57	11.03
2021	15.09	12.30	18.00	19.18	15.95	13.74	15.23	11.75	14.60	11.12
2022	15.29	12.48	19.01	19.79	16.14	14.19	15.37	11.88	14.62	11.21
2023	15.46	12.69	20.22	20.41	16.27	14.68	15.49	12.03	14.68	11.27
2024	15.66	12.86	21.42	20.88	16.37	15.12	15.62	12.16	14.70	11.35
2025	15.86	13.03	22.76	21.29	16.42	15.52	15.75	12.29	14.71	11.43

Table 4.10: Total diabetes prevalence results for both Men and Women using regression models (test data), 2014-2025

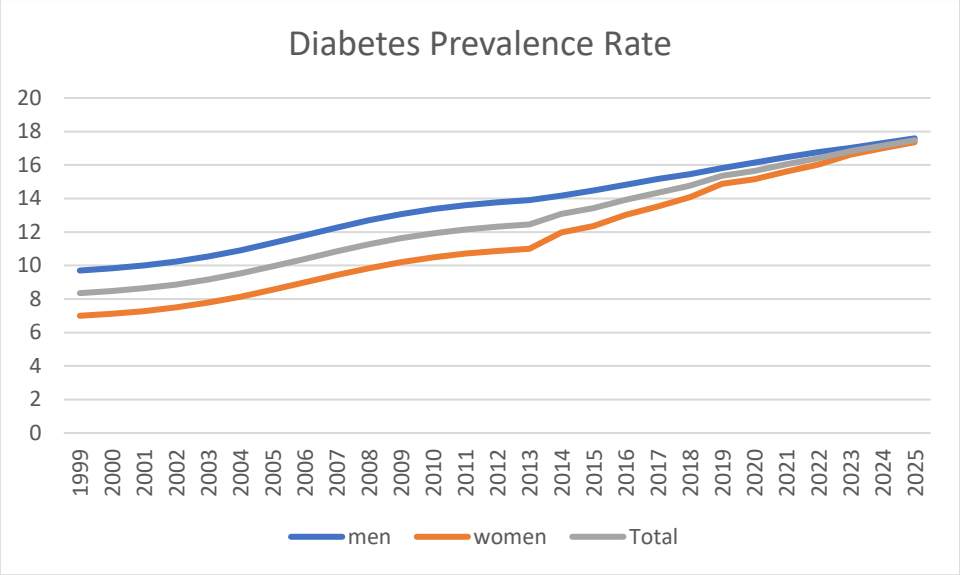

Figure 4.7: Diabetes prevalence estimations for Saudis aged 25-75+, 1999-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	10.22	10.01	9.82	9.35	9.30	5.25	7.81	7.55	7.64	7.99
2015	10.39	10.44	10.10	9.48	9.45	5.32	8.52	8.35	8.09	7.85
2016	10.59	10.91	10.36	9.58	9.59	5.49	9.17	9.97	8.55	9.12
2017	10.79	11.37	10.61	9.66	9.72	5.57	9.75	10.79	9.00	8.98
2018	10.99	11.83	10.88	9.73	9.85	5.74	10.23	11.84	9.45	10.24
2019	11.19	12.32	11.18	9.80	9.98	5.81	10.63	12.15	9.89	10.15
2020	11.35	12.74	11.47	9.86	10.08	5.98	11.05	12.49	10.34	11.41
2021	11.55	13.20	11.84	9.91	10.20	6.05	11.48	12.59	10.80	11.27
2022	11.75	13.69	12.27	9.97	10.32	6.13	11.91	12.66	11.24	11.13
2023	11.95	14.16	12.71	10.02	10.43	6.3	12.34	12.72	11.70	12.40
2024	12.16	14.65	13.16	10.07	10.55	6.37	12.77	12.75	12.15	12.26
2025	12.32	15.06	13.51	10.12	10.66	6.45	13.20	12.78	12.60	12.12

Table 4.11: Diabetes prevalence results for men and women aged 25-34 using regression models (test data), 2014-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	16.84	15.34	16.85	15.23	15.38	10.64	9.65	9.71	9.45	9.47
2015	17.24	15.71	17.05	15.39	15.49	10.75	9.90	9.80	9.69	10.42
2016	17.64	15.94	17.22	15.48	15.53	10.91	9.96	9.87	9.89	10.17
2017	18.08	16.23	17.36	15.53	15.54	11.02	10.18	10.14	10.13	11.12
2018	18.48	16.31	17.47	15.49	15.44	11.18	10.39	10.44	10.34	10.87
2019	18.88	17.29	17.64	15.90	15.82	11.29	10.60	10.84	10.58	11.83
2020	19.28	18.38	17.83	16.40	16.29	11.45	10.85	11.50	10.79	11.58
2021	19.72	19.56	17.99	16.91	16.79	11.56	11.06	11.97	11.03	12.53
2022	20.12	20.54	18.15	17.42	17.27	11.68	11.27	12.42	11.27	13.48
2023	20.52	21.49	18.30	17.92	17.74	11.83	11.52	13.16	11.47	13.23
2024	20.96	22.58	18.40	18.43	18.24	11.95	11.73	13.54	11.72	14.18
2025	21.36	23.38	18.52	18.94	18.69	12.06	11.94	13.84	11.96	15.13

Table 4.12: Diabetes prevalence results for men and women aged 35-44 using regression models (test data), 2014-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	30.66	31.81	31.35	30.37	31.55	27.30	27.88	27.40	27.53	28.73
2015	30.80	32.24	31.93	31.21	31.95	27.33	28.06	27.57	27.57	28.92
2016	30.94	32.67	32.50	32.05	32.35	27.46	28.36	27.88	27.73	29.51
2017	31.08	33.09	33.05	32.86	32.75	27.50	28.55	28.21	27.77	29.70
2018	31.22	33.51	33.54	33.60	33.14	27.53	28.73	28.62	27.80	29.89
2019	31.36	33.93	33.96	34.26	33.54	27.66	29.03	29.24	27.97	30.48
2020	31.51	34.35	34.30	34.84	33.94	27.80	29.33	29.78	28.13	31.07
2021	31.65	34.77	34.57	35.32	34.33	27.83	29.51	30.07	28.16	31.26
2022	31.79	35.20	34.78	35.72	34.73	27.87	29.70	30.27	28.20	31.45
2023	31.93	35.62	34.93	36.05	35.13	28.00	30.00	30.42	28.36	32.04
2024	32.07	36.04	35.04	36.31	35.52	28.03	30.18	30.50	28.40	32.23
2025	32.21	36.46	35.11	36.51	35.92	28.07	30.36	30.55	28.43	32.43

Table 4.13: Diabetes prevalence results for men and women aged 45-54 using regression models (test data), 2014-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	52.52	53.93	54.23	52.69	53.53	41.46	42.21	42.51	41.26	41.39
2015	53.59	55.18	55.66	53.73	54.50	41.96	42.78	42.87	41.86	42.01
2016	54.61	56.32	56.86	54.74	55.31	42.44	44.04	43.82	42.44	42.74
2017	55.60	57.39	57.85	55.72	56.03	42.86	44.84	44.54	42.98	43.37
2018	56.57	58.40	58.69	56.68	56.69	43.28	45.58	45.40	43.51	43.98
2019	57.53	59.39	59.43	57.63	57.31	43.67	46.30	46.33	44.01	44.60
2020	58.47	60.34	60.07	58.58	57.89	44.05	46.86	46.97	44.51	45.18
2021	59.40	61.27	60.67	59.52	58.45	44.43	47.43	47.42	45.00	45.77
2022	60.33	62.19	61.22	60.45	58.99	44.79	47.99	47.68	45.49	46.35
2023	61.25	63.10	61.73	61.38	59.51	45.17	48.70	47.85	45.98	47.00
2024	62.18	64.01	62.22	62.31	60.04	45.54	49.23	47.90	46.46	47.58
2025	63.09	64.90	62.68	63.24	60.54	45.90	49.76	47.92	46.94	48.16

Table 4.14: Diabetes prevalence results for men and women aged55-64 using regression models (test data), 2014-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	52.31	54.60	52.25	50.71	51.65	47.7	52.01	51.58	50.62	51.37
2015	52.79	56.10	53.77	51.48	53.23	49.1	55.15	54.13	52.54	53.34
2016	53.28	57.60	55.32	52.26	54.81	50.5	58.42	56.74	54.47	55.32
2017	53.76	58.97	56.84	53.03	56.43	51.8	61.34	58.77	56.23	57.05
2018	54.28	60.52	58.34	53.80	57.76	53.1	64.04	60.54	57.99	58.79
2019	54.82	61.88	59.75	54.58	59.14	54.4	67.38	63.19	59.91	61.16
2020	55.31	63.36	61.10	55.35	60.75	55.7	69.59	64.72	61.67	62.90
2021	55.79	64.71	62.37	56.12	62.33	57.1	71.60	66.14	63.60	64.88
2022	56.27	66.19	63.52	56.89	63.92	58.4	73.18	66.98	65.36	66.61
2023	56.75	67.54	64.56	57.66	65.41	59.8	74.60	67.76	67.29	68.59
2024	57.23	69.03	65.49	58.44	67.00	61.1	75.72	68.16	69.05	70.31
2025	57.82	70.61	66.27	59.21	68.24	62.3	77.65	69.82	70.80	72.46

Table 4.15: Diabetes prevalence results for men and women aged 65-74 using regression models (test data), 2014-2025

	Men					Women				
Year	MLR	ANFIS	ANN	SVR	BLM	MLR	ANFIS	ANN	SVR	BLM
2014	51.19	53.60	53.61	50.70	55.50	45.95	48.99	50.57	47.83	47.61
2015	52.15	55.90	56.06	51.49	58.19	46.75	50.52	52.97	48.75	48.35
2016	53.10	58.18	58.40	52.27	60.88	47.54	52.03	55.07	49.66	49.08
2017	54.05	60.36	60.41	53.05	63.61	48.34	53.57	56.84	50.57	49.83
2018	54.99	62.38	62.01	53.83	66.18	49.14	55.13	58.28	51.49	50.58
2019	55.94	64.29	63.25	54.62	68.87	49.93	56.66	59.39	52.39	51.30
2020	56.90	66.06	64.15	55.40	71.60	50.73	58.22	60.31	53.31	52.06
2021	57.85	67.72	64.81	56.18	74.29	51.53	59.79	61.07	54.24	52.81
2022	58.80	69.28	65.28	56.96	76.98	52.32	61.32	61.66	55.14	53.53
2023	59.74	70.70	65.61	57.74	79.59	53.12	62.88	62.20	56.06	54.29
2024	60.69	72.10	65.85	58.52	82.28	53.92	64.44	62.66	56.97	55.03
2025	61.65	73.44	66.01	59.31	84.97	54.62	66.06	62.55	57.89	55.79

Table 4.16: Diabetes prevalence results for men and women aged +74 using regression models (test data), 2014-2025

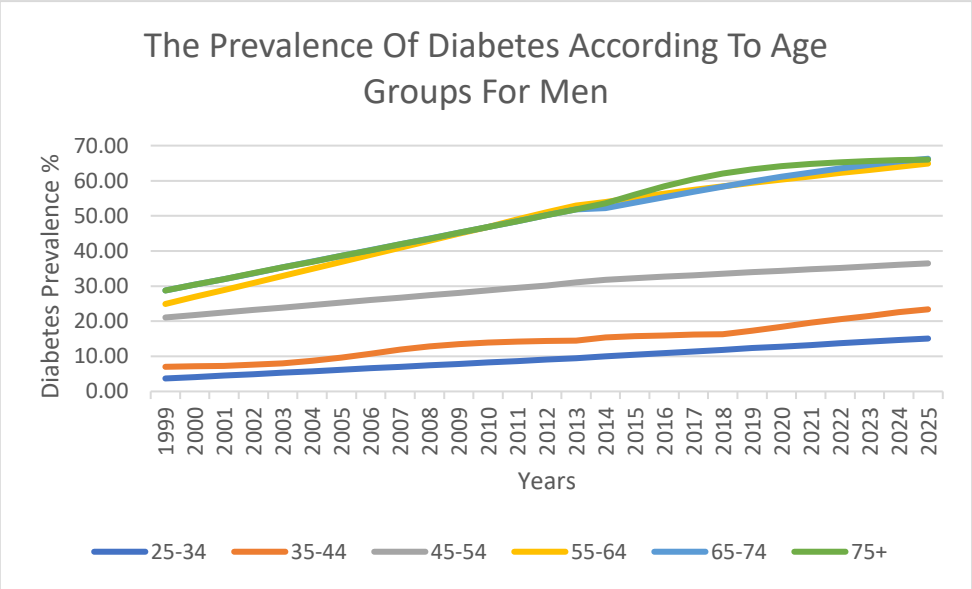

Figure 4.8: Diabetes prevalence estimations for men according to age groups

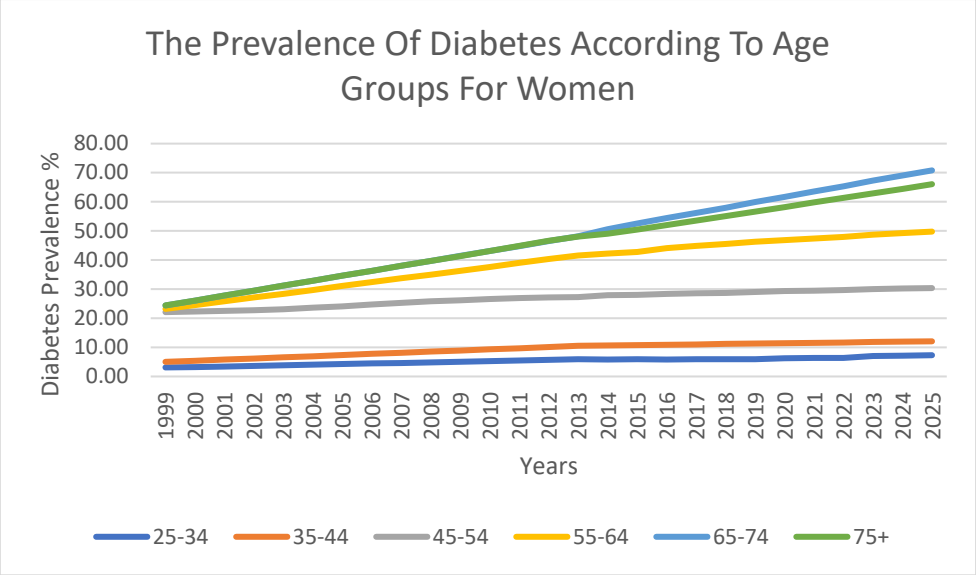

Figure 4.9: Diabetes prevalence estimations for women according to age groups

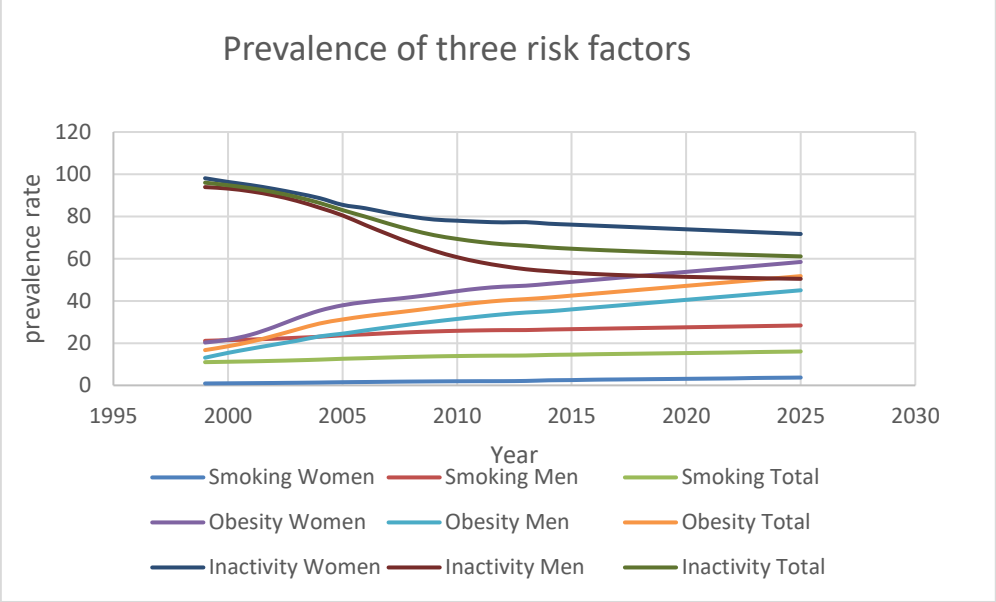

Figure 4.10: Prevalence rates of smoking, obesity, and inactivity for Saudis aged 25-75+, 1999-2025

Models	Men				Women			
	MSE	RMSE	MAPE	R^2	MSE	RMSE	MAPE	R^2
MLR	0.0420	0.2049	0.0150	0.9814	0.0247	0.1571	0.0139	0.9878
ANFIS	**0.0013**	**0.0365**	**0**	**0.9994**	**0.0005**	**0.0239**	**0.0021**	**0.9997**
ANN	0.0081	0.0899	0.0252	0.9964	0.0594	0.2437	0.0137	0.9705
SVR	0.0328	0.1810	0.0147	0.9855	0.0636	0.2522	0.0214	0.9684
BLM	0.0032	0.0564	0.0011	0.9986	0.0392	0.1980	0.0177	0.9806

Table 4.17: Statistical evaluation metrics results for all regression models for both men and women

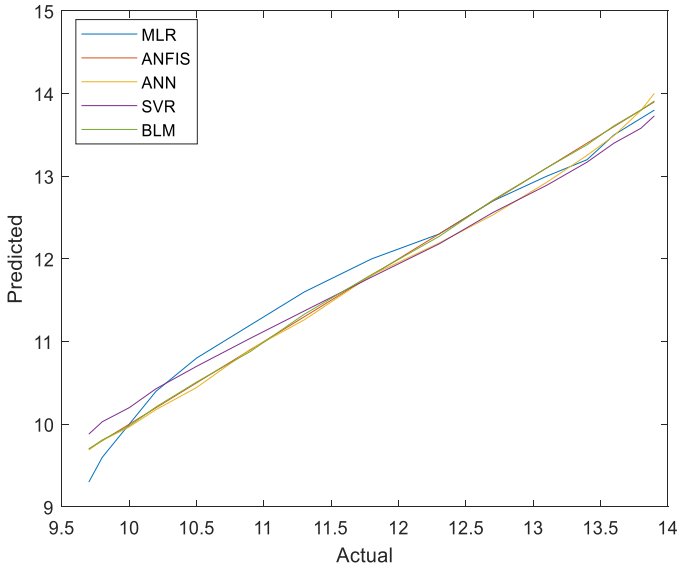

Figure 4.11: Actual data against predicted for the total diabetes prevalence by all regression models (men training data)

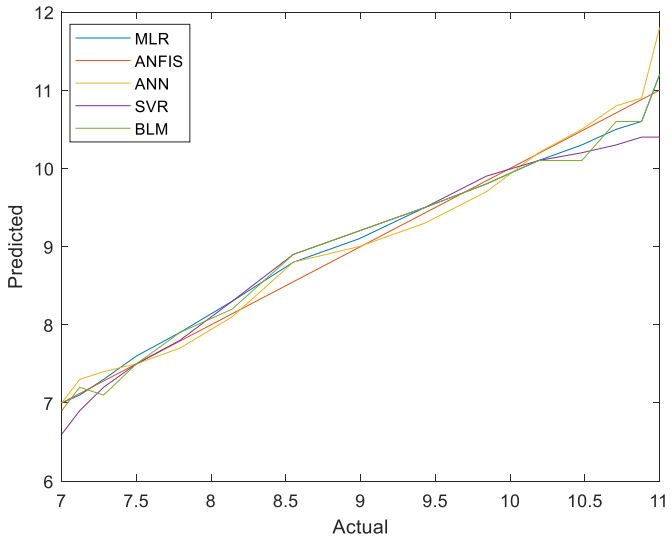

Figure 4.12: Actual data against predicted for the total diabetes prevalence by all regression models (men training data)

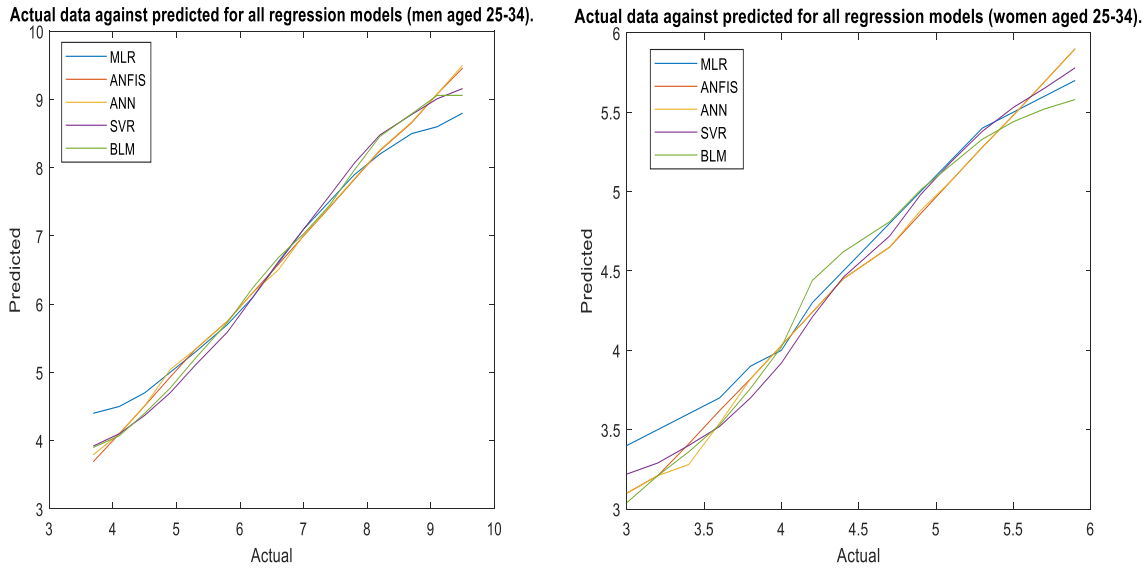

Figure 4.13: Actual data against predicted for all regression models (men & women aged 25-34)

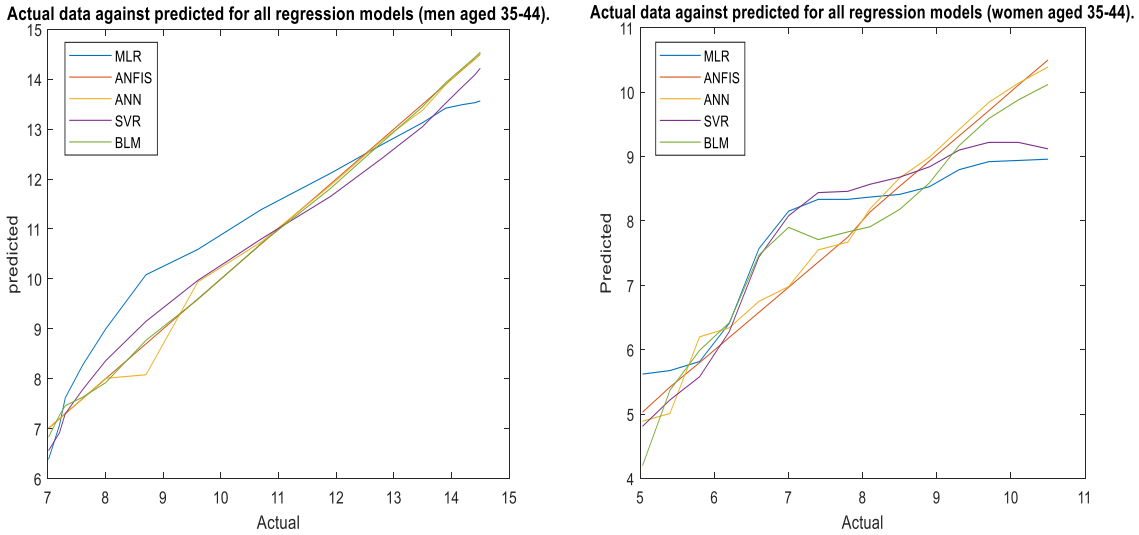

Figure 4.14: Actual data against predicted for all regression models (men & women aged 35-44)

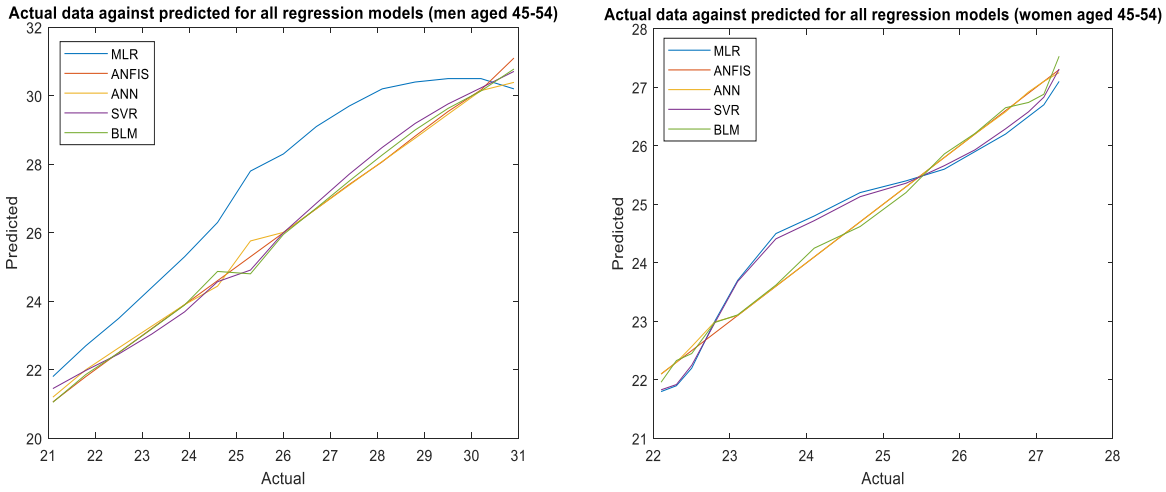

Figure 4.15: Actual data against predicted for all regression models (men & women aged 45-54)

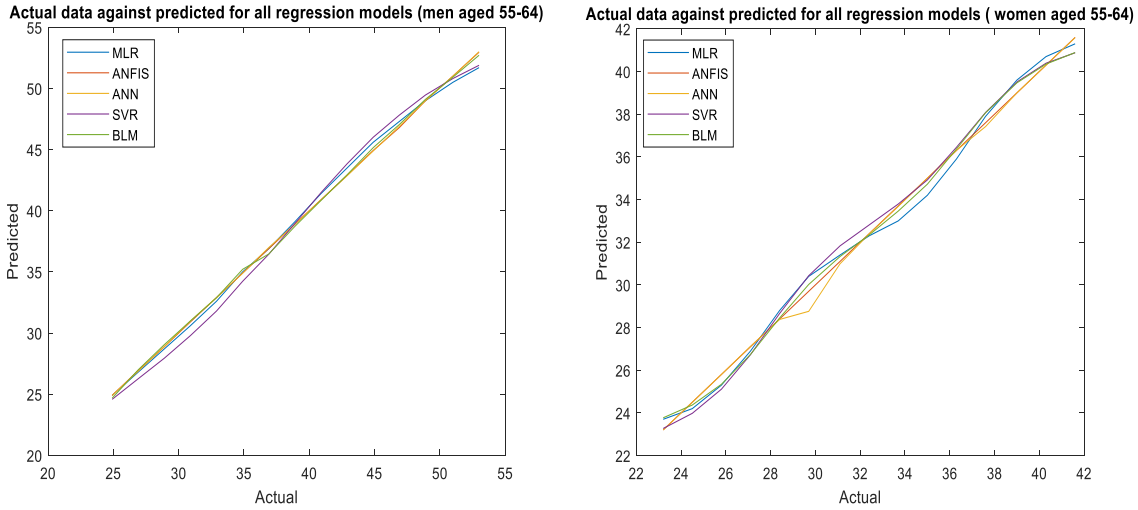

Figure 4.16: Actual data against predicted for all regression models (men & women aged 55-64)

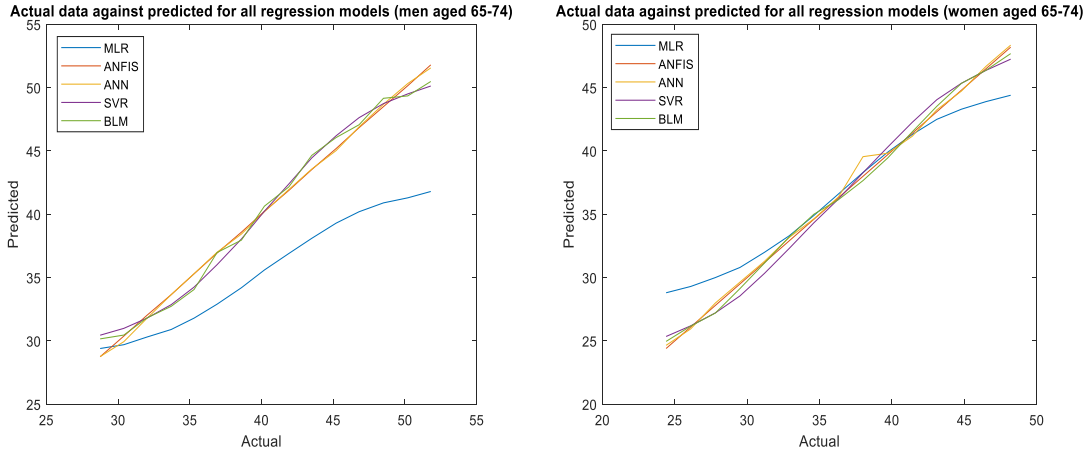

Figure 4.17: Actual data against predicted for all regression models (men & women aged 65-74)

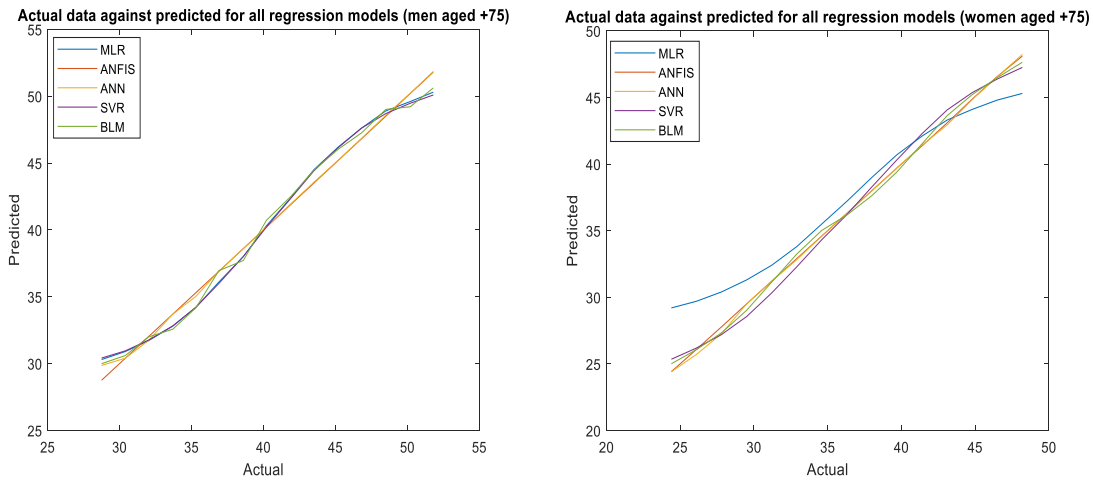

Figure 4.18: Actual data against predicted for all regression models (men & women aged 75⁺)

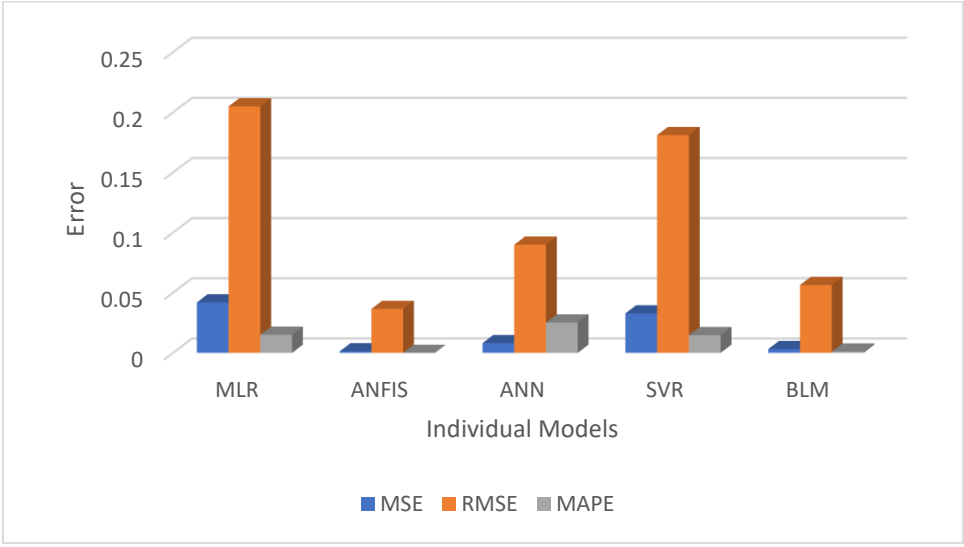

Figure 4.19: Performance metrics of regression models men data

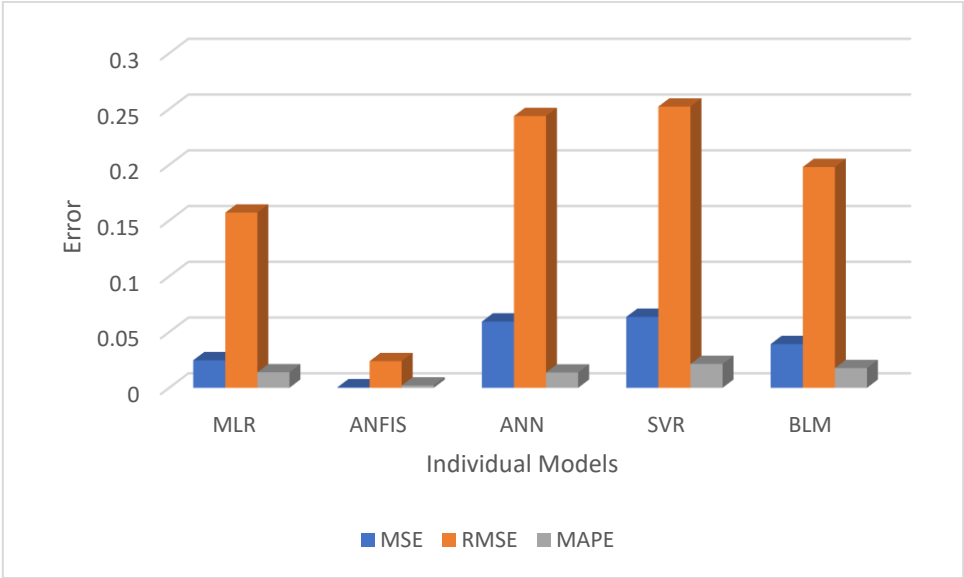

Figure 4.20: Performance metrics of regression models women data

4.5. Summary

This chapter investigated the trends of diabetes prevalence in Saudi adult population using historical diabetes prevalence data along with smoking, obesity, and inactivity data as predictor variables using five regression modelling techniques. It also explored the use of statistical performance measures to evaluate the performance of each model. The results obtained show that there was no huge difference in the performance of the models when using the datasets of men and women; however, ANFIS model was performing well for predicting the total prevalence of diabetes for men and women and for each age group. In order to overcome the disadvantage of each model, a combination technique is a good way to obtained more accurate and reliable model. In the next chapter, different combination approaches are used for the purpose to enhance the performance of the model, then the predictions of the combined model are validated by comparison with the observed results from existing studies.

Chapter 5
Ensemble Methods

5.1. Introduction

Ensemble methods are statistical and computational machine learning techniques that aim to obtain more reliable and accurate predictions in supervised and unsupervised learning problems by combining multiple learning models, instead of choosing the best-performing model [194][195]. In the literature, ensemble methods known by different terms other than “ensemble”, such as “committee”, “combination”, “fusion”, and others, which have been used to represent sets of learning machines that work together to solve a machine learning problem [194][196]. The history dates of integrating multiple models to develop a predictive model (ensemble methods) back to as early as 1977, and the concept was significantly improved during the 1990s [197]. Ensemble methods have been successfully applied in several fields, including medicine, bioinformatics, cheminformatics, image retrieval, and finance. It has been indicated by some empirical studies that combining the outcomes for both regression and classification problems leads to much more accurate results compared to the individual learners [198][199][200][201].

Some empirical studies carried out by machine learning community highlighted the effectiveness of combining the outcomes of multiple models in reducing generalisation, variance, and bias error, mainly because various types of models have different “inductive biases” [197][202]. These errors reduction contributes to improve the predictive performance (e.g., obtaining a lower error in regression or a high classification accuracy) [203]. Ensemble approach is particularly useful in dealing with some situations, such as a limited number of training data, unusually high dimensional patterns, or a large amount of noise [204]. Two different types of models can be used to build an ensemble: homogeneous or heterogeneous models. Homogeneous models are obtained by different implementation of the same learning algorithm. In homogeneous models, one learning algorithm is used in different implementations, such as using different training datasets; in heterogeneous models, multiple learning algorithms are applied on the same dataset, so there are different opinions about the data [205].

In the previous chapter, the principles of regression modelling and the proposed regression models were reviewed, and the evaluation of the performance of each individual model was demonstrated. The next step in their analysis is to examine how they work together, using ensemble techniques to combine the results obtained by each single regression model. In fact, the ensemble is just a machine learning model, whose arguments are the results of all

single models. There are several available ensemble methods for combining multiple models, the most simple and popular of which are majority voting, averaging, weighted averaging, median , product etc [202][204][206]. In this thesis the following ensemble methods have been applied: simple average, weighted average, majority voting, weighted voting, minimum, maximum and consensus combiner.

5.2. Combination Methods

This section briefly describes the main combination methods used in this chapter. This is mainly to highlight their theoretical background by presenting and analysing their mathematical models and demonstrating their strengths and weaknesses.

5.2.1. Min and Max Rules

Minimum and maximum rules are some of the simplest algebraic combiners, whereby functions simply take the minimum or maximum of the single learners' outputs [207]. If we have a set of B single learners $\{h_1, h_2, \ldots\ldots h_B\}$ and the output of h_i for instance x is $h_i(x) \in R$, these learners can be combined as:

$$H(x) = F\{h_1(x), h_2(x), \ldots\ldots\ldots, h_i(x)\} \tag{5.1}$$

where *F* is the simple minimum or maximum function, which can be given as:

$$H(x) = min_{i=1,\ldots B}\{h_i(x)\} \tag{5.2}$$

$$H(x) = max_{i=1,\ldots B}\{h_i(x)\} \tag{5.3}$$

In the literature, a theoretical framework of these rules and other rules such as sum, product, and median, was presented by Kittler et al. [208]. All these rules can be applied to combine models on measurement level; they are very simple, and do not require any further training. There is an obvious disadvantage of min and max combiners which might result in significant reduction in their accuracy. After training the models, if one of them got poor performance, these two combiners may select the minimal or maximal value of the outcome, which would be erroneous. Therefore, for both combiners min and max it is required that all single models are highly accurate in order for them to show good accuracy.

5.2.2. Simple Average

Simple averaging (AVR) rule is among the most basic and popular combination methods for numerical values. This approach is simple to apply as there is no need for previous training [209][210]. This rule combines the outputs of multiple models by taking the average (mean) directly. If we have a set of B single learners $\{h_1, h_2, \dots \dots . h_B\}$ and the output of h_i for instance x is $h_i(x) \in R$, the combined outputs can be defined as:

$$H(x) = \frac{1}{B}\sum_{i=1}^{B} h_i(x) \tag{5.4}$$

The average combiner is the preferred choice in many real applications, because of its efficiency and simplicity. It is also more trusted than min and max rules when compared with them, its outcomes are equally dependent on the outcomes of all models. However, reducing the error is generally hard to obtained since the errors in the ensemble learning are highly correlated as we train the individual learners on the same problem. Another disadvantage is the big variation of the performance between single models, so the AVR performance will probably not be higher than the best single model.

5.2.3. Weighted Average

Weighted averaging (WAVR) is an extension of the simple averaging rule. In this method, the outputs of all models are combined by taking the average with different weights indicating different levels of importance [202]. In general, the combined output can be defined as:

$$H(x) = \sum_{i=1}^{B} w_i \, h_i(x) \tag{5.5}$$

where w_i is the weight for h_i, and the weights w_i are normally supposed to be constrained by $w_i \geq 0$ and:

$$\sum_{i=1}^{B} w_i = 1 \tag{5.6}$$

The possibility of making more accurate model decisions in turn may positively affect ensemble results, whereas less accurate models contribute less to the final results. However, weighted average has a disadvantage relating to over training; some single models have

features allowing models to be over-trained, giving much better results through the training data than through the test data. SVM and Neural Network are examples of these models. This gives a justification of why some good models may get low weights, which will negatively affecting the results of the weighted average combined model. A way to overcome this limitation is by increasing the training set, in order to get training set accuracy equally proportioned to the test set accuracy for all models.

5.2.4. Majority Vote

Majority voting (MAJ) is considered one of the most common ways of voting in statistical analysis. In this method, the predictions of multiple models are combined. The predictions of every single model is represented as a single vote, and the final output is the one that obtains the majority votes of the models [202]. If there are three different models for a specific classification or regression problem [211]: $h_1(x), h_2(x), h_3(x)$ these models can be combined as:

$$H(x) = mode\{h_1(x), h_2(x), h_3(x)\} \tag{5.7}$$

Generally, a majority vote combiner involving votes from multiple learners (models) $h_1, h_2, \dots \dots . h_B$, can be defined as:

$$H(x) = arg \max_i \sum_{j=1}^{B} (h_j(x) = i) \tag{5.8}$$

The majority vote is one of the simple and popular elementary combiners which does not require any further training, and it differs from min, max, and average as it works in the abstract level. However, a limitation of this method is that all models are treated equally without taking into consideration the different features of each model.

5.2.5. Weighted Majority Vote

The weighted majority vote (WMAJ) is one of the most common and widely used combiners, and it is considered as a trainable version of the majority vote [212][213]. If we believe that some of the learners (models) are stronger than others, it is reasonable to weighting the decisions of these models by weighted voting. In this combination method, the votes are multiplied by a weight which can be determined according to the performance of each individual model. Then, by assigning a weight w_i to a learner h_i , we can get the output combined model of the ensemble as:

$$H(x) = arg\max_{j} \sum_{i=1}^{B} w_i h_i^j(x) \quad (5.9)$$

where w_i is the weight for h_i; and the weights w_i are normally supposed to be constrained by $w_i \geq 0$ and:

$$\sum_{i=1}^{B} w_i = 1 \quad (5.10)$$

Weighted majority vote can give better outcomes than both the majority voting and the best single model, and the main task is how to obtain the weights. Selecting the weights when the outputs are independent leads to minimum error for the weighted majority vote. If we got agreement weighted majority of the votes, then the final decision would be quite trustworthy.

5.2.6. Consensus Approach

Consensus method is another combination method where the outputs of different models in the ensemble interact in a cooperative way to reach an agreement on the final decision. By combining decisions made by different predictive models or classifiers, more efficient decision-making can be reached. The idea of the consensus method has been explored by different studies in several fields, such as statistics, web information retrieval, multi-agent systems, and geography [214][215][216][217][218]. The consensus method is different from the other fusion techniques, as it utilises relationships between individual models, and formulates a process similar to the process of decision making into the group of real experts, so that each individual expert can amend its own opinion according to the opinions of other experts in the group. This process is carried out with number of iterations until reaching a stage where no more changes are in the opinions. Based on the good features of this method, it has been chosen to be applied in order to build a good combiner model with the best possible decisions. According to DeGroot, for a group of predictive models or agents K, the consensus might be reached for the K individual models, and it can be mathematically presented as the following [219]:

For i=1, ..., K, let A_i indicate to the prediction which individual i assigns to the input sample Q. The predictions, $A_i, \ldots\ldots, A_k$,will be influenced by the different levels of expertise of the members of the group. It is supposed that, when the individual model knows the decisions made by each member of the group, it might wish to adjust its decision according to the other decisions. Furthermore, it is supposed that when individual i makes this adjustment, this adjusted decision is a linear combination of the predictions $A_i, \ldots\ldots, A_k$. Let B_{ij} indicate the

weight that individual i assigns to A_j while making the decision adjustment. It is supposed that all the B_{ij}'s are non-negative and:

$$\sum_{j=1}^{k} B_{ij} = 1 \tag{5.11}$$

Thus, when individual i is informed of the decisions made by the other members of the group, then his own decision can be adjusted from A_i to:

$$A_{i1} = \sum_{j=1}^{k} B_{ij}\ A_j \tag{5.12}$$

Let $\underline{B}$ indicate the $K \times K$ matrix where its (i ,j)th element is B_{ij}(i=1,……,K; j=1,……..,K). Since the elements are all nonnegative and the sum of the rows is one, $\underline{B}$ is a stochastic matrix. Let $\underline{A}$ and $\underline{A}^{(1)}$ be the vectors whose transposes are $\underline{\acute{A}} = (A_1, \ldots\ldots, A_k)$ and $\underline{A}^{(\acute{1})} = (A_{11}, \ldots\ldots, A_{k1})$. Then we could write the vector of adjusted decisions as:

$$\underline{A}^{(1)} = \underline{B}\ \underline{A} \tag{5.13}$$

The critical step in this process is that the above adjustment is iterated. After being informed of the decisions made by the other members of the group, $A_{11}, \ldots\ldots, A_{k1}$, it is supposed that individual i now adjust its own decision from A_{i1} to:

$$A_{i2} = \sum_{j=1}^{k} B_{ij}\ A_{j1} \tag{5.14}$$

This process will continue in the same way. Let A_{in} indicate the decision of individual i after n adjustments, and $\underline{A}^{(n)}$ indicate the vector whose transpose is $A^{(\acute{n})} = (A_{1n}, \ldots\ldots, A_{kn})$. Consequently:

$$\underline{A}^{(n)} = \underline{B}\ \underline{A}^{(n-1)} = \underline{B}^{(n)}\underline{A}\ ,\ (n = 2, 3, \ldots\ldots) \tag{5.15}$$

These adjustments or changes are supposed to be done indefinitely or until $\underline{A}^{(n+1)} = \underline{A}^{(n)}$ for some n. A definition has been given by DeGroot states that "a consensus is reached if and only if all k elements of $\underline{A}^{(n)}$ converge to the same limit as $n \to \infty$" [219]. This means that a

consensus is reached if and only if there exists a prediction A^* such that $\lim_{n\to\infty} A_{in} = A^*$, i= 1,…..,k. DeGroot emphasizes in his definition that a consensus is reached if and only if each row of the matrix $\underline{B}^n$ converges to the same vector, let we say $\underline{\omega}$ = ($\omega_1, \dots, \omega_k$). It could be said that this is obviously a sufficient condition for reaching a consensus. However, this condition is not necessary as can be shown by this simple example. Assume that $A_1 = A_2 = \cdots = A_k$, so whatever $\underline{B}$ is it will not make any difference since $\underline{A}^{(n)} = \underline{B}^n \underline{A} = \underline{A}$, n= 2, 3,… . So, no matter what weights B_{ij} are used the consensus A_1 can be reached. Reaching a consensus or not, in these two cases it depends not only on $\underline{B}$ (as stated by DeGroot's [219] condition) but also on $\underline{A}$.

5.3. Combination Methods Implementation

After reviewing the proposed combination methods with their mathematical models and their advantages and disadvantages, this section describes their implementation. Four applied methods of combining models do not require any previous training: minimum and maximum rules, average, and majority voting. For these methods simple functions (rules) have been used to aggregate the predictions of all individual regression models (the training data and test data) for both men and women which have been implemented in Chapter 4. The procedure for the ensemble process is shown in Figure 5.1.

The simple functions are min, max, mean, and mode. Min and max take the minimum and maximum predictions among the models, mean takes the average between the models' predictions, and mode finds the majority votes of values predicted by majority of the models. For the other combination methods (weighted average, weighted majority vote, and consensus) a training process is required to calculate the weights and then use them to find the final predictions, as explained by equations (5.5) and (5.9). All the calculations of the combining models have been executed in MATLAB R2018a. Tables 5.1 and 5.2 show the calculated weights for weighted average and weighted majority vote respectively.

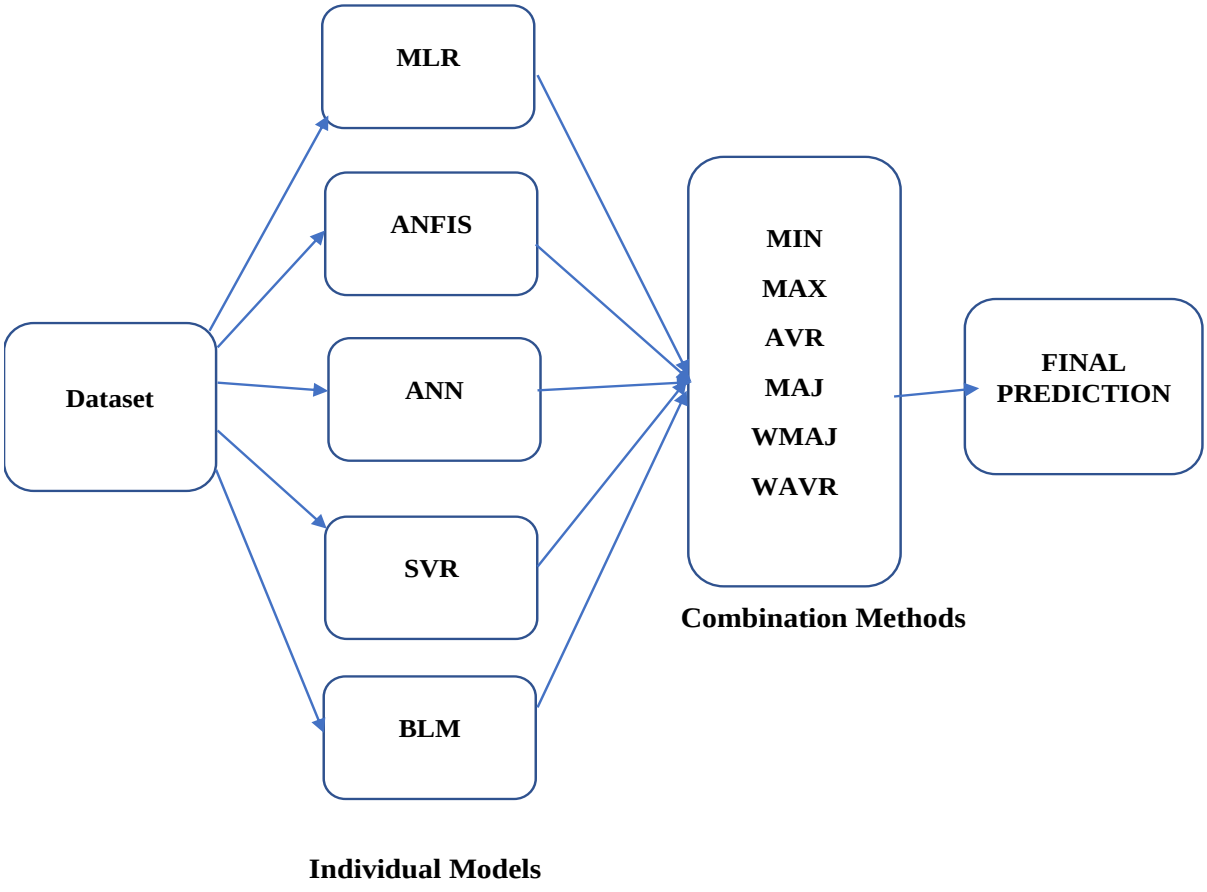

Figure 5.1: Ensemble process of combining multiple individual models

Models	The corresponding weights	
	Men data	Women data
MLR	0.01	0.16
ANFIS	0.19	0.50
ANN	0.11	0.21
SVR	0.22	0.13
BLM	0.47	0.0004

Table 5.1: The corresponding weights of the individual models using WAVR

Models	The corresponding weights	
	Men data	Women data
MLR	1.76	1.86
ANFIS	1.00	0.99
ANN	1.54	1.65
SVR	1.38	1.66
BLM	1.72	1.48

Table 5.2: The corresponding weights of the individual models using weighted majority vote method

5.4. Results and Discussion

This section represents the obtained results from each combination method described in section 5.2. Analysis of these results is undertaken to see how the performance of the ensemble system improves through combination methods, by evaluating their performance according to four important statistical evaluation measures: MSE, RMSE, MAPE, and the coefficient of determination R^2, respectively given in equations (3.2, 3.3, 3.4, 3.5), which were presented in Chapter 3. Seven different combination methods were applied to combine the predictions of the individual regression models proposed in the previous chapter. Four methods were directly used by applying simple rules (minimum, maximum, average, majority vote), and for the remaining three methods (weighted average, weighted majority vote, and consensus method) a training process was required. Tables 5.3 and 5.4 show the results of all combination methods for men training and test datasets respectively.

The results of the combination methods for the women training and test datasets are given in Tables 5.5 and 5.6. These tables present the combination outcomes by each combiner, whereby some of these combiners give good results like those obtained by the best individual model (ANFIS). While it is noted that the worst results are achieved when using the majority vote rule, which are similar to the performance of the minimum rule for both men and women datasets, the results obtained by these two rules are worse than any of the individual models as well, which is attributable to their treatment of all models equally, without considering weak models. Another interesting outcome of the ensemble modelling is that the WAVR has the best combination results, which is akin to the best results achieved by the ANFIS model. Also, the average rule as well as the consensus method were both having good results. The weighted majority vote rule is very close in performance to the average and consensus rules. The maximum rule is sensitive; it performed poorly for the women data, and was only good for the men data, where it is performed slightly better than a weaker individual model. Figures 5.2 and 5.3 show a graph of actual versus predicted values by all the combination methods.

These graphs show the bad performance of the minimum and majority vote combiners and the reasonably good performance of the other combination methods. However, the WAVR outperforms all the combination methods, as it is had a good fit between actual and predicted values. Table 5.7 shows the overall performance of all combination methods, evaluated by the statistical evaluation metrics MSE, RMSE, MAPE, and R^2. Based on the results, it can be observed that the best performance was given by the WAVR, with RMSE = 0.04 and R^2 = 0.99 for men data; and RMSE = 0.07 and R^2 = 0.99 for women data. Among other combination methods, average rule and consensus method both gave the same good accuracy in terms of RMSE and R^2 for both the men and women datasets. Figures 5.4 and 5.5 provide a

comparison of the performance of the combination methods by the evaluation metrics. It is clearly showed that the WAVR had the least error among all the other methods for both men and women datasets.

Table 5.8. presents a comparison of ANFIS regression model and WAVR against other studies of diabetes prevalence in KSA for years from 2015 to 2019. This table clearly provides a further validation of our individual model as well as the combined model, and both achieved good accuracy. Figure 5.6 shows a comparison between ANFIS model and WAVR against observed data from different studies of the total diabetes prevalence for both men and women in KSA for the period 2015-2019. This graph clearly shows that the results obtained by ANFIS model and WAVR are similar to those observed by different studies of diabetes, indicating the good accuracy and validation of the proposed model.

Combination methods						
AVR	WAVR	MAJ	WMAJ	MIN	MAX	CONSENSUS
9.66	9.75	9.70	9.76	9.30	9.88	9.65
9.79	9.87	9.60	9.92	9.60	10.03	9.80
9.98	10.04	10.00	10.12	9.70	10.20	9.98
10.19	10.25	9.70	10.38	9.70	10.43	10.19
10.44	10.50	9.71	10.67	9.71	10.80	10.43
10.76	10.82	9.80	11.07	9.80	11.20	10.76
11.25	11.29	10.65	11.45	10.65	11.60	11.25
11.86	11.87	11.78	11.92	11.78	12.00	11.85
12.28	12.30	12.30	12.32	12.18	12.36	12.29
12.67	12.69	12.70	12.71	12.56	12.71	12.67
13.04	13.05	13.10	13.07	12.89	13.10	13.04
13.30	13.32	13.17	13.35	13.17	13.40	13.31
13.56	13.58	13.40	13.57	13.40	13.70	13.55
13.79	13.80	13.80	13.78	13.58	14.09	13.79
13.87	13.89	13.73	13.89	13.73	14.01	13.86

Table 5.3: Combiners results (Men training data)

Combination methods						
AVR	WAVR	MAJ	WMAJ	MIN	MAX	CONSENSUS
14.18	14.10	14	14.21	13.8	14.5	14.2
14.48	14.32	14	14.48	13.9	15.1	14.46
14.82	14.56	14	14.83	14	15.8	14.82
15.18	14.81	14	15.21	14.1	16.6	15.2
15.45	14.99	14	15.46	14.1	17	15.44
15.81	15.22	15	15.83	14.2	17.7	15.82
16.15	15.44	15	16.15	14.2	18.2	16.16
16.47	15.63	15	16.44	14.3	18.6	16.46
16.77	15.81	15	16.70	14.3	18.8	16.74
17.02	15.94	14	16.94	14.4	18.9	17
17.31	16.10	19	17.22	14.4	19.1	17.3
17.59	16.25	14	17.50	14.5	19.6	17.6

Table 5.4: Combiners results (Men test data)

Combination methods						
AVR	WAVR	MAJ	WMAJ	MIN	MAX	CONSENSUS
6.90	6.95	6.58	6.89	6.58	7	6.90
7.12	7.12	6.89	7.12	6.89	7.26	7.12
7.23	7.29	7	7.23	7.08	7.4	7.23
7.52	7.52	7.49	7.52	7.49	7.6	7.52
7.82	7.80	7.85	7.83	7.7	7.9	7.82
8.22	8.19	8	8.22	8.14	8.31	8.22
8.78	8.68	8.58	8.80	8.55	8.9	8.78
9.10	9.04	8.82	9.11	8.99	9.25	9.10
9.44	9.42	9	9.44	9.25	9.55	9.44
9.79	9.81	9.77	9.79	9.69	9.88	9.79
10.15	10.17	10.06	10.14	10.06	10.19	10.15
10.33	10.43	10.13	10.33	10.13	10.53	10.33
10.59	10.65	10.35	10.58	10.35	10.78	10.59
10.67	10.77	10.4	10.65	10.4	10.88	10.67
11.13	11.13	10.42	11.14	10.42	11.85	11.13

Table 5.5: Combiners results (women training data)

Combination methods						
AVR	WAVR	MAJ	WMAJ	MIN	MAX	CONSENSUS
12.26	12.30	12	12.26	10.70	12.9	12.26
12.68	12.78	13	12.67	10.80	13.4	12.68
13.48	13.65	14	13.44	10.90	14.6	13.48
13.89	14.12	15	13.83	11.00	15.1	13.88
14.30	14.60	15	14.23	11.10	15.7	14.3
14.81	15.03	16	14.69	11.20	16.8	14.78
15.07	15.48	16	14.96	11.40	16.8	15.06
15.46	15.98	17	15.36	11.50	17.3	15.48
15.81	16.39	17	15.68	11.60	17.9	15.82
16.51	17.10	18	16.33	11.70	19	16.5
16.85	17.50	19	16.63	11.80	19.5	16.82
17.20	17.97	20	17.00	12.00	20.1	17.22

Table 5.6: Combiners results (women training data)

Combination methods	Men				Women			
	MSE	RMSE	MAPE	R^2	MSE	RMSE	MAPE	R^2
AVR	0.0033	0.0572	0.0039	0.9986	0.0156	0.1249	0.0107	0.9923
WAVR	**0.0022**	**0.0469**	**0.0032**	**0.9990**	**0.0059**	**0.0766**	**0.0061**	**0.9971**
MAJ	0.1779	0.4218	0.0233	0.9213	0.0925	0.3042	0.0271	0.9541
WMAJ	0.0110	0.1048	0.0077	0.9951	0.0177	0.1331	0.0114	0.9912
MIN	0.2030	0.4506	0.0315	0.9102	0.0789	0.2809	0.0236	0.9609
MAX	0.0399	0.1997	0.0151	0.9824	0.0721	0.2685	0.0185	0.9642
CONSENSUS	0.0033	0.0574	0.0039	0.9985	0.0156	0.1249	0.0107	0.9923

Table 5.7: Statistical evaluation metrics results for all combination methods for both men and women

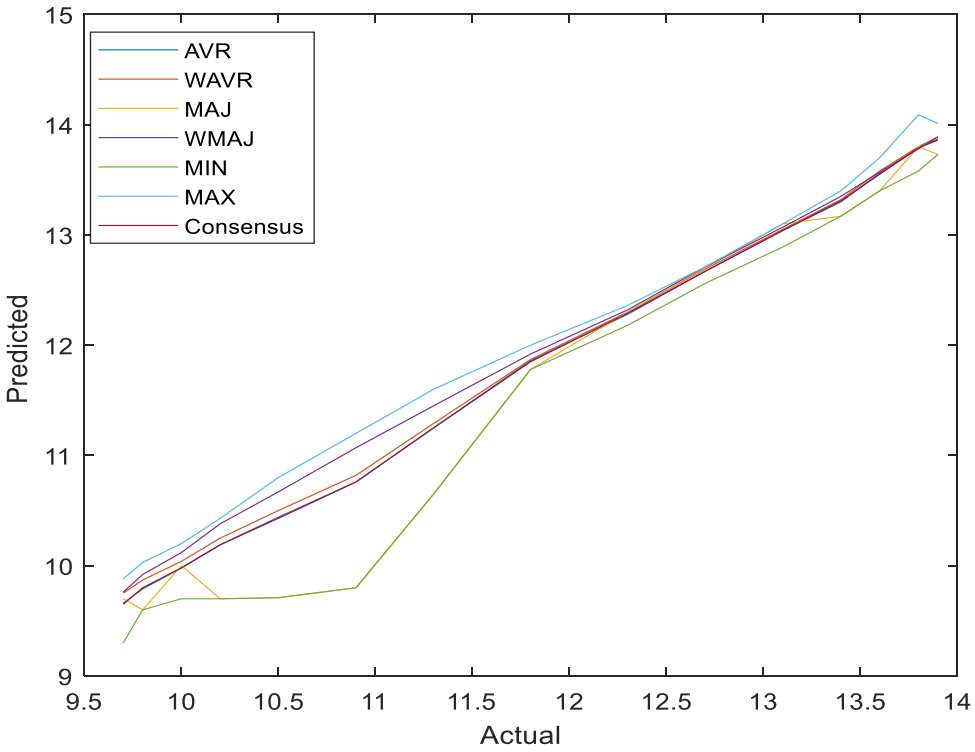

Figure 5.2: Actual data against predicted for all combination methods (men morbidity data 1999-2013)

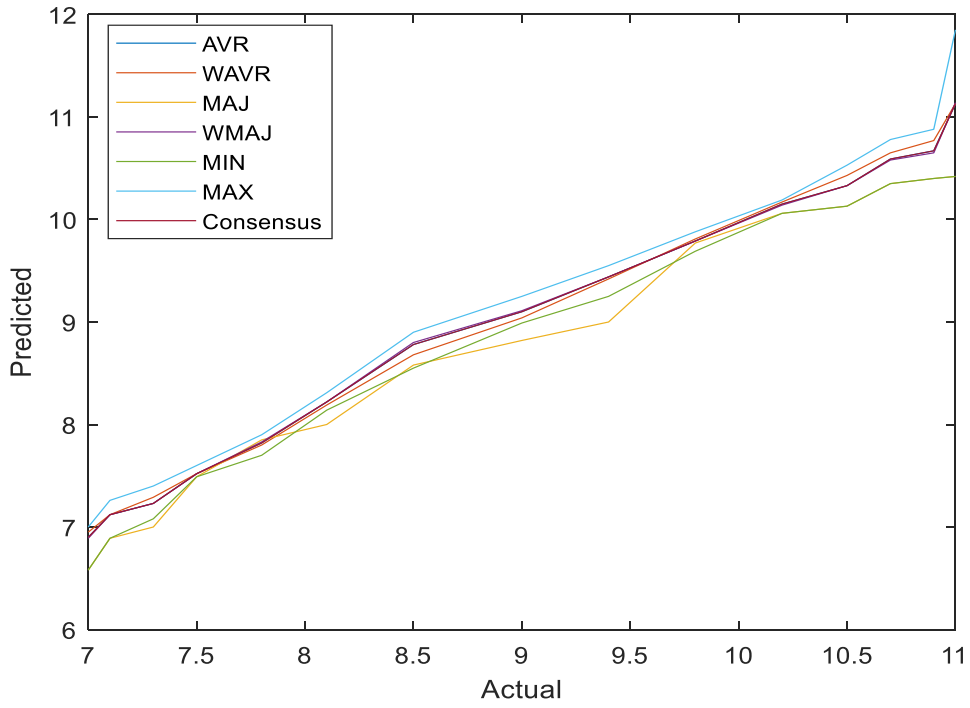

Figure 5.3: Actual data against predicted for all combination methods (women morbidity data 1999-2013)

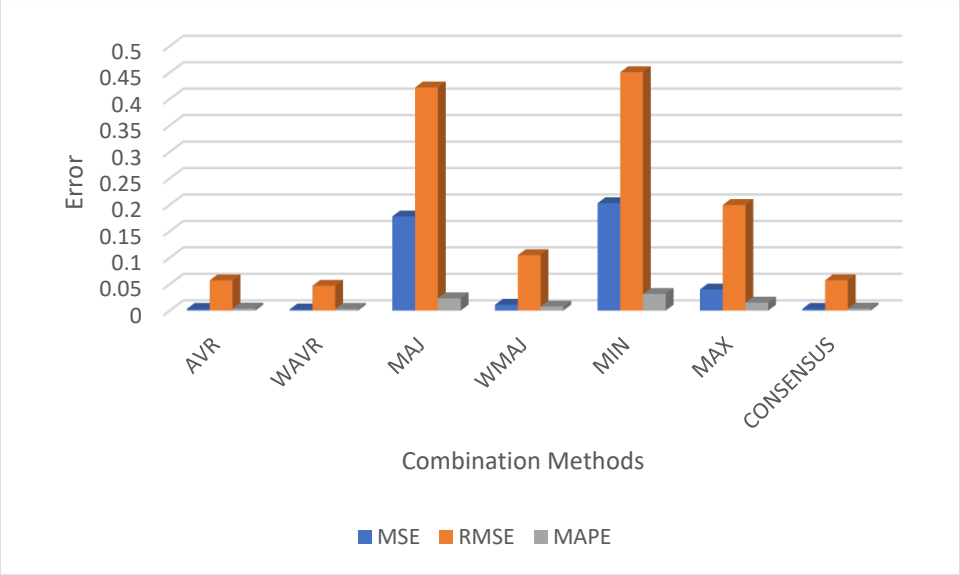

Figure 5.4: Performance metrics of combination methods for men data

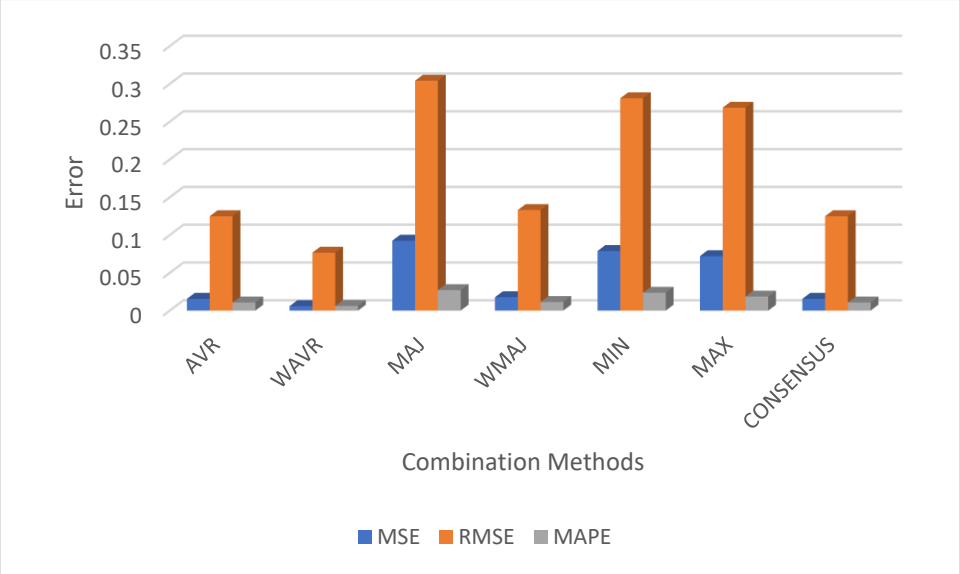

Figure 5.5: Performance metrics of combination methods for women data

Years	Diabetes prevalence estimates by different studies	ANFIS model	WAVR model
2015	*KSA Health Profile*, WHO (2015) Men: 14.8 Women: 11.7 Total: 13.3	Men: 14.4 Women: 12.8 Total:13.6	Men: 14.32 Women: 12.78 Total: 13.55
2016	*Diabetes Country Profile: KSA*, WHO (2016) Men: 14.7 Women: 13.8 Total: 14.3	Men: 14.8 Women: 13.1 Total:13.95	Men: 14.56 Women: 13.65 Total: 14.11
2017	*Family Health Survey*, Saudi General Authority for Statistics (2017) Men: (10.4) *, Men: (14.9) ** Women:(9.8) *, Women:(14.5) ** Total: (10.1) *, Total:(14.7)**	Men: 15.2 Women: 13.8 Total:14.5	Men: 14.81 Women: 14.12 Total: 14.46
2018	*Household Health Survey*, Saudi General Authority for Statistics (2018) Men: (10.3) *, Men: (16.8) ** Women:(9.9) *, Women:(14.2) ** Total: (10.1) *, Total:(15.5) **	Men: 15.5 Women: 14.4 Total:14.95	Men: 14.99 Women: 14.60 Total: 14.79
2019	Diabetes Atlas (9th edition), IDF (2019) Total: 15.8 (10.3-17.7) (95% confidence interval)	Men: 16 Women: 14.8 Total:15.4	Men: 15.22 Women: 15.03 Total: 15.13

* Prevalence rate for Saudi population (15 years and over)

** Prevalence rate for Saudi population (age-adjusted 25 years and over)

Table 5.8: A comparison of ANFIS model and WAVR against other studies of diabetes prevalence in KSA, 2015-2019.

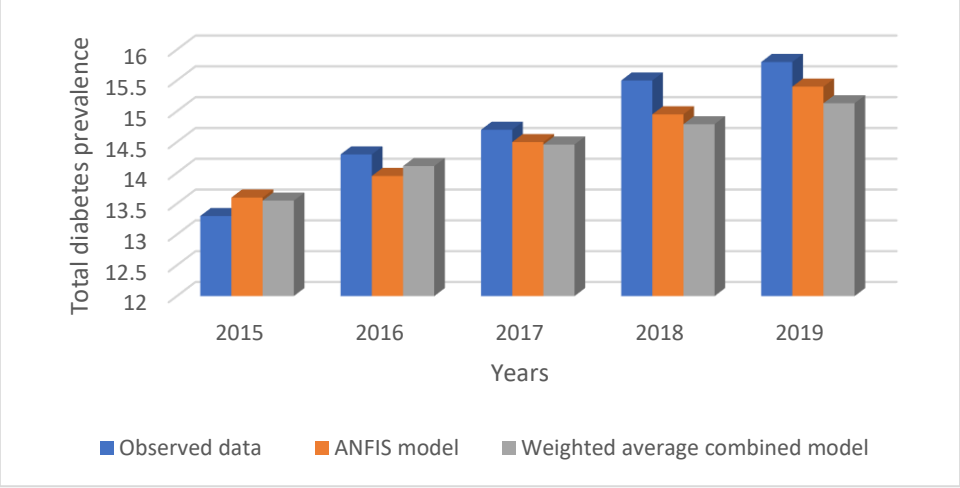

Figure 5.6: Total diabetes prevalence by ANFIS model and WAVR against observed data of diabetes prevalence in KSA, 2015-2019

5.5. Summary

This chapter explored one of the most popular topics of research in the field of machine learning, ensemble and combination methods. Average, weighted average, majority vote, weighted majority vote, minimum, maximum, and consensus approach are the selected methods presented. A brief description of these methods was provided, including their mathematical models, and their advantages and limitations. The results obtained by these combination methods clearly show that the best combiners provide similar results to those obtained by the best individual regression model (ANFIS model). As shown in the above results evaluating the performance of the proposed combination methods, WAVR outperforms all combination methods, with a lower error rate. Furthermore, the performance of the average and consensus combiners was as good as the WAVR in terms of minimal errors. On the other hand, the other combination methods, including minimum majority vote, and maximum, had bad performance for the women dataset, worse than the weakest performance of individual models. In addition, to further validate the best individual model and the best combined model, a comparison between them against observed data from different studies was undertaken, which indicated the good accuracy achieved by these models. The next chapter presents classification modelling, with further discussion and analysis.

Chapter 6
Classification Modelling

6.1. Introduction

This chapter presents the classification modelling approach including LD, SVM, KNN, and NPR classification techniques, with four SVM kernel functions: Linear, Gaussian, Quadratic, and Cubic. Two types of KNN algorithms were employed: Fine KNN and Weighted KNN. These methods were applied to investigate model ability to classify diabetes prevalence rates and the predicted trends of the disease according to the associated risk factors (smoking, obesity, and inactivity). Section 6.2 describes the selected machine learning classification methods, while their implementation is illustrated in Section 6.3. Section 6.4 presents the analysis and discussion of the results of these methods and the comparisons of their performance using accuracy measurement. A summary of the chapter is given in Section 6.5.

6.2. Machine Learning Classification Methods

This section describes the selected machine learning classification methods: Linear SVM, Gaussian SVM, Quadratic SVM, Cubic SVM, Fine KNN, Weighted KNN, Linear Discriminant (LD), and neural net pattern recognition (NPR). This is mainly to present the mathematical background of these methods and to highlight their individual operational characteristics.

6.2.1. Support Vector Machine

SVM is one of the most common used supervised machine learning algorithms, which can be used for both regression analysis and classification tasks; it was applied in the regression modelling in Chapter 4. SVMs are considered as new types of pattern classifiers related with learning algorithms, which can identify patterns and analyse data. They have been successfully applied on various types of applications such as verification and detection, text categorisation and detection, recognition information, recognition of handwritten characters, speakers, and speech verification [220]. SVM can perform linear classification and predict non-linear separable patterns by mapping its inputs into a hyperplane (high-dimensional feature spaces). Minimising the upper bound of the generalisation error is the main aim of SVM by maximising the transaction between the data and the separating hyper plane. The performance of SVMs tends to be accentuated when using new data not included through the training process, because of its fundamental classification principle, which produces new examples into the related class.

In a standard classification case, the components of the dataset include several parameters X_1, X_2,.., X_3 and one or multiple variables for classes C_1, C_2,….., C_P. The objective is to develop a classifier to appoint the inputs (data points) to their classes (C_1, C_2,….., C_P) by using the N data points in the training set. Thus, for every point in the training set $\{x_n\}_{n=1}^{N}$ a class t_n should be predicted where $t_n \in \{-1,1\}, n = 1, \dots, N$. Then the classifier can be defined in the following equation:

$$y(x,w) = \sum_{i=1}^{J} w_i \, \varphi_i(x) + b \tag{6.1}$$

where $w \in R^J$ is the wight vector; $\varphi(.)$ is the transformation function; and b $\in R$ is the constant. If the data space is nonlinearly separable, SVMs use an appropriate mapping (φ) of the input data to a high-dimensional space which will be arranged by the kernel function. The kernel function, defined as follows:

$$K(x,x') = \varphi(x).\varphi(x') \tag{6.2}$$

Separating the hyperplane by kernel solution does not require knowledge of φ, but only of K. Therefore, any kernel function can be used in a larger dimension space. Different kernel functions can be most efficient this depends on the nature of the dataset. The following sections describe the used kernel functions.

6.2.1.1. Linear SVM

The linear kernel function is one of the simplest types of kernel functions and the basic way to use SVM classifier. It is calculated by the inner product of two vectors, x_i and x_j, plus an optional constant, b. Thus, the Kernel algorithms using a linear kernel are often the same as their non-kernel equivalents. The linear kernel provides a signed measure of the similarity between x_i and x_j, where the angle between the two points helps determine their similarity, and can give negative values of K. The following equation shows the linear kernel function:

$$K\langle x_i, x_j \rangle = x_i^T . x_j + b, \tag{6.3}$$

where *b* is a constant.

6.2.1.2. Gaussian SVM

Gaussian SVM kernel function is a type of radial basis function (RBF), which represents the most generalised form of kernelization and the most commonly used kernels. It can be described as a general-purpose kernel, and it is used when there is no previous information regarding the data. Unlike linear kernel, the Gaussian kernel only relies on the Euclidean distance between x_i and x_j, and it also depends on the assumption that similar points in the feature space are close to each other [221]. This final assumption is logical in many situations; thus, the Gaussian kernel is widely used in practice. The Gaussian kernel can be mathematically represented as follows:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{6.4}$$

where $\gamma > 0$ is a given parameter, which must be accurately tuned as it directly affects kernel performance.

6.2.1.3. Quadratic SVM

Quadratic kernel is a common form of polynomial kernel, commonly applied in speech recognition. The computation of this function is less intensive than the Gaussian kernel function and can be alternatively used if using the Gaussian kernel is too expensive. These functions do not generalize well because that higher order kernels tend to overfit the training data. The quadratic function can be mathematically expressed as follows:

$$K(x_i, x_j) = 1\text{-}\frac{\|x_i - x_j\|^2}{\|x_i - x_j\|^2 + b} \tag{6.5}$$

6.2.1.4. Cubic SVM

SVM classification method is helpful when facing a problem of low memory space. SVM can find a hyperplane in multidimensional space which separate the label classes by the best possible way. The cubic SVM classifier is a form where the kernel function of the classifier is cubic, and its mathematical expression can be given as the following:

$$K(x_i, x_j) = (x_i^T x_j + 1)^3 \tag{6.6}$$

6.2.2. K-Nearest Neighbours

KNN is one of the simplest machine learning methods that can be used for regression as well as for classification, but it is commonly used for the classification problems to classify data input into pre-defined classes (k). This method dated back to 1968 when first introduced by Cover and Hart [222]. The KNN method considered as a nonparametric method, this means there is no assumptions will be made on the involved data. Also, this algorithm does not immediately learn from the training data but rather it stores the dataset and at the time of classification it makes an action on the dataset so, it is called a lazy learner algorithm. In this method, the input data includes all the k closest training points in the feature space, where k is an integer.

The data are classified by determining the most common class among the k nearest neighbours. These neighbours are members in the dataset in which this method was first trained and are identified using the distance from the test sample. This means that during the testing, the class which appears most commonly amongst the neighbouring classes of the test sample under observation becomes the class to which this individual test sample belongs. In other words, after the training phase any new data obtained are classified into a similar category to these new data. The accuracy of the KNN classifiers is increased with a decreasing number of neighbours. This leads to increasing the complexity of the classifier model; however, it does not ensure that the new samples will be classified correctly. There are many types of distance function between the samples, the most popular used function is the Euclidean distance, which presented in the following equation:

$$d = \sqrt{\sum_{k=i}^{n} (X_{ik} - X_{jk})^2} \tag{6.7}$$

where X_1 and X_2 are input samples; and k is the number of values in each sample.

There are six types of KNN classifiers available in MATLAB: Fine KNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, and Weighted KNN. Some of these types of KNN algorithms make use of Euclidean distance to determine the nearest neighbours (Fine, Medium, and Coarse KNN algorithms). The Cosine KNN algorithm employs a Cosine distance metric as given in equation (6.8). For Cubic KNN algorithm a cubic distance metric is employed as in equation (6.9). For Weighted KNN algorithm, a distance weight is employed, as in equation (6.10).

$$d = (1 - \frac{x_i \, x_j'}{\sqrt{(x_i \, x_i')(\, x_j \, x_j')}}) \tag{6.8}$$

$$d = \sqrt[3]{\sum_{k=i}^{n} |x_{ik} - x_{jk}|^3} \tag{6.9}$$

$$d = \sqrt{\sum_{k=i}^{n} w_i (X_{ik} - X_{jk})^2} \tag{6.10}$$

A brief description of these algorithms as given in MATLAB is provided below:

- **Fine KNN**: Creates finely detailed distinctions between classes with the number of neighbours set to 1.
- **Medium KNN**: Creates less distinctions than a Fine KNN with the number of neighbours set to 10.
- **Coarse KNN**: Creates coarse distinction between classes, with the number of neighbours set to 100.
- **Cosine KNN**: Employs the cosine distance metric.
- **Cubic KNN**: Employs the cubic distance metric.
- **Weighted KNN**: Uses distance weighting.

Only Fine KNN and Weighted KNN were chosen along with other classifiers to classify our data.

6.2.3. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), also known as Fisher's LDA, is a commonly used statistical machine learning method for data classification and dimensionality reduction problems [223]. The main aim of LDA is to find the best linear approximations of object feature vectors for efficient and sensible use in a variety of classification tasks. The idea behind this method is maximising the ratio of between-class variance to the within-class variance in any specific dataset, in that way ensuring maximal separability. In other words, utilising a linear transformation process for projecting high dimensional feature vectors into a lower dimensional to best separates the data groups [224]. The advantage of LDA is its implementation simplicity, whereby a linear combination of features is used to distinct classes of samples. Furthermore, it is easily managing the case where the within-class frequencies

are unequal, and their performances has been assessed on randomly generated test data. On the other hand, the simplicity of its execution leads to a drawback, particularly if the class differences are low. In this case, LDA assumes the mean as discriminating factor and not the variance, which in turn leads to overfit the data.

Consider a dataset that contains P different class labels and let $X = \{X_1, X_2, \ldots, X_P\}$ denote to the set having these P classes. If the dataset is d-dimensional (apart from the class label), the class matrix Xl_{si} for class $X_i \in X$ can be defined as:

$$Xl_{si} = \begin{bmatrix} c_{11} & c_{12} \ldots\ldots & c_{1d} \\ c_{21} & c_{22} \ldots\ldots & c_{2d} \\ c_{n1} & c_{n2} \ldots\ldots & c_{nd} \end{bmatrix} \tag{6.11}$$

where each row in the class matrix $Xl_{si} \in$ class X_i .

The mean of the class matrix Xl_{si}can be defined as an array containing the means of each column of the class matrix Xl_{si}, represented by:

$$\mu Xl_{si} = [m_{c1}\ m_{c2} \ldots\ldots m_{cd}] \tag{6.12}$$

where m_{ci} indicates the mean of the i^{th} column feature of the class matrix Xl_{si}. Then the overall mean can be defined as the mean of all the class means, presented by:

$$\mu = \frac{1}{P}\sum_{i=1}^{P} \mu Xl_{si} \tag{6.13}$$

Then, the next calculation for the mean corrected class matrix is defined as the matrix whose every column object in each row is subtracted by the corresponding column object of the overall mean (μ):

$$Xl_{si}{}^{mc} = Xl_{si}[c][e] - \mu[e] \tag{6.14}$$

where $c = \{1,2, \ldots\ldots, n\}$ indicates row indices; and $e = \{1,2, \ldots\ldots, d\}$ indicates the column indices of the class matrix Xl_{si}.

Then, the definition of the covariance matrix of the class matrix Xl_{si} is given by:

$$Xl_{si}{}^{cov} = \frac{Xl_{si}{}^{mc^T} * Xl_{si}{}^{mc}}{ni} \quad (6.15)$$

where $Xl_{si}{}^{mc^T}$ represents the transpose; and ni indicates the number of row entries of the mean corrected class matrix $Xl_{si}{}^{mc}$. If the dataset contains P classes, then the pooled group covariance matrix of the dataset can be given as:

$$X = \frac{1}{N}(\sum_{j=1}^{P} n_j \, Xl_{sj}{}^{cov}) \quad (6.16)$$

where $N = \sum_{j=1}^{P} n_j$.

For classifying a new data point $y = \{ y_1, y_2, \ldots\ldots, y_d\}$ into one of the P classes, the LDF for every P class can be defined as:

$$f_i = \mu Xl_{si} X^{-1} y^T - 0.5 \mu Xl_{si} X^{-1} \mu^T Xl_{si} + \ln(q_i) \quad (6.17)$$

where X^{-1} represents the inverse value of the pooled group covariance matrix X; $y^T$is the transpose of input data point vector y; and $\mu^T Xl_{si}$ are the transposes of the class mean vector μXl_{si}. The data point y is appointed to the class with the maximum value of the LDF.

6.2.4. Neural Networks Pattern Recognition

Pattern recognition is one of the most important aspects of computer science, which can be defined as the process of observing a system or event to recognise and distinguish patterns by utilising machine learning methods. It could also be defined as a data analysis approach for classifying the data based on their background or statistical information derived from patterns, and then making reasonable decisions about the categories of the patterns. Its advantages include its ease of use and practical possibility of widespread application. Despite many years of research, the design of a multi-purpose machine for pattern recognition is still a long way off. According to Ross [225], “The more relevant patterns at your disposal, the better your decisions will be”. This is promising news for AI supporters, as it is possible to teach computers how to recognize patterns. In fact, a variety of successful computer programs which have been used in diagnosing diseases, bank credit scoring, and landing airplanes which in some way depend on pattern recognition. There are two possible ways can be used in recognising patterns include classification and cluster. Classification is a supervised learning where a proper class is appointed according to a pattern that is extracted from training datasets. Whereas clustering is unsupervised learning, and it is work by dividing the data into

groups which in turn help for making decisions. In this chapter artificial neural network method have been used for pattern recognition as a classification method.

ANNs can be defined as statistical models with nonlinear way of modelling the data which simulate the way of biological NNs. Statistical pattern-based methods have been the most common and applied in different practices. However, ANN models have gained more popularity and attractiveness due to offering more efficient and successful ways of dealing with pattern recognition problems in many cases. In contrast with conventional pattern methods, ANN can easily handle complex or multi complicated tasks. The conventional methods which used to deal with pattern recognition problems can be divided into three types: statistical, structural, and hybrid methods. However, insufficient results can be obtained by both the statistical and structural methods when they are used to solve the complicated pattern recognition problems only. For example, when applying the structural method, the performance can be poor and unable to handle noise patterns. In the same way, the statistical method is incompetent in dealing with information related to patterns tasks. Therefore, the combination of the two methods was widely accepted by researchers which in turn leads to the hybrid method. However, in present time ANN models can be used instead of the conventional hybrid methods, this is because of the good results obtained in pattern recognition even in more complicated problems [226].

The function of ANN is advantageous compared to other techniques used for pattern recognition, as it has more flexibility with noticeable success. These good features due to the good performance of the network, which is increased using feedback information achieved from the difference between the actual input and the required output. Then, this information will be used to set the communications between the neurons at the input layer which in turn makes the actual outcome consistence with the required one. In addition, the specified algorithms of this method characterised by self- adaptive and self-organizing which add more efficiency of the network in pattern recognition. ANN is a trainable method which can be applied in both classification and regression tasks (it was used as a regression model presented in Chapter 4). An ANN involves interconnected processing units known as neurons, which work together to produce outcomes [227].

The process of training the ANN is iteratively updated to achieve the required MSE, and provide an optimal generalization for test data [225]. One of the most popular neural network families used for pattern classification problems is the feed-forward network, which contains RBF and MLP (supervised). Also, self-organizing map (SOM) is another common network which is mostly applied for feature mapping and data clustering (unsupervised). The ANN structure and activation function are described in detail in Chapter 4.

6.3. Implementation

In Chapter 4, different regression models were developed to describe the development of diabetes disease in the Saudi population, by integrating diabetes data on the population (adults aged > 25 years) and for both genders (men and women), along with the dataset of the behavioural (modifiable) risk factors (smoking, obesity, and inactivity). In this chapter, several classification methods were employed to classify the prevalence rate of diabetes disease into five different classes using different classification methods: LD, SVM with linear kernel, SVM with Gaussian kernel, SVM with quadratic kernel, SVM with cubic kernel, Fine KNN, Weighted KNN, and NPR. All these methods were implemented in MATLAB using Classification Learner App, except the neural net pattern recognition method, which was implemented using Neural Net Pattern Recognition app [228].

The latest versions of MATLAB incorporate a statistics and machine learning toolbox, which includes a great number of machine learning techniques including the Classification Learner app, which provides easy access to several supervised learning methods, which can be used for different applications in the real world, such as image and speech recognition, medical diagnosis, statistical and predictive analytics. In this research, data were prepared for classification by converting the continues values of diabetes dataset into discrete. The morbidity data of diabetes (output) were grouped into five classes for both men and women using (discretize) function in MATLAB.:

- Class 1, low (7% to 10.5%).
- Class 2, medium low (8.55% to 10.9%).
- Class 3, medium (10.71% to 14.8%).
- Class 4, high (13.02% to 16.15%).
- Class 5, extremely high (15.61% to 17.59%).

All the datasets of the three risk factors (inputs) with the labelled classes of diabetes morbidity data (outputs) were fitted into the (classification learner) for the training stage, using the default options for each method.

For SVM classification method, the characteristics of this algorithm regarding the kernel function, kernel scale, and other features are given in Table 6.1. In addition, the information of KNN classification method, such as the number of neighbours, distance metric, and distance weight, are shown in Table 6.2. After that, the pre-trained classifiers were exported to the MATLAB workspace, where they were used to make predictions for data chosen randomly from our data points using the MATLAB function *predictFcn*. This testing stage tests the accuracy of the classification models.

For pattern recognition classification method, the Neural Net pattern recognition app in MATLAB was used, which can be used directly from the Apps tab or by typing (*nprtool*) in the command window. In this app, a two-layer feed-forward network with sigmoid hidden layer and softmax output neurons were used, and the network was trained with scaled conjugate gradient backpropagation. The structure of the network consists of input, hidden, and output layers. Each layer has number of neurons or elements; in our experiment, 3 neurons were used in the input layer (representing the risk factors), and 5 neurons in the output layer (representing the five classes). There is no exact way to define the number of neurons in the hidden layer, and it is commonly chosen according to trial-and-error method to get the best network performance, which in this experiment was achieved with 10 neuros. The performance of all classifiers is evaluated by accuracy. In addition, for the classification learner's methods, further comparisons were conducted between the classifiers according to the prediction speed and training time. All the results are presented and discussed in the next section.

Classifier Characteristics	Model 1	Model 2	Model 3	Model 4
Preset	Linear SVM	Quadratic SVM	Cubic SVM	Medium Gaussian SVM
Kernel function	Linear	Quadratic	Cubic	Gaussian
Kernel scale	Automatic	Automatic	Automatic	1.7
Box constraint level	1	1	1	1
Multiclass method	One-vs-One	One-vs-One	One-vs-One	One-vs-One
Standardize data	true	true	true	true

Table 6.1: Characteristics of SVM classification models

Classifier Characteristics	Model 1	Model 2
Preset	Fine KNN	Weighted KNN
Number of neighbours	1	10
Distance metric	Euclidean	Euclidean
Distance weight	Equal	Squared inverse
Standardize data	true	true

Table 6.2: Characteristics of KNN classification models

6.4. Results and Discussion

This section presents and discusses the results obtained from each classification model discussed in the previous section. The performance evaluation in terms of accuracy of each developed model was determined, and the developed classifiers were compared using the Classification Learner App, according to prediction speed and training time. As mentioned in the previous section, diabetes data was pre-processed by converting the continuous values

into discrete ones, as shown in Table 6.3 for both the men and women datasets. Table 6.4 and Table 6.6 show the resulting classes by the trained models for both men and women, respectively. The pre-trained classifiers tested with samples picked randomly from the training datasets in order to check its classification capability; the testing results are given in Table 6.5 and Table 6.7 for men and women, respectively.

All classification methods were compared according to their accuracy as shown in Table 6.8 and Table 6.9 for the men and women datasets, respectively. For the men data, Gaussian SVM, Weighted KNN, and Neural net pattern recognition models gave the same accuracy (92.6%), which was the highest value achieved among the models. The remaining models also performing well, with similar accuracy of 88.9%.

Table 6.8 also shows the accuracy prediction speed and training time for all models used by classification learner app, the maximum training time was taken by Linear SVM with 10.177 sec, while the Weighted KNN takes the least training time of 1.2459 sec. Thus, the prediction speed (observation/second) is maximum for Weighted KNN with 310 obs/sec, and minimum for Linear SVM with 77 obs/sec.

Table 6.9 provides different accuracies given by models when applying to the women data. The highest accuracy was given by Gaussian SVM, Fine KNN, and Weighted KNN models of 96.3%, followed by neural net pattern recognition with an accuracy of 92.6%. Also, there are three other models have a good performance with the same accuracy of 85.2%: Linear discriminant, Quadratic SVM, Cubic SVM. Lastly, the Linear SVM based model have the lowest accuracy of 77.8%. In addition, in terms of prediction speed and training time, as for the men data, the maximum training time was taken by Linear SVM with 5.4628 sec, whereas the Fine KNN takes the least training time of 1.5032 sec. However, the prediction speed (observation/second) was highest for Weighted KNN with 340 obs/sec, and lowest for Quadratic SVM with 91 obs/sec.

Figures 6.1 to 6.7 present the confusion matrixes of the resulting models from the classification learners app for both men and women datasets respectively. For each model the diagonal green squares represent the correct prediction ratio of the predicted classes over true classes, while the red squares give the incorrect class ratio. The results show the level of the decision ratio between the true value and the predicted value. Figure 6.1 show the confusion matrix of the predicted and true class categories obtained by LD model, for men data there are 3 data sample misclassified, 1 data for each of the classes (class 1, class 2, and class 3). For women data there are 4 data sample misclassified, 2 data for class 2 and 2 data for class 3. So, the overall classification accuracy of LD model is 88.9% and 85.2% for men and women datasets respectively. Figure 6.2 show the maximum misclassification among models which obtained

by LSVM, a total number of 9 data sample misclassified for both men and women data for the classes 1, 2, 4, and 5, which gave an average classification accuracy of 83.4% for men and women datasets. Figures 6.3 and 6.4 demonstrate the confusion matrixes of the predicted and true class obtained by Quadratic SVM and Cubic SVM for men and women respectively, as can be seen 7 data sample misclassified for both models for men and women data for the classes 1, 2, 4, and 5, with classification accuracy of 88.9% for both models for men and 85.2% for both models for women respectively. Figures 6.5, 6.6, and 6.7 present the confusion matrixes of the best performing classifiers, Gaussian SVM, Fine KNN, and Weighted KNN, the misclassification data samples range between 1 to 3 for all models for men and women, with an average classification accuracy of 93.8% for men and women datasets.

From the obtained results using men datasets, the error rate for diabetes prevalence classification was the same for all models (3%) except for Gaussian SVM (2%). For the women data, the achieved results show that the same error rate was given by LD, Quadratic SVM, and Cubic SVM (4%), while Linear SVM gave the maximum error rate of 6%. Moreover, as the best results again were given by Gaussian SVM, Fine KNN, and Weighted KNN models, all indicating the same error rate of 1%.

Figure 6.8 compares the classification results of all models in terms of accuracy for both the men and women datasets. It can be seen that the performance of Gaussian SVM, Fine KNN, and Weighted KNN models was better for women data than for men data, while the performance accuracy of LD, Linear SVM, Quadratic SVM, and Cubic SVM was higher using men data than women data. The NPR model had the same performance accuracy for both men and women datasets.

Years	Morbidity data men	Morbidity after discretizing (men)	Morbidity data women	Morbidity after discretizing (women)
1999	9.7	1	7	1
2000	9.8	1	7.12	1
2001	10.0	1	7.28	1
2002	10.2	1	7.5	1
2003	10.5	1	7.79	1
2004	10.9	2	8.14	1
2005	11.3	2	8.55	2
2006	11.8	2	8.99	2
2007	12.3	2	9.43	2
2008	12.7	3	9.84	2
2009	13.1	3	10.19	2
2010	13.4	3	10.48	2
2011	13.6	3	10.71	3
2012	13.8	3	10.88	3
2013	13.9	3	11	3
2014	14.18	3	11.97	3
2015	14.48	4	12.37	3
2016	14.82	4	13.02	4
2017	15.18	4	13.53	4
2018	15.45	4	14.08	4
2019	15.81	4	14.89	4
2020	16.15	4	15.15	4
2021	16.47	5	15.61	5
2022	16.77	5	16.03	5
2023	17.02	5	16.62	5
2024	17.31	5	17	5
2025	17.59	5	17.36	5

Table 6.3: Morbidity data with discretized classes for men and women, 1999-2025

Classifiers	LD	LSVM	QSVM	CSVM	Gaussian SVM	Fine KNN	Weighted KNN	NPR
1999	1	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1	1
2001	1	1	1	1	1	1	1	1
2002	1	1	1	1	1	1	1	1
2003	1	2	1	1	1	1	1	1
2004	1	2	2	2	2	2	2	2
2005	2	2	2	2	2	2	2	2
2006	2	2	2	2	2	2	2	2
2007	2	2	2	2	2	2	2	2
2008	3	3	3	3	3	3	3	2
2009	3	3	3	3	3	3	3	3
2010	3	3	3	3	3	3	3	3
2011	3	3	3	3	3	3	3	3
2012	3	3	3	3	3	3	3	3
2013	3	3	3	3	3	3	3	3
2014	3	3	3	3	3	3	3	3
2015	4	4	4	4	4	4	4	4
2016	4	4	4	4	4	4	4	4
2017	4	4	4	4	4	4	4	4
2018	4	4	4	4	4	4	4	4
2019	4	4	4	4	4	4	4	4
2020	4	4	4	4	4	4	4	4
2021	5	4	5	5	5	5	5	5
2022	5	5	5	5	5	5	5	5
2023	5	5	5	5	5	5	5	5
2024	5	5	5	5	5	5	5	5
2025	5	5	5	5	5	5	5	5

Table 6.4: Classification models results for men (training data), 1999-2025

Random points	Years	Classifiers							
		LD	LSVM	QSVM	CSVM	Gaussian SVM	Fine KNN	Weighted KNN	NPR
22	2020	4	4	4	4	4	4	4	4
6	2004	1	2	2	2	2	2	2	2
3	2001	1	1	1	1	1	1	1	1
16	2014	3	3	3	3	3	3	3	3
11	2009	3	3	3	3	3	3	3	3
7	2005	2	2	2	2	2	2	2	2
17	2015	4	4	4	4	4	4	4	4
14	2012	3	3	3	3	3	3	3	3
8	2006	2	2	2	2	2	2	2	2
5	2003	1	2	1	1	1	1	1	1
21	2019	4	4	4	4	4	4	4	4
25	2023	5	5	5	5	5	5	5	5
27	2025	5	5	5	5	5	5	5	5
26	2024	5	5	5	5	5	5	5	5
19	2017	4	4	4	4	4	4	4	4
15	2013	3	3	3	3	3	3	3	3
1	1999	1	1	1	1	1	1	1	1
23	2021	5	5	5	5	5	5	5	5
2	2000	1	1	1	1	1	1	1	1
4	2002	1	1	1	1	1	1	1	1
18	2016	4	4	4	4	4	4	4	4
24	2022	5	5	5	5	5	5	5	5
13	2011	3	3	3	3	3	3	3	3
9	2007	2	2	2	2	2	2	2	2
20	2018	4	4	4	4	4	4	4	4
10	2008	3	3	3	3	3	3	3	3
12	2010	3	3	3	3	3	3	3	3

Table 6.5: Classification models results for men (random data)

Classifiers	LD	LSVM	QSVM	CSVM	Gaussian SVM	Fine KNN	Weighted KNN	NPR
1999	1	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1	1
2001	1	1	1	1	1	1	1	1
2002	1	1	1	1	1	1	1	1
2003	1	1	1	1	1	1	1	1
2004	1	2	2	1	1	1	1	1
2005	2	2	2	2	2	2	2	2
2006	2	2	2	2	2	2	2	2
2007	2	2	2	2	2	2	2	2
2008	2	2	2	2	2	2	2	2
2009	2	2	2	2	2	2	2	2
2010	2	3	3	2	2	2	2	2
2011	3	3	3	3	3	3	3	3
2012	3	3	3	3	3	3	3	3
2013	3	3	3	3	3	3	3	3
2014	3	3	3	3	3	3	3	3
2015	3	3	3	3	3	3	3	3
2016	4	4	4	4	4	4	4	4
2017	4	4	4	4	4	4	4	4
2018	4	4	4	4	4	4	4	4
2019	4	4	4	4	4	4	4	5
2020	4	4	4	4	4	4	4	5
2021	5	4	5	5	5	5	5	5
2022	5	5	5	5	5	5	5	5
2023	5	5	5	5	5	5	5	5
2024	5	5	5	5	5	5	5	5
2025	5	5	5	5	5	5	5	5

Table 6.6: Classification models results for women (training data) 1999-2025

Random points	Years	Classifiers							
		LD	LSVM	QSVM	CSVM	Gaussian SVM	Fine KNN	Weighted KNN	NPR
22	2020	4	4	4	4	4	4	4	5
6	2004	1	2	2	1	1	1	1	1
3	2001	1	1	1	1	1	1	1	1
16	2014	3	3	3	3	3	3	3	3
11	2009	2	2	2	2	2	2	2	2
7	2005	2	2	2	2	2	2	2	2
17	2015	3	3	3	3	3	3	3	3
14	2012	3	3	3	3	3	3	3	3
8	2006	2	2	2	2	2	2	2	2
5	2003	1	1	1	1	1	1	1	1
21	2019	4	4	4	4	4	4	4	5
25	2023	5	5	5	5	5	5	5	5
27	2025	5	5	5	5	5	5	5	5
26	2024	5	5	5	5	5	5	5	5
19	2017	4	4	4	4	4	4	4	4
15	2013	3	3	3	3	3	3	3	3
1	1999	1	1	1	1	1	1	1	1
23	2021	5	4	5	5	5	5	5	5
2	2000	1	1	1	1	1	1	1	1
4	2002	1	1	1	1	1	1	1	1
18	2016	4	4	4	4	4	4	4	4
24	2022	5	5	5	5	5	5	5	5
13	2011	3	3	3	3	3	3	3	3
9	2007	2	2	2	2	2	2	2	2
20	2018	4	4	4	4	4	4	4	4
10	2008	2	2	2	2	2	2	2	2
12	2010	2	3	3	2	2	2	2	2

Table 6.7: Classification models results for women (random data)

Classifiers	Accuracy	Prediction speed	Training time
LD	88.9%	93 obs/sec	9.9083 sec
Linear SVM	88.9%	77 obs/sec	10.177 sec
Quadratic SVM	88.9%	100 obs/sec	4.3084sec
Cubic SVM	88.9%	120 obs/sec	3.3928 sec
Gaussian SVM	**92.6** %	120 obs/sec	3.7141 sec
Fine KNN	88.9%	140 obs/sec	3.7141 sec
Weighted KNN	**92.6** %	310 obs/sec	1.2459 sec
NPR	**92.6** %	…….	…….

Table 6.8: Classification outcome information (men data)

Classifiers	Accuracy	Prediction speed	Training time
LD	85.2 %	310 obs/sec	2.2019 sec
Linear SVM	77.8 %	110 obs/sec	5.4628 sec
Quadratic SVM	85.2 %	91 obs/sec	4.6329 sec
Cubic SVM	85.2 %	100 obs/sec	4.2164 sec
Gaussian SVM	**96.3** %	110 obs/sec	3.5432 sec
Fine KNN	**96.3** %	320 obs/sec	1.5032 sec
Weighted KNN	**96.3** %	340 obs/sec	1.6991 sec
NPR	**92.6** %	…….	…….

Table 6.9: Classification outcome information (women data)

Figure 6.1: Linear discriminant model confusion matrixes for men and women data respectively

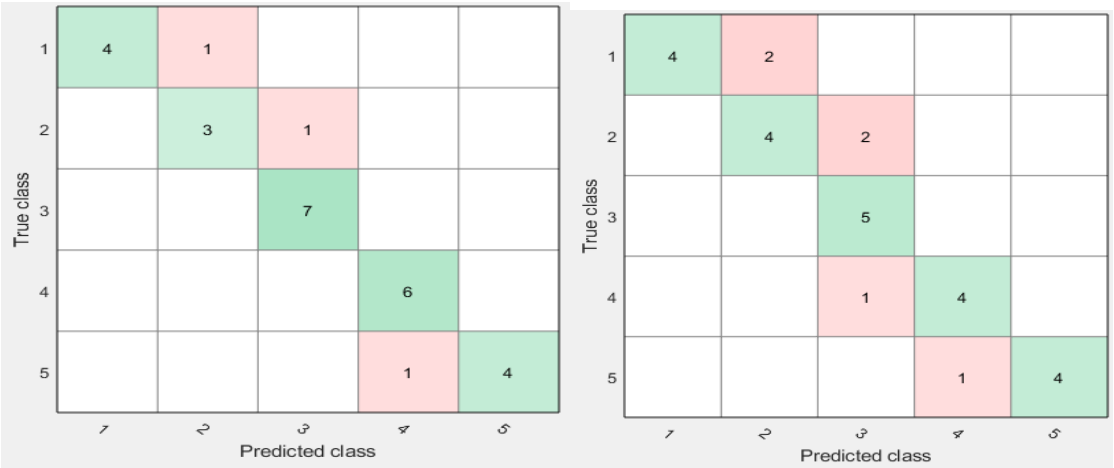

Figure 6.2: Linear SVM model confusion matrixes for men and women data respectively

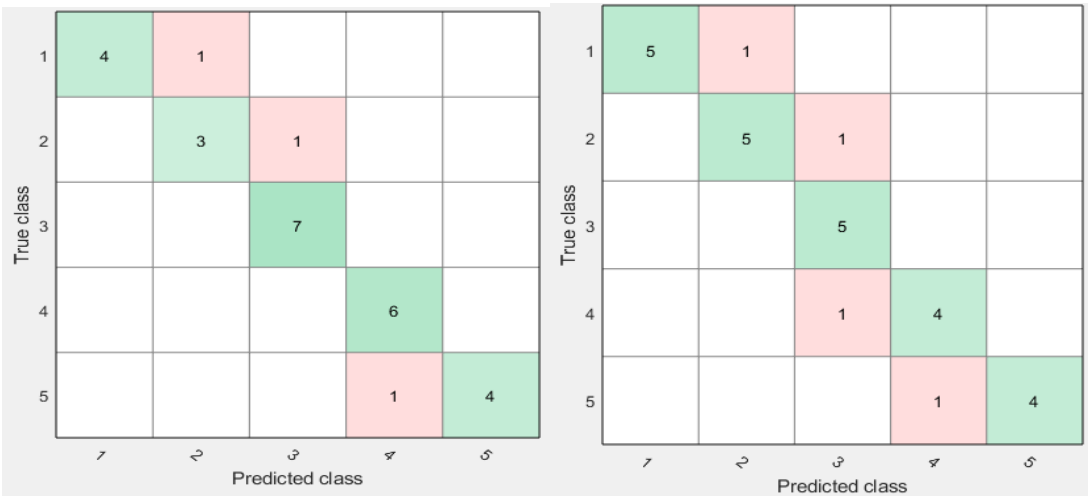

Figure 6.3: Quadratic SVM model confusion matrixes for men and women data respectively

Figure 6.4: Cubic SVM model confusion matrixes for men and women data respectively

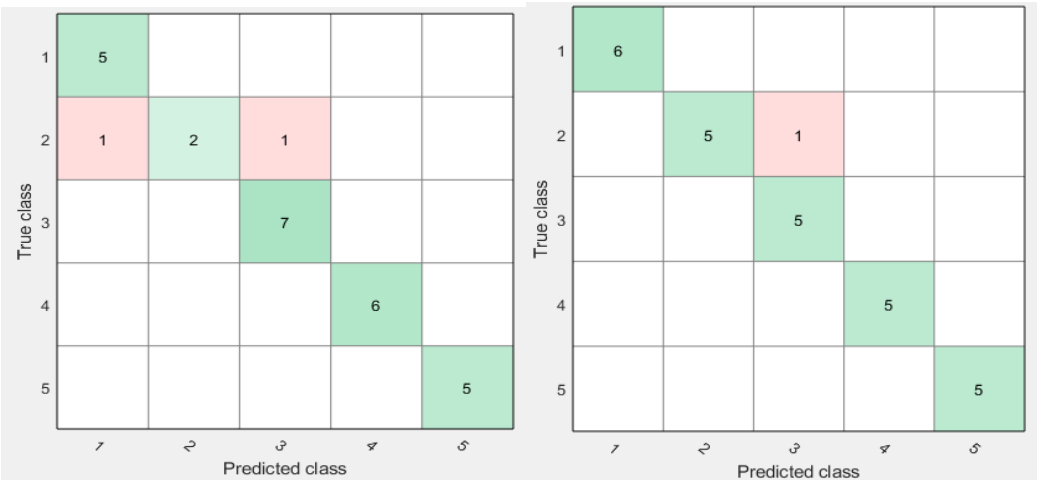

Figure 6.5: Medium Gaussian SVM model confusion matrixes for men and women data respectively

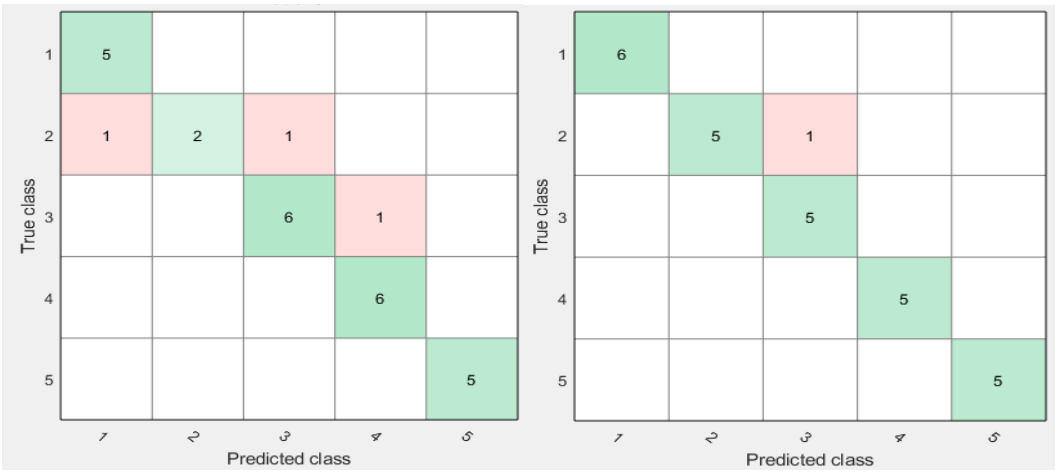

Figure 6.6: Fine KNN model confusion matrixes for men and women data respectively

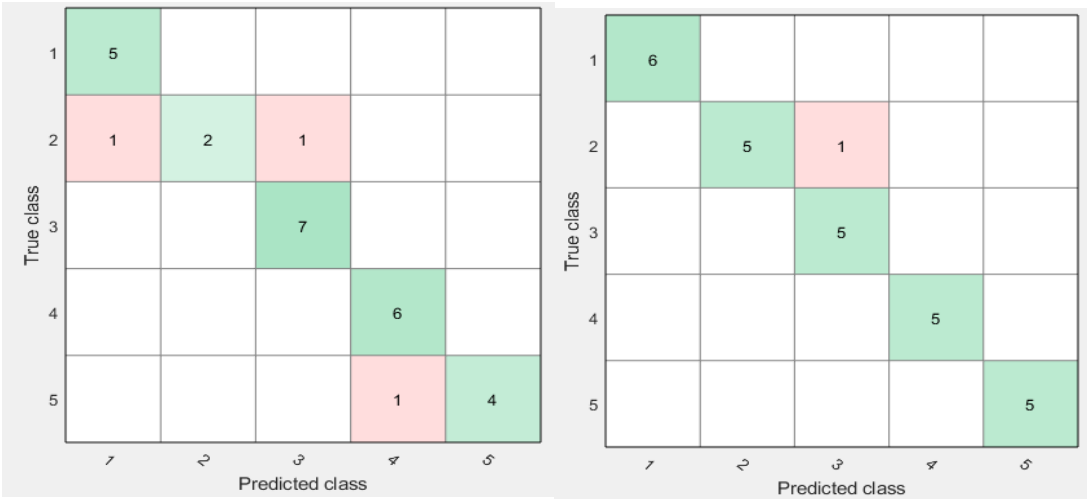

Figure 6.7: Weighted KNN model confusion matrixes for men and women data respectively

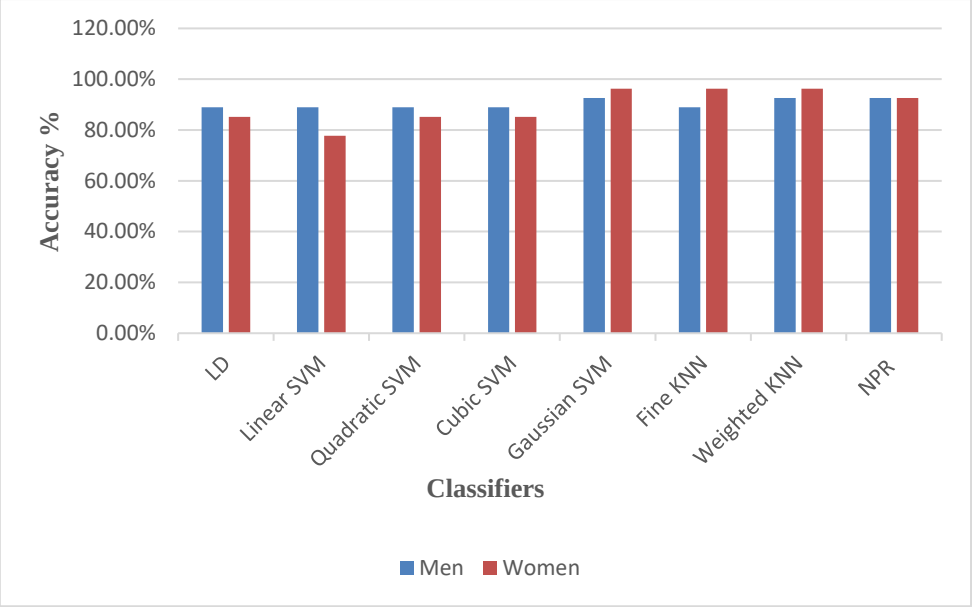

Figure 6.8: Classification results (accuracy) for men and women datasets

6.5. Summary

This chapter presented classification learning models for classifying diabetes prevalence based on Linear discriminant, Linear SVM, Quadratic SVM, Cubic SVM, Gaussian SVM, Fine KNN, Weighted KNN, and Neural Net pattern recognition. All these models were briefly described, and their mathematical equations were presented. The chapter investigated the use of classification methods to group or classify the prevalence level of diabetes disease that associated with behavioural risk factors (smoking, obesity, and inactivity). The proposed methods were applied on the men and women datasets, and their performance was evaluated in terms of accuracy, training time, and prediction speed. The obtained results show that there were slightly differences in the performance of models when using the men and women datasets. All classification methods used by the classification learner app required less than 10 seconds to predict the target. In addition, the experimental results on the predictive performance analysis of the classification models showed that Weighted KNN performed well on the prediction of diabetes prevalence rate, with the highest accuracy and less training time than the other classification methods, for both men and women datasets. In the next chapter another approach of modelling is presented, time series modelling, using Neural Network Time Series app in MATLAB. In this modelling approach three NARX-NN models for each risk factor are developed, and then the estimated values are used to construct ANN model by Neural Network fitting app, to forecast the prevalence rate of diabetes. Also, the developed NARX-NN model and the ANN model applied for regression modelling in Chapter 4 are compared.

Chapter 7
Time-Series Modelling

7.1. Introduction

This chapter presents the third method of modelling, which is time series modelling approach. The aim of this chapter is to examine the flexibility of Neural Network models in time series forecasting by comparing two modelling techniques NARX-NN time series and ANN. Three time series models of the three independent risk factors (smoking, obesity, and inactivity) were developed using NARX model (one-step ahead prediction) to predict the prevalence rate of diabetes for men and women datasets. The training dataset is from 1999 to 2013, and the test dataset is from 2014 to 2025, for both men and women. The performance of the developed models is evaluated using four statistical measures: MSE, RMSE, MAPE, and R-squared.

A brief background of time series modelling is provided in Section 7.2. Time Series and Neural Networks with a description of NARX-NN model are given in Section 7.3. The NARX-NN time series is explained in Section 7.4, and its implementation with ANN models is presented in Section 7.5. Section 7.6 provides the results and the discussion of the modelling outcomes. Lastly, a summary of this chapter is given in Section 7.7.

7.2. Time Series Modelling

Statistical analysis and theoretical developments of time series data started a long time ago with stochastic processes. The work of Yule and Walker in the 1920s and 1930s was the first practical application of autoregressive models to data [229]. In the 1970s, time series models were developed by Box and Jenkins, which were introduced in their classic *Time Series Analysis*, describing the required processes of modelling individual series, such as estimation and forecasting. Since that time, the classical time series (“Box–Jenkins”) models have gained popularity and have been used in many forecasting problems.

A time series can be defined as a sequence of observations measured at a specific discrete or continuous time units. In time series modelling, a model is developed based on the past observations of the studied variables to analyse and describe the underlying relationship between these variables. The developed model can then be used for the prediction of the future [230]. A mathematical representation of a time series prediction problem can be given in the following equation as:

$$\hat{y}_{t+i} = f(y_t, y_{t-1}, y_{t-2}, \dots, y_{t-n}) \quad (7.1)$$

where y is the observed time series with n points; t is the most recent observation; and t- n indicates the most distant observation. At $t + i$, the future value can be predicted by a function f, this function known as the model and it is applied to get the predicted value $\hat{y}_{t+i}$ [173].

During the past few decades, a great deal of attention has been devoted to time series analysis, which has become a popular research topic in different fields such as finance, energy, electricity, and medicine etc. Time series forecasting is an important task in time series analysis, and it considered as a powerful tool for describing a complex system using observed data. ARIMA models have been the most popular and widely used in the time series forecasting domain. These models gained their popularity due to its statistical characteristics as well as the well-known Box–Jenkins approach in the model development procedures. Some studies of time series analysis and models were reviewed in Chapter 2.

Despite the popularity of the traditional ARIMA model, it has a major limitation regarding the pre-assumed linear form of the model, whereby an assumption of linear correlation is made among the time series values, thus this model cannot process nonlinear patterns. Even assuming that linearity is useful and can be a powerful tool in many different fields, it was clearly shown in the early 1980s that the application of the linear models on complex real-world problems is not always sufficiently acceptable [231].

During the last two decades, testing and modelling nonlinearity of time series data has been developed in several fields, and examples of nonlinear time series models include the bilinear, autoregressive conditional heteroscedastic (ARCH), general autoregressive conditional heteroscedastic (GARCH), chaotic dynamics, and the threshold autoregressive (TAR) models. These models are model-driven approaches, where the relation type between the variables needs to be identified, and then the selected model parameters are estimated. Moreover, the recent notable activities in ANN research along with its successful forecasting applications indicated that they can also be effectively used for time series forecasting [232][233].

7.3. Time Series and Neural Networks

As mentioned in the above section, ANNs can be effectively applied to several time series prediction and modelling tasks. Contrary to traditional time series methods, ANN is a self-adaptive, data-driven, nonparametric, nonlinear statistical method, in which only few previous assumptions are required about the models for the studied problems. ANN models can be helpful in dealing with nonlinear systems in the data series that have an unknown functional

relationship. In ANN models, a specific general purpose learning algorithm is used to deal with the process of constructing the relationship between the input and output variables. Using ANN as a modelling technique for time series estimation and forecasting is an essential task in different fields, including statistics. Researchers have been attracted by ANN's features such as the freedom from restrictive assumptions like linearity. Because of the nonlinearity feature of ANN models and their powerful handling of noise, they often outperform traditional linear methods in time series forecast applications. Two types of time series prediction can be made: one-step ahead, or multi-step ahead. In the latter form of prediction, the output of the ANN is commonly fed back to the input. The NARX nonlinear autoregressive with exogenous input can be efficiently used for these types of time series [234][235].

The following section describes the NARX-NN model and its main characteristics, then its implementation and prediction results are compared with the ANN model previously described and used for regression modelling in Chapter 4.

7.4. NARX Neural Network Model

NARX Neural Network is among the most powerful ANN models used in time series modelling, because of its valuable characteristics such as fast convergence in achieving the optimal weights of the network connections between neurons and/or inputs. It is excellent at determining long time dependences compared to conventional RNN, and it is an effective method compared to other ANNs [236][237]. NARX is based on a nonlinear autoregressive neural model with exogenous inputs. In this method the dynamic (time series) of a variable is modelled using its past values and the past values of external input (exogenous inputs). Even though that the NARX NN model is used for short-term series forecasts, multi-step-ahead predictions can be obtained if prior information of the future exogenous inputs is identified. This can be achieved in an iterative process by utilising the output of a one-step ahead prediction as the input for the following prediction. Basically, the NARX network is described as a recurrent dynamic network, which allows connections of feedback that have impact on different layers of the network. The mathematical representation of the NARX model can be given as:

$$y(s+1) = f\begin{pmatrix} y(s), y(s-1), \ldots\ldots, y(s-d_y) \\ w(s), w(s-1), \ldots\ldots, w(s-d_w) \end{pmatrix} \tag{7.2}$$

where s represents the time series; d is the number of the specified input and output delays; y (s + 1) is the NARX output; (y(s), y (s − 1), ⋯ , y(s − d_y)) indicates the prior values of an

exogenous output series; (w(s), w (s − 1), ⋯, w(s − d_w)) contains the variables driving exogenous input series; and f is the mapping function of the neural network.

There are two modes of training the NARX Network model: series-parallel (open loop) and parallel (closed loop). In the first mode, the feedback delayed information is taken from the real values; in the second mode, the estimated outputs are fed back and contained within the output's regressor. It is useful to point out that these modes can also be applied through the prediction stage. These modes are different in one main point related to the time delay line of the output, whether the time delay will use the real output during the training and prediction such in series-parallel or will use the estimated output from the output layer such in the parallel mode. The future value of the time series y(s-1) in the series-parallel is predicted using the past and present values of w(s) and the real past values of the time series y(s). While the prediction in the parallel mode is achieved by using the past and present values of w(s) and the past forecasted values of the time series $\hat{y}$(s). There are two advantages of using series-parallel mode. First, using the real values as input for the feedforward network leads to more accurate results. Second, the architecture of the resulting network is purely feedforward one, and frequently used MLP algorithms can be deployed for training. Therefore, in this study, the series-parallel mode was first applied during the training stage of the NARX-NN model. Following the training stage, the NARX-NN model was converted to parallel mode, which is useful for making the multi-step-ahead prediction.

The initial design of the NARX NN model was as a feedforward backpropagation network. The map function F (.) at the beginning is undetermined and it is approximated through the training stages of the prediction. The MLP is the internal architecture that makes this approximation. MLP can provide a robust structure to learn any form of continuous nonlinear mapping. A simple MLP structure involves three layers: input, hidden, and output layer. Other components include neurons, weights, and activation functions.

During the training stage, the information is directed from the input layer to the output layer. Within each layer, the multiplication process carries out between the input vector and the weights vector to produce the scalar outcome, then the activation function is designed to achieve the neuron output. Furthermore, after presenting all the inputs and target values to the network, the training function used to update the network weights and biases to achieve the best performance output value.

In the case of the NARX NN model, the Bayesian regularisation backpropagation algorithm is mainly used for training the NARX NN network and updating the weight and bias values by Levenberg-Marquardt optimisation. This algorithm reduces the combination of squared errors

and weights and then defines the correct combination to generate a network that performing well and prevents overtraining.

7.5. Implementation

All estimation and forecasting of time series modelling presented in this chapter was implemented using MATLAB's Neural Network Time Series and Neural Net Fitting toolbox. The Neural Network Time Series app can be started directly from the Apps tab, under Machine Learning and Deep Learning, or by typing (*ntstool*) in the MATLAB command window. In this app, a neural network can be trained to solve three types of time series problems: NARX networks, NAR networks, and nonlinear input-output networks. The NARX network type is known as nonlinear autoregressive with exogenous (external) input. This form can be used to predict future values of a time series y(t) from a given past values d of that time series and past values of a second time series x(t). The second type of time series problems is the NAR network, which is called nonlinear autoregressive. This type involved only one series, in which the future values of a time series y(t) are predicted only from past values of that series, given by the following:

$$y(t) = f(y(t-1), \dots, y(t-d)) \tag{7.3}$$

Nonlinear Input-Output Network is the third type of time series problems, similar to NARX model in involving two series: an input series x(t) and an output series y(t). In this type of time series, the values of y(t) can be predicted from previous values of x(t), but without knowledge of past values of y(t). The Nonlinear Input-Output Network model can be given as:

$$y(t) = f(x(t-1), \dots, x(t-d)) \tag{7.4}$$

The NARX model is better in providing predictions than the input-output model, because it employs a further information included in the past values of y(t). However, some applications may not have past values of y(t), so in these situations it would be useful to use the input-output model instead of the NARX model. In this research, NARX-NN model was used for time series modelling. The three behavioural risk factors (smoking, obesity, inactivity) were used as predictors to determine the prevalence rate of diabetes. In the first step, three models for each risk factor have been developed using the (NARX-NN) model (*ntstool*) from Neural time series app in MATLAB using the historical datasets from 1999 to 2013 for each factor. The developed models were then used to predict the data of the three factors from 2014 to 2025.

Figure 7.1 shows the block diagram of the developed NARX-NN model, which was first trained in open-loop mode. The structure of the developed model is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. The structure of this network has been set with 10 neurons in the hidden layer and one delay, and the Levenberg-Marquardt backpropagation algorithm (*trainlm*) was selected as the optimised training method. Figure 7.2 shows how the NARX-NN model transformed from open-loop to closed-loop for making the multi-step-ahead prediction. Figure 7.3 shows the step-ahead prediction network diagram which is used to get predicted timestep values early.

In the next step, the predicted datasets of the three risk factors fed into the Neural Network fitting app (*nftool*) as independent variables, with the morbidity data of diabetes as a dependent variable, as illustrated in Figure 7.4. Finally, after training the NARX-NN model, it was tested using the predicted datasets from time series models of the three factors from 2014 to 2025 to undertake predictions of diabetes prevalence for both men and women. The performance of the NARX-NN model was evaluated using four statistical evaluation measures, these are MSE, RMSE, MAPE, and the coefficient of determination R-squared, and then the obtained results were compared with those of the ANN model whose regression modelling is presented in Chapter 4.

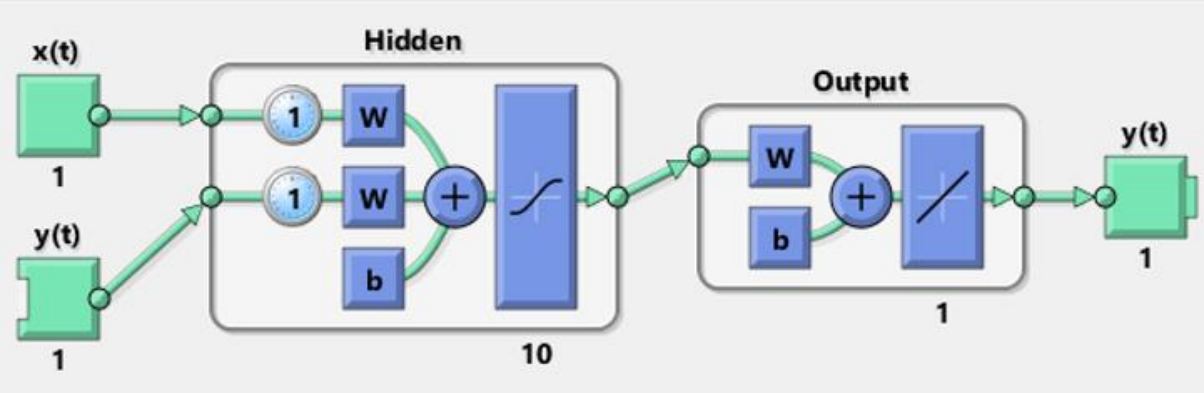

Figure 7.1: NARX neural network time series block diagram

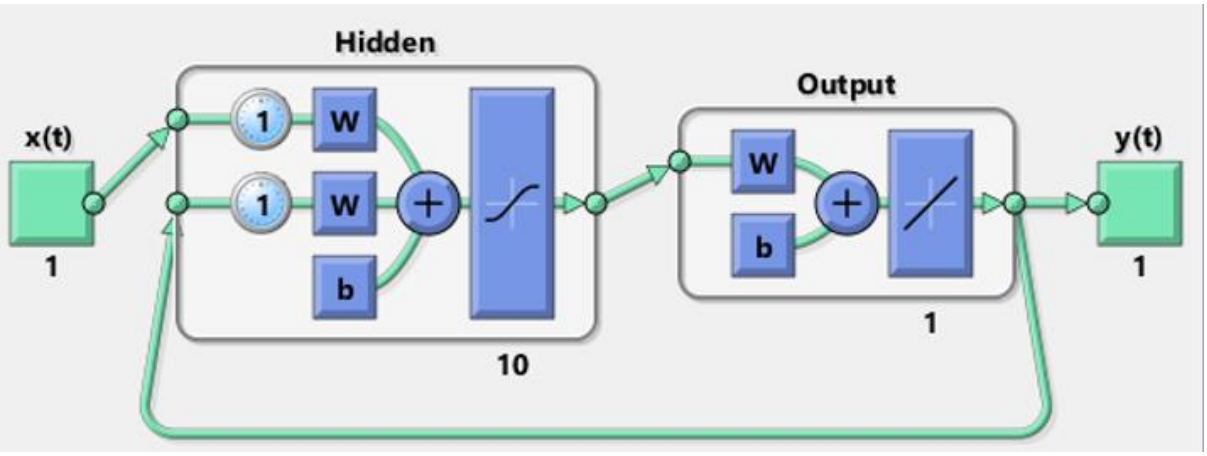

Figure 7.2: NARX neural network time series closed loop block diagram

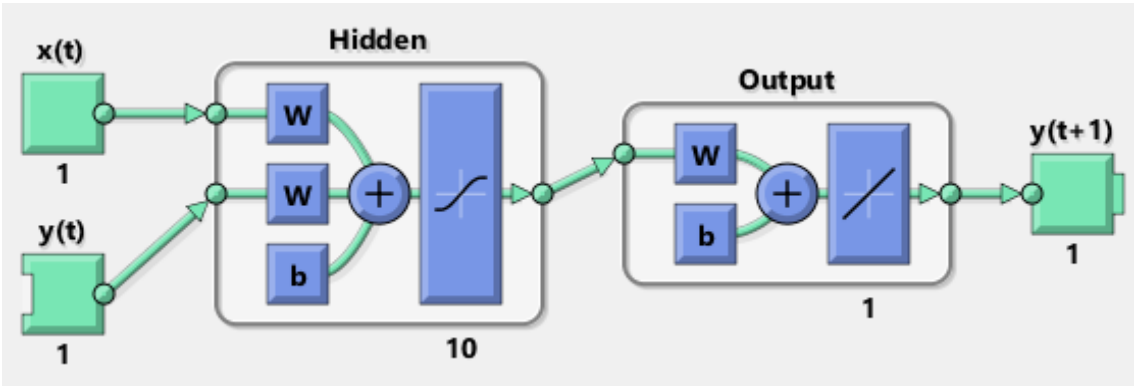

Figure 7.3: NARX neural network time series predict one-step ahead block diagram

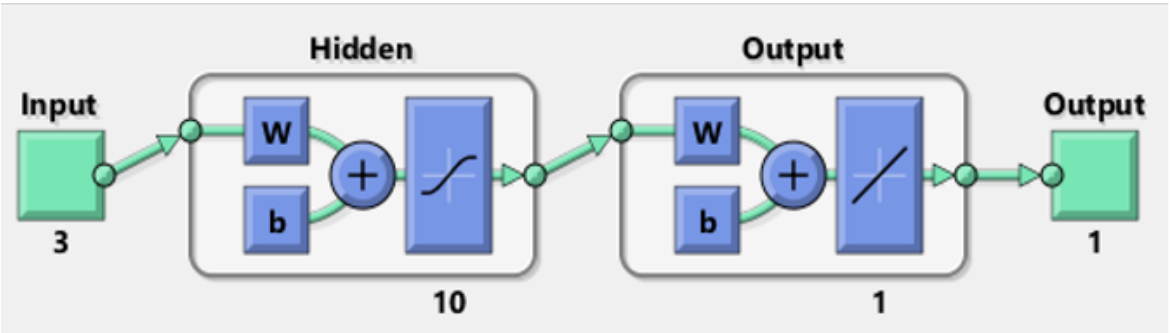

Figure 7.4: Fitting neural network block diagram

7.6. Results and Discussion

This section presents the results obtained by NARX-NN time series modelling for each risk factor (smoking, obesity, and inactivity) for both men and women. The estimated results of diabetes prevalence rate for men and women from the constructed NARX-NN model using time series prediction data are also provided. The developed NARX-NN model is compared with the ANN model developed for regression in Chapter 4. The comparison is carried out by evaluating their performance using four statistical evaluation measures, specifically MSE,

RMSE, MAPE, and the coefficient of determination R^2. Their equations were given previously in Chapter 3.

Table 7.1 and Table 7.2 show the time series modelling results of the three risk factors (smoking, obesity, and inactivity) for both men and women training data (1999-2013) and test data (2014-2025) respectively. In addition, Table 7.3 and Table 7.4 show the results of diabetes prevalence by NARX-NN and ANN models for both men and women using training data (1999-2013) and test data (2014-2025), respectively. The NARX-NN model was successfully estimated the values corresponding to the past (1999–2013) and predicted the values of the future (2014–2025) for each risk factor for both men and women datasets, and then predicted the prevalence rate of diabetes.

Figure 7.5 and Figure 7.6 show graphs of diabetes prevalence by NARX and ANN models, for morbidity data of the period 1999-2013, for men and women respectively. These graphs compare the prevalence rate of diabetes obtained by NARX-NN and ANN models with actual values. It can be clearly shown from these graphs that NARX-NN model outperformed the ANN model when compared relative to the actual values.

Table 7.5 shows the overall performance of both NARX-NN, and ANN models evaluated by the statistical evaluation metrics (MSE, RMSE, MAPE, and R^2). Based on the results, it can be observed that the best performance was given by NARX-NN model, with RMSE = 0.02 for men data and RMSE = 0.1287 for women data, and MAPE= 0.0006 for men and MAPE= 0.0095 for women. All the evaluation metrics show that the performance of the NARX-NN model was better than that of the ANN model in predicting the prevalence of diabetes for both men and women datasets. Furthermore, Figure 7.7 presents a graph that compares the performance evaluation metrics' results of NARX-NN and ANN models for men and women data, indicating the good performance of the proposed neural network time series model.

	Men training data after modelling by neural network time series			Women training data after modelling by neural network time series		
Years	Smoking	Obesity	Inactivity	Smoking	Obesity	Inactivity
1999	21.4	15.5	93.2	0.9	18.0	96.8
2000	21.7	17.3	90.4	1.0	19.6	96.0
2001	22.1	19.6	88.4	1.1	24.1	93.8
2002	22.5	22.1	87.4	1.3	31.8	90.5
2003	22.9	24.1	86.4	1.5	38.9	88.3
2004	23.7	26.1	80.5	1.6	38.9	84.6
2005	24.1	26.3	72.4	1.8	39.5	83.0
2006	24.5	27.0	71.5	1.8	40.4	82.7
2007	25.8	28.7	72.3	1.7	41.8	81.0
2008	26.0	30.7	72.9	1.9	43.2	78.9
2009	25.8	31.5	66.6	1.9	44.6	77.8
2010	26.0	32.3	58.3	2.0	46.0	77.4
2011	26.2	33.7	56.5	2.1	46.8	77.3
2012	26.6	35.2	55.9	2.2	47.4	77.1
2013	26.9	35.9	55.0	2.5	48.2	76.9

Table 7.1: Time series modelling results of the three risk factors for men and women (training data), 1999-2013

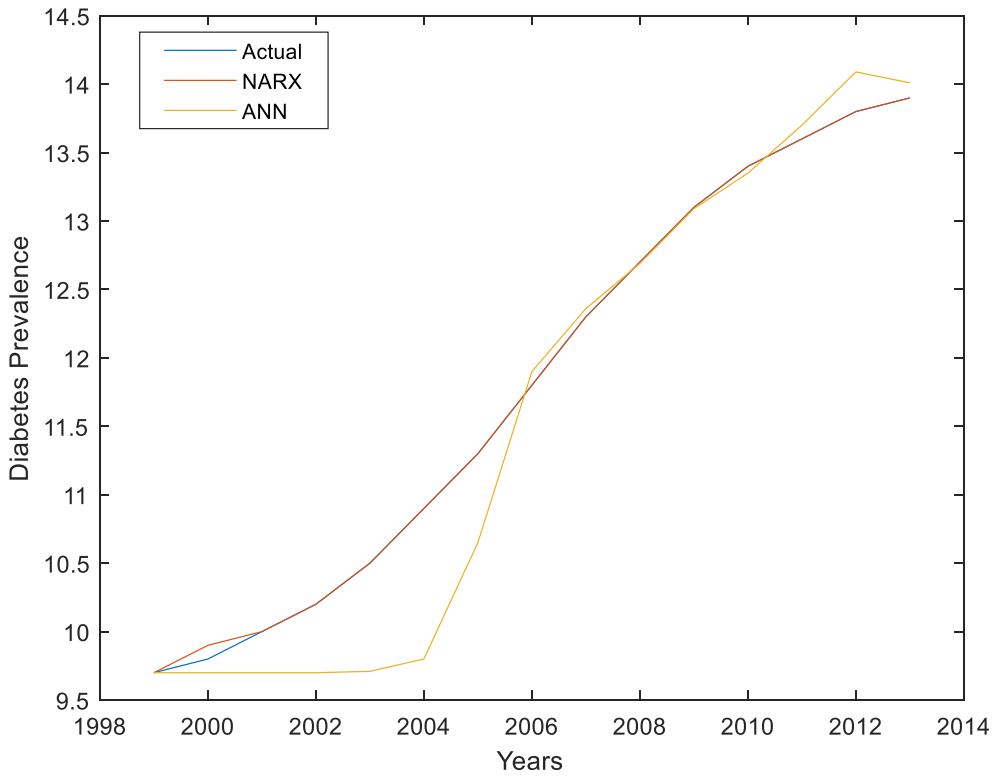

Figure 7.5: Diabetes prevalence by NARX and ANN models (men morbidity data 1999-2013)

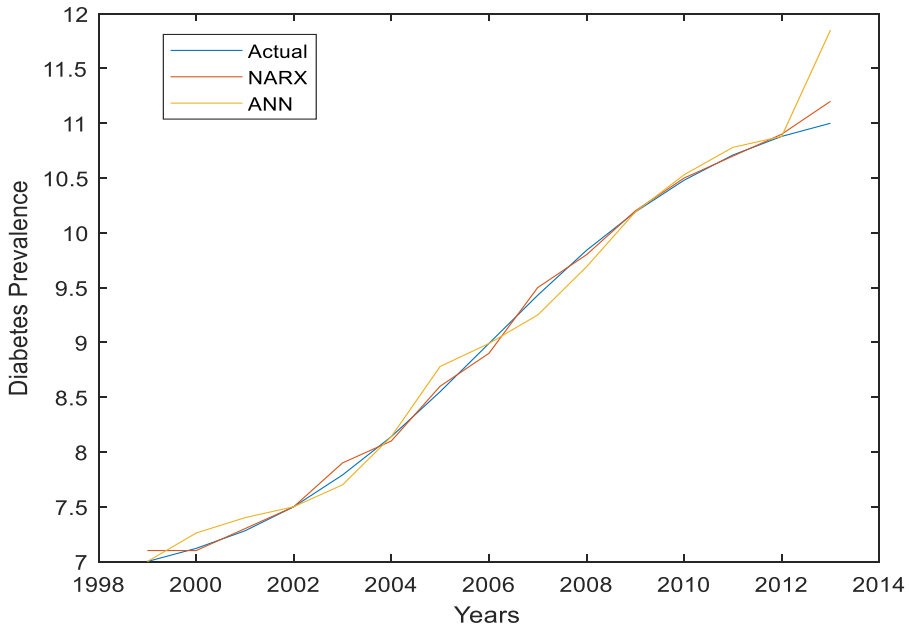

Figure 7.6: Diabetes prevalence by NARX and ANN models (women morbidity data 1999-2013)

	Men testing data after modelling by neural network time series			Women testing data after modelling by neural network time series		
Years	Smoking	Obesity	Inactivity	Smoking	Obesity	Inactivity
2014	27.25	36.46	53.27	2.33	48.99	76.60
2015	27.48	37.06	51.46	2.61	49.78	76.07
2016	27.68	38.09	49.44	2.67	50.66	75.07
2017	27.86	39.66	47.17	3.06	51.63	73.81
2018	27.85	41.76	45.26	3.46	52.64	72.40
2019	28.00	44.19	44.45	3.79	53.64	71.30
2020	28.16	46.60	44.18	4.01	54.55	70.65
2021	28.34	48.65	43.31	4.13	55.32	70.25
2022	28.56	50.19	42.00	4.20	55.92	70.05
2023	28.72	51.21	40.93	4.21	56.33	69.93
2024	29.11	51.83	40.30	4.24	56.56	69.88
2025	29.53	52.17	40.02	4.25	56.65	69.85

Table 7.2: Time series modelling results of the three risk factors for men and women (test data), 2014-2025

Years	Men		Women	
	NARX-NN	ANN	NARX-NN	ANN
1999	9.7	9.7	7.0	7
2000	9.9	9.7	7.1	7.26
2001	10.0	9.7	7.3	7.4
2002	10.2	9.7	7.7	7.5
2003	10.5	9.71	8.2	7.7
2004	10.9	9.8	8.2	8.14
2005	11.3	10.65	8.6	8.78
2006	11.8	11.9	8.9	8.99
2007	12.3	12.36	9.5	9.25
2008	12.7	12.69	9.9	9.69
2009	13.1	13.09	10.2	10.19
2010	13.4	13.35	10.5	10.53
2011	13.6	13.7	10.7	10.78
2012	13.8	14.09	10.8	10.88
2013	13.9	14.01	10.9	11.85

Table 7.3: Diabetes Prevalence results by NARX-NN and ANN models for men and women (training data), 1999-2013

Years	Men		Women	
	NARX-NN	ANN	NARX-NN	ANN
2014	14.69	14.5	10.74	11.5
2015	15.06	15.1	11.94	11.7
2016	15.73	15.8	12.16	12.3
2017	16.72	16.6	13.19	13.1
2018	17.70	17	14.87	14.1
2019	18.15	17.7	15.84	15.5
2020	18.34	18.2	16.08	15.9
2021	18.42	18.6	16.14	16.4
2022	18.44	18.8	16.16	16.8
2023	18.45	18.9	16.16	17
2024	18.42	19.1	16.17	17.1
2025	18.37	19.2	16.17	17.2

Table 7.4: Diabetes Prevalence results by NARX-NN and ANN models for men and women (test data), 2014-2025

Evaluation metrics	NARX-NN model		ANN model	
	Men	Women	Men	Women
MSE	**0.0006**	**0.0166**	0.0081	0.0595
RMSE	**0.02**	**0.1287**	0.0899	0.2437
MAPE	**0.0006**	**0.0095**	0.0252	0.0137
R^2	**0.99**	**0.99**	0.99	0.97

Table 7.5: Statistical evaluation metrics results

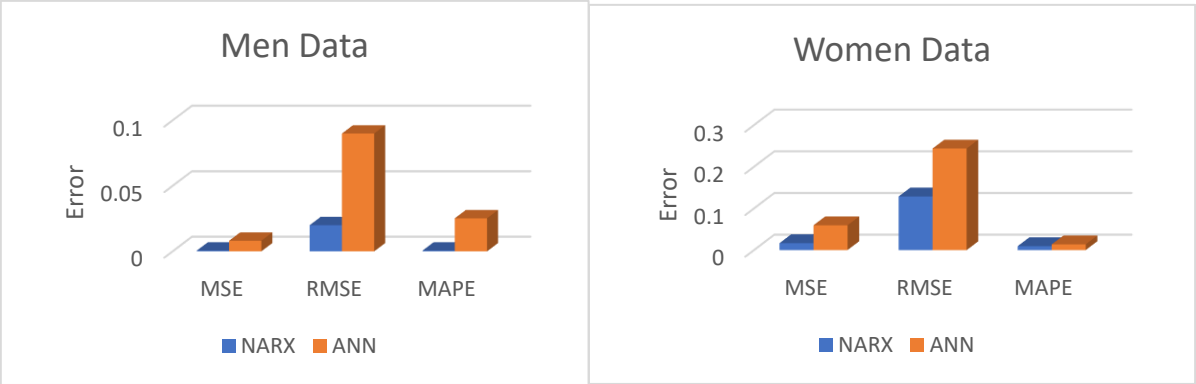

Figure 7.7: Performance evaluation metrics results of NARX-NN and ANN models for men and women data

7.7. Summary

This chapter presented the NARX-NN model developed for the three behavioural risk factors (smoking, obesity, and inactivity) for both men and women datasets employing Neural Network Time Series app in MATLAB. The simulated data by time series approach were used to construct ANN model using Neural Network Fitting app in MATLAB for predicting the prevalence rate of diabetes for men and women. The chapter compared between this model and the ANN model previously used in regression modelling in Chapter 4. This comparison was based on evaluating their performance according to different statistical metrics. It was indicated by these evaluations that the performance of the developed NARX-NN model using time series data in this chapter outperformed the ANN model applied in Chapter 4. The obtained results indicated that using time series modelling was successful in estimating the values corresponding to the past and predicting the values of the future for each risk factor for both men and women datasets, and then in predicting the prevalence rate of diabetes.

Chapter 8
Conclusions and Future Work

8.1. Conclusions

The primary aim of this thesis was to study trends in the prevalence rate of diabetes along with the related behavioural risk factors among Saudis to predict future disease burden using mathematical modelling, by applying different machine learning techniques. To that end, various regression and classification methods were developed and integrated to enhance model prediction performance. The design and implementation of the proposed predictive models was discussed and examined. In order to give better results and develop a powerful predictive model, seven ensemble methods were used to combine the predictions of the individual models. This helped in overcoming the disadvantages of each individual model by pooling attributes in combination with other models. The efficiency of the developed models in predicting diabetes on men and women datasets for different age groups was compared, both in terms of their intrinsic performance in this study, and with observed data from other studies.

The statistical metrics including RMSE, MSE, MAPE, R-squared, and classification accuracy were used to evaluate the performance efficiency of the developed models. Very competitive results obtained by the ANFIS regression model and WAVR, which makes them the preferred models in terms of accurate outputs. The experimental results on the predictive performance analysis of the classification models show that Weighted KNN performed well on the prediction of diabetes prevalence rate, with the highest accuracy and less training time compared to other classification methods, for the datasets of both men and women.

Regarding the time series modelling, the proposed NARX-NN model outperformed the standard ANN based predictor model with the lowest error, as presented in Chapter 7. Good results were obtained in comparison with ANN model, which indicates that the NARX neural network model can successfully use its output feedback loop to improve its predictive performance in time series predictions of diabetes prevalence and the related risk factors.

These models were applied on diabetes data obtained from KSA, where the level of diabetes prevalence is predicted to increase, affected by the increased of population and aging, along with significantly rising levels of different DM risk factors. The obtained results from this research indicate that demographic as well as behavioural risk factors significantly contributed to the increased level of diabetes, with a significance level of 0.05; however, smoking, obesity, and physical inactivity were the most significant factors. In addition, the highest prevalence of

diabetes from all the age groups was found in the population aged 55-74 years old. The highest growth in diabetes prevalence was in men aged 55-64 years, which showed an increase in prevalence from 24.91% in 1999 to 53% in 2013. The highest prevalence of diabetes for women was in the group aged 65-74 years, indicating an increase from 24.4% in 1999 to 48.2% in 2013. It was clear from these findings that diabetes prevalence was higher among men than women.

Furthermore, the prevalence of the three risk factors, including obesity, smoking, and physical inactivity, varied according to gender. The prevalence of smoking was higher among men than women (21.1% against 0.9%, respectively). Women had a higher prevalence rate of obesity than men (20.3% against 13.1%). Also, the prevalence rate of physical inactivity was higher among women than men (98.1% against 93.9%).

Diabetes disease is an epidemic expected to rapidly increase in KSA, which indicates the need to develop a national surveillance system to examine trends and risk factors of diabetes disease on a regular basis. The national transformation plan (Vision 2030) launched by the government of KSA stipulates the key objective of developing the economy, which can be achieved by investing in control and prevention strategies of non-communicable diseases, including diabetes, enabling Saudis to remain active and economically productive for more years, with reduced burdens in health costs. Enormous contributions and efforts have been made by the government of KSA to develop the population's health and living standards, but the obtained results in this thesis highlight the urgency need for seek more strategies and take further actions on diabetes control to overcome with this disease, reduce its complications (or at least delay them), increase individuals' healthy life expectancy, reduce health expenditure, and expand national productivity and economic growth.

8.2. Future Work

For future research directions, it could be useful to examine the impact of using big datasets with more variables. This should include a family record of diabetes, smoking status (whether they were active or previous smokers, job demands (considering physical activity levels and working hours), medications, and current or previous health status. Also, investigating the effect of including more risk factors for diabetes in KSA, such as diet or blood pressure, or even including different categories of risk factors (e.g., non-modifiable risk factors), such as family history and gestational diabetes. The integration of different risk factors might help to obtain more precise predictions of diabetes prevalence and thus contribute to supporting health policy makers with more plans and options for intervention. Moreover, examining the use of machine learning methods to predict the risk of developing any complications of

diabetes such as nephropathy, retinopathy, and cardiovascular disease. This would help to maintain the quality of life of diabetic individuals and reduce the rising burden of diabetes on healthcare budget.

The structure of the model could also be improved to include information to estimate the future health and economic burden associated with diabetes and the related complications using measures such as quality adjusted life years (QALYS) and provide estimations of the total direct and indirect disease costs. Furthermore, this modelling work of diabetes could be enhanced and developed to study other types of chronic diseases such as high blood pressure, heart problems, etc. Additionally, the proposed models could be further used to estimate specific morbidity and mortality outcomes from other diabetes-related disease such as cardiovascular disease, chronic kidney disease, by using the estimated future cardiometabolic risk factors' outcomes to estimate the impact of diabetes on those diseases. Furthermore, one could explore the use of the latest modelling technologies, such as deep learning and reinforcement learning, to study the prevalence of diabetes disease in KSA or in any of the Gulf Cooperation Council Countries (GCC), as they share many similarities in environments, lifestyle, social and cultural habits (including diabetes risk factors).

References

[1] J. Moini, *Epidemiology of Diabetes*. Elsevier Science, 2019.

[2] D. Mellitus, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 29, p. S43, 2006.

[3] D. T. Jamison *et al.*, *Disease Control Priorities in Developing Countries*. World Bank Publications, 2006.

[4] W. H. O. 2016, “Diabetes Country Profiles 2016.”

[5] M. Rewers and R. F. Hamman, “Risk factors for non-insulin-dependent diabetes,” *Diabetes Am.*, vol. 2, pp. 179–220, 1995.

[6] I. D. F. D. Atlas, “International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium,” 2019.

[7] J. M. Ekoé, M. Rewers, R. Williams, and P. Zimmet, *The Epidemiology of Diabetes Mellitus*. Wiley, 2008.

[8] L. Guariguata, D. R. Whiting, I. Hambleton, J. Beagley, U. Linnenkamp, and J. E. Shaw, “Global estimates of diabetes prevalence for 2013 and projections for 2035,” *Diabetes Res. Clin. Pract.*, vol. 103, no. 2, pp. 137–149, 2014.

[9] B. Zhou *et al.*, “Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants,” *Lancet*, vol. 387, no. 10027, pp. 1513–1530, 2016.

[10] “Erratum: Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants (The Lancet (2016) 387(10027) (1513–1530)(S0140673616006188)(10.1016/S0140-6736(16)00618-8)),” *The Lancet*, vol. 389, no. 10068. p. e2, 2017.

[11] M. C. Weinstein *et al.*, “Modeling for health care and other policy decisions: uses, roles, and validity,” *Value Heal.*, vol. 4, no. 5, pp. 348–361, 2001.

[12] M. C. Weinstein *et al.*, “Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies,” *Value Health*, vol. 6, no. 1, p. 9—17, 2003.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques ,” *Frontiers in Genetics* , vol. 9. 2018.

[14] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, “Predictive models for diabetes mellitus using machine learning techniques,” *BMC Endocr. Disord.*, vol. 19, no. 1, pp. 1–9, 2019.

[15] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[16] T. A. Assegie and P. S. Nair, “The Performance of different machine learning models on diabetes prediction,” *Int. J. Sci. Technol. Res.*, vol. 9, no. 01, 2020.

[17] L. J. Muhammad, E. A. Algehyne, and S. S. Usman, “Predictive supervised machine learning models for diabetes mellitus,” *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–10, 2020.

[18] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, “A Machine Learning Approach to Predicting Diabetes Complications,” in *Healthcare*, 2021, vol. 9, no. 12, p. 1712.

[19] M. A. R. Refat, M. Al Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, “A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach,” in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2021, pp. 654–659.

[20] D. O. Abegunde, C. D. Mathers, T. Adam, M. Ortegon, and K. Strong, “The burden and costs of chronic diseases in low-income and middle-income countries,” *Lancet*, vol. 370, no. 9603, pp. 1929–1938, 2007.

[21] H. King *et al.*, “Global Burden of Diabetes, 1995-2025: Prevalence, numerical estimates, and projections,” *Diabetes Care*, vol. 21, no. 9, pp. 1414–1431, 1998.

[22] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, “Global Prevalence of Diabetes,” *Diabetes Care*, vol. 27, no. 5. pp. 1047–1053, 2004.

[23] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet, “Global estimates of the prevalence of diabetes for 2010 and 2030,” *Diabetes Res. Clin. Pract.*, vol. 87, no. 1, pp. 4–14, 2010.

[24] G. Danaei *et al.*, “National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2· 7 million participants,” *Lancet*, vol. 378, no. 9785, pp. 31–40, 2011.

[25] N. A. Rosenthal, “Infections, Chronic Disease, and the Epidemiological Transition: A New Perspective,” *Clin. Infect. Dis.*, vol. 61, no. 3, pp. 489–490, Aug. 2015.

[26] J. Frenk, J. L. Bobadilla, J. Sepuúlveda, and M. L. Cervantes, “Health transition in

middle-income countries: New challenges for health care," *Health Policy Plan.*, vol. 4, no. 1, pp. 29–39, 1989.

[27] A. Esteghamati *et al.*, "Third national Surveillance of Risk Factors of Non-Communicable Diseases (SuRFNCD-2007) in Iran: methods and results on prevalence of diabetes, hypertension, obesity, central obesity, and dyslipidemia," *BMC Public Health*, vol. 9, p. 167, 2009.

[28] C. S. Cockram, "The epidemiology of diabetes mellitus in the Asia-Pacific region," *Hong Kong Med. J.*, vol. 6, no. 1, pp. 43–52, 2000.

[29] World Health Organization, "WHO South-East Asia," *WHO South-East Asia J. Public Heal.*, vol. 5, no. 2, pp. 77–173, 2016.

[30] A. T. Kharroubi, "Diabetes mellitus: The epidemic of the century," *World J. Diabetes*, vol. 6, no. 6, p. 850, 2015.

[31] International Diabetes Federation, *IDF Diabetes Atlas Eighth Edition 2017*. 2017.

[32] W. E. Winter and M. R. Signorino, *Diabetes mellitus: pathophysiology, etiologies, complications, management, and laboratory evaluation: special topics in diagnostic testing*. Amer. Assoc. for Clinical Chemistry, 2002.

[33] A. D. American Diabetes Association, "Diagnosis and classification of diabetes mellitus.," *Diabetes Care*, vol. 33 Suppl 1, no. Supplement 1, pp. S62-9, 2010.

[34] R. I. G. Holt, C. S. Cockram, A. Flyvbjerg, and B. J. Goldstein, *Textbook of Diabetes: Fourth Edition*. 2010.

[35] J. Munden and L. W. & Wilkins., *Diabetes mellitus : a guide to patient care.* Philadelphia : Lippincott Williams & Wilkins, c2007.

[36] J. A. K. Carrier, *Managing long-term conditions and chronic illness in primary care: a guide to good practice*. 2009.

[37] P. Arleta Rewers, MD, "Acute Metabolic Complications in Diabetes," in *Diabetes in America, 3rd Edition*, 2016.

[38] "Hypoglycemia(low blood glucose)," *American Diabetes Association*, 2016. [Online]. Available: http://www.diabetes.org/living-with-diabetes/treatment-and-care/blood-glucose-control/hypoglycemia-low-blood.html?loc=lwd-slabnav.

[39] C. K.-Z. M. . Edited by Peter Manu M.D., *Handbook of Medicine in Psychiatry Second Edition*. .

[40] A. H. Goroll and A. G. Mulley, “Primary care medicine : office evaluation and management of the adult patient.” Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 2009.

[41] R. Bilous and R. Donelly, “Handbook of Diabetes, 4th Edition,” *http://www.diabetesincontrol.com/handbook-of-diabetes-4th-edition-excerpt-6-epidemiology-and-aetiology-of-type-2-diabetes/*, p. Accessed May 2018, 2014.

[42] C. M. F. Matthew D. McEvoy, *Advanced Perioperative Crisis Management*. 2017.

[43] Margaret A. Powers, *Handbook of Diabetes Medical Nutrition Therapy*. 1996.

[44] R. D. Rudy Bilous, *Handbook of Diabetes 4TH Edition*. .

[45] D. M.-W. Barry J. Goldstein, *Type 2 Diabetes: Principles and Practice, Second Edition*. 2007.

[46] J. M. O. Derek LeRoith, Simeon I. Taylor, *Diabetes Mellitus: A Fundamental and Clinical Text 3RD Edition*. 2003.

[47] Rajeev Chawla, *Complications of Diabetes*. 2012.

[48] S. I. Ahmad and SpringerLink (Online service), “Diabetes An Old Disease, a New Insight,” *Advances in Experimental Medicine and Biology,*. p. XXXIII, 485 p., 2013.

[49] By Health and Administration Development Group, *Diabetes Management: Clinical Pathways, Guidelines, and Patient Education*. Aspen Publishers, 1999.

[50] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, “Epidemiology of Diabetes and Diabetes-Related Complications,” *Phys. Ther.*, vol. 88, no. 11, pp. 1254–1264, 2008.

[51] V. L. Katch, W. D. McArde, and F. I. Katch, *Essentials of Exercise Physiology*, no. 1. 2011.

[52] W. V. Bolie, “Coefficients of normal blood glucose regulation,” *J Appl Physiol.*, vol. 16, no. 5, pp. 783–788, 1961.

[53] E. Ackerman, J. W. Rosevear, and W. F. McGuckin, “A mathematical model of the glucose-tolerance test,” *Phys. Med. Biol.*, vol. 9, no. 2, p. 203, 1964.

[54] R. Bergman, Y. Ider, C. Bowden, and C. Cobelli, “Quantitative estimation of insulin sensitivity,” *Am. J. Physiol. Endocrinol. Metab.*, vol. 236, no. 6, pp. E667-677, 1979.

[55] G. Toffolo, R. N. Bergman, D. T. Finegood, C. R. Bowden, and C. Cobelli, “Quantitative estimation of beta cell sensitivity to glucose in the intact organism. A minimal model of

insulin kinetics in the dog," *Diabetes*, vol. 29, no. 12, pp. 979–990, 1980.

[56] R. N. Bergman, L. S. Phillips, and C. Cobelli, "Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose.," *J. Clin. Invest.*, vol. 68, no. 6, pp. 1456–67, 1981.

[57] R. N. Bergman, "Pathogenesis and prediction of diabetes mellitus: lessons from integrative physiology.," *Mt. Sinai J. Med.*, vol. 69, no. 5, pp. 280–290, 2002.

[58] R. N. Bergman, D. T. Finegood, and M. Ader, "Assessment of insulin sensitivity in vivo," *Endocr. Rev.*, vol. 6, no. 1, p. 45, 1985.

[59] A. Boutayeb and A. Chetouani, "A critical review of mathematical models and data used in diabetology," *Biomed. Eng. Online*, vol. 5, 2006.

[60] C. Cobelli, G. Federspil, G. Pacini, A. Salvan, and C. Scandellari, "An integrated mathematical model of the dynamics of blood glucose and its hormonal control," *Math. Biosci.*, vol. 58, no. 1, pp. 27–60, 1982.

[61] C. Dalla Man, R. A. Rizza, and C. Cobelli, "Meal simulation model of the glucose-insulin system," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 10, pp. 1740–1749, 2007.

[62] C. Dalla Man, D. M. Raimondo, R. A. Rizza, and C. Cobelli, "GIM, Simulation Software of Meal Glucose—Insulin Model," *J. Diabetes Sci. Technol.*, vol. 1, no. 3, pp. 323–330, 2007.

[63] J. T. Sorensen, "A physiologic model of glucose metabolism in man and its use to design and assess improved insulin therapies for diabetes." Massachusetts Institute of Technology, 1985.

[64] P. R.S., D. I. I. I. F.J., R. S. Parker, and F. J. Doyle, "Control-relevant modeling in drug delivery," *Adv Drug Deliv Rev*, vol. 48, no. 2–3, pp. 211–228, 2001.

[65] S. M. Lynch and B. W. Bequette, "Model predictive control of blood glucose in type I diabetics using subcutaneous glucose measurements," in *Proceedings of the 2002 American Control Conference (IEEE Cat. No.CH37301)*, 2002, vol. 5, pp. 4039–4043 vol.5.

[66] T. V. and M. E. W. Hovorka, R., V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. OrsiniFederici, T. R. Pieber, H. C. Schaller, L. Schaupp, "Nonlinear Model Predictive Control of Glucose Concentration in Subjects with Type 1 Diabetes," *Physiol. Meas.*, vol. 25, pp. 905–920, 2004.

[67] R. Hovorka *et al.*, “Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT,” *Am. J. Physiol. - Endocrinol. Metab.*, vol. 282, no. 5, pp. E992–E1007, 2002.

[68] M. E. Wilinska, L. J. Chassin, H. C. Schaller, L. Schaupp, T. R. Pieber, and R. Hovorka, “Insulin kinetics in type-1 diabetes: Continuous and bolus delivery of rapid acting insulin,” *IEEE Trans. Biomed. Eng.*, vol. 52, no. 1, pp. 3–12, 2005.

[69] R. Basu *et al.*, “Use of a novel triple-tracer approach to assess postprandial glucose metabolism,” *Am. J. Physiol. Metab.*, vol. 284, no. 1, pp. E55–E69, 2003.

[70] L. Magni *et al.*, “Model Predictive Control of Type 1 Diabetes: An Model Predictive Control of Type 1 Diabetes: An in Silico Trial,” *J Diabetes Sci Technol J. Diabetes Sci. Technol.*, vol. 1, no. 6, 2007.

[71] I. Ajmera, M. Swat, C. Laibe, N. Le Novére, and V. Chelliah, “The impact of mathematical modeling on the understanding of diabetes and related complications,” *CPT: Pharmacometrics and Systems Pharmacology*, vol. 2, no. 7. 2013.

[72] C. B. Landersdorfer and W. J. Jusko, “Pharmacokinetic/pharmacodynamic modelling in diabetes mellitus,” *Clinical Pharmacokinetics*, vol. 47, no. 7. pp. 417–448, 2008.

[73] B. Topp, K. Promislow, G. deVries, R. M. Miura, and D. T. Finegood, “A model of beta-cell mass, insulin, and glucose kinetics: pathways to diabetes.,” *J. Theor. Biol.*, vol. 206, no. 4, pp. 605–619, 2000.

[74] A. De Gaetano *et al.*, “Mathematical models of diabetes progression.,” *Am. J. Physiol. Endocrinol. Metab.*, vol. 295, no. 6, pp. E1462-79, 2008.

[75] T. Hardy, E. Abu-Raddad, N. Porksen, and A. De Gaetano, “Evaluation of a mathematical model of diabetes progression against observations in the diabetes prevention program,” *Am. J. Physiol. Metab.*, vol. 303, no. 2, pp. E200–E212, 2012.

[76] J. Nie, *Muscle biomarkers of type 2 diabetes disease progression in Goto-Kakizaki rats*. State University of New York at Buffalo, 2010.

[77] P. M. Clarke *et al.*, “A model to estimate the lifetime health outcomes of patients with Type 2 diabetes: The United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68),” *Diabetologia*, vol. 47, no. 10, pp. 1747–1759, 2004.

[78] E. Mueller, S. Maxion-Bergemann, D. Gultyaev, S. Walzer, and R. Bergemann, “PDB33 EAGLE—DIABETES MODEL: BASIC FEATURES AND INTERNAL VALIDATION OF SIMULATING LONG-TERM DIABETIC OUTCOMES AND RELATED COSTS,” *Value*

Heal., vol. 7, no. 6, p. 745, 2004.

[79] D. M. Eddy and L. Schlessinger, “Archimedes: A trial-validated model of diabetes,” *Diabetes Care*, vol. 26, no. 11, pp. 3093–3101, 2003.

[80] A. J. Palmer *et al.*, “Computer modeling of diabetes and its complications: a report on the Fourth Mount Hood Challenge Meeting.(ADA WORKGROUP REPORT)(Conference notes),” *Diabetes Care*, vol. 30, p. 1638, 2007.

[81] D. M. Eddy and L. Schlessinger, “Validation of the Archimedes Diabetes Model,” *Diabetes Care*, vol. 26, no. 11, pp. 3102–3110, 2003.

[82] A. G. Mainous *et al.*, “Impact of the population at risk of diabetes on projections of diabetes burden in the United States: an epidemic on the way,” *Diabetologia*, vol. 50, no. 5, pp. 934–940, May 2007.

[83] M. Brändle and W. H. Herman, “The CORE Diabetes Model,” *Curr. Med. Res. Opin.*, vol. 20, no. sup1, pp. S1–S3, Jan. 2004.

[84] A. Bagust, P. K. Hopkinson, W. Maier, and C. J. Currie, “An economic model of the long-term health care burden of Type II diabetes,” *Diabetologia*, vol. 44, no. 12, pp. 2140–2155, Dec. 2001.

[85] R. Kahn, “Guidelines for computer modeling of diabetes and its complications,” *Diabetes Care*, vol. 27, no. 9. pp. 2262–2265, 2004.

[86] S. Petrou and A. Gray, “Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting,” *BMJ*, vol. 342, Apr. 2011.

[87] E. Mueller *et al.*, “Development and validation of the economic assessment of glycemic control and long-term effects of diabetes (EAGLE) model.,” *Diabetes Technol. Ther.*, vol. 8, no. 2, pp. 219–236, 2006.

[88] A. Neumann, L. Lindholm, M. Norberg, O. Schoffer, S. J. Klug, and F. Norström, “The cost-effectiveness of interventions targeting lifestyle change for the prevention of diabetes in a Swedish primary care and community based prevention program,” *Eur. J. Heal. Econ.*, vol. 18, no. 7, pp. 905–919, Dec. 2017.

[89] H. Squires and P. Tappenden, “Mathematical modelling and its application to social care Improving the evidence base for adult social care practice The School for Social Care Research.”

[90] S. Saha, U.-G. Gerdtham, and P. Johansson, “Economic Evaluation of Lifestyle Interventions for Preventing Diabetes and Cardiovascular Diseases,” *International*

Journal of Environmental Research and Public Health , vol. 7, no. 8. 2010.

[91] U. Siebert, “When should decision-analytic modeling be used in the economic evaluation of health care?,” *European Journal of Health Economics*, vol. 4, no. 3, pp. 143–150, 2003.

[92] W. H. Herman, “Diabetes Modeling,” *Diabetes Care*, vol. 26, no. 11, pp. 3182–3183, 2003.

[93] U. Siebert *et al.*, “State-transition modeling: A report of the ISPOR-SMDM modeling good research practices task force-3,” *Med. Decis. Mak.*, vol. 32, no. 5, pp. 690–700, 2012.

[94] A. D. M. Briggs, J. Wolstenholme, T. Blakely, and P. Scarborough, “Choosing an epidemiological model structure for the economic evaluation of non-communicable disease public health interventions,” *Popul. Health Metr.*, vol. 14, no. 1, p. 17, 2016.

[95] L. Schlessinger and D. M. Eddy, “Archimedes: a new model for simulating health care systems--the mathematical formulation,” *J Biomed Inf.*, vol. 35, no. 1, pp. 37–50, 2002.

[96] E. DM, L. Schlessinger, and R. Kahn, “CLinical outcomes and cost-effectiveness of strategies for managing people at high risk for diabetes,” *Ann. Intern. Med.*, vol. 143, no. 4, pp. 251–264, Aug. 2005.

[97] M. F. Faruque and I. H. Sarker, “Performance analysis of machine learning techniques to predict diabetes mellitus,” in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–4.

[98] R. Patil and S. Tamane, “A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, p. 3966, 2018.

[99] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, “An improved artificial neural network model for effective diabetes prediction,” *Complexity*, vol. 2021, 2021.

[100] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes prediction using ensembling of different machine learning classifiers,” *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[101] N. Abdulhadi and A. Al-Mousa, “Diabetes Detection Using Machine Learning Classification Methods,” in *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 350–354.

[102] A. Oleiwi, L. Shi, Y. Tao, and L. Wei, “A comparative analysis and risk prediction of diabetes at early stage using machine learning approach,” *Int. J. Futur. Gener. Commun. Netw.*, vol. 13, no. 3, pp. 4151–4163, 2020.

[103] K. Kantawong, S. Tongphet, P. Bhrommalee, N. Rachata, and S. Pravesjit, “The Methodology for Diabetes Complications Prediction Model,” in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, 2020, pp. 110–113.

[104] A. Dagliati *et al.*, “Machine learning methods to predict diabetes complications,” *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 295–302, 2018.

[105] M. S. Islam, M. K. Qaraqe, and S. B. Belhaouari, “Early Prediction of Hemoglobin Alc: A novel Framework for better Diabetes Management,” in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 542–547.

[106] A. Earnest, M. I. Chen, D. Ng, and L. Y. Sin, “Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore,” *BMC Health Serv. Res.*, vol. 5, no. 1, pp. 1–8, 2005.

[107] Q. Liu, X. Liu, B. Jiang, and W. Yang, “Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model,” *BMC Infect. Dis.*, vol. 11, no. 1, pp. 1–7, 2011.

[108] E. D. Dan, O. Jude, and O. Idochi, “Modelling and forecasting malaria mortality rate using Sarima models (a case study of Aboh Mbaise general hospital, Imo State Nigeria),” *Sci. J. Appl. Math. Stat.*, vol. 2, no. 1, pp. 31–41, 2014.

[109] M. Villani, A. Earnest, N. Nanayakkara, K. Smith, B. De Courten, and S. Zoungas, “Time series modelling to forecast prehospital EMS demand for diabetic emergencies,” *BMC Health Serv. Res.*, vol. 17, no. 1, pp. 1–9, 2017.

[110] T. Singye and S. Unhapipat, “Time series analysis of diabetes patients: A case study of Jigme Dorji Wangchuk National Referral Hospital in Bhutan,” in *Journal of Physics: Conference Series*, 2018, vol. 1039, no. 1, p. 12033.

[111] S. F. U. C. S. M. Group, *Modelling in Healthcare*. American Mathematical Society, 2010.

[112] E. A. Bender, *An Introduction to Mathematical Modeling*. Dover Publications, 2012.

[113] W. J. Meyer, *Concepts of Mathematical Modeling*. Dover Publications, 2012.

[114] C. M. Rutter, A. M. Zaslavsky, and E. J. Feuer, “Dynamic microsimulation models for health outcomes: a review,” *Med. Decis. Making*, vol. 31, no. 1, pp. 10–18, 2011.

[115] R. Davies, P. Roderick, and J. Raftery, “The evaluation of disease prevention and treatment using simulation models,” *Eur. J. Oper. Res.*, vol. 150, no. 1, pp. 53–66, 2003.

[116] E. S. Huang, A. Basu, M. O’grady, and J. C. Capretta, “Projecting the future diabetes population size and related costs for the US,” *Diabetes Care*, vol. 32, no. 12, pp. 2225–2229, 2009.

[117] J. A. Critchley and S. Capewell, “Why model coronary heart disease?,” *Eur. Heart J.*, vol. 23, no. 2, pp. 110–116, 2002.

[118] A. B. Shiflet and G. W. Shiflet, *Introduction to Computational Science: Modeling and Simulation for the Sciences - Second Edition*. Princeton University Press, 2014.

[119] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing, 2020.

[120] X. Sun and T. Faunce, “Decision-analytical modelling in health-care economic evaluations,” *Eur. J. Heal. Econ.*, vol. 9, no. 4, pp. 313–323, 2008.

[121] T. L. S. J. M. Alexander, *Thinking with models: Mathematical Models in the Physical, Biological, and Social Sciences*. RWS Publications, 2015.

[122] O. Saidi *et al.*, “Forecasting Tunisian type 2 diabetes prevalence to 2027: validation of a simple model,” *BMC Public Health*, vol. 15, no. 1, p. 104, 2015.

[123] M. Jit and M. Brisson, “Modelling the epidemiology of infectious diseases for decision analysis,” *Pharmacoeconomics*, vol. 29, no. 5, pp. 371–386, 2011.

[124] M. T. Halpern, B. R. Luce, R. E. Brown, and B. Geneste, “Health and economic outcomes modeling practices: a suggested framework,” *Value Heal.*, vol. 1, no. 2, pp. 131–147, 1998.

[125] M. Adibuzzaman, P. DeLaurentis, J. Hill, and B. D. Benneyworth, “Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2017, pp. 384–392, Apr. 2018.

[126] M. Dixon-Woods, S. McNicol, and G. Martin, “Ten challenges in improving quality in healthcare: lessons from the Health Foundation’s programme evaluations and relevant literature,” *BMJ Qual. Saf.*, vol. 21, no. 10, pp. 876–884, 2012.

[127] S. Piri, D. Delen, T. Liu, and H. M. Zolbanin, “A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble,” *Decis. Support Syst.*, vol. 101, pp. 12–27, 2017.

[128] F. A. Sonnenberg, M. S. Roberts, J. Tsevat, J. B. Wong, M. Barry, and D. L. Kent, “Toward a peer review process for medical decision analysis models,” *Med. Care*, pp. JS52-JS64, 1994.

[129] C. L. Chiang, *Statistical Methods Of Analysis*. World Scientific Publishing Company, 2003.

[130] N. Yokota *et al.*, “Predictive models for conversion of prediabetes to diabetes,” *J. Diabetes Complications*, vol. 31, no. 8, pp. 1266–1271, 2017.

[131] M. D. Rossetti, *Simulation Modeling and Arena*. Wiley, 2015.

[132] G. A. Bekey and B. J. Kogan, *Modeling and Simulation: Theory and Practice: A Memorial Volume for Professor Walter J. Karplus (1927–2001)*. Springer US, 2012.

[133] B. P. Zeigler, A. Muzy, and E. Kofman, *Theory of Modeling and Simulation: Discrete Event & Iterative System Computational Foundations*. Elsevier Science, 2018.

[134] J. A. Sokolowski and C. M. Banks, *Principles of Modeling and Simulation: A Multidisciplinary Approach*. Wiley, 2011.

[135] D. Cvetković, *Modeling and Computer Simulation*. IntechOpen, 2019.

[136] K. Velten, *Mathematical Modeling and Simulation: Introduction for Scientists and Engineers*. Wiley, 2009.

[137] G. Allaire, G. Allaire, A. Craig, and D. M. A. Craig, *Numerical Analysis and Optimization: An Introduction to Mathematical Modelling and Numerical Simulation*. OUP Oxford, 2007.

[138] A. J. Palmera *et al.*, “Validation of the CORE Diabetes Model against epidemiological and clinical studies,” *Curr. Med. Res. Opin.*, vol. 20, no. sup1, pp. S27–S40, 2004.

[139] M. C. Chubb and K. H. Jacobsen, “Mathematical modeling and the epidemiological research process,” *Eur. J. Epidemiol.*, vol. 25, no. 1, pp. 13–19, 2010.

[140] D. Walker and J. A. Fox-Rushby, “Allowing for uncertainty in economic evaluations: qualitative sensitivity analysis,” *Health Policy Plan.*, vol. 16, no. 4, pp. 435–443, 2001.

[141] “Saudi Health Interview Survey Results,” 2013.

[142] A. S. Warsy and M. A. El Hazmi, “Diabetes mellitus, hypertension and obesity-common multifactorial disorders in Saudis,” 1999.

[143] “WHO STEPwise Approach to NCD Surveillance, Country-Specific Standard Report, Saudi Arabia;,” 2005.

[144] M. M. Al-Nozha *et al.*, “Obesity in Saudi Arabia.,” *Saudi Med. J.*, vol. 26, no. 5, pp. 824–829, 2005.

[145] J. S. Jarallah, K. A. Al-Rubeaan, A. R. A. Al-Nuaim, A. A. Al-Ruhaily, and K. A. Kalantan, “Prevalence and determinants of smoking in three regions of Saudi Arabia,” *Tob. Control*, vol. 8, no. 1, pp. 53–56, 1999.

[146] R. Ramezani, M. Maadi, and S. M. Khatami, “A novel hybrid intelligent system with missing value imputation for diabetes diagnosis,” *Alexandria Eng. J.*, vol. 57, no. 3, pp. 1883–1891, 2018.

[147] H. Turabieh, M. Mafarja, and S. Mirjalili, “Dynamic Adaptive Network-Based Fuzzy Inference System (D-ANFIS) for the Imputation of Missing Data for Internet of Medical Things Applications,” *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9316–9325, 2019.

[148] E. Dogantekin, A. Dogantekin, D. Avci, and L. Avci, “An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS,” *Digit. Signal Process.*, vol. 20, no. 4, pp. 1248–1255, 2010.

[149] H. Temurtas, N. Yumusak, and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8610–8615, 2009.

[150] K. Kayaer and T. Yildirim, “Medical diagnosis on Pima Indian diabetes using general regression neural networks,” in *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, 2003, vol. 181, p. 184.

[151] E.-L. Silva-Ramírez and J.-F. Cabrera-Sánchez, “Co-active neuro-fuzzy inference system model as single imputation approach for non-monotone pattern of missing data,” *Neural Comput. Appl.*, pp. 1–24, 2021.

[152] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[153] C. K. Enders and D. L. Bandalos, “The relative performance of full information maximum likelihood estimation for missing data in structural equation models,” *Struct. Equ. Model.*, vol. 8, no. 3, pp. 430–457, 2001.

[154] V. Kuppusamy and I. Paramasivam, “Grey fuzzy neural network-based hybrid model for missing data imputation in mixed database,” *Int J Intell Eng Syst*, vol. 10, no. 2, pp. 146–155, 2017.

[155] P. Koikkalainen, “Neural networks for editing and imputation,” in *DataClean 2002 Conference, Jyväskylä (Finland)*, 2002, pp. 1013–1035.

[156] W. Wei and Y. Tang, “A generic neural network approach for filling missing data in data mining,” in *SMC’03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, 2003, vol. 1, pp. 862–867.

[157] S. Nordbotten, “Neural network imputation applied to the Norwegian 1990 population census data,” *J. Off. Stat.*, vol. 12, pp. 385–402, 1996.

[158] I. B. Aydilek and A. Arslan, “A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm,” *Inf. Sci. (Ny).*, vol. 233, pp. 25–35, 2013.

[159] J. Kim and N. Kasabov, “HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems,” *Neural networks*, vol. 12, no. 9, pp. 1301–1319, 1999.

[160] S. Chavan, K. Shah, N. Dave, S. Mukherjee, A. Abraham, and S. Sanyal, “Adaptive neuro-fuzzy intrusion detection systems,” in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, 2004, vol. 1, pp. 70–74.

[161] S. Sun, “A survey of multi-view machine learning,” *Neural Comput. Appl.*, vol. 23, no. 7, pp. 2031–2038, 2013.

[162] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning,” *Cited on*, p. 33, 2009.

[163] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[164] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[165] G. Bonaccorso, *Machine Learning Algorithms*. Packt Publishing, 2017.

[166] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York, 2013.

[167] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Wiley, 2021.

[168] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, 2013.

[169] C. C. Aggarwal, *Data Classification: Algorithms and Applications*. Taylor & Francis, 2014.

[170] J. K. Mandal and D. Bhattacharya, *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*. Springer Singapore, 2019.

[171] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne, “Machine learning strategies for time series forecasting,” in *European business intelligence summer school*, 2012, pp. 62–77.

[172] T. C. Mills, *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*. Elsevier Science, 2019.

[173] A. Nielsen, *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O’Reilly Media, 2019.

[174] H. Madsen, *Time Series Analysis*. CRC Press, 2007.

[175] A. Lapedes and R. Farber, “Nonlinear signal processing using neural networks: Prediction and system modelling,” 1987.

[176] P. J. Werbos, “Generalization of backpropagation with application to a recurrent gas market model,” *Neural networks*, vol. 1, no. 4, pp. 339–356, 1988.

[177] D. Wallach and B. Goffinet, “Mean squared error of prediction as a criterion for evaluating and comparing system models,” *Ecol. Modell.*, vol. 44, no. 3, pp. 299–306, 1989.

[178] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer New York, 2013.

[179] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[180] P. J. Lavrakas, *Encyclopedia of Survey Research Methods*. SAGE Publications, 2008.

[181] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists: 50 Essential Concepts*. O’Reilly Media, 2017.

[182] P. M. Swamidass, Ed., “MAPE (mean absolute percentage error)MEAN ABSOLUTE

PERCENTAGE ERROR (MAPE) BT - Encyclopedia of Production and Manufacturing Management," Boston, MA: Springer US, 2000, p. 462.

[183] D. A. Swanson, J. Tayman, and T. M. Bryan, "MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts," *J. Popul. Res.*, vol. 28, no. 2, pp. 225–243, 2011.

[184] Y. Dodge, *The Concise Encyclopedia of Statistics*. Springer New York, 2008.

[185] T. P. Ryan, *Modern Regression Methods*. Wiley, 2008.

[186] L. Rutkowski, *Flexible Neuro-Fuzzy Systems: Structures, Learning and Performance Evaluation*. Springer US, 2006.

[187] M. N. M. Salleh, N. Talpur, and K. Hussain, "Adaptive Neuro-Fuzzy Inference System: Overview, Strengths, Limitations, and Solutions BT - Data Mining and Big Data," 2017, pp. 527–535.

[188] W. Suparta and K. M. Alhasa, *Modeling of Tropospheric Delays Using ANFIS*. Springer International Publishing, 2015.

[189] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress, 2015.

[190] R. Sałat and K. Sałat, "The application of support vector regression for prediction of the antiallodynic effect of drug combinations in the mouse model of streptozocin-induced diabetic neuropathy," *Comput. Methods Programs Biomed.*, vol. 111, no. 2, pp. 330–337, 2013.

[191] W. Koehrsen, "Bayesian Linear Regression in Python: Using Machine Learning to Predict Student Grades Part 2," *towards data science*, 2018. [Online]. Available: https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e%0D.

[192] B. M. Wilamowski and J. D. Irwin, *Intelligent Systems*. CRC Press, 2018.

[193] H. M. Al-Hazzaa, "Physical inactivity in Saudi Arabia revisited: a systematic review of inactivity prevalence and perceived barriers to active living," *Int. J. Health Sci. (Qassim).*, vol. 12, no. 6, p. 50, 2018.

[194] M. Re and G. Valentini, "1 Ensemble methods: a review 3," 2012.

[195] J. Wang, *Encyclopedia of Data Warehousing and Mining, Second Edition*. Information Science Reference, 2008.

[196] G. Valentini and F. Masulli, “Ensembles of learning machines,” in *Italian workshop on neural nets*, 2002, pp. 3–20.

[197] L. Rokach, *Pattern Classification Using Ensemble Methods*. World Scientific, 2010.

[198] M. O. Elish, T. Helmy, and M. I. Hussain, “Empirical study of homogeneous and heterogeneous ensemble models for software development effort estimation,” *Math. Probl. Eng.*, vol. 2013, 2013.

[199] A. V Kelarev, A. Stranieri, J. L. Yearwood, and H. F. Jelinek, “Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare,” in *2012 15th International Conference on Network-Based Information Systems*, 2012, pp. 441–446.

[200] G. Zorluoglu and M. Agaoglu, “Diagnosis of breast cancer using ensemble of data mining classification methods,” *Int. J. Oncol. Cancer Ther.*, vol. 2, 2017.

[201] J. Liu, L. Wang, L. Zhang, Z. Zhang, and S. Zhang, “Predictive analytics for blood glucose concentration: an empirical study using the tree-based ensemble approach,” *Libr. Hi Tech*, 2020.

[202] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.

[203] J. Brownlee, *Ensemble Learning Algorithms With Python: Make Better Predictions with Bagging, Boosting, and Stacking*. Machine Learning Mastery, 2021.

[204] A. J. C. Sharkey, “Linear and order statistics combiners for pattern classification,” in *Combining artificial neural nets*, Springer, 1999, pp. 127–161.

[205] O. Okun, *Applications of Supervised and Unsupervised Ensemble Methods*. Springer Berlin Heidelberg, 2009.

[206] L. I. Kuncheva, “A theoretical study on six classifier fusion strategies,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, 2002.

[207] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Springer New York, 2012.

[208] J. Kittler, “Combining classifiers: A theoretical framework,” *Pattern Anal. Appl.*, vol. 1, no. 1, pp. 18–27, 1998.

[209] S. B. Kotsiantis and P. Pintelas, “Selective averaging of regression models,” *Ann. Math. Comput. Teleinformatics*, vol. 1, no. 3, pp. 65–74, 2005.

[210] C. A. Shipp and L. I. Kuncheva, “Relationships between combination methods and

measures of diversity in combining classifiers," *Inf. fusion*, vol. 3, no. 2, pp. 135–148, 2002.

[211] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, p. 100203, 2019.

[212] L. I. Kuncheva and J. J. Rodríguez, "A weighted voting framework for classifiers ensembles," *Knowl. Inf. Syst.*, vol. 38, no. 2, pp. 259–275, 2014.

[213] M. Ponti, *Combining Classifiers: From the Creation of Ensembles to the Decision Fusion*. 2011.

[214] M. H. DeGroot, "Reaching a Consensus," *J. Am. Stat. Assoc.*, vol. 69, no. 345, pp. 118–121, Jan. 1974.

[215] K. B. Shaban, O. A. Basir, K. Hassanein, and M. Kamel, "Information fusion in a cooperative multi-agent system for web information retrieval," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, 2002, vol. 2, pp. 1256–1262 vol.2.

[216] K. Shaban, O. A. Basir, M. Kamel, and K. Hassanein, "Intelligent information fusion approach in cooperative multiagent systems," in *Proceedings of the 5th Biannual World Automation Congress*, 2002, vol. 13, pp. 429–434.

[217] P. Gramatica, E. Giani, and E. Papa, "Statistical external validation and consensus modeling: a QSPR case study for Koc prediction," *J. Mol. Graph. Model.*, vol. 25, no. 6, pp. 755–766, 2007.

[218] M. Marmion, M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller, "Evaluation of consensus methods in predictive species distribution modelling," *Divers. Distrib.*, vol. 15, no. 1, pp. 59–69, 2009.

[219] R. L. Berger, "A necessary and sufficient condition for reaching a consensus using DeGroot's method," *J. Am. Stat. Assoc.*, vol. 76, no. 374, pp. 415–418, 1981.

[220] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, no. Nov, pp. 45–66, 2001.

[221] M. Fischetti, "Fast training of support vector machines with Gaussian kernel," *Discret. Optim.*, vol. 22, pp. 183–194, 2016.

[222] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, and S. Abdulwahed, "Diabetes classification based on KNN," *IIUM Eng. J*, vol. 21, no. 1, pp. 175–181, 2020.

[223] K. Torkkola, “Linear discriminant analysis in document classification,” in *IEEE ICDM workshop on text mining*, 2001, pp. 800–806.

[224] C. H. Park and H. Park, “A comparison of generalized linear discriminant analysis algorithms,” *Pattern Recognit.*, vol. 41, no. 3, pp. 1083–1097, 2008.

[225] J. K. Basu, D. Bhattacharyya, and T. Kim, “Use of artificial neural network in pattern recognition,” *Int. J. Softw. Eng. its Appl.*, vol. 4, no. 2, 2010.

[226] O. I. Abiodun *et al.*, “Comprehensive review of artificial neural network applications to pattern recognition,” *IEEE Access*, vol. 7, pp. 158820–158846, 2019.

[227] H. Kukreja, N. Bharath, C. S. Siddesh, and S. Kuldeep, “An introduction to artificial neural network,” *Int J Adv Res Innov Ideas Educ*, vol. 1, pp. 27–30, 2016.

[228] “Matlab, ‘Statistics and machine learning toolbox,’” 2018. [Online]. Available: https://mathworks.com/products/statistics.html,.

[229] R. S. Tsay, “Time Series and Forecasting: Brief History and Future Research,” *J. Am. Stat. Assoc.*, vol. 95, no. 450, pp. 638–643, Apr. 2000.

[230] W. Palma, *Time Series Analysis*. Wiley, 2016.

[231] G. K. Jha and K. Sinha, “Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India,” *Neural Comput. Appl.*, vol. 24, no. 3, pp. 563–571, 2014.

[232] G. K. Jha, P. Thulasiraman, and R. K. Thulasiram, “PSO based neural network for time series forecasting,” in *2009 International Joint Conference on Neural Networks*, 2009, pp. 1422–1427.

[233] S. F. Crone, M. Hibon, and K. Nikolopoulos, “Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction,” *Int. J. Forecast.*, vol. 27, no. 3, pp. 635–660, 2011.

[234] D. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. Wiley, 2001.

[235] J. M. P. Menezes Jr and G. A. Barreto, “Long-term time series prediction with the NARX network: An empirical evaluation,” *Neurocomputing*, vol. 71, no. 16–18, pp. 3335–3343, 2008.

[236] E. Diaconescu, “The use of NARX neural networks to predict chaotic time series,” *Wseas Trans. Comput. Res.*, vol. 3, no. 3, pp. 182–191, 2008.

[237] A. Tatli and S. Kahvecioğlu, “NARX neural networks based time series prediction for amount of airworthiness time,” in *2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO)*, 2016, pp. 130–134.