Type 2 Diabetes Mellitus Automated Risk Detection Based on UAE National Health Survey Data: A Framework for the Construction and Optimization of Binary Classification Machine Learning Models Based on Dimensionality Reduction

أتمتة تقصي خطر الإصابة بمرض السكر من النوع الثاني بناء على بيانات المسح الصحي الوطني لدولة الإمارات العربية المتحدة: إطار عمل لبناء و تحسين نتائج نماذج التصنيف الثنائي المبني على تعلم الآله من خلال تقنيات تقليل الأبعاد للبيانات المستخدمة

by

MOHAMED SALIM AL SHUWEIHI

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

November 2020

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

Machine Learning (ML) saw a great increase in general and domain specific research. ML in bioinformatics and epidemiology in particular grew drastically, powered by the proliferation of Electronic Medical Records (EMR) in healthcare systems worldwide and the efficiency of new programmatic and computational tools supporting Artificial Intelligence (AI) application. This research motivated by the unprecedented increase in diabetes and specifically Type 2 Diabetes Miletus (T2DM), proposes two significant contributions. The first is a comprehensive ML framework for the construction of diagnostic binary classification high accuracy models to predict T2DM in the United Arab Emirates based on STEPS style National Health Survey. The second major contribution is the design and construction of a Logistic Regression (LR) ML binary classification model with an accuracy of 87% and F1-score of 89%. A special consideration was given to data pre-processing and dimensionality reduction such Chi Squared (CS) and Recursive Feature Elimination (RFE) to improve progressively the proposed models performance. LR with the reduced feature set using the intersection between CS and RFE proved to be the best model among the tested algorithms. This model can be used in a clinical setting as a decision support system or for public health awareness as an informal risk prediction system. Many people can find difficulty accessing diagnostic healthcare services for many reasons including, but not limited to economical and regional factors. ML based informal diagnostic and decision support systems can provide a first line of detection to alert patients about potential disease risk. Early alert of T2DM risk using a free ML tool can help the patients and healthcare workers to manage the disease as early as possible, reducing the risk of complication and financial burden.

ABSTRACT (Arabic)

شهد التعلم الآلي زيادة كبيرة في البحث العام والبحوث الخاصة بالمجالات المتخصصة مثل المعلوماتية الحيوية وعلم الأوبئة على وجه الخصوص بشكل كبير، مدعومًا بانتشار السجلات الطبية الإلكترونية في أنظمة الرعاية الصحية في جميع أنحاء العالم وكفاءة الأدوات البرمجية والحاسوبية الجديدة التي تدعم تطبيق الذكاء الاصطناعي. و يقترح هذا البحث بدافع الزيادة غير المسبوقة في مرض السكري وخاصة مرض السكري من النوع الثاني مساهمتين مهمتين. الإسهام الأول هو تقديم إطار عمل للتعلم الآلي شامل لبناء نماذج تشخيصية عالية الدقة لتصنيف ثنائي لتوقع داء السكر من النوع الثاني في دولة الإمارات العربية المتحدة مبني على بيانات المسح الصحي الوطني على طريقة STEPS. المساهمة الرئيسية الثانية هي تصميم وبناء نموذج تصنيف ثنائي للانحدار اللوجستي بدقة تبلغ 87٪ ودرجة F1 بنسبة 89٪. تم إيلاء اعتبار خاص للمعالجة المسبقة للبيانات وطرق تقليل الأبعاد مثل Chi Squared و الإلغاء التكراري للأبعاد لتحسين أداء النماذج المقترحة تدريجيًا. أثبت الإنحدار اللوجستي مع مراعاة تقليل الأبعاد للبيانات المستخدمة باستخدام التقاطع بين Chi Squared و الإلغاء التكراري للأبعاد أنه أفضل نموذج بين الخوارزميات المختبرة. يمكن استخدام هذا النموذج في بيئة سريرية كنظام لدعم القرار أو للتوعية بالصحة العامة كنظام غير رسمي للتنبؤ بالمخاطر و يمكن أن يجد العديد من الأشخاص صعوبة في الوصول إلى خدمات الرعاية الصحية التشخيصية لأسباب عديدة بما في ذلك ، على سبيل المثال لا الحصر ، العوامل الاقتصادية والإقليمية لذا يمكن للأنظمة غير الرسمية للتشخيص ودعم اتخاذ القرار القائمة على التعلم الآلي أن توفر خط الكشف الأول لتنبيه المرضى حول مخاطر المرض المحتملة و بناء عليه يمكن أن يساعد التنبيه المبكر لمخاطر الإصابة بداء السكر باستخدام أداة التعلم الآلي المقترحة المجانية المرضى والعاملين في الرعاية الصحية على إدارة المرض في أقرب وقت ممكن ، مما يقلل من مخاطر المضاعفات والعبء المالي.

DEDICATION

To my wife, for her unwavering support and unconditional motivation through stressful times and long hours of diligent research. Your robustness, thoughts and intelligence is what makes me want to ceaselessly pursue research and knowledge as an integral part of our life.

Thank you.

ACKNOWLEDGMENTS

Prof. Khaled Shaalan, Head of Informatics Programme in BUID, for not just being the dissertation supervisor, but a mentor and a teacher throughout the program. Thanking you for your care and continuous and overall passion for research and pursue of advancement of the field's knowledge through research and publication.

Dr. Nawal Al Mutawa, Consultant Endocrinologist in Ministry of Health and Prevention, for the dedication of hours to review the initial experiment design and expert input with relation to the intermediate and final findings, which supported the informatics work with valuable clinical insight.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
1R	One Rule Algorithm
2-h PG	Two-Hour plasma glucose
Abbreviation	Definition
ACCORD	Action to Control Cardiovascular Risk in Diabetes
ADA	American Diabetes Association
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASC	Abu Dhabi Statistics Centre
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
CADS	Computer Aided Diagnosis systems
CAPI	Computer Assisted Personal Interview
CDSS	Clinical Decision Support Systems
CEG	Clarke Error Grid
CKD	Chronic Kidney Disease
CkNN	Class-wise k Nearest Neighbours
CNN	Convolutional Neural Network
CS	Chi Squared
CSII	Continuous Subcutaneous Insulin Infusion
DHA	Dubai Health Authority
DIARETDB1	Standard Diabetic Retinopathy Database
DM	Diabetes Mellitus
DM	Data Mining
DSC	Dubai Statistics Centre
DT	Decision Tree
EMR	Electronic Medical Records
ESRD	End-Stage Renal Disease
FCSA	Federal Competitiveness & Statistics Authority
FPG	Fasting Plasma Glucose
GBM	Gradient Boosting Machine
GCP	Good Clinical Practice
GDM	Gestational Diabetes Mellitus
GDP	Gross Domestic Product
GE	Grammatical Evolution
GFR	Glomerular Filtration Rate
GPC	Gaussian Process Classification
HAAD	Health Authority of Abu Dhabi
HTML	Hypertext Markup Language
IDF	International Diabetes Federation
IoT	Internet of Things
J48	J48 Decision Tree
kNN	k Nearest Neighbours

LDA	Linear Discriminant Analysis
LEA	Lower Extremity Amputation
LOOCV	Leave One Out Cross Validation
LR	Logistic Regression
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MDI	Multiple Daily Injections
MENA	Middle East and North Africa
ML	Machine Learning
ML-BEC	Machine Learning Bagging Ensemble Classifier
MLP	Multilayer Perceptron
MOHAP	Ministry of Health and Prevention
MOHAP REC	MOHAP Research Ethics Committee
MOHAP SARC	MOHAP Statistics and Research Centre
NB	Naïve Bayes
NCD	Noncommunicable Diseases
NGSP	National Glycohemoglobin Standardization Program
NHANES	National Health and Nutrition Examination Survey
NHLBI	National Heart, Lung, and Blood Institute
NHS	National Health Survey
NLP	Natural Language Processing
NPV	Negative Predictive Value
OGTT	Oral Glucose Tolerance Test
OHE	One Hot Encoder
PHI	Protected Health Information
PPV	Positive Predictive Value
QDA	Quadratic Discriminant Analysis
ReLU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-Sampling Technique
STEPS	STEPwise Approach to Noncommunicable Disease Risk Factor Surveillance
SVM	Support Vector Machines
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
t-SNE	t-distributed Stochastic Neighbour Embedding
UAE	United Arab Emirates
UAE STEPS NHS	United Arab Emirates STEPS National Health Survey
UNPD	United Nation Population Division
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
WHO EMRO	WHO Regional Office for the Eastern Mediterranean

Chapter 1. Introduction

1.1. Background

Managing and improving public health and societal epidemiology are pivotal objectives dictating national health strategies worldwide, and the United Arab Emirates (UAE) is no exception to this (Vision2021, 2020). The World Health Organization (WHO) committed to the support of the world governments' national multisectoral policies to tackle the planning and prevention of Noncommunicable Diseases (NCD). This is done through such initiatives as 2011 Political Declaration on NCDs. NCDs like cardiovascular diseases, cancer, diabetes and chronic respiratory diseases are at the center of improving the public health and epidemiological planning for national health strategy and research. Although most NCDs are preventable through appropriate prevention and screening policies, unmanaged they can be extremely detrimental to the national healthcare and personal lifestyle of individuals. NCDs affect the patient chronically and thus require substantial effort and cost to manage the disease and its complications if it remained unmanaged or managed at an advanced stage of developing it.

Diabetes Mellitus (DM) or as it is commonly being referred to as diabetes is such a devastating chronic metabolic disease, which is characterized by the insistent increased glucose in the blood, mainly due to the deficiency in the production of the insulin hormone by the body or inability to process the produced insulin effectively (IDF, 2019). DM in particular can cause a devastating impact on the nation's public health and individuals from chronic NCDs. This impact can be measured on how the prevalence of Diabetes can affect the economy nationally and worldwide, due to cost associated with managing the patients afflicted by the disease. The cost impact although vital, but it can be augmented by the very negative impact on the diabetes patient lifestyle and ability to function normally in the society leading to disability, loss of

income and even death.

The UAE considered prevention of NCDs as a main strategic multisectoral objective as dictated by the national strategy and designed the indicators that can help in improving the management of the disease in the nation (Vision2021, 2020). There are three main types of diabetes, which constitute the majority of diabetes patients globally. There is Type 1 Diabetes Mellitus (T1DM), which is characterized by the inability of the pancreas to produce any or very limited amount of insulin because of β-cell total loss, causing the body to struggle in the management of glycaemic events and thus deal with the related complication of such events (American Diabetes Association, 2020a). The second and most prominent type is Type 2 Diabetes Mellitus (T2DM), which is related to the body not being able to manage glucose due to β-cell decreased secretion (American Diabetes Association, 2020a). And finally, there is Gestational Diabetes Mellitus (GDM), Which is mainly related to the diabetes occurring during pregnancy and can subside by the end of pregnancy or remain in certain cases and can be associated with certain effects on the newly born child (IDF, 2019). In all types of diabetes, if remained uncontrolled, will impact the patient life very negatively resulting in disability or even death.

This research is mainly concerned with T2DM, due to its high prevalence and its preventable and treatable nature. It is in particular considered as a disease that can be greatly affected by individual's lifestyle and therefore allows for various measures to be introduced in the society to reduce and prevent its prevalence. T2DM is characterized by it is being highly dependent on early detection and management of risk factors. This intervention unfortunately in almost one of every two patients does not happen on time, due to the fact that the undiagnosed diabetes patients are detected with the disease, thus being called the silent killer.

The past few years saw a tremendous progress in the utilization of artificial intelligence (AI) in the management of the various main sectors and functions of healthcare. Chief among these functions is Computer Aided Diagnosis systems (CADS) and Clinical Decision Support Systems (CDSS) among other highly varies use cases. Such systems are meant to act as fact based learned systems that can objectively provide the healthcare provider or physician with insight into the current patient state or predict a future event. However, such systems to be able to capture the complexity of some of the diseases such as a complicated disease like T2DM, has to employ Machine Learning (ML) at its core. ML is subset of AI that is meant with the construction and optimization of learned models, which are able to provide expert level in the intended task based on existing relevant data. Although the field is not new, it saw great leap in research and deployment in all fields and disciplines and more specifically in this case the field of bioinformatics in the past years because of many factors.

The reasons ML saw such an increase in interest in both research and deployment in the field of bioinformatics are many, but most important among them is the tremendous adoption of Electronic Medical Records (EMR) in global healthcare. This wide adoption, manifested in many forms with almost all patients' data being stored in both structured and unstructured forms to enable ease of access and connected and uninterrupted management of the patient. The United Arab Emirates (UAE) went through this transformation in transitioned to EMR in almost all government healthcare facilities through many initiatives such as Wareed in the Ministry of Health and Prevention (MOHAP) and Malaffi in Health Authority of Abu Dhabi (HAAD). This in turn gives the indication that the healthcare system in UAE is both ready for being a great source of data for ML based Bioinformatics solutions and a candidate for the deployment of such solutions.

Considering the complicated nature of T2DM that causes its pathogenesis and risk

factors to be highly intertwined and difficult to assess and correlate on time, and the ongoing advancement in EMR adoption in the country, ML is considered a very viable candidate to capture and predict T2DM states and events. This is expected to enable physicians to introduce proper interventions at ideal stages of the disease to avoid the life threatening complications that can result from the misdiagnosis and delayed of missed intervention. This of course can in turn improve the quality of life of the disease patients, alleviate the burden of cost on the country and healthcare system and highly contribute in understanding the disease and its complicated variables, which are not easy to manage using traditional consultation and treatment means.

1.2. **Motivation**

T2DM is increasingly becoming an alarming burden on individuals and on nations across the world. Great initiatives are being adopted by the global healthcare bodies to educate and raise awareness about the disease, however the prevalence of the disease does not seem to respond to such initiatives as the global data from the WHO and International Diabetes Federation (IDF). Which led to the work in this research to move towards being able to classify and predict the onset of the disease from existing patients' data. Research and employment of validated informal T2DM assessment tools is encouraged to detect the onset of the disease among asymptomatic adults (American Diabetes Association, 2020a), which mainly motivates this work.

1.2.1. **Diabetes Prevalence as a Global and National Health Risk**

TM2D is causing a debilitating constant pressure on healthcare systems and accounts for about 90% to 95% of diabetes cases globally (IDF, 2019). So many factors are causing the disease to be hard to diagnose at an early stage and thus prevent timely delivery of proper management. Developed and developing countries are struggling to reduce the prevalence of

the disease as measures taken recommended by the WHO and IDF are countered by the complicated nature of the disease, which makes it undetectable throughout the early stages of its incidence. To measure the prevalence of diabetes globally, IDF (2019) resorted to diverse data sources with the majority being peer-reviewed related publication from around the world. WHO STEPwise surveys and methods of surveillance were also considered, but with a level of scrutiny due to observed issues with estimation.

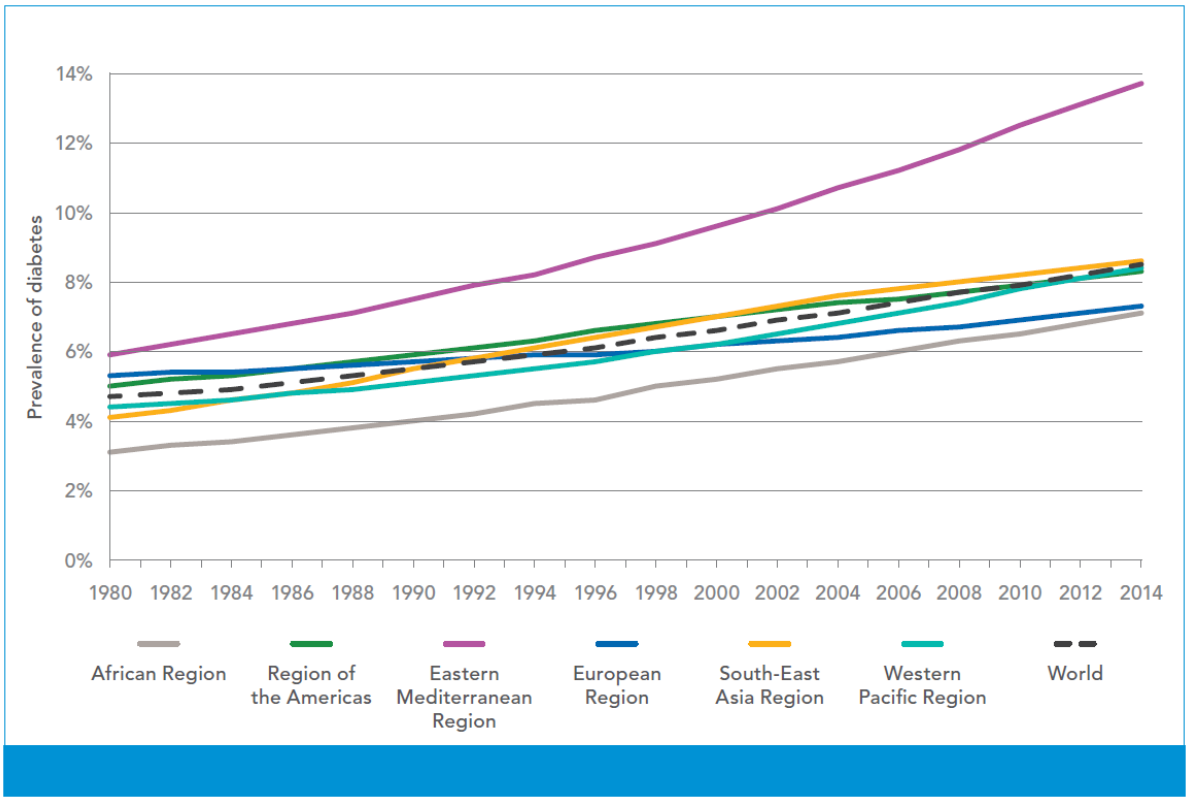

Figure 1. Trends of diabetes prevalence showing general increase globally and a highly uncontrolled increase in the Eastern Mediterranean region (WHO, 2016, p.27)

WHO predicts that diabetes prevalence trends are on the rise as in figure 1. IDF (2019) provides an estimate for the year 2019 and further predicts the prevalence of diabetes globally in 2030 and 2045. These future projection of diabetes prevalence in 2030 and 2045 were based on generalised liner regression estimates by age and sex. The future estimates are based on the United Nation Population Division (UNPD) to account for population projections with

consideration to factors such as levels of urbanisation and diabetes distribution among the age structures (IDF, 2019).

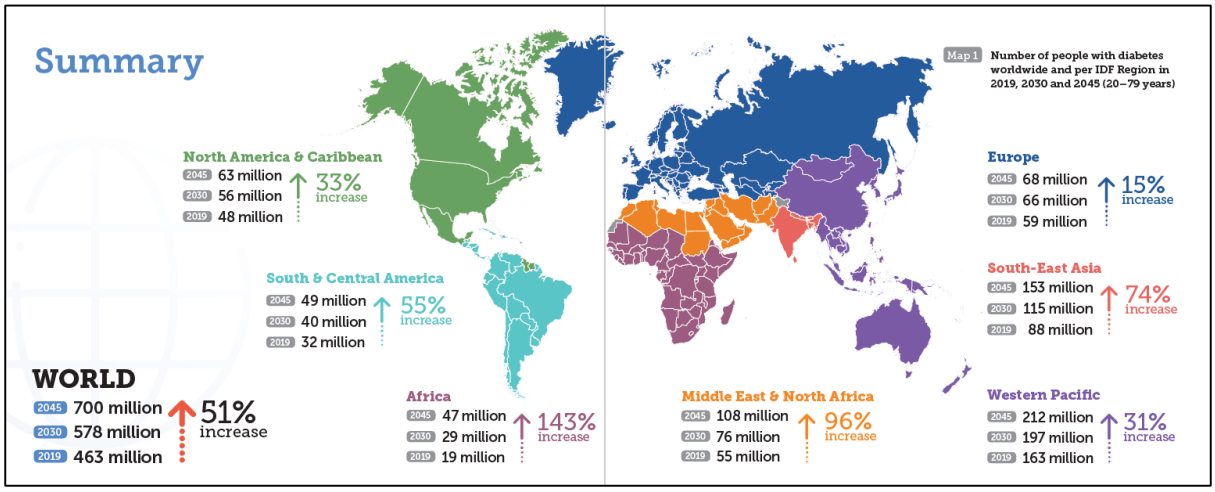

Figure 2. Global Estimated and Projected Diabetes Prevalence in 2019, 2030 and 2045 (IDF, 2019, p.4-5)

Diabetes is seriously impacting developed and developing countries in great scale, with future projections delivering a major call for action due to the great increase in the disease prevalence estimated. Currently, there are 463 million adults affected by diabetes globally, which estimated at around 9.3% of the adult population around the globe (IDF, 2019). Alarmingly, diabetes prevalence is projected to grow to 578 million or 10.2% of adult population by 2030, while further inflate to 700 million adults in 2045 or 10.9% of the global adult population aged 20 to 79 years as per the prevalence map in figure 2. Further to this, Middle East and North Africa (MENA) region is estimated to top IDF regions in age-adjusted prevalence at around 12%. MENA region is expected to reach a 96% increase in diabetes prevalence by 2045, only second to Africa at around 143% increase. Cases in MENA are expected to rise among adult population from 55 million in 2019 to 76 million in 2030 and subsequently to 108 million in 2045. UAE is one of the countries impacted by the alarming increase in diabetes prevalence in MENA region and current estimates in 2019 show a prevalence of around 15.4% among adult population, which accounts for around 1,223,400 of

the 7,925,700 total adult population in the country (IDF, 2019). In the UAE lack of physical activity and poor lifestyle choices are still a major challenge to maintaining target glycaemic levels for patients, and one in two patients most likely will have good control of the disease (Abuelmagd et al., 2018).

Diabetes is also associated with high rate of mortality rate across the globe (Sun & Zhang, 2019). Premature mortality is mainly caused by the various and extremely detrimental complications from diabetes and hyperglycaemia to the body functions (WHO, 2016). About half of diabetes related deaths occur before the age of 70 according to WHO (2016), dealing further socioeconomic pressure on the economy of the most impacted countries by the disease. Globally, in 2019 4.2 million deaths due to diabetes and its related comorbidities among adults aged between 20 and 79 (IDF, 2019). In the United States alone, such premature death rates because of diabetes cost about USD 19.9 billion from the total cost USD 90 billion of annual indirect cost to the economy of the United States. As evident in table 1, MENA comes in the second place in IDF regions in the rate of premature mortality resulting from diabetes with 53.3% of diabetes mortality occurring under the age of 60, constituting 223.3 thousand deaths in 2019. In the UAE around 2090 diabetes related deaths were recorded in 2019in adult population aged between 20 and 79.

IDF Region	Number of deaths due to diabetes before the age of 60 years (thousands)	Proportion of deaths due to diabetes occurring before the age of 60 years (%)
World	1,945.1 (1,528.7–2,525.3)[i]	46.2
AFR	267.6 (157.4–461.8)	73.1
MENA	223.3 (131.0–281.1)	53.3
SEA	592.3 (499.5–713.5)	51.5
NAC	132.7 (106.4–151.1)	44.0
SACA	105.8 (90.6–126.8)	43.5
WP	477.1 (428.3–590.7)	37.7
EUR	146.2 (115.5–200.3)	31.4

IDF: International Diabetes Federation; AFR: Africa, EUR: Europe; MENA: Middle East and North Africa; NAC: North America and Caribbean; SACA: South and Central America; SEA: South-East Asia; WP: Western Pacific

i 95% confidence intervals are reported in brackets..

Table 1. Proportion of Adults (20-79 years) to Die from Diabetes in 2019 before the Age of 60 Years, Globally and by IDF Regions, Ranked by the Proportions of Deaths Due to Diabetes before the age of 60 years (IDF, 2019, p.55)

1.2.2. Diabetes Complications and Comorbidities

Managing the complications resulting from high blood glucose associated with diabetes is key to the wellbeing and quality of life to the disease sufferer (WHO, 2016). On the other hand, failure in the management and control of the complications and comorbidities of the disease, can prove to be extremely detrimental to the patient and eventually fatal. For various reasons that upcoming sections will cover in more detail, almost half of diabetes patients remain undiagnosed. Thus, they become increasingly exposed to the risk of developing microvascular and macrovascular diabetes related complications (American Diabetes Association, 2020a). Although, microvascular complications vary greatly and can affect many bodily functions and organs, diabetic retinopathy, neuropathy and nephropathy are among the most evident and debilitating comorbidities (Dagliati et al., 2018). These morbidities can tremendously affect the wellbeing of the patient and can cause among many other, blindness, disability and kidney failure. Diabetic Retinopathy was responsible for 1.9% of moderate to severe loss of vision cases in 2019 and 2.6% of blindness globally in the same year (WHO,

2016). Diabetes causes 80% of End-Stage Renal Disease (ESRD) based on data from 54 countries (WHO, 2016).

Such long-term complication can manifest as early as in the first 5 years and its impact can be felt more intensely in developing countries. This can be attributed to many reasons, but mainly to lack of routine testing accessibility and means to control hyperglycaemia events. Even access to medications to treat severe hyperglycaemia can be limited by poverty and less developed medical systems (IDF, 2019). Cardiovascular diseases comprises the majority of diabetes related morbidity and even deaths (IDF, 2019). Thus diabetes can increase the risk cardiovascular diseases by 50% if diabetes diagnosis threshold occurred as indicated by IDF (2019). Among cardiovascular diseases, diabetes causes an increased risk of coronary heart disease by 160% and cardiovascular diseases death by 132%. Finally, diabetes and due to the onset of non-healing foot ulcers, can lead to significant increase in the risk of Lower Extremity Amputation (LEA) as diabetes patients are 10 to 20 times more at risk of LEA from other undiagnosed cohorts. In many cases traditional existing screening guidelines to such complications are found to be ineffective in countering and dealing with the progress such complications (Formosa et al., 2019).

1.2.3. **Diabetes Cost of Care and Pressure on Local Healthcare Systems**

From the previous sections it is clearly evident that the debilitating impact of high blood glucose due to diabetes, can put an extreme burden on the individual, family and society. Diabetes cost can be classified into 2 categories, direct and indirect cost of the disease. Direct cost can be attributed to the cost of healthcare services and means to treat, prevent and handle the disease chronic nature pressure on the healthcare system. On the other hand, indirect cost is associated with the burden affecting the economy as a result of inability of the diabetes patient to contribute in the economy of the nation due to severe progression in the disease,

debilitating disability or death. Also, reduced productivity, absenteeism and general inability to act productively in work setting contribute in indirect diabetes cost (IDF, 2019). Reduced access to diagnostic testing for the undiagnosed and access to medication to diagnosed diabetes patients are some of the main reasons for delayed diagnosis and prevented timely control.

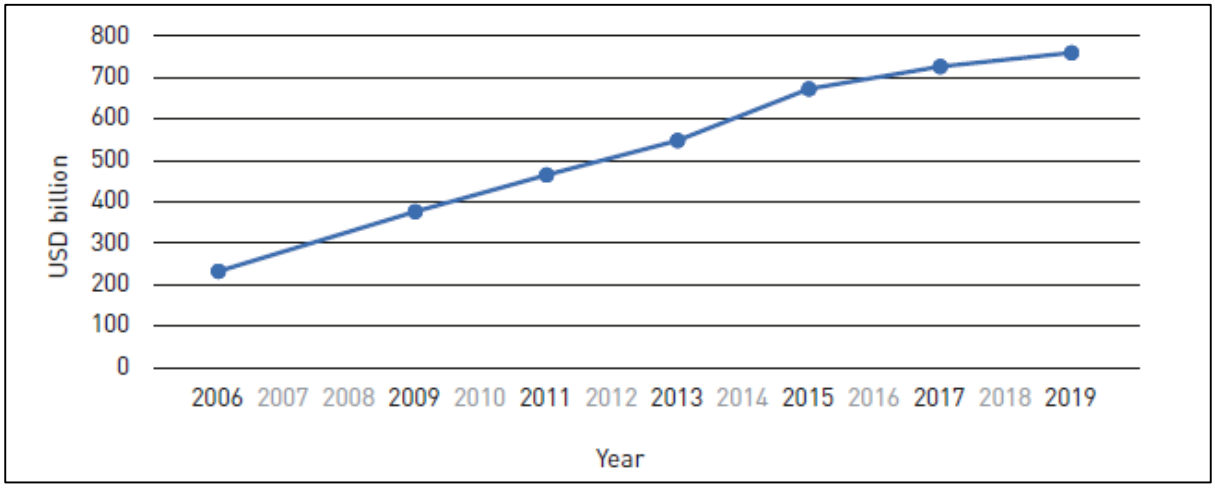

Figure 3. Total Diabetes-Related Health Expenditure for Adults (20-79) with Diabetes (IDF, 2019, p.56)

Globally, it is estimated that direct diabetes costs were estimated at USD 760 billion in 2019 according to IDF (2019) projection as seen in figure 3. This estimate is expected to rise to USD 845 billion by 2045. Diabetes direct costs in MENA region constitute 15.2% of the total health expenditure, making it the second highest regions in diabetes direct cost ration to total health expenditure according to IDF (2019). The UAE spends also significant amount from the health expenditure on diabetes, which accounts for about mean expenditure of AED 5,282 per person with diabetes. While total healthcare expenditure on diabetes in the UAE in 2017 is estimated at AED 6.3 billion. These numbers are forecasted to rise consistent with the dramatic increase in diabetes prevalence in MENA region in the coming few decades. This in turn calls for urgent intervention in the form of novel solutions to reduce progression to such advanced stages of the disease and burden the healthcare in the country considering the unsustainable healthcare resources and the pressure that can place on the country's Gross

Domestic Product (GDP) as evident in figure 5. Another complicating factor for the proper planning of human and physical healthcare resources, is the rapid increase in the UAE population, which in many cases is affected by economic factors and not natural growth from child birth as per figure 4 . This makes the anticipation of such surges in healthcare demands, hard to predict and plan accordingly for.

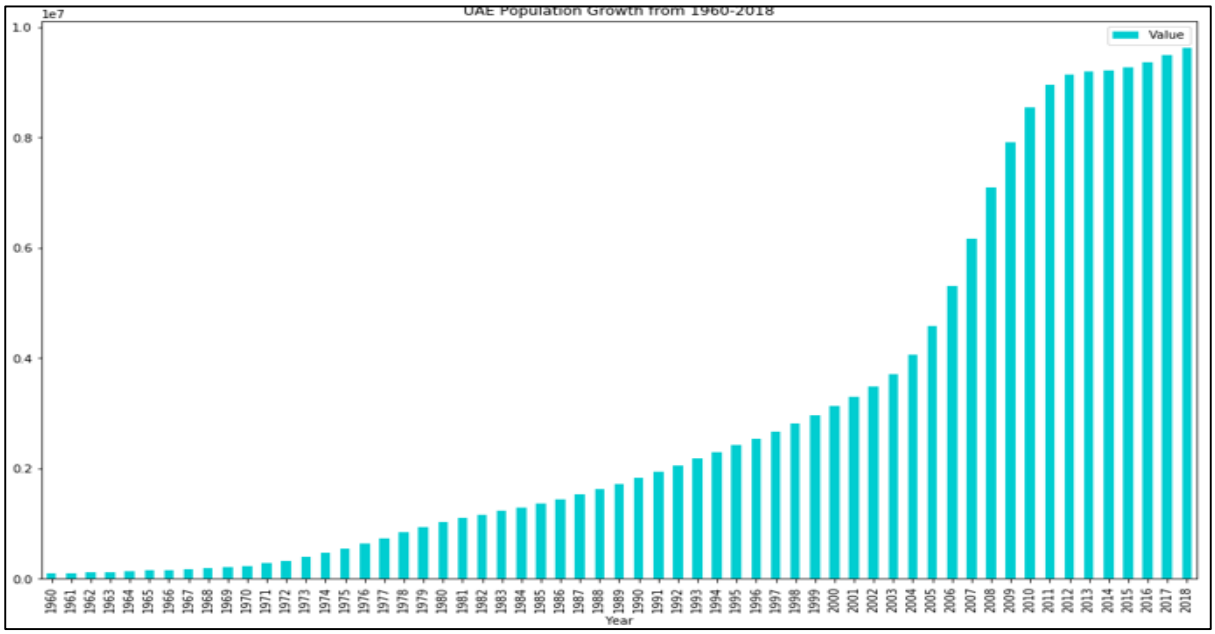

Figure 4. Population Growth of the UAE 1960 to 2018 (based on data from OCHA, 2020)

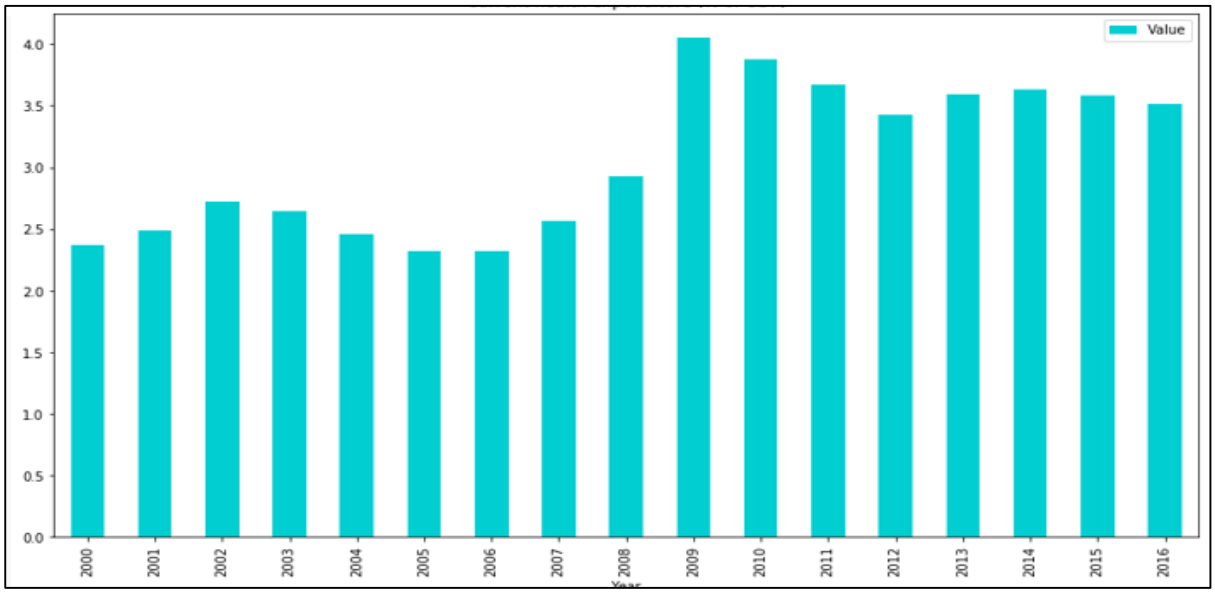

Figure 5. UAE Health Expenditure as % of the GDP (based on data from OCHA, 2020)

1.2.4. **Clinical Diabetes Classification and Diagnostic Research**

Among the many medical functions researched, medical diagnosis is a main focus of a wide spectrum of research. Medical diagnosis is meant to be performed by a qualified physician, based on observed and measured set of symptoms and results to determine what best explains these symptoms. This diagnosis is based on extensive knowledge and experience that must exist in the treating physician to infer workable and treatment decisions that will improve the patient condition (Gunčaret al., 2018). Access to such expertise and means of assessment may greatly differ in availability from one patient to another. Although clinical guidance for the classification and diagnosis through laboratory tests is established as seen in figure 6, there is yet a margin for these tests and measurements to not be as accurate as they should be. Diabetes and ability to predict or early detect it is hindered by the complexity surrounding its risk factors and environmental influencing variables.

The main diagnostic tests for diabetes are Fasting Plasma Glucose (FPG), 75-g Oral Glucose Tolerance Test (OGTT), Two-Hour plasma glucose (2-h PG) and HbA1c. American Diabetes Association (2020b) contends that the mentioned diagnostic tests can provide a fairly accurate diagnosis with HbA1c being superior to the others in many cases, it is not without drawbacks. Factors such as the need for strict adherence to National Glycohemoglobin Standardization Program (NGSP) compliant method to ensure the accuracy of the test can reduce accessibility to the test. Additionally, HbA1c cannot be considered as much as valid is as a single way to test for the disease because it represents the average glycaemia across the past months. So, it is recommended that the treating physician consider and typically when results are close to thresholds, a combination of the tests, also increasing the cost and physical burden on the patient. Other reasons such as the possibility to misdiagnose T1DM as T2DM and vice versa are common and require further research in a more comprehensive profiling of

the patient several personal and clinical variables. Other factors such as ethnicity, ESRD or even pregnancy can distort the results due to the changes it may cause to the body and hormones.

This complexity of the disease can result in patients not exhibiting symptoms and the inability to understand and specify the time of the disease onset, further exposes the population to the risk of progressing through the disease stages and developing the considerably debilitating and deadly complications (IDF, 2019). Hence, the need for less invasive informal tools, such as ML diagnostic systems are a very viable solution to such restrictions and obstacles to diagnostics and classification through traditional clinical means (Jayanthi et al., 2017).

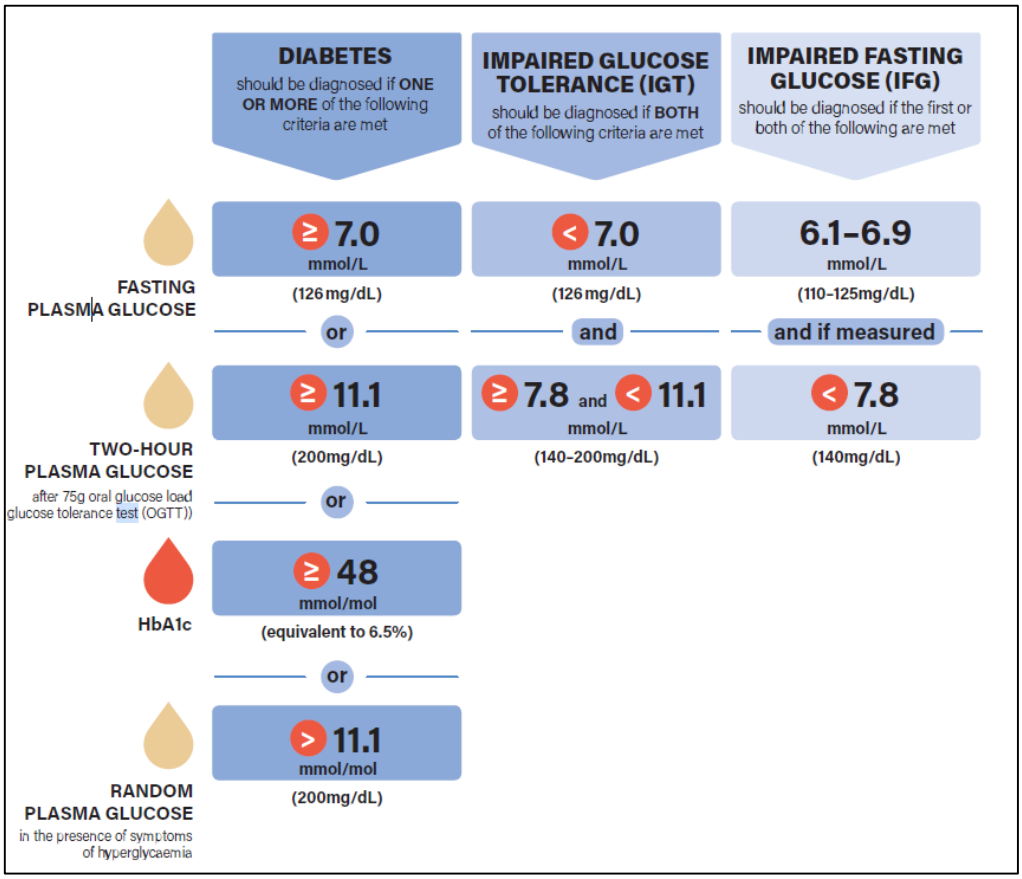

Figure 6. Modified Diagnostic Criteria for Diabetes (IDF, 2019, p.12)

1.3. **Problem Statement**

Having defined the major impacts of diabetes, which are detrimental to personal and public health in the previous sections, it is clear that novel, accessible and cheaper informal ways for diagnosis and classification of diabetes patients are greatly necessary to be researched. To facilitate this, ML is a great candidate to such use case due to its proven effectiveness in delivering highly learned, expert diagnosis systems based on the unique predictive features for the particular disease or clinical outcome. Thus, this work will attempt to propose a ML based binary classification for diabetes risk framework, based on learning from the UAE 2017-2018 National Health Survey (NHS). The complexity surrounding diabetes pathogenesis and risk of being undiagnosed, makes it greatly relevant to study the most important predictive features that influence the classification of people according to risk of diabetes.

1.4. **Contribution**

This research will attempt to deliver an empirical and highly validated detailed framework to source data from STEPS style national health surveys, identify most valid target variables for the clinical surveillance intended and build and optimize a ML binary classification model suitable for the identification of increased risk of diabetes among the UAE population. The research will demonstrate rigorous examination of more than 200 predicting variables based on the UAE NHS and will use state of the art algorithm based techniques to ensure improved features selection and dimensionality reduction of the final models. The models performance evaluation will consider leading metrics suitable to measure the performance of such models, and will further attempt to verify the final models using training/test split techniques and other methods for improved validation. Such empirical techniques are further verified by a domain expert endocrinologist, to make sure that the research results are relevant and clinically validated.

1.5. Dissertation Organization

The research attempts to provide a comprehensive overview of the design and building of a diagnostic ML model to classify people at risk of diabetes. It considers being as detailed and informative about the motivation behind such novel investigation and how such work can contribute in the overall health and wellbeing of the country and generally anywhere in the world. The dissertation is logically structured in 6 chapters as per the following:

1.5.1. Chapter 1. Introduction

The first chapter provides a logical and concise background about the researched question about why does T2DM requires a more generalized and a wide scope data-based ML assessment tools. The non-trivial nature of the motivation is illustrated in this chapter using data from leading literature in the field. The chapter also defines the problem considered by this research and the novel contribution and results validation of the outcomes produced.

1.5.2. Chapter 2. Literature Review

Chapter 2 will go more in depth in specific research about ML based diagnostic system and more specifically, ML research in diabetes assessment, diagnosis, classification and surveillance. The researched literature will be analysed based on the motivation behind it, data used methods and techniques and results. The research will consider all types of diabetes to be as inclusive to techniques and approaches as possible and account for the possibility of extending any mentioned research to another type of diabetes or another clinical outcome.

1.5.3. Chapter 3. Data Sourcing, Structure and Ethics

This chapter will go into detail about the sourcing of the data, measures taken to access and load the data and the major pre-processing activities done to ensure the data readiness for the ML modelling process. A general overview will be given about the STEPS or STEPwise

surveillance approach for NCDs and how it is relevant to the UAE NHS, which the source of data in this research.

1.5.4. Chapter 4. Technology Used in the Research

Chapter 4 will discuss the main ML technology used in the study including the algorithms considered in the research, programming languages and software packages used and the hardware the experiments were performed on.

1.5.5. Chapter 5. Proposed Automated T2DM Diagnosis Machine Learning Based Framework

This chapter will discuss the proposed empirical framework that will comprise of the major activities from sourcing and loading the data, through pre-processing and eventually to feature selection and performance evaluation. The process will suggest a ML pipeline that will start by experimenting with various estimators and algorithms and end up with suggesting the best performing model and features for clinical setting deployment or further extending to other clinical outcomes. Based on the developed framework and constructed ML pipeline, all the resulting models will be listed, described and evaluated, with each model being presented with detailed feature set and results across the approved metrics and validation methods. Extensions and future work will be highlighted in this chapter too.

1.5.6. Chapter 6. Limitations and Conclusion

This chapter will bring the work performed in the previous 2 chapters together and will attempt to finalize the proposed framework based on the experiments results. Possible extensions that can use the methodology in this research or scale up the usage of the best performing models in other clinical targets. Finally, limitation that were observed from the various aspects of the study will be discussed and measures to improve on them for future work will be listed where applicable. The main findings based on the detailed motivation and

problem statements will be highlighted with close up remarks based on the main results and discussion points in the previous sections.

Chapter 2. Literature Review

2.1. Preface

ML research and applications have been exponentially growing in the past decades with highly noticeable progress in the past 10 years (Gunčaret al., 2018). Many factors can be attributed to this progress, but few main advances on other fronts, proved to be essential for the field to advance and grow from theory and limited proof of concept application. The proliferation of big data in almost every field in life is a major driver for such progress as data is the cornerstone in any potentially useful learned model regardless of the field it is applied to. The adoption of information systems across all aspects of life and the capture of almost every event pertaining to the users and processes of these field made data and more specifically learning from data by ML systems as a highly viable field for real life implementations. Further to this, current machines are highly capable to handle many of previously impossible data processing tasks without access to very powerful computers or even supercomputers. Current ML tools and packages are optimized to make the best usage of local or distributed computing tasks. Even basic machines with reasonable memory and processing power can handle very serious application of ML with possibly millions of examples and features. The reduction in computing expense, is a very important enabler in the current progress of the field. Such concepts as the extreme influx of data in rapid velocity and differing variety, is highly evident in the healthcare sector due to the advent of EMR as the main enabler of provision of timely and greatly improved quality of care.

ML in healthcare has been extensively researched due to the aforementioned enablers across the world in almost all fields of medical care with existing viable data. Data coming from healthcare systems varies greatly and this variety can also greatly impact the methodology and tools used to make non-trivial AI a very serious consideration for that specific task. Data

from EMR can be assigned to two main categories: structured data and unstructured data. Structured data is mainly related to tabular structured data that can be assigned to clear categories or label and stored in tables under a clear data type. While unstructured data, can be considered as any free text that is related to the EMR of a patient and cannot be stored easily under a clearly defined data type. Both, types are considered as a very rich source for ML modelling data that can handle complex predictive and classification use cases.

In the past years unstructured data saw great focus due to the advancement of Natural Language Processing (NLP) as a very powerful domain to source rich predictive variables from clinical free text in many languages besides English (AlShuweihi et al., 2020). NLP in bioinformatics and ML learning is highly relevant to extracting model features from EMRs clinical notes and reports (Gronsbell et al., 2019). NLP most prominent objectives in bioinformatics are text based diagnostic predictive models (Roch et al., 2015), identification of clinical risk and prognosis (Doan et al., 2016) and symptoms surveillance (Tvardik et al., 2018). Other types of data, which were once thought of as not viable for accessible ML use, are now highly viable for such use (Liu et al., 2019; Wang et al., 2018). Imaging data, streamed monitoring data and many other forms of streamed and static data are now widely researched for non-trivial healthcare outcomes improvement use cases (Ker et al., 2019; Lu and Chu, 2017).

However, EMRs are not the only source such rich health data. Health surveys and national surveillance is also considered as a viable research domain for supervised and unsupervised learning uses of ML. NHSs around the globe are considered as a rich cross sectional sources of advanced prevalence analysis (Wang et al., 2017) and predictive diagnostic ML modelling (Somasundaram & Alli, 2017). This research used the data from UAE STEPS

based NHS to produce learned binary classification model for prediction diabetes risk. The sourcing and details of this data source is going to be discussed in more detail in Chapter 3.

2.2. Literature Synthesis and Results Discussion

Research in ML and how it can support clinical outcomes for diabetes care guidelines in diagnosis, prevention, detection of risk factors and patients' cohort clustering among other novel and under-explored diabetes research opportunities. Research of ML and predictive modelling in the disease is wildly researched, and mostly targets existing healthcare outcomes issues such as increased cost of care, preventing complications and deaths, improve quality of care and reduce waste in both data in EMRs and other resources.

2.2.1. A Machine Learning-Based Framework to Identify Type 2 Diabetes through Electronic Health Records

2.2.1.1. Motivation and Study Design

Zheng et al. (2017) attempted research T2DM identification from EMRs with emphasis on improving recall and expansion of included patient cohort. Their research tries to avoid the restrictive criteria usually imposed in the sample selection process using genotype-phenotype associations. This research is proposing a data-driven framework to identify T2DM patients and control subjects from EMRs with minimal loss of research subjects. The sample considered for this research consist of 300 cases, with 161 confirmed T2DM patients, 60 control subjects and 79 undiagnosed subjects from regional EMR distributed repositories in the period from 2012 and 2014. Manual expert identification and labelling of samples has been used as a standard in similar studies, however this human expert involvement showed certain levels of limiting ability to widen the study cohorts. To ensure this, a three level criteria was put in place to select controls where subjects, who must satisfy at least two of these criteria: positive diabetes laboratory test results, diabetes prescribed medication and a diagnosis of diabetes.

2.2.1.2. **Discussion, Results and Limitations**

The framework suggests constructing the ML model feature set from the EMRs. These features include data engineered from demographic details, diagnosis information, prescription records and laboratory results. The frame work also, attempts to improve the feature engineering through dimensionality reduction and summarization of the feature set over three levels, with the first set consists of 107 features and the second of 33 and the smallest feature set consist of 5 features. Zheng et al. (2017) selected a number of wildly utilized and proved classifiers to measure the performance of the T2DM identifying framework. Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), k Nearest Neighbours (kNN), Support Vector Machine (SVM) and J48 are considered to be the classifiers to be evaluated for the model as presented in table 2. The majority of the models achieved significantly performance results than the baseline. Most of the classifiers with reduced dimensionality score above 90% accuracy while the average Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was around 98%. This demonstrates improved performance over the state of the art performance at AUC-ROC of 71%. However, this study shows still some of the issues exhibited by similar studies that the features design and engineering still requires human expert involvement. Additionally, the total number of subjects in the dataset if small compared to the average datasets used to produce similar clinical classifying models.

Classifiers	Feature Sets	Accuracy	Sensitivity	Specificity	Precision	AUC
Expert Algorithm	–	0.84	0.78	1.00	1.00	0.71
LR	**#107**	0.86 (0.06)	0.90 (0.09)	0.84 (0.10)	0.70 (0.11)	0.88 (0.07)
	#33	0.91 (0.04)	0.98 (0.03)	0.88 (0.06)	0.77 (0.07)	0.92 (0.03)
	#5	**0.99 (0.01)**	**1.00 (0)**	0.98 (0.01)	0.95 (0.03)	0.99 (0.01)
NB	**#107**	0.94 (0.05)	0.98 (0.03)	0.93 (0.07)	0.85 (0.11)	0.98 (0.02)
	#33	0.91 (0.07)	1.00 (0)	0.88 (0.10)	0.79 (0.15)	1.00 (0)
	#5	0.96 (0.03)	**1.00 (0)**	0.94 (0.05)	0.87 (0.09)	**1.00 (0)**
RF	**#107**	0.98 (0.01)	1.00 (0)	0.97 (0.02)	0.94 (0.05)	1.00 (0)
	#33	0.98 (0.01)	1.00 (0)	0.97 (0.02)	0.94 (0.05)	1.00 (0)
	#5	0.98 (0)	0.98 (0.03)	0.98 (0.01)	0.95 (0.03)	**1.00 (0)**
kNN	**#107**	0.83 (0.06)	0.87 (0.05)	0.81 (0.08)	0.65 (0.09)	0.91 (0.01)
	#33	0.94 (0.05)	0.98 (0.03)	0.92 (0.08)	0.84 (0.12)	0.98 (0.02)
	#5	0.97 (0.03)	**1.00 (0)**	0.96 (0.04)	0.90 (0.08)	0.99 (0.01)
SVM	**#107**	0.96 (0.04)	0.95 (0.03)	0.96 (0.04)	0.91 (0.10)	0.96 (0.03)
	#33	0.97 (0.02)	0.97 (0.04)	0.97 (0.02)	0.93 (0.06)	0.97 (0.02)
	#5	0.98 (0.01)	0.95 (0.03)	**0.99 (0.01)**	**0.98 (0.03)**	0.97 (0.02)
J48	**#107**	0.98 (0.02)	1.00 (0)	0.97 (0.02)	0.93 (0.05)	0.98 (0.01)
	#33	0.97 (0.02)	0.97 (0.04)	0.97 (0.02)	0.94 (0.05)	0.99 (0.01)
	#5	0.97 (0.03)	0.95 (0.03)	0.97 (0.03)	0.94 (0.07)	0.98 (0.03)

The bold values indicate the best models in terms of accuracy, sensitity, specificity, precision and AUC. The significance values are inappropriate for them.

Table 2. Comparison of different classifiers and the expert algorithm (baseline), measured by their average performance (and standard deviation) in cross-validation (Zheng et al., 2017, p.125)

2.2.2. Hypoglycaemic events Prediction and Prevention of T1DM Using Machine Learning

2.2.2.1. Motivation and Study Design

Glycaemic control of glucose levels in the blood are in the center of prevention and treatment of diabetes. Much of the research around blood glucose is associated with hyperglycaemia events, but Vehí et al. (2020) stipulated that hypoglycaemia events especially for T1DM patients can be predicted using ML modelling. Hypoglycaemia can be closely associated with intensive insulin therapy and increased intake of insulin by patients with severe insulin deficiency such as T1DM. T1DM patients have to tightly monitor their daily insulin intake that can be administered in two main ways: by manual administering Multiple Daily Injections (MDI) or through more automated Continuous Subcutaneous Insulin Infusion (CSII). Vehí et al. (2020) sought to utilize ML prediction and classification in the core of a four algorithm based framework to improve safety and glycaemic levels throughout the day and different scenarios as seen in table 3. The framework is supposed to utilize a dedicated algorithm to handle a different ML task. Vehí et al. (2020) proposes to use Grammatical Evolution (GE) to predict mid-term continuous of glycaemia levels, SVM to predict

hypoglycaemic events through postprandial periods, Artificial Neural Networks (ANN) to forecast hypoglycaemic events overnight and finally Data Mining (DM) to profile diabetes management scenarios.

Objective	Window	Methodology	Database
Continuous prediction	1 h	Grammatical evolution	1, 2, and 3
Postprandial prediction	4 h	Support vector machine	1
Nocturnal prediction	6 h	Artificial neural network	2
Daily profiles classification	24 h	Data mining	1 and 3

Table 3. Summary of objectives, databases, and algorithms of the models that compose the system (Vehí et al., 2020, p.5)

2.2.2.2. Discussion, Results and Limitations

The proposed system different ML tasks are trained on one or a combination of three databases of consented glycaemic events data. The performance of the models is evaluated using various metrics including accuracy, sensitivity, specificity and Matthews Correlation Coefficient (MCC). The systems scored an average accuracy of 86.1% for mid-term continuous prediction. It scored and average accuracy of 80.1% for the nocturnal hypoglycaemic events prediction. As a whole, the proposed predictive solution performed well, however not without limitations that may impact its generalizability. One of the main limitations of the constructed system is that employs heterogeneous and mixed databases for each model built. The interaction between these specific databases and the merged data can impact the ability to produce a predictive model without overfitting. However, such ML implementation is very important as decision support for the chronic patient of diabetes, who has to live with the disease for the reminder of their life is vital to the patient's safety and quality of life.

2.2.3. Predictive Models for DM using ML Using Gradient Boosting Machine and LR algorithms

2.2.3.1. Motivation and Study Design

Lai et al. (2019) motivated by the need to predict diabetes risk based on data from Canadian patients EMRs, researched constructing a predictive ML with high accuracy and sensitivity. Compared to Zheng et al. (2017), Lai et al. (2019) focused on constructing the model features based on demographic and laboratory results data, rather than an extensive feature set. The model is trained on 13,309 patients with records of 215,544 clinical visits. Lai et al. (2019) main predictive models are constructed using Gradient Boosting Machine (GBM) and LR algorithms. To account for cohort imbalance in the dataset between diagnosed and control subjects, both class weight and adjusted threshold methods were used. The performance of these models is also compared with RF and Decision Tree (DT) models using AUC-ROC and sensitivity metrics.

2.2.3.2. Discussion, Results and Limitations

Lai et al. (2019) models achieved well, with GBM scoring the highest AUC-ROC score of 84.7% predictive accuracy and sensitivity of 71.6%, while LR scoring relatively similarly at 84% AUC-ROC and sensitivity of 73.4% as in figure 7. The models were evaluated k10 fold cross validation. The diabetes predictive models by Lai et al. (2019) performed as well and some cases better than state of the art performance exhibited by some other similar models on the same cohort. Based on the performance of the models, they can be deployed as a core decision support tools in online diabetes detection applications. Considering the homogeneity of the data, this constructed models may exhibit best performance if tested on similar Canadian or North American population.

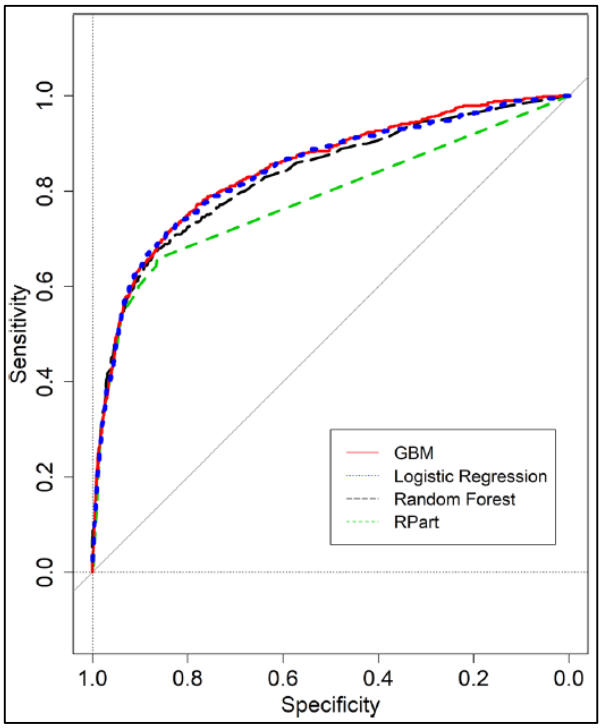

Figure 7. Information gain measure from diabetes classifiers (Lai et al., 2019, p.4)

2.2.4. An IoT-Based Glucose Level Monitoring System Using ANN

2.2.4.1. Motivation and Study Design

As previously discussed, diabetes monitoring and management is key to coexisting with the disease. However, patients in most cases have to endure invasive blood tests or capillary tests throughout the day to achieve adequate monitoring of blood glucose levels. To assist patients monitoring their glycaemia, while reducing the discomfort of undergoing invasive or costly tests, Alarcón-Paredes et al. (2019) presented a novel implementation of a combination of the Internet of Things (IoT) and ML data-driven predictive capability. The solution is built using a Raspberry Pi Zero microcomputer board, and a camera to capture and analyse fingertip images using Tensorflow. Tensorflow is a Python-based machine learning package that will be used to construct the ML model based on histograms derived from the fingertip images, transferred from the microcomputer. The model is then used on a mobile device to aid the patient in predicting and controlling glycaemic intake. The model is constructed using Artificial

Neural Network (ANN) based on data captured from 514 healthy participants. The training data used to train the ANN model consisted of 514 histograms, averaging at 12 histogram per subject. The ANN structure consists of 256 input nodes, two hidden layers consisting of 1024 nodes and a single output layer that predicts the glycaemic concentration level of the body. The hyperparameters of the model are configured as trained at 100 epochs using ADAM to minimize error.

2.2.4.2. **Discussion, Results and Limitations**

To evaluate the performance of the model, Clarke Error Grid (CEG) and the Mean Absolute Error (MAE) metrics are used. The model performance achieved 90.32% on CEG and 10.37 MAE, which indicate comparative performance to other work in the literature. This implementation showed that ML models can be impeded in real life IoT implementations and that existing open-source technology can support low cost and less invasive means for diabetes patient to manage the disease and improve safety. Having said that, such suggested wearable solution to be considered as a convenient alternative to current methods, must be refined and be less obstructive and bulky than the prototype proposed in this study. Such improvements are expected considering the improvement of microcomputers and sensors that can be considered as data sources for ML models.

2.2.5. **Comparing Gaussian Process Classification, Linear Discriminant Analysis and Quadratic Discriminant Analysis for Classification of DM Data**

2.2.5.1. **Motivation and Study Design**

Maniruzzaman et al. (2017) also identified the importance of classification of diabetes cases in the planning and management of the disease treatment and prevention efforts. But, they also considered the non-linearity and complexity of the most common predicting variables of diabetes. Such complexity and unclear relations between to covariates, necessitates utilizing

a complex technique that can accommodate according to it hyperparameters and customization of kernels to handle the specificity and complexity of diabetes classification scenarios. From existing ML literature for the classification of diabetes, Maniruzzaman et al. (2017) considered three classifiers and compared them with a novel implementation of Gaussian Process Classification (GPC). This classifier is experimented with across three different kernels to assess its performance in comparison with other classifiers. Polynomial, radial and linear kernels are chosen to act as the main kernels for GPC model proposed by Maniruzzaman et al. (2017). The performance of GPC is compared with Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and NB. Maniruzzaman et al. (2017) used the Pima Indian diabetes dataset, which one of the most researched in the field of ML applications on diabetes, which makes it a good choice for benchmarking results. The dataset includes 768 subjects' data, with 268 diabetes patients and 500 undiagnosed controls.

2.2.5.2. **Discussion, Results and Limitations**

To evaluate the GPC and other benchmarking models, Maniruzzaman et al. (2017) accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and AUC-ROC. The study also considered experimenting with different cross validation folds and reported the results for all the models measured across both k5 and k10 folds. The GPC model performed better than the other models tested with the configuration with the radial kernel being the highest performing configuration. GPC with radial kernel model achieved accuracy of 81.44% when validated using k5 validation and better performance at 81.97% when validated using k10 fold cross validation. LDA, QDA and NV achieved accuracies of 77.86%, 74.60% and 76.20% respectively when evaluated using k10 fold cross validation as reported in table 4. This shows that that the Gaussian Process proposed can account for fitting complex linear and non-linear functions better than other models that may

not have such complexity and traceability. The work by Maniruzzaman et al. (2017) introduced a new less experimented with algorithm and managed to demonstrate an improved performance compared to more traditional classifiers. However, the proposed model complexity can also contribute in its being less explainable and limit the ability for customization through selecting the most suitable kernel for the ML objective.

K	Classification models	Performance measure				
		ACC	SE	SP	PPV	NPV
5	LDA	77.20%	87.94%	51.94%	77.46%	64.00%
	QDA	74.60%	85.20%	49.44%	77.14%	61.54%
	NB	76.20%	83.69%	57.30%	77.27%	56.67%
	GPC	**81.44%**	**88.65%**	**62.04%**	**80.88%**	**67.86%**
10	LDA	77.86%	87.23%	50.00%	79.31%	57.89%
	QDA	76.56%	87.23%	46.67%	78.95%	55.00%
	NB	77.57%	85.11%	53.33%	80.00%	54.55%
	GPC	**81.97%**	**91.79%**	**63.33%**	**84.91%**	**62.50%**

Table 4. Diabetes classification across the evaluated models, using k5 and k10 folds cross validation (Maniruzzaman et al., 2017, p.31)

2.2.6. Microvascular Complications in T2DM: Retinopathy, Neuropathy and Nephropathy ML Prediction

2.2.6.1. Motivation and Study Design

Considering that complications from diabetes is the main detrimental factor to the quality of life and even mortality of a diabetes patient, researching ML-based diagnosis and prognosis of the onset of complications is a very worthwhile ML objective. As demonstrated in Chapter 1 of this research, diabetes and consequently the lack of proper and timely control of blood glycaemia is the leading cause for the advancing to such serious comorbidities. The research of the diagnosis and prediction of diabetes complications is mostly hindered by the general tendency for diabetes patient to be asymptomatic for an extended period at the early stages of the disease. Stemming from this urge to investigate the possibility to provide a highly

accurate prediction of the onset of diabetes complications, Dagliati et al. (2018) developed a ML prediction modelling pipeline to achieve this objective. Dagliati et al. (2018) aims to predict the onset of three of the most prominent and problematic microvascular complications in T2DM: retinopathy, neuropathy and nephropathy. The research is based on the data of 1000 patients from Italy's healthcare system. Considering that this research is crucial to the management of the complications progress, Dagliati et al. (2018) attempted constructing the predictive models to predict the onset of the complications during three temporal thresholds of 3, 5 and 7 years. Based on literature review and center profiling of the sources of the data, four classifiers were considered for the construction of the predictive models, LR, RF, SVM and NB. Two versions of each classifier were models, a version with normal sampling from the data and a version with a balanced sampling of the patients and control.

2.2.6.2. **Discussion, Results and Limitations**

Dagliati et al. (2018) research is significant in that it considers varied imputing, sampling and ML classifying strategies across different temporal thresholds. The resulting models were extensively evaluated using the most relevant metrics including accuracy, sensitivity, and specificity, PPV, NPV, MCC and AUC-ROC. In general, all the balanced versions of the models performed better AUC-ROC than the standard versions of the models except for the NB models were no considerable improvement was detected. Across the models for the different complications and the temporal thresholds for each complication prediction, there was no clear difference in the prediction accuracy. Certain models performed better the farthest the prediction window was, while others performed better for the nearest projection threshold as demonstrated in table 5. Considering the evaluation observations, Dagliati et al. (2018) proposed to consider the balanced LR model to predict the complications at the 3 years threshold. Dagliati et al. (2018) deduced also that prediction of the complications onset within

5 or 7 years do not provide viable clinical value due to the length of the period and impractical introduction of intervention for such period in the future.

To the credit of Dagliati et al. (2018), the decision to study the prediction of most prevalent T2DM complications using a methodical ML pipeline is very valuable and timely. The study considered evaluating the models systematically using Leave One Out Cross Validation (LOOCV) adding in increased layer of validity to the study. The study is also very critical in identifying most important risk factors for each complication through the usage of Nomograms for LR models at each of the temporal threshold suggested. The study may have been more inclusive if some additional variables were included in the models such as albumin-creatinine values to be considered in the modelling and indicators importance.

Retinopathy							
Year	Accuracy	Sensitivity	Specificity	PPV	NPV	MCC	AUC
3	0.777	0.820	0.730	0.771	0.785	0.552	0.808
5	0.743	0.790	0.685	0.758	0.723	0.478	0.769
7	0.666	0.606	0.745	0.760	0.587	0.348	0.726
Nephropathy							
3	0.647	0.652	0.642	0.680	0.613	0.293	0.701
5	0.693	0.750	0.616	0.723	0.649	0.368	0.734
7	0.686	0.714	0.643	0.750	0.600	0.353	0.721
Neuropathy							
3	0.746	0.783	0.707	0.743	0.750	0.490	0.799
5	0.680	0.667	0.697	0.725	0.635	0.362	0.714
7	0.727	0.688	0.780	0.807	0.652	0.463	0.769

Table 5. Complications prediction classifiers performance for balanced LR models (Dagliati et al., 2018, p.300)

2.2.7. Diabetic Retinopathy Early Detection Using Machine Learning Bagging Ensemble Classifier

2.2.7.1. Motivation and Study Design

Complementing Dagliati et al. (2018) research of the prediction of the onset of multiple types of diabetes related complications, Somasundaram and Alli (2017) elected to construct a model focusing of the prediction retinopathy. Retinopathy is a leading microvascular complication that can result from having diabetes. It develops due to prolonged periods of uncontrolled hyperglycaemia and hypertension, leading to significant damage to the retinal capillaries resulting in most cases in serious vision impairment and even blindness. The early detection of retinopathy is key in managing it and planning the best treatment to avoid progress to severe stages of the disease. This entails constant screening that can prove costly and time consuming to the patients. The main source of screening and predictive features for ML retinopathy classifiers are fundus images of the eye. However, these images still poses significant challenge in the construction for such predicting models due to the high dimensionality of the extracted features from these images. Somasundaram and Alli (2017) proposed an ensemble classifier that can handle the high dimensionality issue of the retina images and provide fast accurate classification performance. The model uses Machine Learning Bagging Ensemble Classifier (ML-BEC) and t-distributed Stochastic Neighbour Embedding (t-SNE) to enable a robust ML-based feature extraction and dimensionality reduction. The model is trained on the Standard Diabetic Retinopathy Database (DIARETDB1).

2.2.7.2. Discussion, Results and Limitations

The ML classification model proposed by Somasundaram and Alli (2017) is evaluated using several metrics including specificity, sensitivity, accuracy detection rate and classification time. Across the metrics, ML-BEC ensemble classifier to detect diabetic

retinopathy from retinal images, performed better than four other existing models as in figure 8. This generally proves that in most cases ensemble classifiers perform better than standalone classifiers. Having said that, the described pipeline of the model seems very specific to the task and requires further evaluation on other databases to ensure the consistency of the performance and avoidance of overfitting. Additionally, performance of the proposed models needs to be further benchmarked with state of the art performance of similar multivariate classifiers.

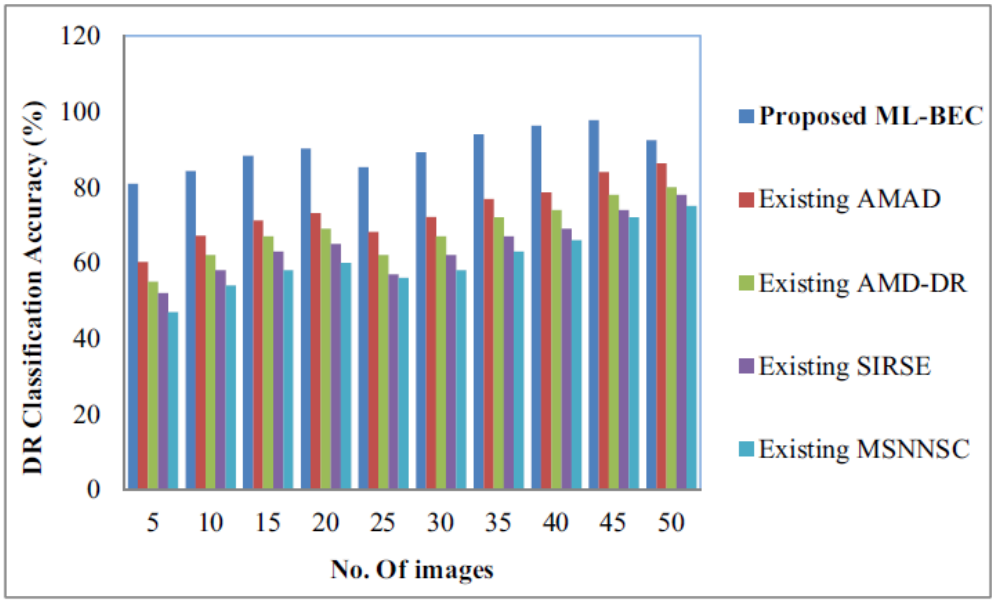

Figure 8. Model accuracy measured for the proposed ML-BEC, compared to other existing retinopathy classification models (Somasundaram & Alli, 2017, p.9)

2.2.8. Application of SVMs Modelling for Prediction of DM

2.2.8.1. Motivation and Study Design

Diabetes is a silent disease that can cause severe morbidity and damage to the person's health even before being diagnosed with it. Almost one in every two patients of diabetes are undiagnosed, greatly increasing their chances of contracting permanent comorbidities of the disease. To mitigate this research into the identification and classification of normal, individuals with prediabetes and diabetes patients. Classifying people as early as prediabetes indications are detected is vital in providing a timely management of the disease, increasing

chances of recovery and complications avoidance greatly. Yu et al. (2010) sought to research this through their proposed SVM based ML classifier. The main objective of the model is to identify patients with diabetes and individuals, who are in prediabetes stage or with no diabetes. Yu et al. (2010) proposed two models based on two classification schemes as in figure 6. The first scheme is designed to determine diagnosed or undiagnosed diabetes patients as a class and people with prediabetes or no diabetes. The other scheme is built to classify undiagnosed diabetes or prediabetes in a class and people with no diabetes in another class. The research is based on cross-sectional data from 1999-2004 National Health and Nutrition Examination Survey (NHANES), which the United States equivalent to national health survey. To determine the Subjects classification, standard diabetes classification guidelines were used, placing the dataset cases into four diabetes related target classifications: diagnosed diabetes, undiagnosed diabetes, prediabetes and no diabetes. The predictive features are almost similar for both models, with slight inclusion of additional variables for the second scheme.

2.2.8.2. **Discussion, Results and Limitations**

The models were evaluated using train/test split and k10 fold cross validation, where a test set was not used in the training and left to be used for the testing to avoid model bias. As observed in Maniruzzaman et al. (2017), k10 fold tend to be ideal and provide the best results k-fold cross validation is used. AUC-ROC, sensitivity, specificity, PPV and NPV were used to evaluate the classifiers. The models for both schemes were configured with four different kernel functions to determine the best performing kernel configuration for the classification objectives. Linear, polynomial, radial and sigmoid kernels were tested, with the radial kernel achieving the best for the scheme I at AUC-ROC of 83.47% and linear kernel performing the best for scheme II at 73.18% AUC-ROC. Based on observed performance, SVM performed acceptably for the two proposed classification schemes. Comparing the SVM model

performance with LR, the performance, did not result in significant difference, putting SVM as a very valid choice for similar binary classification objectives alongside prominent binary classification algorithms such as LR. Yu et al. (2010) is structured and generalizable and targets very important segment of diabetes patients, the prediabetes and undiagnosed patients. This work was integrated into an online predictive tool to measure the public risk of the disease, proving that such tools can be integrated into public or specialized bioinformatics solution as decision support systems. This research can also benefit from experimenting with other classifiers on the same target to evaluate and recommend the best estimator for the objectives.

Diagnostic category	Definition	N	Classification Scheme I	Classification Scheme II
Diagnosed diabetes	Answered "yes" to question "Have you ever been told by a doctor or health professionals that you had diabetes?"	1,266	Cases	Excluded from analysis
Undiagnosed diabetes	Answered "no" to question "Have you ever been told by a doctor or health professionals that you had diabetes?" AND Fasting plasma glucose level ≥ 126 mg/dl	195	Cases	Cases
Pre-diabetes	Fasting plasma glucose level 100-125 mg/dl	1,576	Non-cases	Cases
No diabetes	Fasting plasma glucose level <100 mg/dl	3,277	Non-cases	Non-cases

Notes: Total number of the cases for classification scheme I = 1461
Total number of the non-cases for classification scheme I = 4853
Total number of the cases for classification scheme II = 1709
Total number of the non-cases for classification scheme II = 3206

Table 6. Description of target classifications from NHANES data (Yu et al., 2010, p.3)

2.2.9. AI Based Prediction of Diabetic Nephropathy Progression

2.2.9.1. Motivation and Study Design

Another widely prevalent complication that affects diabetes patients, is Chronic Kidney Disease (CKD). Diabetes and hypertension and many cases a combination of both is responsible for about 80% of ESRD cases (IDF, 2019). Diabetic nephropathy is a leading cause of CKD among other causes such as, polyneuropathic bladder dysfunction, urinary tract infections, macrovascular angiopathy or hypertension. Diabetes patients are also 10 times more susceptible to develop ESRD. Considering the serious and debilitating impact on the wellbeing of diabetes patient stemming from developing diabetic nephropathy, early detection of this complication is key to treat and improve the quality of life of the patient. Preventing T2DM is

also another fundamental strategy to avoid CKD. Motivated by this, Makino et al. (2019) proposed a diabetic nephropathy ML predictive model based on NLP feature extraction from EMR and longitudinal data, to predict the complication aggravation overtime. The main contribution in Makino et al. (2019) work is the AI based feature extraction of structured data, text unstructured data and longitudinal data from the EMRs of 64,059 T2DM patients. The predictive model is build using 3,073 predictive features extracted using deep learning's Convolutional Neural Network (CNN). The Predictive ML model utilizes LR to predict the aggravation of the diabetic nephropathy within six months from a reference point.

2.2.9.2. **Discussion, Results and Limitations**

The ML diabetic nephrology model by Makino et al. (2019) is evaluated using k5 fold cross validation and the model. To avoid overfitting to fine tune the model performance, L2-regularization is also applied to the model. To achieve a better feature explanation, stepwise method is adopted. The predictive model classification is binary and predict the patient case as stable or aggravated. The constructed models are evaluated using accuracy and AUC-ROC. Seven different models were experimented with, differing in the level of breadth of features and their sources. The model with the most basic profile features achieved 54.8% accuracy and 56.2% AUC-ROC. While the model with complete feature set extracted from the various data sources in EMRs, score higher in both metrics at 70.1% accuracy and 74.3% AUC-ROC. This reflects that the novel and in depth CNN based feature extraction is correlated with improved predictive performance of target point future complication aggravation ML models. Such models are in the core of ML research that aims to increase early detection and prevention rates and therefore improve patients' health and quality of life. CKD and diabetic nephropathy in particular is causing substantial aggravation to the diabetes patient health can lead to hemodialysis and eventually to ESRD. Considering the expanded feature extraction process

and high dimensionality, the research had issues with obtaining uniform longitudinal data such as laboratory tests and medication. Another factor that need to be considered when evaluating the work in this study that the data that the model was based on is from a single medical source. Thus, to ensure a higher level of validation, replicating the model on data from other sources can give a better indication of how generalizable the model is.

2.2.10. Prediction of Nephropathy in T2DM: an Analysis of the ACCORD trial applying machine learning techniques

2.2.10.1. Motivation and Study Design

Determined to further understand the progression of diabetic nephropathy and the most important risk factors of the complication during multiple temporal thresholds, Rodriguez-Romero et al. (2019) comprehensively researched this using ML techniques. As stated earlier, predicting the onset of diabetes complications through the screening for identifying risk factors and biomarkers, is the main way to control and revert these complications. Diabetic nephropathy is extremely perilous illness with complicated pathogenesis and progression timeline that can vary widely based on the individual detection stage, intervention administered, biomarkers screened and many other physiological and clinical factors. Rodriguez-Romero et al. (2019) main contribution is highly important association between prediction window and the most important risk factor indicating the presence of diabetic nephropathy as per the study design in figure 9. This is evident in the study design that seeks to distribute risk factors based on importance across eight time prediction windows. The first two time windows consist of the first year first and second half of the year. The reminder of the windows consist of years two to seven. Based on this target, ML binary classification models are constructed based on data from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) study, which is sponsored by the Biologic Specimen and Data Repository

Information Coordinating Center of the National Heart, Lung, and Blood Institute (NHLBI). The dataset includes 10,251 adult subjects with cardiovascular events biomarkers.

The dataset subjects are classified with two labels according to the screening for the onset of diabetic nephrology. 6,777 subjects, who developed diabetic nephrology or 66% of the set and 3474, who were not diagnosed with nephrology. The data was separated into eight sets based on the prediction windows assessed, and Synthetic Minority Over-Sampling Technique (SMOTE) was used to handle class imbalance in the created sets. The classifying algorithms used and evaluated in the study are One Rule (1R), RF, J48 Decision Tree (J48), LR, Sequential Minimal Optimization (SMO), which a variation of SVM and NB. Finally, InfoGain method from Waikato Environment for Knowledge Analysis (WEKA), was used to assess the most important risk factors or features across the prediction windows.

2.2.10.2. **Discussion, Results and Limitations**

To evaluate the models' accuracy using AUC-ROC and sensitivity using k10 fold cross validation. Generally, the models achieved acceptable performance of AUC-ROC above 70%. 1R demonstrated a consistent performance across the different time window models of above 70% AUC-ROC. NB achieved better classifying performance at 83% AUC-ROC. RF achieved an AUC-ROC of about 100%, while LR scored above 81% AUC-ROC. Most of the classifiers including RF, LR, J48 and SMO performed very well during the early prediction windows, but LR achieved above average performance during the middle periods. The study is valuable and its scope is very relevant to the prediction and management of one of the most serious complications of T2DM. It highlights important interactions between the observed developed risk factors of diabetic nephropathy and the stages of nephropathy the patient is in. Nephropathy incidence in T2DM is characterized by five main stages, which are indicative of

the progression of the complication and its associated biomarkers Rodriguez-Romero et al. (2019).

Indications of first stage of nephropathy includes hypertrophy and renal hyperfiltration. The second stage is characterized by being mostly asymptomatic with no noticeable clinical signs of nephropathy. Stage three exhibits the onset of microalbuminuria, transitioning to macroalbuminuria in stage four. Severe worsening of renal functions, indicates progression to stage five of nephropathy and onset of ESRD, diminishing the chances of recovery tremendously. Uncovering the interactions of risk factors, prediction time windows and observed nephropathy stage are vital to introduce therapeutic intervention and possible revert the condition. The research indicated that low-density lipoprotein, urinary creatinine, urinary albumin, Glomerular Filtration Rate (GFR), potassium, cholesterol, and urinary albumin to creatinine ratio are the most significant risk factors for diabetic nephropathy over seven years period.

These findings are very crucial, and done highly methodically, utilizing many strategies to ensure that error is minimized and the findings are generalizable. This research can be expanded from binary classification of patients developing or not developing diabetic nephropathy to other classes that can also consider reverting the condition or stage of the complication. Another issue was that the dataset used consisted mostly of traditional biomarkers, which are commonly associated with diabetic nephropathy. This limited the ability of the research to find novel risk factors or interactions, which may not have been as obvious in the assessment of the progression of the disease before. A future prospective data collection of such data can facilitate this research further.

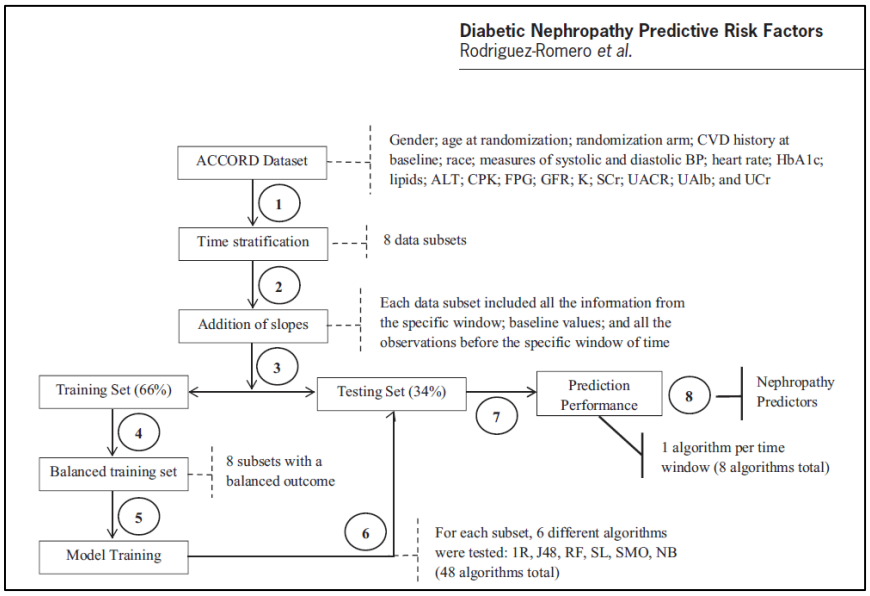

Figure 9. Model development of diabetic nephropathy risk factors based ML predictive model (Rodriguez-Romero et al., 2019, p.521)

2.2.11. Classwise k Nearest Neighbor for the Classification of DM

2.2.11.1. Motivation and Study Design

Real healthcare data is infamous for missing values. This can be attributed to many aspects including the complexity of individual cases, lack of care requirements standardization in the same facility or between different healthcare facilities, inadequate database design and data validation. This is a reality that should not allow data to be fully rejected for ML research purposes as it may be too valuable to be discarded, but new strategies and innovative ways to overcome such challenge must be studied further. Techniques to ensure handling missing values in data analytics varies and is mostly dictated by the target outcome of the research or the impact of records or variables on the overall value of the dataset. Christobel and Sivaprakasam (2013) attempted to propose a strategy to account for missing values in the Pima Indian diabetes dataset. As previously noted, the Pima dataset is frequently researched dataset for ML and statistical applications relevant to diabetes classification. However, the dataset is characterized by having multiple features with high rate of missing values. Christobel and Sivaprakasam (2013) proposed using mean substitution technique of data imputation to replace

missing values with the variable mean. Additionally, Christobel and Sivaprakasam (2013) resorted to using min/max scaling to normalize the data ranges to achieve a more uniform processing by the ML model. Finally, Christobel and Sivaprakasam (2013) proposed a variation kNN, called Class-wise k Nearest Neighbors (CkNN), which a variation of the popular ML classifier. kNN is one of the most popular ML algorithms used for prediction and classification and even clustering purposes. CkNN considers calculating a class-wise distance for each class label, assigning a class label to the lowest distance. Christobel and Sivaprakasam (2013).

2.2.11.2. **Discussion, Results and Limitations**

Christobel and Sivaprakasam (2013) research attempts to compare the classification performance of both the kNN and CkNN across various metrics. The metrics used to evaluate the models are sensitivity, specificity, accuracy, error and time. The models are evaluated using 10-fold cross validation to account for possibility of overfitting. Generally, CkNN proved to be the better model with higher results across all the metrics compared to kNN. CkNN scored 78.16% accuracy compared to 71.84% for kNN. The error rate also was reduced for CkNN at 21.84% compared 28.16% for kNN. The improvement achieved by introducing the data imputing and normalization combined with the variation to the kNN model is significant as seen in table 7. This work can benefit from testing the proposed techniques to other classifiers or hybrid models of multiple algorithms.

Algorithm	Sensitivity	Specificity	Accuracy	Error	Time
kNN	57.34	79.88	71.84	28.16	0.17
CkNN	61.84	87.38	78.16	21.84	0.19

Table 7. Comparative evaluation of the kNN and CkNN models (Christobel & Sivaprakasam, 2013, p.399)

2.3. Conclusion

Fairly Recent and divers of ML modelling for diabetes prediction and classification id reviewed. The emphasis in the selection of the reviewed work is to review as many varying use cases, which dictates the implementation of different algorithms, optimization techniques and data pre-processing strategies. The aim of the review was to identify the most viable algorithms and classifiers for the classification target of this study that was motivated by the factors discussed in chapter 1. The criteria used to select the methodologies to seek in this research was based on best performing classifiers for binary classification tasks and most explainable and responsive to optimization. Literature that based its modelling on various sources of data was reviewed including longitudinal EMR data (Makino et al., 2019), cross-sectional NHS data (Yu et al., 2010) and imgae data (Somasundaram & Alli, 2017).

Noticeably, certain classifiers reoccurred in the majority of the review literature, and were considered as state of the art classifiers in most of the cases. LR was predominant with best classifying performance in multiple studies and across different objectives (Dagliati et al., 2018; Makino et al., 2019; Zheng et al., 2017). Probabilistic classifiers such as NB were also considered with success (Dagliati et al., 2018; Rodriguez-Romero et al., 2019; Zheng et al.,

2017). SVM and its kernel-based variations demonstrated also viable predictive performance (Yu et al., 2010). kNN and kNN-based models such CkNN also exhibited good classifying potential (Christobel & Sivaprakasam, 2013). Other ensemble and neural based classification methods were used, but lacked the explainability of the aforementioned classifiers. Consequently, this research elected to apply LR, NB, SVMs and kNN as the main classifiers to predict the clinical objective of the study. Dimensionality reduction techniques will also be considered as a main aspect of the contribution in this study due to the demonstrated improvement on ML models performance (Zheng et al., 2017). The modelling of these classifiers will not be standalone, but as a part of a fully featured framework that will consider the aspects that will ensure the best input for the ML models (Rodriguez-Romero et al., 2019; Zheng et al., 2017).

Chapter 3. Data Sourcing, Structure and Ethics

The source and quality of clinical data is directly related to the efficacy of bioinformatics research. ML models' construction and performance is dictated by the data taxonomy and availability of relevant features for the ML objective sought to be achieved. The previous chapter provided varied novel sourcing of clinical relevant data. Clinical data can be sourced from two predominant sources, structured and unstructured EMR longitudinal data and cross-sectional NHS data. These two sources of data, which are the source of a very high rate of ML research in bioinformatics, come with their own advantages, disadvantages and challenges. However, technology and adequate data handling strategies can derive reasonable outcomes from any well curated set of clinical data. The following sections will cover in detail the data used in this research to build the ML predictive models to predict high risk of T2DM among UAE population, based on the 2017-2018 UAE Non-Communicable Disease Risk Factors STEPS National Health Survey (NCDs STEPS NHS).

3.1. Data Sourcing

This work is meant to study and evaluate the construction of robust validated diagnostic predictive models to predict the risk of T2DM based on UAE NCDs STEPS NHS. UAE NCDs STEPS NHS is a cross-sectional national survey based on the WHO STEPwise non-communicable disease risk factor surveillance approach (MOHAP, 2018a; MOHAP, 2018b). This type of national surveys can facilitate the surveillance and planning of community's major NCDs and their relevant risk factors. The STEPS approach to NHSs can allow for flexibility for national entities to explore adding locally relevant core and optional modules to support the national healthcare objectives. The UAE STEPS NHS was funded by MOHAP and with participation of the Federal Competitiveness & Statistics Authority (FCSA), Dubai Health Authority (DHA), Department of Health - Abu Dhabi, Dubai Statistics Centre (DSC), Abu

Dhabi Statistics Centre (ASC), WHO Regional Office for the Eastern Mediterranean (WHO EMRO).

The dataset is maintained by the Statistics and Research Centre (SARC) in MOHAP. To contribute in novel bioinformatics related research in the UAE, this study sought to research a local substantial dataset, with not much evident ML research on NCDs and T2DM in particular. To acquire the data, the research question and clinical and technical motivations were presented officially to the MOHAP Research Ethics Committee (REC). Upon fulfilling all the requirements that validate the aim and usage of the dataset, it was deemed that the research question and domain warrants access to the STEPS NHS dataset.

3.2. **Data Structure**

The STEPS NHS dataset consist of the data of 8,214 above 18 years old adults from 10,000 randomly selected households. The UAE STEPS NHS is a STEPwise survey, which is structured in a stepping through approach to questions, where the questionnaire branches out to questions specific to the responses of other questions. For instance, the questionnaire can ask the participant if they ever had a certain NCD screened by a professional healthcare worker and based on the answer can go further into the result of the screening or the date and duration of the result. The survey steps through three main levels or steps of data: questionnaire answers, physical measurements and biochemical measurements. This survey structure of data sourcing can provide principally a good representation of the current state of the participant with regard to general lifestyle and specific NCD risk factors. WHO support STEPS NHS as part of its commitment to guide global governments' policies, strategies and plans to control and prevent NCDs. Governments around the globe considered the prevention and management of NCDs as a priority in response to 2011 Political Declaration on NCDs. Additionally, due to the standardization of the core instrument of the STEPS survey and general guideline for

conducting it, it can serve as a worthwhile cross-national comparison and benchmarking tool between implementing nations.

The UAE is divided regionally into seven emirates: Abu Dhabi the capital, Dubai, Sharjah, Ajman, Ras Al Khaimah, Umm AL Quwain and Fujairah. 9,000 households were interviewed for the survey. The sampling was done on the emirate level into sample frames, and eventually into a unified national sample frame as per table 8. The survey was overseen by a panel of public health, epidemiology and statistics experts to ensure compliance to the standards governing national surveillance field surveys. The sampling design considered representing households and individuals from across the seven emirates, making the dataset representative of sociodemographic structures of these regions and the UAE as a whole. To account for the citizenship disparity among the population, stratified sampling was adopted when sampling from each emirate. Only non-institutional population were included in the STEPS NHS.

House Holds				
No. of Clusters	**Total**	**Non-Emirati**	**Emirati**	**Emirate**
300	3000	1800	1200	AUH
300	3000	1830	1170	DXB
146	1460	1010	450	SHJ
64	640	440	200	AJM
44	440	220	220	UAQ
86	860	430	430	RAK
60	600	240	360	FUJ
1000	10,000	5970	4030	TOTAL

Table 8. Sample size spread across the UAE emirates by cluster (MOHAP, 2018b, p.21)

The survey is conducted in randomized sampling of a single individual of a household to have for a face to face interview and another random member of the household to conduct a survey. The questionnaire main component is the behavioural assessment component, which consist of five main sections. Sociodemographic characteristics such as age, education and marital status. Work history and benefits, covers the respondent occupational status and reasons for working or not being employed. Risk factors and preventative health behaviour section, which contain questions about risk factors such as activity, diet and alcohol and tobacco consumption. Health state description, which assesses the individual perception of their health with regard to mobility, sleep, energy, etc. And finally, chronic conditions and health services coverage, which is mainly concerned with NCDs such as hypertension, hyperglycaemia, hypercholesterolemia and other chronic diseases related factors. The survey was conducted using a mobile version of the questionnaire in the form of Computer Assisted Personal Interview (CAPI).

This work considered deriving all variables that will be used to construct the ML models from the values assigned to each item in the STEPS NHS instrument. This includes responses to questions, physical measurement value and biomedical result. The dataset received was in tabular forms, where the questionnaire answers and values are encoded as columns and the participants records represented as rows. The dataset characteristics and observations will be detailed in section 3.4 of this study.

3.3. **Data Ethics and Privacy**

This research adhered to the highest standards of Good Clinical Practice (GCP) in the requesting, handling and conducting the research on the data of the human participants involved in the study. As part of the commitment to the handling of the researched data, the author conducted the necessary training on GCP and was certified by NIDA Clinical Trials

Network as a qualified researcher to handle human participants' clinical data. MOHAP REC is the review board that reviewed the research protocol and ensured that this research is in compliance with the necessary standards protecting the safety and confidentiality of the subjects in the dataset. This commitment to the protection and ethical handling of the researched data is evidenced in two main aspects. All human participants involved in the UAE STEPS NHS and included in the dataset signed consents authorising the use of their data for the intended research purposes with the survey owners. This consent is transferred to researchers, who conduct this research on the data, thus ensuring that this work is fully consented to by the involved participants. The second measure to protect the privacy of the participants' Protected Health Information (PHI) through full de-identification of personally identifiable data is done as per the standard governing the anonymization of the subjects' data. Furthermore, all participants in this research are adults and no direct interaction with the participants was attempted in anyway. Finally, this research is conducted without any conflict of interest or incentive to or from any party involved in the study.

Chapter 4. Technology Used in the Research

This study involves advanced research of ML supervised learning modelling that requires adequate knowledge and skill in specialized data science technology. This section will describe the algorithms and software ecosystem that was used to conduct the ML research throughout the proposed framework. Utilizing open-source packages and software is given a special attention in this study to illustrate that such involved and non-trivial research on ML applications can be done fully using open-source accessible tools. In the following a description of the ML algorithms used in the study, Programming language and software and hardware used to conduct all the stages described in this work.

4.1. Classification Algorithms

ML algorithms are in the core of any non-trivial application of predictive bioinformatics modelling. Thus, the understanding and proper selection of the most suitable algorithm to the clinical objective researched is crucial. Considering that the aim of this study is to construct a ML T2DM classification models, the classifiers with most relevant and best classifying performance in the reviewed literature were considered in this research. Due to the certain characteristics mostly related to the high dimensionality of the dataset, certain classifiers such as DT and RF were not included in the final list of tested ML algorithms.

4.1.1. Logistic Regression (LR)

LR models are statistical linear classification models, which are based linear regression, but forces the function at zero instead of returning the sum of the weighted values to predict continuous values (Müller & Guido, 2016). LR models uses logistic sigmoid function to force the predicted outcome to be between one and zero, which makes them capable and in many cases ideal for binary classification objectives. LR models although classically proved to perform very well in binary classification tasks, they can be extended to achieve multi-class

classification. LR models separates data points by a line or plane that classify the points into a predicted class. LR can handle ML binary classification objectives highly adequately, making them ideal for many diagnostic tasks, where the expected outcome is the prediction of the onset of a disease or not. Linear models are fast to train and classify and are inexpensive computationally. This makes LR a viable option for large datasets. Also, linear models including LR are easily interpretable and the performance can be more explainable than other algorithms.

4.1.2. **Support Vector Machines (SVMs)**

SVMs is a linear maximum margin algorithm that can perform wide ML tasks such as classification and regression in addition to density estimation and other ML uses (Sammut & Webb, 2017). SVMs models are constructed to find the hyperplane in high dimension feature spaces. It can classify binary and multi class dependent variables through achieving the maximum margin between the predicted classes' points. SVMs can be customized to use specific kernels for improved performance based on the ML objective. The use of such kernels as linear, polynomial, radial and sigmoid kernels were demonstrated to perform differently, improving the generalized SVMs performance depending on the task handled (Yu et al., 2010).

4.1.3. **k Nearest Neighbours (kNN)**

kNN is an instance or lazy learning based classification algorithm and is considered one of the most interpretable and simple to construct algorithms among ML classifiers (Müller & Guido, 2016). This simplicity stems from the way the model attempts to classify points based on the labels assigned to the closest points. The number of close points is determined by the k value assigned to the model, hence the name kNN. The point label is decided by majority vote of the classification of the closest points. The model can be used for binary and multi-class classification. The simplicity and interpretability of the kNN algorithm makes a suitable option

for clinical classification use cases, hence its selection to be one of the algorithms to be tested on the STEPS NHS dataset.

4.1.4. **Naïve Bayes (NB)**

NB is a probabilistic algorithm that is based on Bayes theorem that assumes independence between predictive features (Sneha and Gangil, 2019). NB models tend to be faster computationally to be trained than linear models, but this comes at the expense of being relatively less accurate (Müller & Guido, 2016). NB efficiency is a result of the way it learns parameters through performing individual per class statistics for each predictive variable. NB has three main variations that can perform differently and with varying levels of accuracy based on the predicted target type. These variations are Gaussian NB, Bernoulli NB, and Multinomial NB, which can function optimally for different types of predictive targets such as continuous data prediction, binary classification and text based classification. The robustness of NB models can also contribute in considering it for clinical diagnostic and risk factor prediction ML tasks.

4.2. **Software**

This research aims to perform advanced non-trivial ML and data analysis tasks on a considerable dataset. To achieve this, Python programming language is selected as the main language for all data science operations that will be performed throughout the ML pipeline to construct the STEPS NHS predictive model. Python and its robust data science ecosystem of highly capable open-source libraries is becoming one of the leading choices in general data science tasks and especially in ML operations. The following sections will depict the main libraries used in this work to achieve the intended purposes. Microsoft Excel was used for additional data review purposes.

4.2.1. **Python – 3.6**

Python is a general purpose interpreted programming language (Rossum, 2020). It has been created with code readability and ease of use being primary to its design ethos. Python garnered substantial attention and following by the data science community for combining its scripting general purpose code with domain specific data science languages such MATLAB and R (Müller & Guido, 2016). Python can handle all ML tasks highly adequately at varying levels of complexity and computation demand. Packages specific for data loading, pre-processing data, ML modelling, visualization and constructing complete ML pipelines. To conduct this research, python 3.6 and all the used libraries were installed on a dedicated virtual environment.

4.2.2. **Scikit-Learn – 0.22.1**

Scikit-learn is one of the leading ML python libraries, only rivalled in the past few years by TensorFlow (Müller & Guido, 2016). Scikit-learn is an open source library that offers state of the art ML algorithms and advanced data and model preparation functionalities and methods. Scikit-learn is used by major around the globe such as J. P. Morgan, Spotify, Evernote, Booking.com and many more (Sklearn, 2020). Scikit-learn is used to construct, fit and evaluate all the models in this research.

4.2.3. **Pandas – 0.25.1**

Pandas is another Python open source library, which pivotal to any serious data preparation, pre-processing and wrangling before initiating ML modelling activities. Pandas main data representation object is the DataFrame, which can thought of as a table object to hand tabular data in a similar manner as spreadsheets (McKinney, 2012). It can handle almost all kinds of data types individually for each column. Pandas is the main tool used in this research to load the original data and conduct almost all of the data exploration operations. Much of the data structure changes such as expanding variables, adding calculated variables,

changing data types and removing columns and rows among much more highly involved wrangling operations.

4.2.4. **Jupyter Notebook – 6.0.1**

Jupyter Notebook is an extremely robust and convenient open source web based tool to conduct the processing and modelling in the browser. The interactive environment that can handle Python code, images, visualizations, Hypertext Markup Language (HTML) and text. All the code in this research is documented in a Jupyter notebook, creating a single source for editing, reviewing revising and publishing of the research work done on the data.

4.2.5. **Matplotlib – 3.1.1 and Seaborn – 0.9.0**

Matplotlib and Seaborn are the primary scientific plotting packages for Python and are used in this research for exploratory plotting of the data in various forms. Scatterplots, normal distributions and histograms were created using these packages throughout the work in this research. Both libraries are open source and their visualization is directly integrated in Jupyter Notebook for this research purposes.

4.3. **Hardware**

This research ML development is done on a Windows 10 machine with the following specifications:

Processor: Intel® Core™ i7-8550U CPU @ 1.80GHz 1.99 GHz

Installed RAM: 16.0 GB

Storage: 1 TB HDD

System Type: 64-bit Operating System x64-based processor

Chapter 5. Proposed Automated T2DM Diagnosis Machine Learning Based Framework

As conclusively demonstrated in the past chapters, ML is becoming increasingly a cornerstone in the development and research of AI based models, designed to handle varied and involved clinical diagnostic and predictive tasks. In the core of any biomedical ML solution, are frameworks that encompass the main activates to construct non-trivial models, from the decision about the clinical target being researched, all the way through the evaluation and deployment of the constructed models. In this chapter, a full description of a proposed ML framework for clinical prediction based on independent and dependent variables in a STEPS type health survey. The framework will progress through the model design and ending with proposing the best model for the task at hand. The following is a detailed description of the activities involved in building the proposed framework. Figure 10 gives a full overview of the contributed T2DM binary classification framework in this research.

1. **Specifying the T2DM Diagnostic Clinical Target for the ML Research**
    a. National Epidemiological Targets and Research
    b. Global Epidemiological Targets and Research
2. **Sourcing Relevant Predictive Data (This Study used United Arab Emirates STEPS-based National Health Survey Data)**
    a. Longitudinal: Electronic Medical Records
    b. Cross-Sectional: National Health Surveys
3. **Data Loading and Pre-processing**
    a. Data Decoding
    b. Data Cleansing
    c. Data Imputing
    d. One Hot Encoding
4. **Data Exploration and Statistical Summarization**
    a. Crosstab Summarization
    b. Normal Distribution Plotting
    c. Scatterplot Correlative Analysis
    d. Histogram Multilevel Analysis
5. **ML T2DM Binary Classification Diagnostic Model Construction**
    a. Determining the Dependent (Target) Variable
    b. Select the Classification Algorithms
        i. Logistic Regression (LR)
        ii. Naïve Bayes (NB)
        iii. k Nearest Neighbors (kNN)
        iv. Support Vector Machines (SVMs)
    c. Apply Dimensionality Reduction Techniques
        i. Chi Squared (CS) Feature Selection
        ii. Recursive Feature Elimination (RFE) Feature Selection
        iii. CS/RFE Intersection Feature Selection
    d. Evaluate the ML Models using Prediction Performance Metrics
        i. Accuracy
        ii. F1-Score
        iii. Sensitivity
        iv. Precision
    e. Rank the ML models Based on Performance
    f. Extend the Best Performing Models to Other Diagnostic Clinical Support Systems
        i. Informal T2DM Risk Assessment Tools
        ii. Clinical Decision Support Systems
    g. Research Most Important Features for T2DM Risk and Interrelations With Clinical Domain Experts

Figure 10. Machine Learning (ML) Type 2 Diabetes Mellitus (T2DM) Diagnostic Binary Classification Model based on Progressive Dimensionality Reduction of Feature Sets Frame Work

5.1. Specifying the NCD Clinical Target for the Biomedical ML Research

Selection of a non-trivial, timely and epidemiologically relevant ML research question is the main target of this research. To achieve this, considerable research was done to identify most pressing health objectives in the country and globally and found out as demonstrated in section 1.2 that NCDs and diabetes in particular are leading causes of morbidity and economic burden on healthcare systems. This led to considering the research of the viability of ML to classify T2DM patients based on cross-sectional data of a representative sample of the UAE society. To fulfil this clinical target, this study used the UAE STEPS NHS data as received from MOHAP to construct binary classifying models to target this research question.

5.2. Data Loading and Preparation

5.2.1. Data Loading

The STEPS NHS received data was in CSV format as sourced in detail in chapter 3. The dataset was loaded and converted into a DataFrame in Pandas. The DataFrame is inspected for initial data completeness purposes and checked for shape, which showed that the dataset has 8,120 participants' records as rows and 204 survey items responses as columns. All the columns are identified by the survey code for that particular item and all the responses are encoded too. This encoding can be beneficial for statistical analysis of the data, but it causes a barrier to clear data understanding and exploration. Also, the dataset columns contain many documentation and administrative columns that require further management.

5.2.2. Data preparation, Cleansing and Pre-Processing

Data preparation for initial understanding of the features in the dataset is indispensable and is the foundation for further more technical steps such as data visualization and model features pre-processing. Cleansing of the STEPS NHS data started with decoding all the dataset

variables by changing all the columns names to a title representative of the survey item data as demonstrated in the sample in table 9.

Encoded Title	Decoded Title
A1000_1	time_h
A1000_2	time_m
A1001 (C1)	interviewer_1male_2female
A1002 (C2a)	dob_d
A1002 (C2b)	dob_m
A1002 (C2c)	dob_y

Table 9. Sample of the STEPS NHS Data Variables Names Decoding Process

The next step involved decoding the numeral codes of the responses of all the categorical questions in the survey as per the sample in table 10 for the highest education question.

Ecoded Values	Decoded Values
0	never_educated
1	preprimary_education
2	primary_education
3	lower_secondry_education
4	upper_secondry_education
5	post_secondary_non_tertiary_education
6	short_cycle_tertiary_education
7	bachelors_or_equivalent_level
8	masters_or_equivalent_level
9	doctoral_or_equivalent_level
10	not_elsewhere_classified

Table 10. Sample of the Decoding Process of Categorical Questions in the Survey

After fully decoding the dataset to representative titles and values, defining the initial feature set is made much more convenient and clear. The definition of the initial dataset for data exploration involves the removal of all columns that does not provide any informational or statistical value. These variables include survey administrative variables such as the survey personnel details such as the interviewer gender. Date and time data related to the survey

administration are removed too. All data about the devices and tools used in the survey were also omitted from the analyzed dataset. Any columns, which sum up or aggregate available values in the survey are considered redundant and also removed. Any unidentified columns with no definition in the questionnaire key were removed.

The dataset was reviewed manually to assess the variables that are very specific and if another more generalized variable is available and can be a better predictive feature. An example of such features the existence of variables that measure the frequency of physical activity measured by days, hours and minutes. Having a very wide spread distribution of number of minutes is of limited predictive value compared to number of days. The number of days as predictive feature will be much more less spread and potentially will result in a more significant importance. Another issue with the dataset, is the inclusion of an additional column for the same variable only to contain "don’t know" responses. All these columns were removed also for being redundant.

ML models expect a uniform numerical inputs to be able to demonstrate stable and accurate performance. To achieve this all categorical variable were expanded to per value dummies variables using Scikit-learn's One Hot Encoder (OHE) pre-processing. This results in adding a binary column contain a one or zero based for all the options included in the original column. These binary matrices enable ML models to deal with discrete variable much more predictably. Consequently and after adding the new dummy variables, all the original categorical variables were removed.

Additional variables were added to the dataset based on calculated and summarized values of existing columns. Examples of this include, condensing the three diastolic and systolic blood pressure reading into a single value for each by taking the mean of the three

values. The same is applied to the other measurements, which were taken more than once such as the pulse reading of the participant. Other continuous measurements were summarized into new variables and then discretized into more clinically relevant values such deriving the Body Mass Index (BMI) from the weight and height measurements.

One of the most important pre-processing tasks is dealing with missing values and this issue is highly prevalent in generally in health surveillance surveys, but it is very pronounced in the STEPS approach to health survey. This is mostly due to the stepping in and hierarchical nature of the tool, where a whole section of the questions or measurements can be skipped based on an answer to a certain question. This research considered this very serious issue with the data and how it can impact the final model and ability to have a relevant feature set for ML modelling. There were no initial removal of any variable due to missing values, but based on the selected dependent variable, records were dropped during the modelling process, but not during the data exploration. To account of much of the missing data, missing values were substituted with zero during the OHE process. During the modelling process, all records with missing values with no imputing substitution were dropped.

All the above cleansing and pre-processing steps in addition to manual review of a domain expert consultant endocrinologist resulted in the final ML ready feature set of 212 features to be used in the modelling. This level of dimensionality is going to be experimented with during the final model fine-tuning to improve the overall performance through dimensionality reduction strategies.

5.3. **Data Exploration and Statistical Summarization**

Considering the significance of the UAE STEPS NHS dataset, extensive data exploration was performed on it. The data analysis done on the data set is for two purposes, statistical summarization and assess features soundness and correlation among the dataset

variables before moving to the modelling stage. To analyze the data, Pandas was used exclusively for preparing the dataset and the subsets. Additionally, pandas was used for grouping and crosstabing as the basis for visualizations. The visualization of the data is done using Matplotlib and Seaborn. All coding and visualization is done in Jupyter Notebook.

5.3.1. **Descriptive Statistics**

The survey dataset contains data of 3,944 male and 4266 female participants. The female participants in the survey constitute 52% of the subjects in the dataset. The dataset subjects fall into four age groups ranging from 18 to above 60 years old. The vast majority of the subjects in the survey data fall in the 30 to 44 age group, with low representation of individuals above 60 as in figure 10. This shows that the dataset is fairly balanced gender-wise, but there noticeable lack of representation of senior citizens, and considering that NCDs are highly detrimental to that age group, this may need to be improved on in the coming versions of the survey.

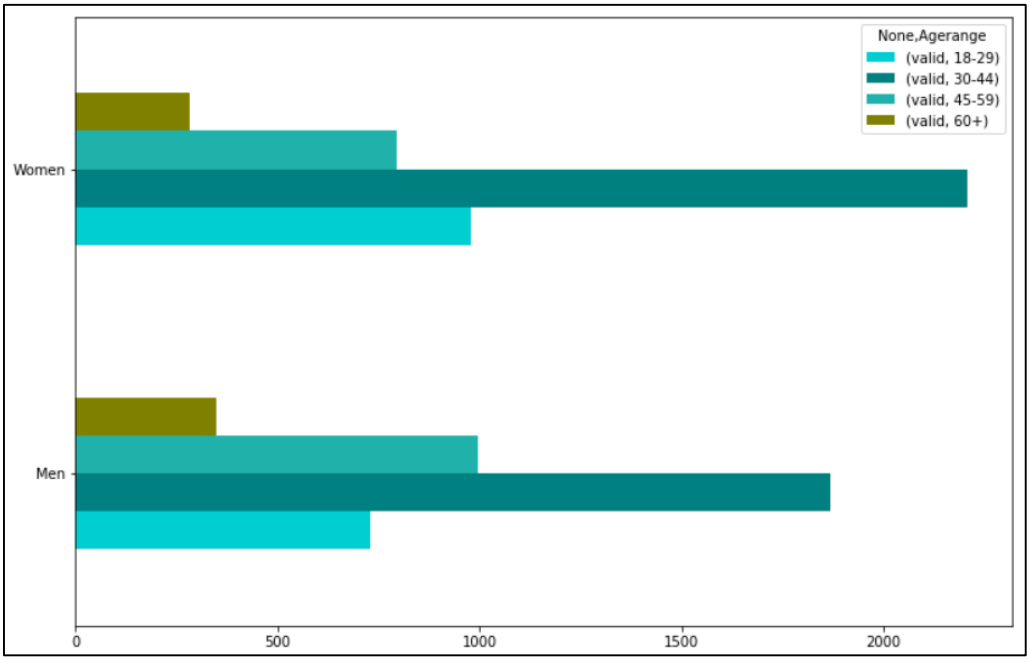

Figure 11. Distribution of the UAE STEPS NHS data by Gender and Age Group

The physical measurements available in the dataset can be a valuable source to identify health trends among the country's population. Crosstab analysis of major physical measurements and laboratory tests are aggregated by average across the males and females in the dataset as evident in the following figures.

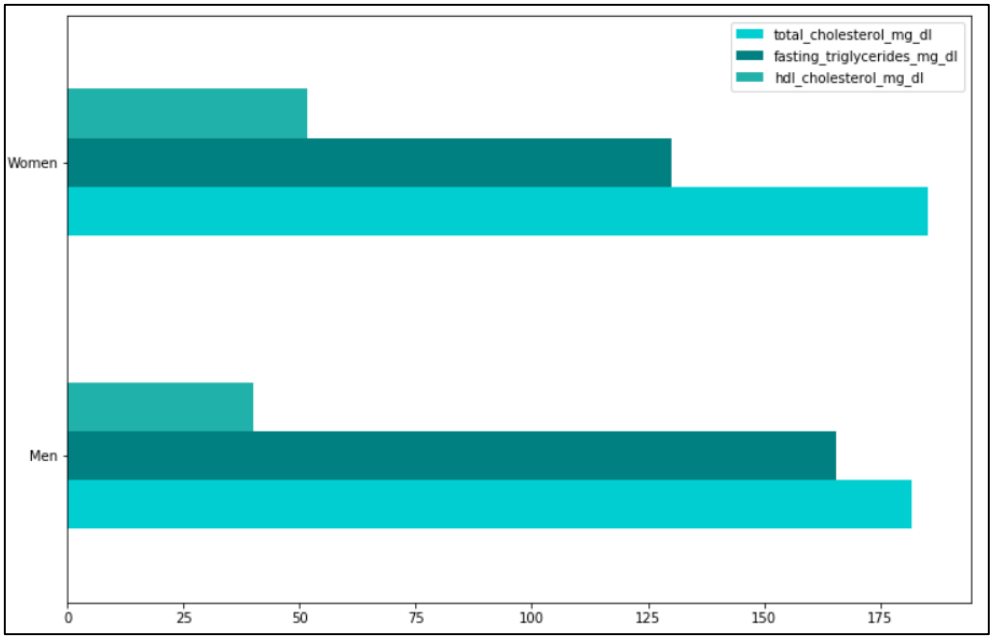

Figure 12. Total Cholesterol, Fasting Triglycerides and High Density Cholesterol (HDL) means for males and females in the dataset

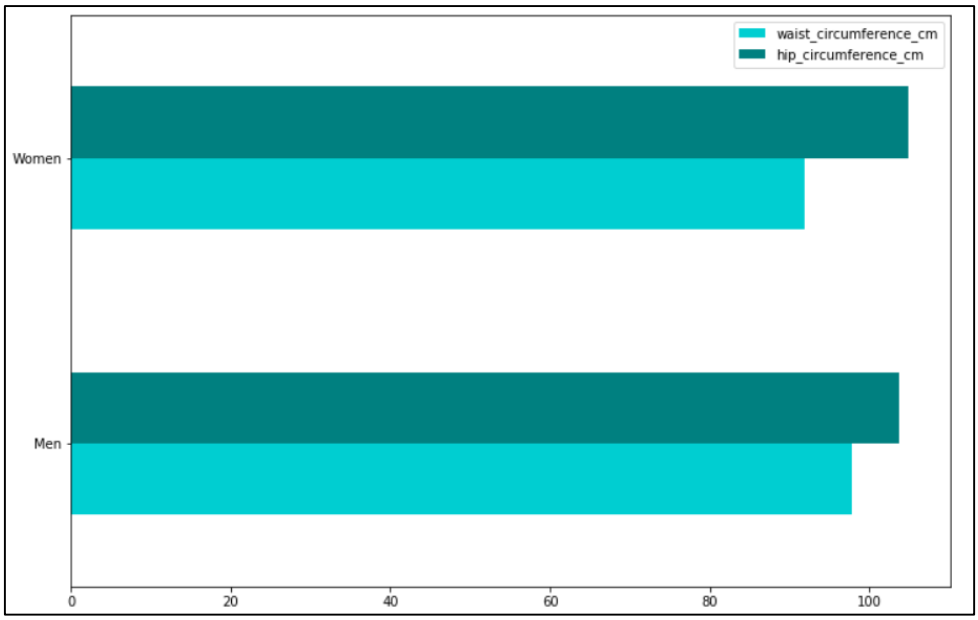

Figure 13. Waist and hip circumferences means for males and females in the dataset

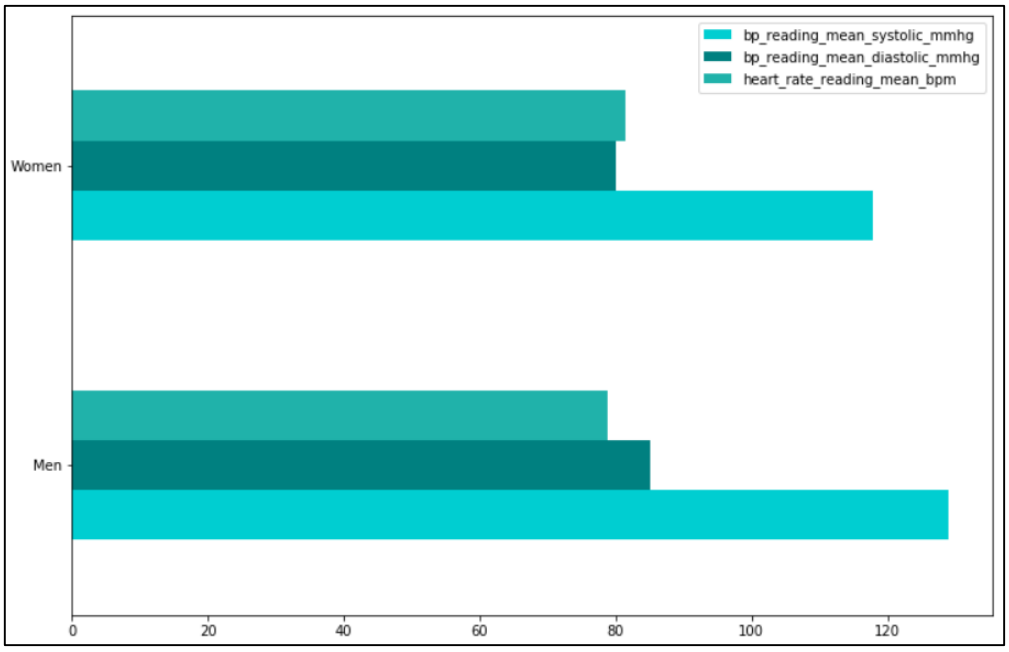

Figure 14. Blood pressure (systolic), blood pressure (diastolic) and heart rate means for males and females in the dataset

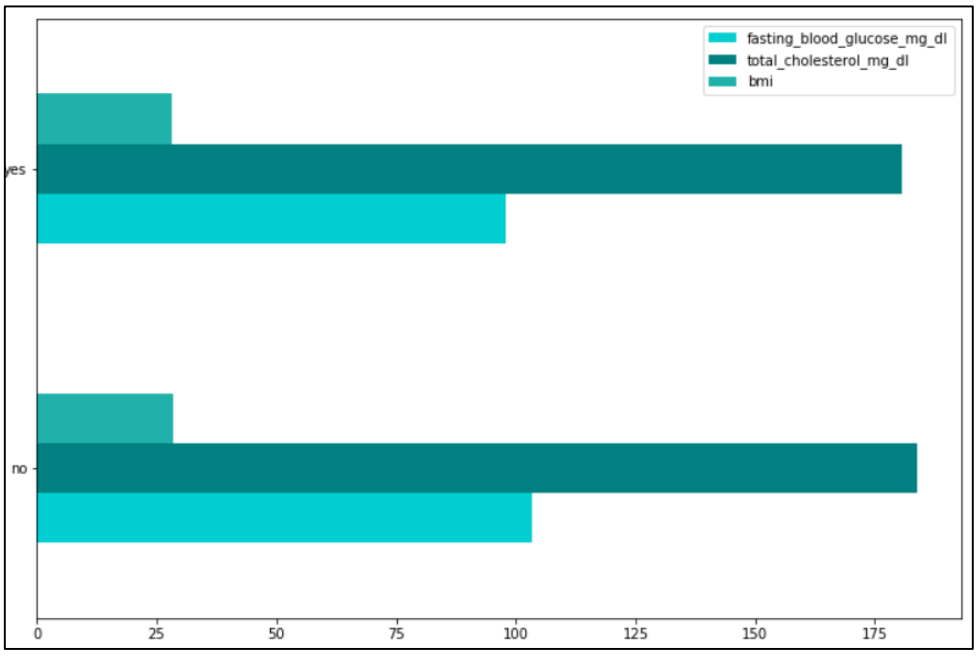

Figure 15. Fasting blood glucose, total cholesterol and BMI means for those, who do vigorous leisure activities and those, who do not

Complex correlations can be plotted from the UAE STEPS NHS data because it conations rich cross-sectional data including physical measurements, activity data, dietary data and laboratory tests' results for many biochemical measurements as demonstrated in figure 15.

An example of possible visualization of that wide range of data is a correlation between BMI levels and the range levels for total cholesterol of the subjects included in the dataset. This can provide good evidence for certain assumptions about the public health in the country, or may disagree with other assumptions and consequently guide the policy for this specific health indicator.

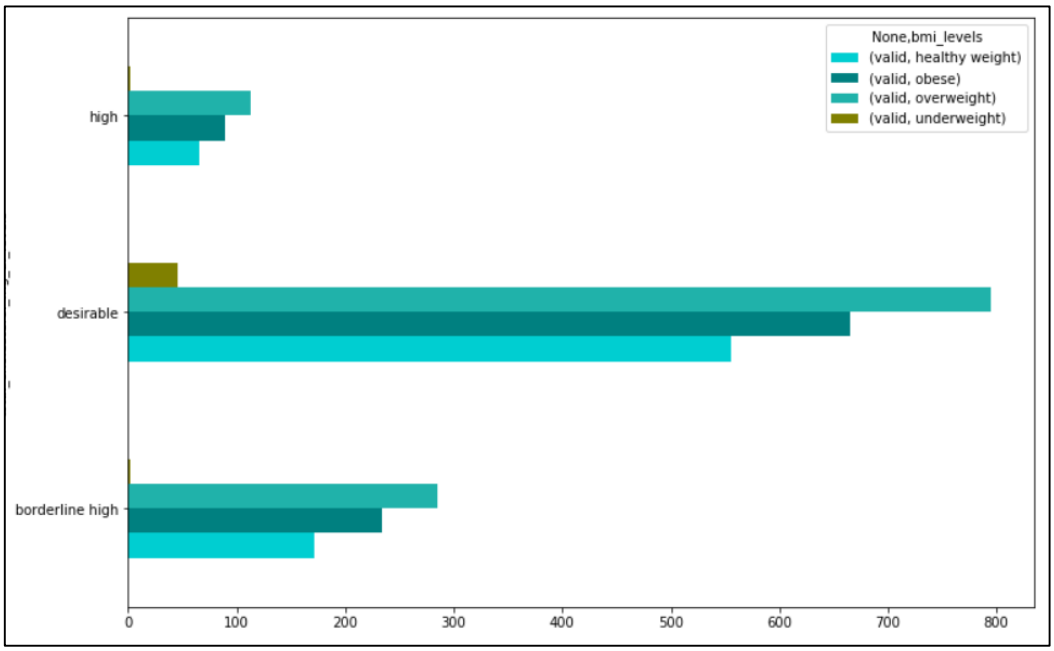

Figure 16. BMI ranges distributed across total cholesterol ranges to give an idea about cholesterolemia and healthy weight

Correlation scatter plots were also analyzed for potential relations and trends between the physical, biochemical and behavioral variables in the dataset.

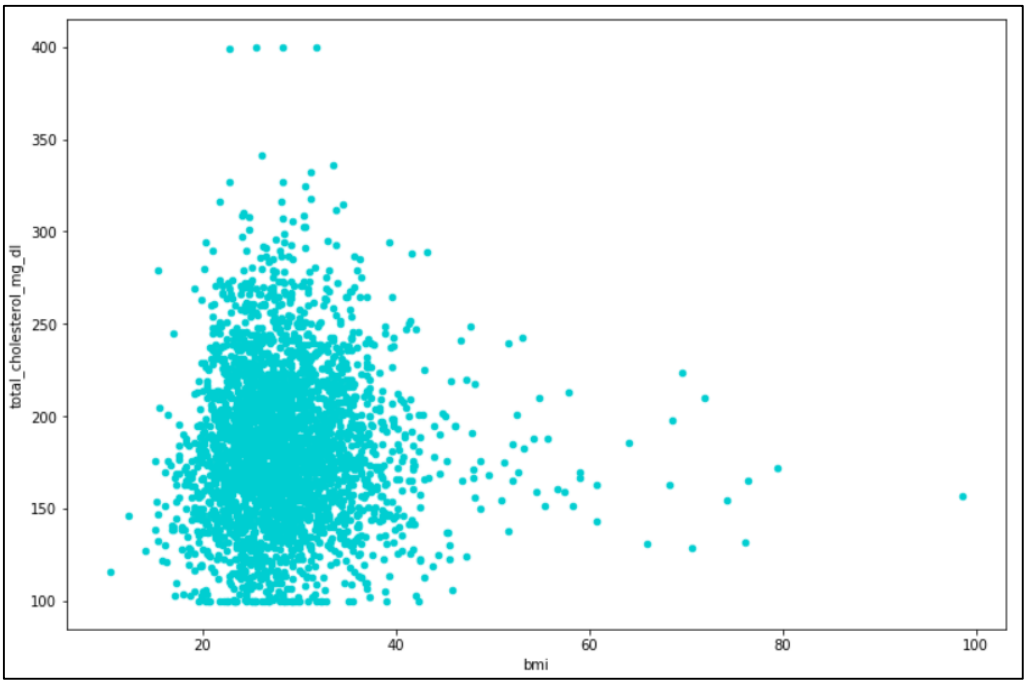

Figure 17. BMI ranges distributed across total cholesterol ranges to give an idea about cholesterolemia and healthy weight

Most of the correlation plots analyzed for this study did not exhibit a linear correlation between the variables in the dataset. This gives a more evidence to the previously asserted complexity of diabetes related risk factors and how ML can detect these difficult to find correlations.

5.3.2. Continuous Variables Normal Distribution Analysis

The UAE STEPS NHS dataset contains substantial number of numerical variables, which indicates to a wide range of measurements and results. Studying the distribution of these values can assist in understanding insight about the sampling process, uncover outliers and most importantly find variables, which can be disadvantageous to the ML construction. The following are samples of the many normal distribution plots analyzed in this study.

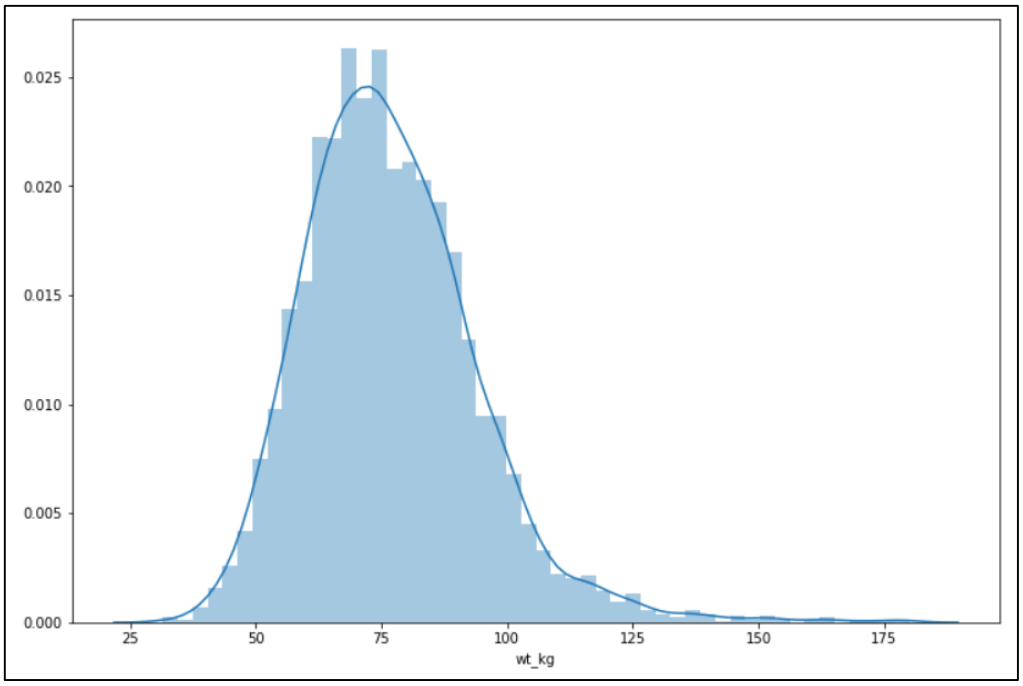

Figure 18. Normal distribution plot of the weight variable data in UAE STEPS NHS

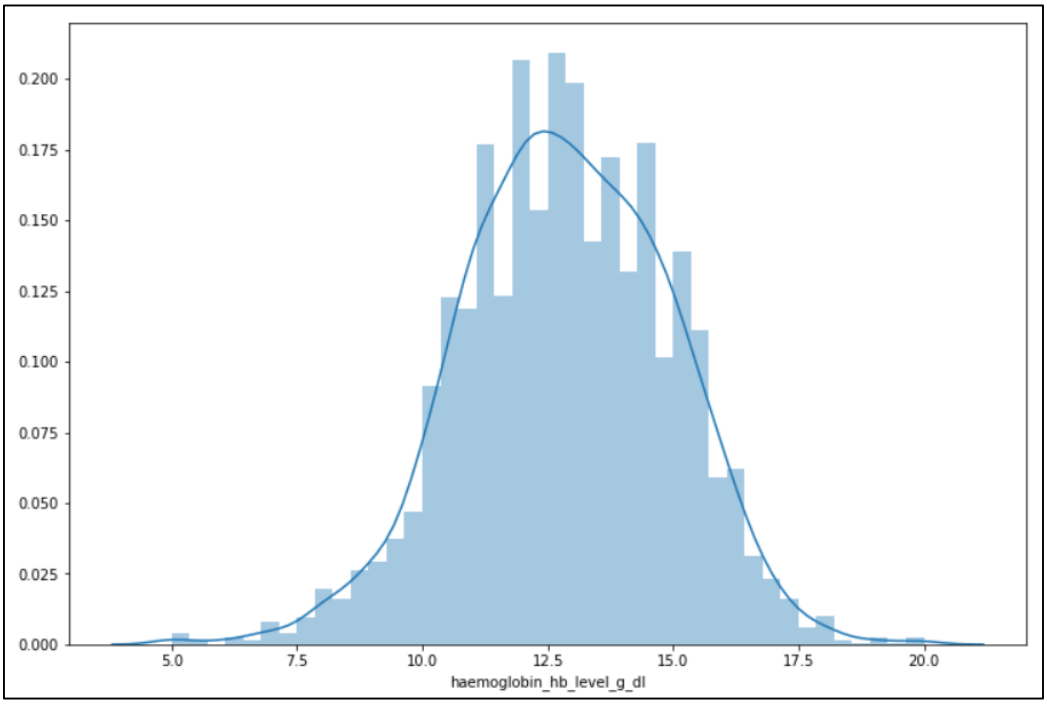

Figure 19. Normal distribution plot of the Hemoglobin level variable data in UAE STEPS NHS

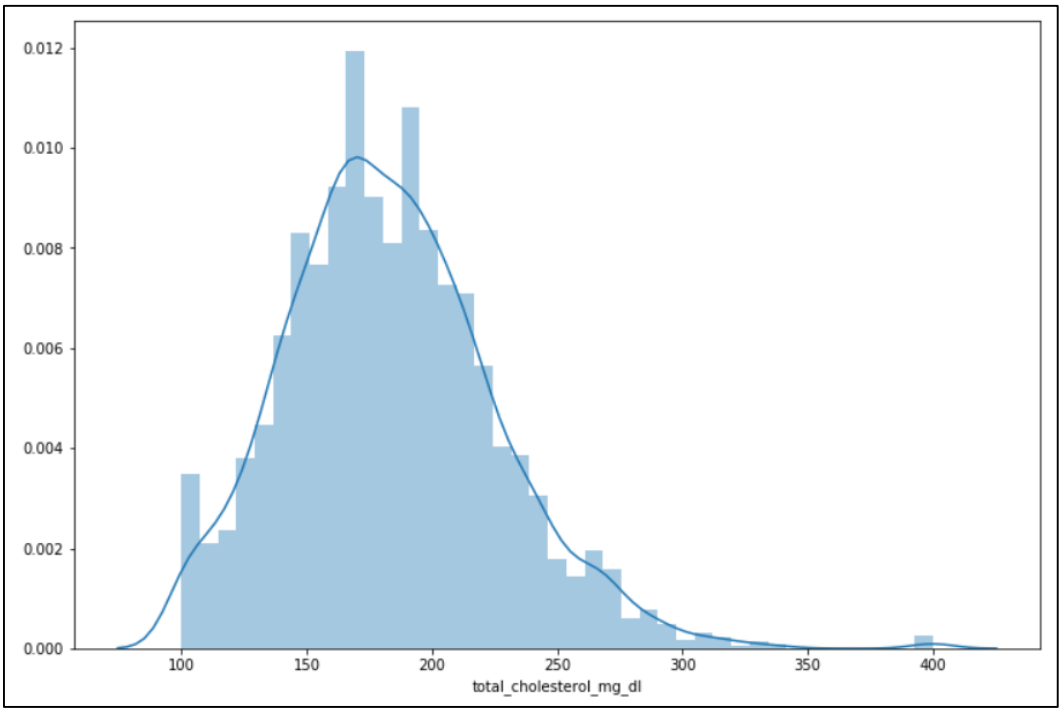

Figure 20. Normal distribution plot of the total cholesterol level variable data in UAE STEPS NHS

The analysis of the dataset normal distribution uncovered a problematic issue about two of the most important T2DM tests, HbA1c and fasting blood glucose as per figure 20 and 21. The extremely skewed distribution in Hba1c was identified to be the reason for including zeros and missing values in addition to actual results and results outside the range. These issues and considering that HbA1c to be useful must comply with strict standards as specified in section 1.2.4. Considering this, Hba1c was removed from the final feature set. The same is applicable to fasting blood glucose. The following normal distribution plots indicates the abnormal skewness of the two variables. This reflects that plotting and assessing variables is highly important in both considering important variables and removing possibly corrupted variables. This can also assist in the collection of the same variables in the future.

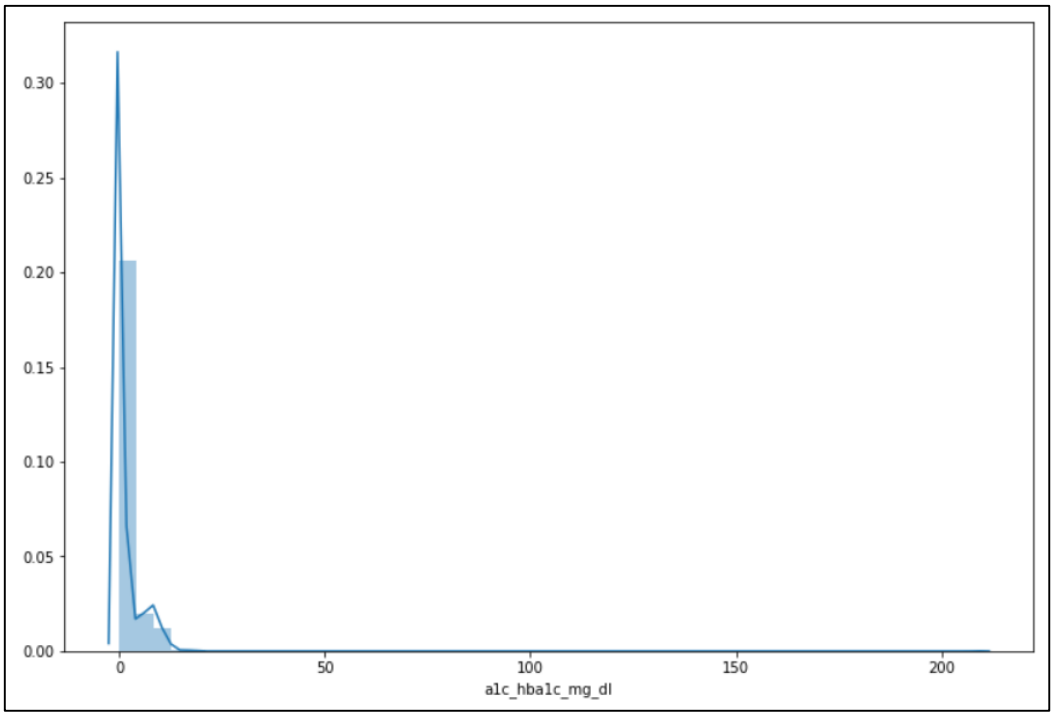

Figure 21. Normal distribution plot of the HbA1c variable data in UAE STEPS NHS, showing abnormal distribution that led to finding it not suitable to be a feature in the analysis

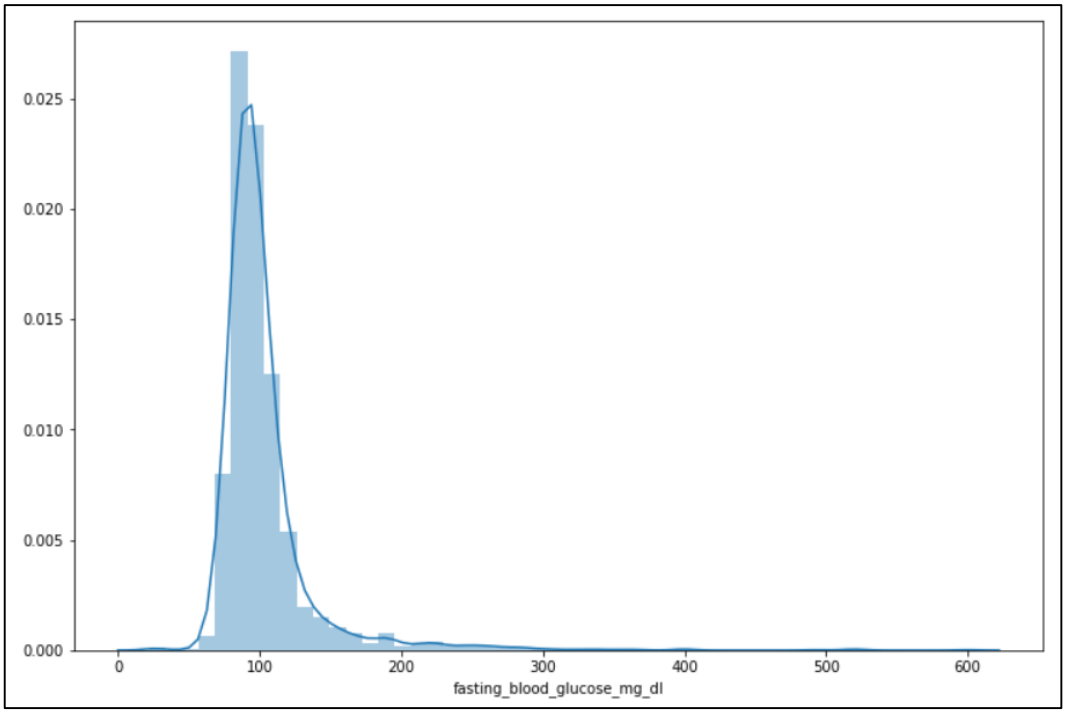

Figure 22. Normal distribution plot of the fasting blood glucose variable data in UAE STEPS NHS

5.4. ML T2DM Binary Classification Diagnostic Model Construction

5.4.1. Determining the Dependent (Target) Variable

ML tasks are based primarily on the methodology and techniques used in the construction and fitting of the trained data to intended models. As decided in the aim and task design, the primary objective of the constructed models will be to act as supervised learning ML binary classification models to determine the individuals with high risk of T2DM from the UAE STEPS NHS data. To determine the dependent variable that will contain the outcome classes, three main variables were considered. The three variables are: the patients have ever been diagnosed with diabetes, the patient have been diagnosed with diabetes in the past 12 months and the patient is taking diabetes medication in the past two week. After deliberation with the domain expert in the study and reviewing the variable potential, it has been decided that the variable classifying the participants according to them being diagnosed with diabetes in the past 12 months is the most suitable. This is because of the limited predictive value of the variable stating that the patient ever being diagnosed with diabetes and that they are taking diabetes medication now. These two variables have high bias and very imbalanced in the distribution of the binary classes of yes and no. Considering STEPS nature of the survey, asking someone, who answered that they are diagnosed with diabetes recently if they are taking medications for it will certainly yield a high rate of positive response.

To avoid this bias and due to the balance and number of responses to the variable related to being diagnosed with diabetes in the past 12 months, it has been selected as the target variable and it has been assigned with one and zero codes for the binary classification training of the model. The variable has 1,123 not null records with the distribution of 715 responses as yes and 408 responses as no. This shows that this variable is not greatly skewed and has acceptable representation to be considered as the outcome variable. Compared to the ever

diagnosed with diabetes has 5,011 not null records, but 3888 negative responses and 1,123 positive responses, raising the risk of bias in the training of the data.

5.4.2. **Experiment Design**

This study will contribute highly in proposing a novel approach of model evaluation and qualification that consider two main factors to improve the models. This research considers feature engineering as the most important factor in optimizing the constructed models and thus, it will consider developing a control model with the full feature set and consequently demonstrate the power of ML based dimensionality reduction to improve the performance of the optimized models. Three dimensionality reduction techniques will be tested in this research: Chi Squared (CS), which is a filter based method, Recursive Feature Elimination (RFE) method, which is a wrapper based method and finally the combination of the two methods to find the common features among the two reduced sets through intersection. This method will be called CS/RFE Intersection. As discussed in section 4.1, four ML classification algorithms were selected to act as the main experimental classifiers in this research the selection rationale was discussed and mostly based on reviewed literature and the statistically fit for purpose nature of the selected classifiers.

5.4.3. **Role of Feature Engineering**

The role of ML based feature selection to reduce the dimensionality of high dimensional feature sets is crucial and can significantly improve ML model predictive performance as it will be demonstrated in the evaluation of the proposed models. To demonstrate the importance of ML based feature selection, three approaches were selected. The first approach used is applying CS on the dataset to filter a k number of features based on the highest filtered features by importance. The k assigned to the method was 50 to select the 50 most important predictive features from the set. The second method used to also define another ML assisted feature

selection set, is the wrapper based RFE method and similarly, k was selected as 50 to select another set most important 50 features. Finally, a function was declared to intersect the two sets and find the features repeated in both lists as most important and define a new set of most important features. The two methods intersection list, consisted of 21 features. To demonstrate the improvement on the constructed models, the experiment will test the performance of LR, SVM, kNN and NB based on the full feature set (211 features), RFE feature set (50 features) and CS/RFE intersection feature set (21 features) as per table 11.

Algorithm	Features Size	Dimensionality Reduction	Validation Method
K Nearest Neighbors (KNN) - No Dimensionality Reduction	211	No Dimensionality Reduction	Train / Test Split
Logistic Regression (LR) - No Dimensionality Reduction	211	No Dimensionality Reduction	Train / Test Split
Naïve Bayes (NB) - No Dimensionality Reduction	211	No Dimensionality Reduction	Train / Test Split
Support Vector Machines (SVM) - No Dimensionality Reduction	211	No Dimensionality Reduction	Train / Test Split
K Nearest Neighbors (KNN) - RFE Only	50	RFE	Train / Test Split
Logistic Regression (LR) - RFE Only	50	RFE	Train / Test Split
Naïve Bayes (NB) - RFE Only	50	RFE	Train / Test Split
Support Vector Machines (SVM) - RFE Only	50	RFE	Train / Test Split
K Nearest Neighbors (KNN) - Chi Square and RFE	21	CS / RFE Intersection	Train / Test Split
Logistic Regression (LR) - Chi Square and RFE	21	CS / RFE Intersection	Train / Test Split
Naïve Bayes (NB) - Chi Square and RFE	21	CS / RFE Intersection	Train / Test Split
Support Vector Machines (SVM) - Chi Square and RFE	21	CS / RFE Intersection	Train / Test Split

Table 11. List of the proposed binary classification models across four algorithms: kNN, LR, SVM and NB. Three models of each algorithm were created based on the dimensionality strategy used

12 ML binary classification models to classify potential T2DM patients within 12 months based on the data from UAE STEPS NHS were created. Three models of each algorithm were designed and experimented with to demonstrate the classification efficacy and responsiveness to the dimensionality reduction strategies.

5.5. **ML Models Performance Evaluation**

Empirical evaluation of the performance of ML models is the basis for any further research or extension to the proposed design, accordingly this research used the most relevant applicable evaluation metrics to evaluate and the proposed methodology and T2DM predictive models. To ensure that the evaluation is done according to structured and evidence based manner, the final dataset of 368 records was divided into two sets to separate training cases

from testing cases. This split is done to avoid the memorization phenomenon where models tend to memorize training sets and perform poorly on unseen data. A training set of 301 records and a testing set of 67 records were created and ensured balanced distribution to avoid bias. The metrics used in the study to evaluate the models are accuracy, F1-Score, sensitivity and precision. The metrics will be based on the models predictions compared to ground truths using True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Due to the aim to create a decision support model that can alert to risk, accuracy and F1-Score are selected as the main ranking metrics. There will be minimal risk on individuals, who will be flagged falsely as they will undergo further testing to ensure that they healthy. Table 5.2 lists the metrics and formulas used to calculate their scores.

Metric	Formula
Accuracy	(TP + TN) / (TP+TN+FP+FN)
F1-Score	2 * ((precision * sensitivity) / (precision + sensitivity))
Sensitivity	TP / (TP + FN)
Precision	TP / (TP + FP)

Table 12. Metrics used to evaluate the 12 proposed binary classification models in the study

Based on the performance evaluation of the 12 models, the scores demonstrate that LR has the highest classification accuracy among the tested algorithms regardless of the dimensionality reduction technique applied. LR scored accuracies of 70% when no dimensionality reduction is used, 81% when RFE is used and scored the best among all the tested models at 87% when CS/RFE intersection dimensionality reduction technique was used. The results also demonstrate a clear advantage of introducing ML based dimensionality reduction strategies. The other algorithms responded differently to the changing feature sizes and dimensionality reduction technique. kNN was the second best performing classifier when no dimensionality reduction technique is used and all the 211 features were considered at 64%.

NB was second best at 79% accuracy when RFE feature selection is used. SVM performed second best at 84% accuracy when CS/RFE intersection method is used. NB accuracy was highly reduced when no dimensionality reduction was used at 42%, which demonstrate reduced viability to be used under similar conditions of high dimension and low cases dataset.

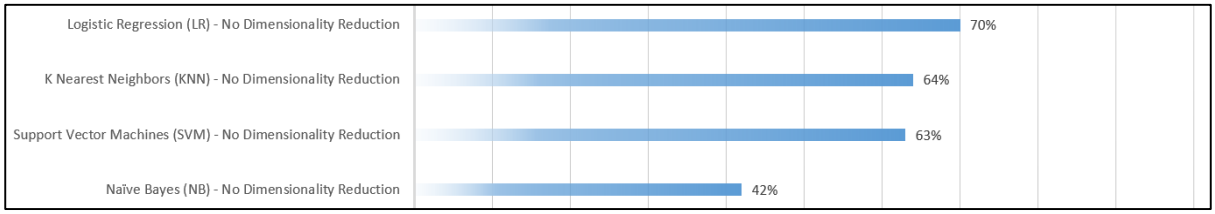

Figure 23. Accuracy of the models when no dimensionality reduction technique is used. LR is highest at 70%

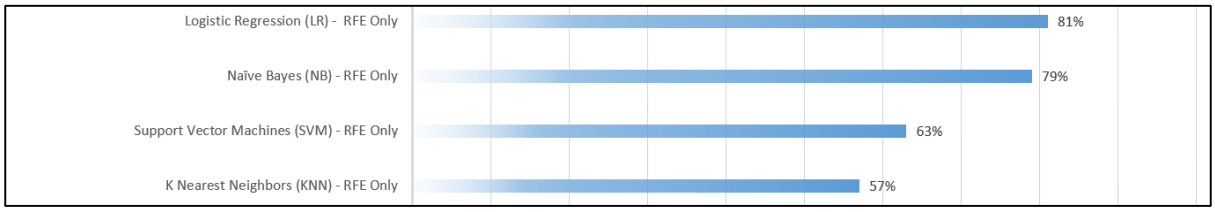

Figure 24. Accuracy of the models when RFE dimensionality reduction technique is used. LR is highest at 81%

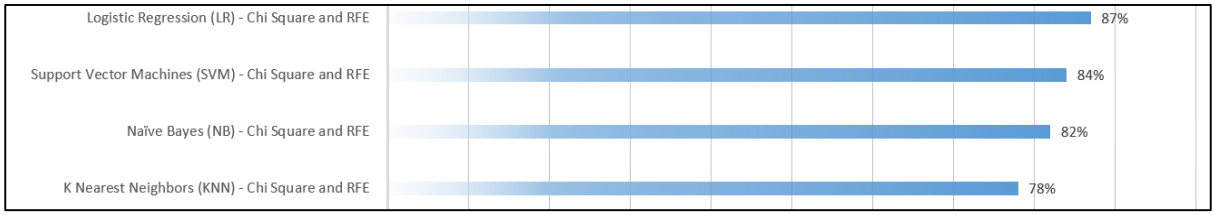

Figure 25. Accuracy of the models when CS/RFE intersection dimensionality reduction technique is used. LR is highest at 87%

F1-Score analysis also confirm the superiority of LR for the T2DM binary classification objective. Across all the dimensionalities, LR scored the highest harmonic means of 69% when no dimensionality reduction is used, 84% when RFE technique is used and scored the highest overall F1-score of 89% when CS/RFE intersection strategy is used.

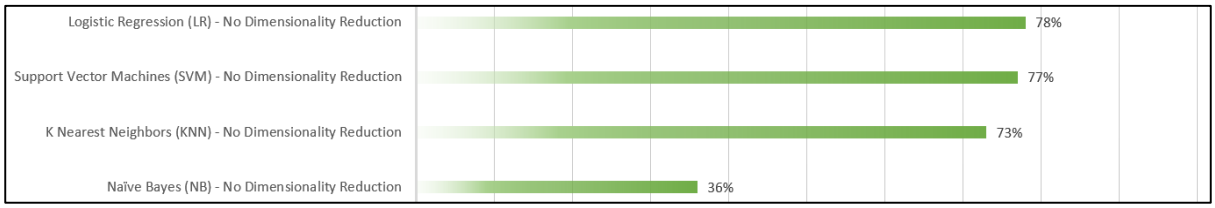

Figure 26. F1-Score of the models when no dimensionality reduction technique is used. LR is highest at 78%

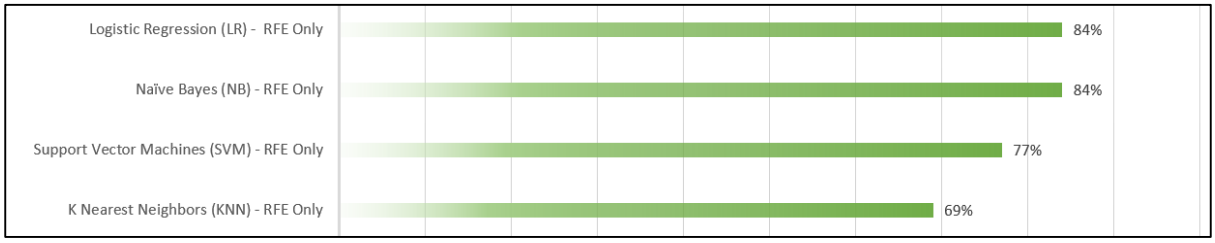

Figure 27. F1-Score of the models when RFE dimensionality reduction technique is used. LR is highest at 84%

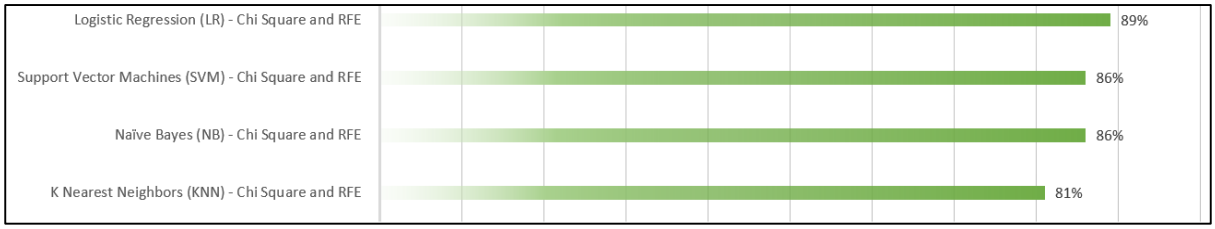

Figure 28. F1-Score of the models when CS/RFE intersection dimensionality reduction technique is used. LR is highest at 89%

The systematic evaluation of the constructed predictive models showed very promising results for the ML T2DM diagnostic classification objective. Starting from a high dimensional feature set and going through rigorous data cleansing, selection, training and evaluation, proved that this proposed frame work can be utilized for serious ML based bioinformatics research on national health surveillance cross-sectional datasets. The validation method used ensured that the metrics will measure the generalizability of the models adequately with reduced risk of overfitting. To the core of the progressive improvement of the models is dimensionality reduction and importance based feature selection.

5.6. **ML Models Deployment and Extensions**

Based on the construction of the T2DM binary classification models and consequent evaluation, LR with CS/RFE intersection dimensionality reduction exhibited close to state of the art performance at 87% accuracy and 89% F1-score, ML proved to be a promising option for non –invasive clinical diagnostic solutions. This qualifies this model to act as a core to an EMR related decision support system. The variation of the variables used in the model can highlight patients with high risk of diabetes that cannot be detected using conventional profiles. Also, the model can act as a highly necessary informal awareness tool to guide the public through a web application or a mobile app about their predicted risk for diabetes. This prediction is advisory and experimental of course, but it can alert the public to seek medical consultation at an early stage of the disease. Such informal tools that can be accessible to unprivileged or remote community members can be very crucial in saving lives by being persistently available and at a very low or no cost. Taking this into consideration, this research field of non-invasive and affordable advisory ML diagnostic tools should be a priority for research in domain specific AI and epidemiology.

The framework proposed to build highly accurate medical classification models based on national health surveillance surveys, can be used to construct models for any other NCD as it is not designed to construct the models purely for a specific disease. As a considered future research, a comparative analysis of the quality of models constructed based on the proposed framework is to be tested on EMR data. This will give an indication of the most suitable data source for ML predictive modelling to be considered by the country's health strategists.

Chapter 6. Limitations and Conclusion

6.1. Limitations

6.1.1. Reduced Dataset due to the Survey Design

The main limitation for this study is the contraction of the dataset size due to missing values. This study attempted to reduce the impact of the decrease of the size due to missing value by using data imputing strategies such as using dummy variables with negative value for categorical missing variables. However, no imputing technique was used to handle continuous variables to avoid over influencing the predictability of the data considering the already small data size. The reduction of the dataset is primarily due to the nature of the data being based on STEPS based health surveys. The STEPS health survey approach as proposed by WHO is very valuable for comparative and standardization purposes. However, it can highly affect the completeness of the data if used in a very rigid manner. The UAE STEPS NHS data is vastly affected by missing values due to this, which reduced the chance for having a more representative community assessment and a more generalizable dataset. Another issue that can affect health surveys is that they are prone to repeated errors because of the use of the same method to measure or test a certain value in the survey. This can manifest in many ways such as using the same uncalibrated or faulty device for a wide number of survey subjects or repeat the wrong process across many respondents. This inherent issue with health surveys can cause propagation of the error and cause the variable to be unusable and wasted. This was demonstrated in this research when the HbA1c was detected to be not accurately measured, which affected the whole variable measurement and caused it to be removed from the final dataset. EMR data can avoid such repeated errors as they are frequently calibrated, done by different people and with different instruments.

Another factor that may affect the quality of health survey based data, is that sometimes the respondent may not have a record of the question asked such as the amount of salt in mg used per week or in a meal. This leads to sometimes guessing that over or underestimate the values collected. However, knowing these assessments are better that not having any indication at all about the variable measured. Also, as evident in the data exploration section, the selection of the survey subjects was highly skewed towards one age group. This can hinder the ability to represent the NCDs prevalence across the community and also the ability to develop balanced and stratified ML models. NCDs have different patterns across the different age groups, thus for the STEPS NHS to be representative of the NCDs prevalence and be more useful for prediction use cases, a more diversified sampling protocol must be adopted for the upcoming version of the survey. The sampling frames in the UAE STEPS NHS gave high importance on stratifying the sampled subjects based on nationality, however there are other more representations that were supposed to be given a higher attention. Factors such age, risk to certain NCDs and genetic disposition must be also considered in coming versions of UAE STEPS NHS. This will improve the completeness and predictability of the models based on the datasets generated from the survey. Another factor that may have impacted the ability to construct a more aware model, is lack of to access some of the NHS modules such as data related to tobacco and alcohol consumption, mental health and some other NCD related data. The inaccessibility to this may have prevented uncovering of important predictive features for diabetes classification.

It is highly recommended that STEPS NHS are done with the highest quality control in practice. The STEPS NHS is a demanding survey that can result in hundreds of variables, so considering evaluation and validation of each variable to the highest standard is going to produce much higher quality datasets for ML and other uses. Another very important factor, is

the need to consider taking biochemical measurement even from healthy participants, who answer negatively to STEPS NCD trigger questions. This will expand the dataset cohort and will make available a balanced set of cases of health and people with the disease.

6.1.2. **Excluded Classification Algorithms**

On the design level, the initial high dimensionality of the dataset, prevented the experimentation with certain viable classification techniques such as DT and RF. ST and to lesser extent RF, are highly explainable classification models and in smaller dimensions can give a very clear picture about the hierarchical classification of the features in the dataset. Another highly relevant classification is ANNs and neural based classification models. As demonstrated in the literature reviewed, ANNs can be extremely capable of achiving a good generalization of complex feature sets such as imaging and natural language. However this research elected to not consider the inclusion of ANN in the experimented models due to many reasons.

This research attempted to experiment on modelling the binary classification on ANNs, but the results were average at best and some cases even below of the other linear and probabilistic classifiers used in this study. Multilayer Perceptron (MLP) feed-forward neural networks were used with varies hyperparameters such as using Rectified Linear Unit (ReLU) activation and different layer and neurons sizes (Müller & Guido, 2016). Considering the computational expense that resulted from the MLP classifier, it was not justifiable in comparison with the high performing simpler models considered in the study. The other reason is the aim to produce explainable models that can be adjusted according to the customization of the dataset and respond predictably to dimensionality reduction and feature selection methods. The black box nature of ANNs can reduce this level of explainability, which is available in the other models in this study that can even outperform ANN considering the task

and the dataset size. This also supported by literature where complex tasks such as predicting diabetes onset from ECG readings showed excellent convergence (Swapna et al., 2018), simple binary classification on limited data showed average performance by ANN models (Nguyen et al., 2019). This is also supported by literature where complex tasks such as predicting diabetes onset from ECG readings showed excellent convergence (Swapna et al., 2018), simple binary classification on limited data showed average performance by ANN models (Nguyen et al., 2019). The poor performance of ANNs and MLP for the study objective and dataset, may be attributed to the reduced number of cases in the final dataset, preventing it from converging, so further testing is required on a larger dataset to determine if the reason behind the underperformance is related to the size of the dataset.

6.2. **Conclusion**

ML research in bioinformatics and more specifically in NCDs and public health saw tremendous attention in the past years. Many factors such as the proliferation of EMR systems, the unprecedented increase in the prevalence of diabetes and its accompanying complications, motivated this research to consider ML for the early detection of undiagnosed T2DM based on cross-sectional data from UAE STEPS NHS. The early detection of diabetes is the key to controlling the disease and reducing the personal quality of life and financial burdens on patients and the healthcare sector. Many people can find difficulty accessing diagnostic healthcare services for many reasons including, but not limited to economical and regional factors. ML based diagnostic and decision support systems can provide a first line of detection to alert patients about potential disease risk. This will guide the public to seek medical help at an early stage of the disease and counter the serious complications, which can result from delayed diagnosis. STEPS NHSs are potentially a very viable source for NCDs ML based research as determined by the high accuracy models based on UAE STEPS NHS presented in

this research. NHS STEPS datasets provide a good cross-sectional representation of the community public health variables that may not be easily found in EMR. This standardized and rounded assessment of the various physical, clinical and behavioral variables of members of the society makes STEPS NHS an ideal candidate for epidemiological research and public health strategy guidance.

This work proposed a comprehensive framework to find a clinical diagnostic query, which in this case was the prediction of T2DM risk for the UAE community. The framework is based on literature review and ML learning modelling best practice. It considered highly rigorous variables pre-processing effort to ensure that the dataset used in the modelling of predictive classifiers is clean and optimized as possible. Systematic analysis of the data was done to select the dependent target variable with best representation and least missing values. In addition to the systematic pre-processing of the data and cleansing effort, dimensionality reduction and ML based feature selection proved to be very effective in vastly improving the accuracy and F1-score of the binary classification ML models. Out of the algorithms used, LR performed the best across all feature sets, showing that LR is ideal for the proposed objective of binary diagnostic classification. The best dimensionality reduction technique based on the performance tested, is CS/RFE intersection technique.

Although STEPS NHS can be a very useful source of data to be experimented with and research in ML, it comes also with many issues that can be handled with proper quality assurance during the fieldwork of the survey. Also, the survey administrators are responsible to demonstrate all the validation methods and standards used to ensure the highest compliance of the collected measurements with the clinical guidelines. Although EMR data can be more accurate than NHS, EMR data still suffers from being disjointed, scattered, unstandardized and not fully representative of the individual lifestyle. Ensuring that most important NHS variables

are integrated in EMR and NHS being conducted to the highest clinical standards will ensure having a rich validated data from both sources that can feed ML biomedical systems and medical diagnostic predictive modelling research. This will consequently support the construction of highly accurate and validated ML diagnostic models capable of assisting the healthcare system in the country and improve the healthcare outcomes and improve the community members' quality of life.

References

Abuelmagd, W., Afandi, B., Håkonsen, H., Khmidi, S., & Toverud, E. L. (2018). Challenges in the management of Type 2 Diabetes among native women in the United Arab Emirates. Diabetes research and clinical practice, 142, 56-62.

Alarcón-Paredes, A., Francisco-García, V., Guzmán-Guzmán, I. P., Cantillo-Negrete, J., Cuevas-Valencia, R. E., & Alonso-Silverio, G. A. (2019). An IoT-Based Non-Invasive Glucose Level Monitoring System Using Raspberry Pi. Applied Sciences, 9(15), 3046.

AlShuweihi, M., Salloum, S. A., & Shaalan, K. (2020). Biomedical Corpora and Natural Language Processing on Clinical Text in Languages Other Than English. Recent Advances in Intelligent Systems and Smart Applications, 295, 491.

American Diabetes Association. (2020a). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2020. Diabetes Care, 43(Suppl 1), S14-S31

Christobel, Y. A., & Sivaprakasam, P. (2013). A new classwise k nearest neighbor (CKNN) method for the classification of diabetes dataset. International Journal of Engineering and Advanced Technology, 2(3), 396-200.

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., ... & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. Journal of diabetes science and technology, 12(2), 295-302.

Doan, S., Maehara, C. K., Chaparro, J. D., Lu, S., Liu, R., Graham, A., ... & Ohno-Machado, L. (2016). Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes. Academic Emergency Medicine, 23(5), 628-636.

Dwivedi, A. K. (2018). Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Computing and Applications, 30(12), 3837-3845.

Formosa, C., Chockalingam, N., & Gatt, A. (2019). Diabetes foot screening: Challenges and future strategies. The Foot, 38, 8-11.

Gronsbell, J., Minnier, J., Yu, S., Liao, K., & Cai, T. (2019). Automated feature selection of predictors in electronic medical records data. Biometrics, 75(1), 268-277.

IDF. (2019). Idf diabetes atlas ninth edition. International Diabetes Federation.

Jayanthi, N., Babu, B. V., & Rao, N. S. (2017). Survey on clinical prediction models for diabetes prediction. Journal of Big Data, 4(1), 26.

Ker, J., Bai, Y., Lee, H. Y., Rao, J., & Wang, L. (2019). Automated brain histology classification using machine learning. Journal of Clinical Neuroscience, 66, 239-245.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19(1), 1-9.

Liu, X., Zhou, Y., & Wang, Z. (2019). Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. Journal of Visual Communication and Image Representation, 60, 1-15.

Lu, W., Li, Z., & Chu, J. (2017). A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. Computers in Biology and Medicine, 83, 157-165.

Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., ... & Saitoh, E. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Scientific reports, 9(1), 1-9.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer methods and programs in biomedicine, 152, 23-34.

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".

MOHAP. (2018a). Uae national health survey report 2017-2018. Ministry of Health and Prevention.

MOHAP. (2018b). Non-communicable disease risk factor survey (steps) data book for uae 2017-2018. Ministry of Health and Prevention.

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., ... & Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 182, 105055.

OCHA. (2020). United arab emirates datasets. Retrieved 15.04.2020, from https://data.humdata.org/group/are

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Roch, A. M., Mehrabi, S., Krishnan, A., Schmidt, H. E., Kesterson, J., Beesley, C., ... & Schmidt, C. M. (2015). Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. Hpb, 17(5), 447-453.

Rodriguez-Romero, V., Bergstrom, R. F., Decker, B. S., Lahu, G., Vakilynejad, M., & Bies,

R. R. (2019). Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques. Clinical and translational science, 12(5), 519-528.

Sammut, C., & Webb, G. I. (2017). Encyclopedia of machine learning and data mining. Springer.

Sklearn. (2020). Scikit-learn user guide. Scikit-learn developers, Release 0.22.2.

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 13.

Somasundaram, S. K., & Alli, P. (2017). A machine learning ensemble classifier for early prediction of diabetic retinopathy. Journal of Medical Systems, 41(12), 201.

Sun, Y. L., & Zhang, D. L. (2019). Machine Learning Techniques for Screening and Diagnosis of Diabetes: a Survey. Tehnički vjesnik, 26(3), 872-880.

Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4(4), 243-246.

Tvardik, N., Kergourlay, I., Bittar, A., Segond, F., Darmoni, S., & Metzger, M. H. (2018). Accuracy of using natural language processing methods for identifying healthcare-associated infections. International journal of medical informatics, 117, 96-102.

Vasilakos, A. V., Tang, Y., & Yao, Y. (2016). Neural networks for computer-aided diagnosis in medicine: A review. Neurocomputing, 216, 700-708.

Vehí, J., Contreras, I., Oviedo, S., Biagi, L., & Bertachi, A. (2020). Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. Health Informatics Journal, 26(1), 703-718.

Vision2021. (2020). World-class healthcare. Retrieved 15.04.2020, from https://www.vision2021.ae/en/national-agenda-2021/list/world-class-circle

Wang, L., Gao, P., Zhang, M., Huang, Z., Zhang, D., Deng, Q., ... & Zhou, M. (2017). Prevalence and ethnic pattern of diabetes and prediabetes in China in 2013. Jama, 317(24), 2515-2523.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77, 34-49.

WHO. (2016). Global report on diabetes. World Health Organization.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), 16.

Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.