



Arab American University
Faculty of Graduate Studies

**Raised Blood Pressure and Sugar Detection and Prediction
Models Using Machine Learning Algorithms: As A Screening
Tool for Hypertension and Diabetes**

By

Marwa Mustafa Owess

Supervisor

Dr. Amani Yousef Owda

Co-Supervisor

Dr. Majdi Owda

**This thesis was submitted in partial fulfillment of the
requirements for the Master's degree in Data Science and
Business Analytics**

July/ 2024

Thesis Approval

Raised Blood Pressure and Sugar Detection and Prediction Models Using Machine Learning Algorithms: As A Screening Tool for Hypertension and Diabetes

By

Marwa Mustafa Owess

This thesis was defended successfully on 6/7/2024 and approved by:

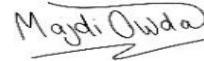
Committee members

Signature

1. Dr. Amani Owda: Supervisor

Handwritten signature of Amani Owda in black ink, enclosed in a hand-drawn oval.

2. Dr. Majdi Owda: Co- Supervisor

Handwritten signature of Majdi Owda in black ink, enclosed in a hand-drawn oval.

3. Dr. Huthaifa Ashqar: Internal Examiner

Handwritten signature of Huthaifa Ashqar in black ink.

4. Dr. Mohammad Jubran: External Examiner

Handwritten signature of Mohammad Jubran in blue ink.

Declaration

I certify that the thesis, "Raised Blood Pressure and Sugar Detection and Prediction Models Using Machine Learning Algorithms: As A Screening Tool for Hypertension and Diabetes" is an entirely original work of mine, and hasn't been submitted for another degree or certification, and was completed for the Masters in Data Science and Business Analytics at Arab American University - Palestine.

The Name of The Student: Marwa Mustafa Mohammad Owess

ID: 202112687

Signature: *Marwa Mustafa Owess*

Date: 16/10/2024

Acknowledgments

Completing this study could not have been possible without the expertise and guidance of Dr. Amani Owda and Dr. Majdi Owda, my thesis advisors. Therefore, I would like to express my sincere gratitude to them, for their valuable support throughout this endeavor, and for generously sharing their expertise and knowledge during their conducted academic courses that we attended with them over the past two years.

I also owe a debt of gratitude to the Arab American University – Palestine (AAUP), the institution I attended, for fostering an environment conducive to growth and learning. Additionally, I extend my appreciation to the Professors of the Data Science and Business Analytics Master's program.

My heartfelt gratitude goes out to my family, especially my parents because without them, none of this would have been possible. Their encouragement and sacrifices enabled me to achieve all I have today.

I wish to acknowledge the essential support provided by the Palestinian Ministry of Health, and the World Health Organization for providing free access to the data subset used in this study from the original STEPS Dataset (Palestine 2022). Also, the Palestinian National Institute of Public Health and fieldworkers for their efforts in collecting the STEPS dataset.

Abstract

High blood pressure and raised blood sugar are key risk factors for many chronic Non-Communicable Diseases (NCDs). Raised blood sugar is a strong indicator of prediabetes or diabetes mellitus. Similarly, high blood pressure is considered a sign of hypertension, which is identified as a key risk factor for developing heart and cardiovascular diseases. Interestingly, hypertension and diabetes mellitus are the top common global NCDs affecting the adult population not only the elderly. Recently, the prevalence of diabetes and hypertension has been increasing at a faster rate, especially in developing countries. The primary concern associated with these diseases is the potential for serious health complications to occur if it is not diagnosed early or not managed properly, which may progress poorly and lead to disabilities. Therefore, timely detection and screening of diabetes and hypertension is considered a crucial factor in treating and controlling those diseases and averting their progression into severe health consequences. Population screening for high blood pressure and raised blood sugar aims to identify individuals at risk before symptoms appear, enabling timely intervention and potentially improved health outcomes. However, implementing large-scale screening programs can be expensive, requiring testing, follow-up, and management resources, potentially straining healthcare systems. Given the above facts, this study presents supervised machine-learning models to detect and predict raised blood pressure and sugar health conditions. The proposed prediction models utilize the related risk factors that are shared by the two health conditions of high blood pressure and raised blood sugar. These common risk factors involve age, body mass index, eating habits, physical activity, history of other diseases, and fasting blood sugar, obtained from the dataset of the STEPwise study of NCDs risk factors, collected from adults in the Palestinian community. The NCDs risk factors gathered by the STEPS dataset were used as input for building the prediction models, which were trained using various types of supervised-learning classification algorithms including Random Forest, XGBoost, Decision Tree, and Multilayer Perceptron. Based on the experimental results, the proposed models performed the best predictive power by employing the Random Forest algorithm, yielding an accuracy of 98.05%, and 94.76% for the raised blood sugar and blood pressure detection models respectively. Additionally, the experimental results for the models implemented using the other classifiers are promising. The raised blood pressure and sugar detection models can be improved by incorporating multiple separate measurements of fasting blood sugar, and blood pressure that are taken on various days, so it can be used as a highly efficient and

accurate diagnosis tool for diabetes, and hypertension, not only for screening purposes. In addition, it can be extended for determining the classification of the fasting blood sugar whether it is normal, impaired, or raised, and classifying the type of blood pressure disorders.

Table of Contents

Thesis Approval.....	I
Declaration.....	II
Acknowledgments	III
Abstract.....	IV
List of Tables	IX
List of Figures.....	X
List of Appendices.....	XIII
List of Abbreviations	XIV
List of Publications	XV
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Problem Statement and Motivation	3
1.3 Study Objectives.....	5
1.4 Significance of the Study.....	6
1.5 Contribution to Knowledge	7
1.6 Thesis Arrangement.....	8
Chapter 2: Background and Literature Review	10
2.1 Introduction	10
2.2 Hypertension.....	12
2.3 Diabetes	14
2.4 Using Machine Learning in Healthcare.....	17
2.5 Using Machine Learning in Detecting Blood Pressure and Glucose Metabolism Disorders.....	19
2.5.1 Using Machine Learning in Detecting Glucose Metabolism Disorders.....	22

2.5.2 Using Machine Learning in Detecting Blood Pressure Disorders.....	27
2.6 List of Challenges in Adopting Machine Learning Models in Healthcare.....	36
2.6.1 Current Problems and Limitations in Hypertension and Diabetes Risk Prediction Models.....	37
2.7 Research Gap.....	38
2.8 Conclusion.....	39
Chapter 3: Methodology.....	41
3.1 Introduction	41
3.2 Used Technology.....	41
3.3 Data Collection and Dataset Description.....	42
3.4 Data Cleaning	48
3.4.1 Merging the Dataset	49
3.4.2 Missing Values	49
3.4.3 Mapping Features Values	50
3.4.4 Checking Outliers.....	50
3.4.5 Features Engineering and Elimination	51
3.5 Exploratory Data Analysis	55
3.5.1 Main Characteristics of the Population Sample in the STEPS Dataset.....	55
3.5.2 Exploring Variables Relationships Using Bivariate Analysis.....	57
3.5.3 Analysis of Key Features of Interest	70
3.6 Data Preprocessing	73
3.6.1 Determining Target Feature.....	74
3.6.2 Features Selection.....	74
3.6.3 Features Scaling.....	81
3.6.4 Dataset Oversampling.....	81

VIII

3.7 Machine Learning Models.....	83
3.7.1 Machine Learning Classifiers.....	84
3.7.2 Experimental Settings.....	85
3.8 Models Performance Evaluation Criteria	86
3.9 Conclusion.....	88
Chapter 4: Results and Discussion	89
4.1 Introduction	89
4.2 Raised Blood Sugar Detection Model	89
4.3 Raised Blood Pressure Detection Model.....	91
4.4 Discussion.....	94
4.5 Conclusion.....	95
Chapter 5: Conclusion and Future Work.....	96
5.1 Conclusion.....	96
5.2 Study Obstacles and Mitigation Methods.....	97
5.3 Future Work.....	98
References	99
Appendices	137
المخلص.....	138

List of Tables

Table 2.1 Research findings related to machine learning prediction model for glucose metabolism disorders.....	25
Table 2.2 Research findings related to machine learning prediction models for blood pressure.....	34
Table 3.1 The Features of STEPS Dataset and their Description.....	44
Table 3.2 The Accepted Ranges of STEPS Dataset Variables.....	51
Table 3.3 Configuration of RBS & RBP Models Parameters	86
Table 4.1 The results of Performance Metrics for the Raised Blood Sugar Prediction Models	89
Table 4.2 The Results of Performance Metrics for the Raised Blood Pressure Prediction Models	92

List of Figures

Figure 1.1 Thesis Structure.....	9
Figure 2.1 Heart and Blood Vessels in Normal Case vs. High Blood Pressure Case (Shrestha, 2023).....	13
Figure 2.2 Type 2 diabetes mellitus risk factors and complications (Tiwary et al., 2023)	16
Figure 3.1 Main Phases of the Raised Blood Pressure and Sugar Detection Models	41
Figure 3.2 WHO-5 Well-Being Questionnaire.....	47
Figure 3.3 Four-Item Patient Health Questionnaire (PHQ-4)	47
Figure 3.4 Jenkins Sleep Scale	48
Figure 3.5 The Applied Steps for Cleaning STEPS Raw Dataset	48
Figure 3.6 Age Distribution of Participants in the STEPS Dataset	56
Figure 3.7 The Distribution of Participants in the STEPS Dataset by Sex	56
Figure 3.8 The Distribution of Participants in the STEPS Dataset by Age Group and Sex	57
Figure 3.9 Comparison of BMI Rates by Participants' Gender	58
Figure 3.10 Comparison of BMI Class by Age and Sex	59
Figure 3.11 The Relationship Between Weight and Systolic Blood Pressure.....	59
Figure 3.12 The Distribution of Participants by Smoking Status and Gender	60
Figure 3.13 Blood Sugar Level Classifications by BMI	61
Figure 3.14 Blood Pressure Measurements' Classifications by BMI.....	62
Figure 3.15 Comparison of Blood Sugar Between Diabetic and Nondiabetic Groups ..	63
Figure 3.16 Comparison of Blood Sugar Between Hypertensive and Non-hypertensive Groups	64
Figure 3.17 The Relationship Between Blood Pressure and Blood Sugar Levels	65

Figure 3.18 Comparison of the Raised Blood Sugar Prevalence Between Female and Male Groups	66
Figure 3.19 Comparison of the Raised Blood Pressure Prevalence Between Female and Male Groups	66
Figure 3.20 The Prevalence of Diabetes Among Participants by their BMI Class	67
Figure 3.21 The Prevalence of Hypertension Among Participants by their BMI Class.	68
Figure 3.22 Comparison of the Sugar Intake Between the Raised and Not Raised Blood Sugar Groups	69
Figure 3.23 Comparison of the Salt Intake Between the Raised and Not Raised Blood Pressure Groups	69
Figure 3.24 The Distribution of Participants by the Levels of Fasting Blood Sugar	70
Figure 3.25 The Distribution of Participants by the readings of Systolic Blood Pressure	71
Figure 3.26 The Distribution of Participants by the readings of Diastolic Blood Pressure	72
Figure 3.27 The Preprocessing Methods Applied on STEPS Dataset.....	73
Figure 3.28 Variables Correlation Matrix	76
Figure 3.29 Feature Importance using Univariate Feature Selection (Chi-square test) - RBS Target Feature	77
Figure 3.30 Feature Importance using Univariate Feature Selection (Chi-square test) - RBP Target Feature	78
Figure 3.31 Feature Importance using Recursive Feature Elimination based on Random Forest - RBS	79
Figure 3.32 Feature Importance using Recursive Feature Elimination based on Random Forest – RBP.....	80
Figure 3.33 Distribution of Cases by the Level of Blood Sugar Before (A) and After (B) Applying Oversampling	82

Figure 3.34 Distribution of Cases by the Blood Pressure Before (A) and After (B) Applying Oversampling	82
Figure 3.35 The Workflow of the Proposed Prediction Models for Raised Blood Pressure and Sugar	83
Figure 4.1 Raised Blood Sugar Prediction Models Performances Comparison using ROC AUC.....	90
Figure 4.2 Comparison of Performance of the Raised Blood Sugar Detection Models Using Different Classifiers by Confusion Matrix.....	91
Figure 4.3 Raised Blood Pressure Prediction Models Performances Comparison using ROC AUC	92
Figure 4.4 Comparison of Performance of the Raised Blood Pressure Detection Models Using Different Classifiers by Confusion Matrix.....	93

List of Appendices

Appendix (1) Use of STEPS Dataset Authorization Letter	137
--	-----

List of Abbreviations

ANN	Artificial Neural Network
BMI	Body Mass Index
BP	Blood Pressure
CNN	Convolutional Neural Network
CVD	Cardio Vascular Disease
DBP	Diastolic Blood Pressure
DT	Decision Tree
ECG	Electrocardiogram
EEG	Electroencephalography
EHR	Electronic Health Records
FBS	Fasting Blood Sugar
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HBA1C	Glycated Hemoglobin
HDL	High-Density Lipoprotein Cholesterol.
IDF	International Diabetes Federation
JSS	Jenkins Sleep Disturbances
KNN	K- Nearest Neighbor
LDL	Low-Density Lipoprotein Cholesterol.
LR	Logistic Regression
MLP	Multi-Layer Perceptron
NCD	Non-Communicable Disease
RBP	Raised Blood Pressure
RBS	Raised Blood Sugar
RF	Random Forest
SPB	Systolic Blood Pressure
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TRG	Triglyceride
WHO	The World Health Organization
WHR	Waist-to-Hip Ratio

List of Publications

The following paper has reported some of the work related to this thesis:

M. M. Owess, A. Y. Owda, M. Owda, & S. Massad, "Supervised Machine Learning-Based Models for Predicting Raised Blood Sugar", International Journal of Environmental Research and Public Health, Vol. 21, Page 840, 21(7), 840, Jul. 2024, doi: 10.3390/IJERPH21070840.

Chapter 1: Introduction

1.1 Overview

NCDs are identified as the leading cause of death globally and are responsible for 74% of the total mortalities worldwide, claiming the lives of 41 million persons annually (The World Health Organization, 2023e). Moreover, 17 million individuals, which represents around 41% of these deaths are premature and are less than 70 years old. The incidence of these premature fatalities is much more noticeable in low and middle-income countries where they account for 86% of premature mortality. Noncommunicable diseases are recognized as chronic diseases, meaning that they often persist throughout the life of the patient diagnosed with them (Bernell & Howard, 2016). These figures are considered alarming emphasizing the necessity of preventing and treating the NCDs. To meet this objective, the various types of NCDs, their causes, factors, symptoms, consequences, and treatment methods must be carefully understood.

For a comprehensive understanding of chronic NCDs, first, it needs to be mentioned that NCDs are not infectious and develop in individuals over time, they don't affect people suddenly (Ackland et al., 2003). NCDs comprise cardiovascular diseases (such as coronary heart disease, stroke, hypertension), diabetes, cancer, and chronic respiratory diseases (Prakash, 2017).

The incidence of chronic diseases is due to multiple reasons either genetic, physiological, environmental, behavioral reasons, or a combination of several of them (Al-Hadlaq et al., 2022). Risk factors related to behavioral habits of individuals, such as using tobacco products, low levels of physical activity, non-balanced diet, and drinking alcohol all have a strong and direct impact on developing NCDs (Ezzati & Riboli, 2013). Nevertheless, these habits are considered modifiable risk factors, meaning that managing these factors can mitigate the risk of developing noncommunicable diseases among adults (Budreviciute et al., 2020). Furthermore, the development of NCDs may be attributed to environmental risk factors that can be in the form of chemical, external physical, biological, or work-related factors (Dhimal et al., 2021). WHO stated that around 23% of yearly mortalities around the globe are due to environmental risk factors including polluted air and water, poor hygiene and sanitation, noise, climate change,

radiation, and exposure to toxic substances (The World Health Organization, 2016, 2024b). Nearly two-thirds of these deaths were due to NCDs that were associated with causes related to environmental risk factors (The World Health Organization, 2024b). Air pollution has been identified as the strongest contributor among the environmental risk factors associated with mortalities from non-communicable diseases (Neira & Prüss-Ustün, 2016). Nonetheless, similar to behavioral risk factors, environmental risk factors are deemed as manageable and can often be preventable (Zarocostas, 2006). In sum, all of these modifiable risk factors often lead to metabolic disorders in the human body system, represented in four key conditions linked with NCDs including raised blood pressure, overweight and obesity, raised blood glucose, and raised cholesterol (Esmailnasab et al., 2012; Kiani et al., 2023; The World Health Organization, 2023f).

In light of this, it has been found that in terms of deaths attributed to metabolic factors worldwide, the top risk factor is raised blood pressure, which represents 19% of global mortalities, followed by raised blood glucose levels and then overweight and obesity (Institute of Health Metrics and Evaluation, 2022; The World Health Organization, 2023e). These findings emphasize the importance of paying attention to and monitoring metabolic disorders associated with raised blood pressure and raised blood sugar levels, which are directly correlated with hypertension and diabetes (Haffner, 2006a; Kyung et al., 2006). Failure to manage these factors properly could result in the progression of chronic conditions (Haffner, 2006b; L. Tuck & B. Corry, 2010). Among these chronic conditions hypertension and diabetes have been identified by evidence in many studies as major prevalent NCDs affecting adults worldwide through the last decades, along with cancer, stroke, and chronic lung disease (Jobe et al., 2024; Peng et al., 2024). There have been numerous studies that have concluded that an increased rate of people around the globe are living with a greater number of morbid conditions in comparison to earlier generations, affecting life expectancy as a direct impact of hypertension and diabetes (Franco et al., 2005; Kang et al., 2017; Laditka & Laditka, 2015; Wang et al., 2023a, 2023b). Across the world, over the last 20 years, there was a double increase in the number of people diagnosed with hypertension aged between 30 and 79, from 648 million in 1990 to 1278 million in 2019, as has the number of people with diabetes, which rose from 108 million in 1980 to 422 million by 2014 (The World Health Organization, 2023c; Zhou et al., 2021). This rapid increase in the incidence of hypertension and diabetes was

very noticeable in low and middle-income countries, with a rate of 80% of reported cases diagnosed with NCDs in developing countries (Kobashi et al., 2024; Mocumbi, 2024; Reddy, 2002). It is expected that this skyrocketed increase in the prevalence of chronic diseases will burden the healthcare and economic systems in low and middle-income countries (Allotey et al., 2014; Gheorghe et al., 2014). Consequently, the poverty rates are predicted to escalate, due to the costs spent by households for the treatment of chronic diseases and their possible health complications (N. Gupta et al., 2021; J et al., 2018). Interestingly, some types of NCDs can be controlled efficiently, detected, and treated at primary healthcare levels, also even avoided if prevention measures are followed, such as diabetes, and cardiovascular diseases including hypertension, and ischemic heart diseases (Demaio et al., 2014; Lenfant, 2001; Miranda et al., 2008; Ndubuisi, 2021; Rohwer et al., 2021; Uwimana Nicol et al., 2018). In contrast, certain NCDs are much more complicated in terms of detection, treatment, and incurred costs, in addition to their higher risk, such as cancer, and chronic respiratory diseases (Atun et al., 2013; Kabir et al., 2022; Nojilana et al., 2014; Subramanian et al., 2018).

Accordingly, it can be concluded that hypertension and diabetes are major NCDs affecting people and causing high mortality rates worldwide, where this health situation can be manageable and preventable through early detection of raised blood pressure and raised blood sugar as their associated metabolic syndromes.

1.2 Problem Statement and Motivation

Hypertension “the silent killer” and diabetes diseases are global health problems that are considered the key major NCDs that cause deaths worldwide (The World Health Organization, 2024c; Uthman et al., 2022). In general, NCDs are considered the main cause of mortalities for over 40 million people per year (The World Health Organization, 2019). In addition, high blood pressure and diabetes are among the most common NCDs affecting adults, with the possibility to affect young as well, however, the prevalence of these diseases increases with age (Centers for Disease Control and Prevention, 2017).

In hypertension cases, people can suffer from high blood pressure without being aware, due to not feeling or recognizing the symptoms, for this reason, hypertension is called the silent killer (Alnasir, 2008; Haldar, 2013). Headaches, irregular heartbeats, nosebleeds, blurring in vision, and ear buzzing are usually the most common symptoms that occur in

patients with raised blood pressure (Balwan & Kour, 2021). Asymptomatic high blood pressure that may go undetected can pose a serious threat to a person's health as it can advance to acute hypertension and lead to convulsion, brain bleeding, coma, and ultimately death (Vaughan & Delanty, 2000).

In diabetes, there are many signs and symptoms, including frequent urination, thirst, fatigue, an increased feeling of hunger, losing weight, and changes in vision (Clark et al., 2007a). Besides these main symptomatic signs of hypertension and diabetes, they can also cause more complicated health issues such as heart attack, kidney failure, eye problems, and stroke, which place a significant strain on the healthcare systems (Stamler, 1991; Taylor & Ward, 2003; Zheng et al., 2017).

Suffering from diabetes and hypertension is attributed to many common risk factors including unhealthy lifestyle behaviors like lack of physical activity, unbalanced diets, obesity, smoking tobacco products, and drinking Alcohol, besides genetic factors (Flack & Adekola, 2020; Soomro & Jabbar, 2024). Studies demonstrate a strong interrelationship between diabetes and high blood pressure, as these two chronic conditions frequently co-occur as comorbidities (Haffner, 2006b; Jobe et al., 2024; Kyung et al., 2006; L. Tuck & B. Corry, 2010). This association has significant public health implications, highlighting the need for integrated care and management strategies to prevent or ameliorate these diseases by early detection and addressing their related risk factors (Asif, 2014; Samadian et al., 2016).

Thus, early detection of the metabolic disorders associated with raised blood pressure and raised blood sugar that characterize hypertension and diabetes is a key component of responding to these health problems (Bauer et al., 2014a; Rampal, 2017). This supports timely diagnosis, providing proper treatment, and the needed healthcare management, to mitigate the advancement of serious chronic diseases (Bauer et al., 2014b; Bromfield & Muntner, 2013). Furthermore, controlling their related risk factors and metabolic disorders can primarily prevent the disease itself not only the subsequent health consequences (Lombard, 2013). Therefore, to prevent these disorders from developing into chronic and complicated health conditions, early screening for diabetes, hypertension, and their associated metabolic disorders is crucial for timely control and management. Screening is the process of rapidly identifying potential diseases or

abnormalities in individuals without symptoms through tests, examinations, or other procedures (Wilson et al., 1968). The purpose of screening tests is to differentiate those who probably have a disease from those who likely do not. However, it is important to note that screening is not intended to provide a definitive diagnosis. Individuals with positive or suspicious results from screening tests must be referred to their physicians for further diagnosis and necessary treatment. Diagnosis, on the other hand, is a more comprehensive procedure that confirms the presence of a disease or condition. Diagnosis involves using symptoms, medical history, and tests, so healthcare professionals can identify a disease definitively, and guide patients to proper treatment (Brody & Waters, 1980).

Based on the preceding evidence, this study presents a data-driven solution that utilizes advanced data science techniques for developing machine learning-based models for the detection and prediction of metabolic disorders related to raised blood sugar and pressure, using their associated risk factors for learning these models. The proposed raised blood pressure and sugar detection models in this study can serve as a screening tool for identifying potential hypertension and diabetes cases.

1.3 Study Objectives

This study uses supervised machine learning algorithms to present prediction models to detect raised blood pressure and raised blood sugar. The objectives of this study are summarized as follows: -

- Explore the health conditions related to raised blood pressure and sugar disorders, including contributing factors, consequences, detection methods, and diagnosis criteria from clinical and medical perspectives.
- Comprehensive review of the literature concerning the machine learning models for predicting metabolic disorders related to raised blood sugar and high blood pressure including the type of datasets used, the variables used as predictors and outcomes, the employed algorithms and techniques, and evaluation of their performance.
- Determine the predictive variables for building the proposed prediction models, by identifying the risk factors for developing raised blood pressure and sugar.

- Design the framework and workflow for building the raised blood pressure and sugar prediction models proposed in this study, by identifying the needed phases and components.
- Validate the effectiveness of using a local dataset comprising all primary and secondary factors leading to the development and onset of raised blood pressure and raised blood sugar in implementing prediction models for these health conditions. The dataset was collected from a population survey for studying the risk factors of NCDs among adults in Palestine.
- Examine different supervised machine algorithms to find the best algorithm to utilize in the proposed prediction models.
- Assess the performance of the prediction models implemented using different machine learning algorithms by evaluating the classification metrics.
- The main objective of this study is to build machine learning-based models that can be used in screening for hypertension and diabetes by detecting individuals with raised blood pressure or blood sugar disorders.

1.4 Significance of the Study

In general, there are great benefits of incorporating data science techniques in the healthcare domain, as it has a great impact in providing undetected findings and understanding relationships between different variables. This can be used as a basis for constructing smart models to automate the diagnosis and detection processes in a computerized approach that might be difficult to accomplish by a human. The proposed study aims to facilitate the screening process of high blood pressure and raised blood sugar syndromes through incorporating machine learning techniques, which serve as a proactive approach to detecting hypertension and diabetes cases, thereby supporting the prevention and suitable management of these health concerns.

The significance of the proposed models in this study is mainly represented in:

- Offering an automated screening for hypertension and diabetes as major NCDs affecting adults, especially in low and middle-income countries. Screening is a major component in the intervention strategies designed to respond to the

increasing incidence rates of NCDs and their mortalities (Christe et al., 2020; Hewson et al., 2024; Prihanti et al., 2022; Solomons et al., 2017).

- The proposed raised blood pressure and blood sugar prediction models offer a low-cost solution for screening practices for hypertension and diabetes. Such modes can be integrated into electronic health records and health information systems.
- Offering the capability to assist healthcare providers in identifying potential future diabetic or hypertensive chronic patients easily and timely, to aid them with the necessary care, preventive measures, and health management methods they need to monitor their health. This might be achieved simply by raising the necessary awareness on how to manage the modifiable risk factors promptly, thus mitigating the risk of developing new hypertension and diabetes cases (Almomani et al., 2021; Anand et al., 2018; Legesse et al., 2022). In this way, individuals can reduce their likelihood of developing hypertension or diabetes by managing the causative risk factors in the early stages of development.
- Ultimately, utilizing this proactive approach of predicting raised blood pressure and sugar in healthcare strategies can enhance overall health and well-being, prevent disability, decrease the burden on healthcare providers, and strengthen the health systems.

1.5 Contribution to Knowledge

- The key contribution of the raised blood sugar and pressure prediction models proposed in this study is its training on a unique first-hand and comprehensive dataset, considering both the number of features and observations, that combines all the relevant risk factors of diabetes and hypertension. The used dataset for training the models includes gender, age, BMI, smoking status, physical activity level, alcohol consumption level, stress level, sugar and salt intake, medical history, blood pressure, blood lipids, and physical measurement variables. This dataset was collected from a study designed particularly to study the risk factors of NCDs, which was used for the first time for modeling purposes, achieving significantly higher accuracy when compared to previous work presented in this field. This research addressed the challenges faced by prior proposed models that

were demonstrated on frequently used datasets containing a small set of feature issues, in addition to data quality issues (Jaiswal et al., 2021; Zhu et al., 2021).

- The proposed raised blood pressure and sugar detection models can serve as a supportive tool to aid in screening for diabetes and hypertension, by identifying adults with raised blood pressure and sugar levels. Thus, facilitating referrals for further diagnosis, and enabling early detection.
- Notably, the raised blood pressure and sugar prediction models proposed in this study offer the advantage of reducing the time and costs used in other standard approaches for hypertension and diabetes screening, especially in developing countries, as the target features in this study are based on the readings of systolic and diastolic blood pressure measurements, and fasting blood sugar test, which are considered low-cost, accessible, and effective methods for hypertension, and diabetes screening.
- The results generated by the proposed models in this study can prompt individuals to take necessary preventive measures and manage risk factors associated with high blood pressure, and diabetes, including adjusting their lifestyle and monitoring their health. As well as this, it identifies potential cases, for undergoing additional tests for timely detection and treatment.
- The proposed raised blood pressure and sugar detection models will assist in improving the public health of communities and their members, by decreasing the rates of premature mortalities resulting from health consequences and diseases associated with high blood pressure, and raised blood sugar, reducing the expenditures of treating these diseases and their complications, and using the financial allocations and human resources in health sectors to cover and treat more difficult health problems.

1.6 Thesis Arrangement

This chapter presented an introduction about the study, the related facts, the research problem and motivation, the study objectives, the significance of the study, and the study's contribution to knowledge. The following chapters in this study are organized as follows:

- **Chapter Two: Background and Related Work.**

This chapter presents background about the research topic, with a detailed literature review of the previous works related to the proposed study.

- **Chapter Three: Methods and Implementations.**

This chapter presents the description of the proposed models, and the used methodology in detail for implementing the proposed study with all the stages, from the data collection stage to obtaining the model outcomes stage.

- **Chapter Four: Results and Discussion.**

This chapter presents the main findings and the experimental results of the constructed models and discusses these results, the performance, and the accuracy achieved after implementing the models.

- **Chapter Five: Conclusion, Future Work, and Recommendations.**

This chapter provides a conclusion about the study, including a summary of the main findings, the key achievements, and an evaluation of the limitations of the study. Also, discuss the future works for enhancing or extending the work done in the study and provide recommendations.

Figure 1.1 Outlines the chapters and main sections of this thesis.

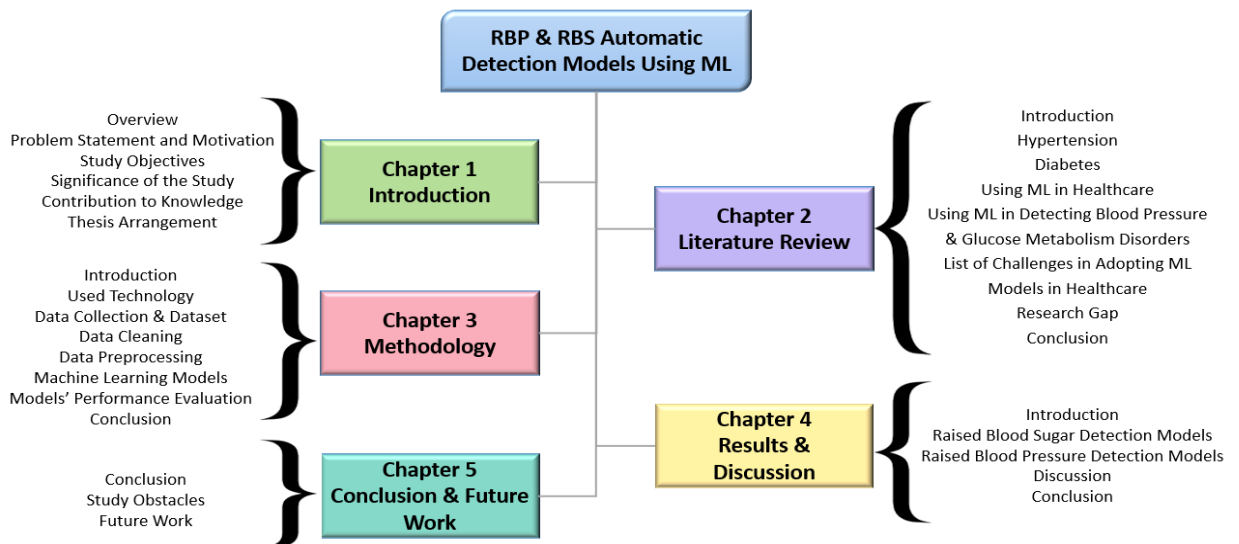


Figure 1.1 Thesis Structure

Chapter 2: Background and Literature Review

2.1 Introduction

This chapter presents a comprehensive literature review of the automatic detection and prediction models for raised blood pressure and sugar using machine learning algorithms. It explores the causative risk factors that contribute to developing these metabolic disorders. In addition to discussing the role of emerging new technological techniques in improving the healthcare sector, particularly in the detection of blood pressure and glucose metabolism disorders, and the associated NCDs, by investigating the related works and studies that have been conducted previously in this area.

Over the past decades, the world has seen a rapid increase in the prevalence of NCDs, coupled with a significantly increased rate of deaths due to their associated risks, and injuries (The World Health Organization, 2023a). Additionally, in 2023 NCDs were identified as the leading cause of death worldwide, accounting for three-quarters of all deaths, and this rate is expected to increase (The World Health Organization and others, 2013). Therefore, NCDs were recognized as a main challenge in the 2030 Agenda for Sustainable Development issued by the United Nations, which set a global goal to reduce the mortality rate worldwide from noncommunicable diseases by one-third by the year 2030, by applying the prevention and treatment protocols (Huck, 2023).

Globally, diabetes, chronic pulmonary disease, cancer, and cardiovascular diseases including blood vessels and heart diseases, hypertension, and stroke are the most common NCDs affecting people around the world (The World Health Organization, 2022a). NCDs are attributed to many risk factors such as physical inactivity, unhealthy diet, unhealthy weight, tobacco smoking, and Alcohol consumption, besides other factors such as air pollution, which are all modifiable (Di Cesare, 2019). Most types of NCDs share four major metabolic disorders including raised blood pressure, hyperglycemia (high blood glucose levels), hyperlipidemia (high levels of fat in the blood), overweight, and obesity (The World Health Organization, 2023g). The leading metabolic disorders that have been identified as causing deaths are raised blood pressure and raised blood sugar (hyperglycemia) (Pan American Health Organization & The World Health Organization, 2021). Raised pressure in the blood vessels is the main abnormality that characterizes

hypertension disease (Schiffrin, 2020). Diabetes is primarily characterized by raised glucose levels in the bloodstream as a metabolism dysfunction (Banday et al., 2020). High blood pressure was recognized as a main risk factor for developing much more complicated chronic diseases and health consequences like heart failure, coronary thrombosis, cardiovascular disease, stroke, and ischemic heart disease (R. Gupta & Xavier, 2018; Kavishe et al., 2015; Kjeldsen, 2018). In addition, add to the fact that diabetes is a chronic disease itself, diabetics are at greater risk of developing cardiovascular diseases, chronic renal disease, cancer, and other communicable diseases such as COVID-19, influenza, and tuberculosis (American Diabetes Association Professional Practice Committee et al., 2024; Jeong et al., 2020; Koye et al., 2018; M.-R. Lee & et al, 2017; Ling et al., 2020). Being a hypertension or diabetes patient is considered risky somehow, due to many reasons, first people might be hypertensive or diabetic patients without being aware, second these diseases are considered chronic, in addition, non-controlling these diseases can lead to serious health complications, such as kidney failure, brain stroke, blindness, damaging blood vessel, clots in heart and blood vessels, nerve damage, then affecting the sensory system, and amputation of limbs (Kifle et al., 2022; Tomic et al., 2022). Consequently, evidence shows that raised blood sugar and high blood pressure are common risk factors for severe comorbidities including retinopathy, neuropathy, nephropathy, non-congenital blindness, chronic renal failure, coronary artery disease, arteriosclerosis obliterans, cardiovascular disease (Chi & Lee, 2022; Long & Dagogo-Jack, 2011; X. Meng et al., 2021; Okosun et al., 2001). This explains why NCDs pose a burden on health systems, both in terms of financial expenditures and demands on healthcare staff (Kankeu et al., 2013; Peer et al., 2020). Another concern regarding these diseases is that they have become an inevitable matter for most people of middle age (Lin et al., 2017).

Research shows that addressing the NCDs, their metabolic disorders, and their risk factors needs collaboration beyond the health sector alone. For an effective response, it is essential to engage key sectors, such as finance, trade, education, agriculture, transportation, and foreign affairs. Their related health threats will continue to incur high costs without urgent and comprehensive action (Pan American Health Organization, 2022).

In response to the integrated multisectoral approach that aims to mitigate NCDs, their causative factors, and their consequences, some researchers and experts from the information technology field worked in collaboration with experts in the healthcare field, to present computerized models for the prediction and detection of hypertension, diabetes or their related metabolic disorders represented in raised blood pressure and glucose metabolisms (Abhari et al., 2019; Kaur et al., 2023). These models had been constructed by studying and analyzing the associated risk factors using data mining and machine learning techniques, as a contribution to strengthening health systems, supporting health workers, assisting them in diagnosis and decision-making processes, as well as improving the health of individuals in society. So, this chapter aims to investigate how previous studies utilized machine learning techniques to contribute to the prevention of NCDs, by using different methods, algorithms, and data types. In addition to investigating the feasibility of implementing automatic detection models by using data from a cross-sectional study.

2.2 Hypertension

High blood pressure, which is known as hypertension is a medical condition that causes the blood pressure to be higher than normal, which happens when the force of blood flow in the body's arteries is too high (World Health Organization, 2024). Globally, around 1.3 billion adults aged 30 to 79 years are diagnosed with hypertension, two-thirds of whom live in countries with low or middle-incomes (The World Health Organization, 2023d). The blood pressure is measured as two numbers, systolic and diastolic blood pressure. The systolic component is defined as the force of blood in the blood vessels that occurs as a result of pumping the blood from the heart out to the body (Magder, 2018; Ogedegbe & Pickering, 2010). The diastolic component of blood pressure is the force that occurs on the vessels after pumping the blood when the heart is at rest in between (Magder, 2018; Ogedegbe & Pickering, 2010). The hypertension guidelines stated by WHO are adopted to evaluate blood pressure, whether it falls in the identified normal range of blood pressure or not. These guidelines stated that hypertension is recognized in case any of the blood measurement components equals to or is higher than the normal threshold, which is 140 mmHg for systolic pressure and 90 mmHg for diastolic pressure (World Health Organization (WHO), 2023). Figure 2.1 shows a simple illustration for understanding the effect of high blood pressure on the heart work mechanism (Shrestha, 2023).

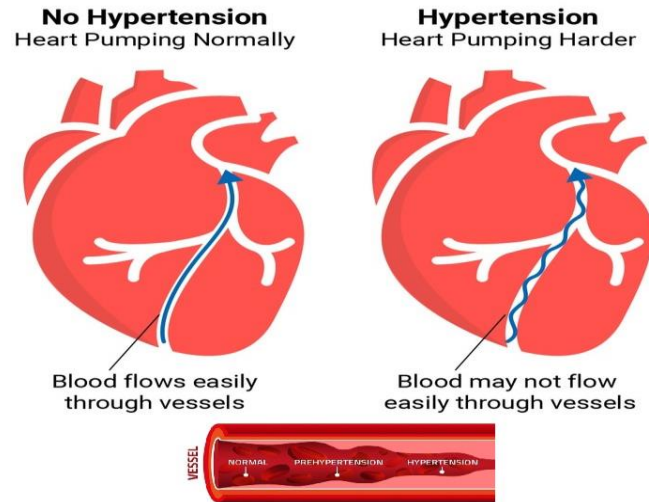


Figure 2.1 Heart and Blood Vessels in Normal Case vs. High Blood Pressure Case (Shrestha, 2023)

The main symptoms that high blood pressure usually causes are severe headaches, chest pain, dizziness, breathing difficulties, nausea, vomiting, hazy vision, anxiety, confusion, ear tinnitus, nosebleeds, as well as irregular heart rhythms, so it's recommended to measure your blood pressure in case of the suffering of any from these symptoms (Q. Li et al., 2023). However, in most cases, high blood pressure does not always cause symptoms, as according to the statistics of WHO nearly 46% of adult hypertension patients do not know about their condition because they don't suffer from hypertension symptoms (World Health Organization (WHO), 2023). Hence, the best method of finding out if you have high blood pressure is to take your blood pressure, especially if one has a family history of blood pressure (Hodgkinson et al., 2011). Among adults to detect the presence of hypertension, at least two readings on two different days are needed, if the systolic blood pressure in both readings is higher than 140 mmHg and/or the diastolic component is higher than 90 mmHg, it is diagnosed as hypertension case (Sakhaee et al., 2017).

Hypertension is attributed to many risk factors either genetic or lifestyle risk factors including the person's age, a family history of hypertension, being diagnosed with other chronic diseases such as diabetes or kidney disease, smoking tobacco products, physical inactivity, harmful consumption of alcoholic drinks, and unhealthy diets like diets high

in salt, processed and fast foods, fat-saturated meals, and low-fiber foods, and overweight (Meher et al., 2023; Pinto & Martins, 2017).

Hypertension can lead to serious health complications, such as heart damage since excessive pressure can lead to hardening of the arteries, which reduces the oxygen and blood flow to the heart, and thus leads to chest pain and problems in heart muscle due to lack of oxygen supply. Blocking the blood flow for a long time damages the heart more, which causes abnormal heartbeat and heart failure that can lead to sudden death. Another possible complication as a result of hypertension is a stroke that is caused by the lack of blood and oxygen supply to the brain due to the blocked arteries. Furthermore, hypertension can damage the kidneys and cause kidney failure (Abiodun et al., 2024; Desai et al., 2021; Marques da Silva et al., 2019). Hypertension is considered a serious health condition that needs follow-up and treatment to control its related symptoms and to avoid any future complications. Moreover, hypertension is considered one of the strongest associated factors with cardiovascular diseases and one of the leading causes of premature death globally which is responsible for around 7.5 million deaths which represents 12.8% of all deaths worldwide (The World Health Organization, 2023b).

However, the symptoms of high blood pressure and its resulting complications can be prevented or at least mitigated by early detection, controlling the associated risk factors, and providing the proper treatment (Flack et al., 2003). Hypertension treatment is not limited to medication only, it can be prevented or ameliorated by changing bad lifestyle habits like stopping smoking, eating healthy diets, doing physical exercises, losing weight, and avoiding stress and anxiety. In addition to measuring the blood pressure regularly, taking the prescribed medication and treating other medical conditions for managing high blood pressure (Almeida et al., 2016; Flack et al., 2003; Pearce & Furberg, 1994). So, the earlier the prevention or treatment procedures are provided, the better in controlling the health situation and contributing to one of the health goals that aims to decrease the prevalence of hypertension globally within the next years.

2.3 Diabetes

Diabetes is a chronic disease characterized by either insufficient insulin production from the pancreas or ineffective insulin use by the body, thus causing the glucose levels in the blood to be raised and irregular (Rahman et al., 2021). Globally, around 422 million

people are diagnosed with diabetes, and approximately 1.5 million deaths yearly worldwide are attributed to diabetes and its associated complications (The World Health Organization, 2023c). There are many signs and symptoms that indicate high blood sugar and the possibility that a person has diabetes, including frequent urination, thirst, constant hunger, weight loss, vision changes, and fatigue (Clark et al., 2007b).

Diabetes is classified into four main types: type 1, type 2, gestational diabetes, and prediabetes (Adler et al., 2021; Alberti & Zimmet, 1998; Thomas & Philipson, 2015). A person with Type 1 diabetes usually has it as a child, a teen, or in their twenties, but it may occur at any age. Generally, type 1 diabetes occurs when your pancreas does not make insulin, so those with type 1 diabetes need to take insulin every day because their pancreas does not produce insulin (Akil et al., 2021; Monaghan et al., 2015). The Centers for Disease Control and Prevention (CDC) estimates that up to 10% of diabetics have this type of diabetes (*Diabetes Quick Facts / Basics / Diabetes / CDC*, 2024). Type 2 diabetes can also occur at any age, but it is more common in people over 40 (Galicia-Garcia et al., 2020). Almost 90% to 95% of people with diabetes have type 2 diabetes (Stumvoll et al., 2005). It occurs when your pancreas doesn't produce enough insulin, or if your body isn't metabolizing it well (Galicia-Garcia et al., 2020). Though type 2 diabetes has historically affected mainly adults, recently its prevalence in children and teenagers has been expanding (Dabelea et al., 2014; Perng et al., 2023). Gestational diabetes occurs when women who do not have diabetes before pregnancy develop it during pregnancy (Plows et al., 2018). It normally disappears after the baby is born, but it increases the risk of developing type 2 diabetes later in life (Volkova et al., 2021). In the case of prediabetes, the blood sugar levels in this stage are higher than they should be, but not yet high enough to get diagnosed as a type 2 diabetes patient (Ferrannini et al., 2011; Weiss et al., 2017). This condition is termed impaired blood sugar (Amelia & Luhulima, 2020). There is a high risk that the person will develop type 2 diabetes in case of having prediabetes, as its name implies (Bergman, 2013).

The main risk factors responsible for causing type 2 diabetes are very common with risk factors for high blood pressure, mainly unhealthy lifestyle behaviors such as lack of physical activity, unbalanced diets, obesity, smoking tobacco products, and drinking alcohol, along with family history and being diagnosed with other chronic diseases. Figure 2.2 illustrates some risk factors and complications associated with type 2 diabetes

(Tiwary et al., 2023). People diagnosed with diabetes have a higher risk of developing severe health consequences like heart attack, renal failure, stroke, periodontal (gum) disease, and damaging blood vessels in the heart, eyes, and nerves which can lead to vision loss, and problems in the foot that can be developed later to disability (Alqadi, 2024; Tiwary et al., 2023; Williams & Airey, 2002).

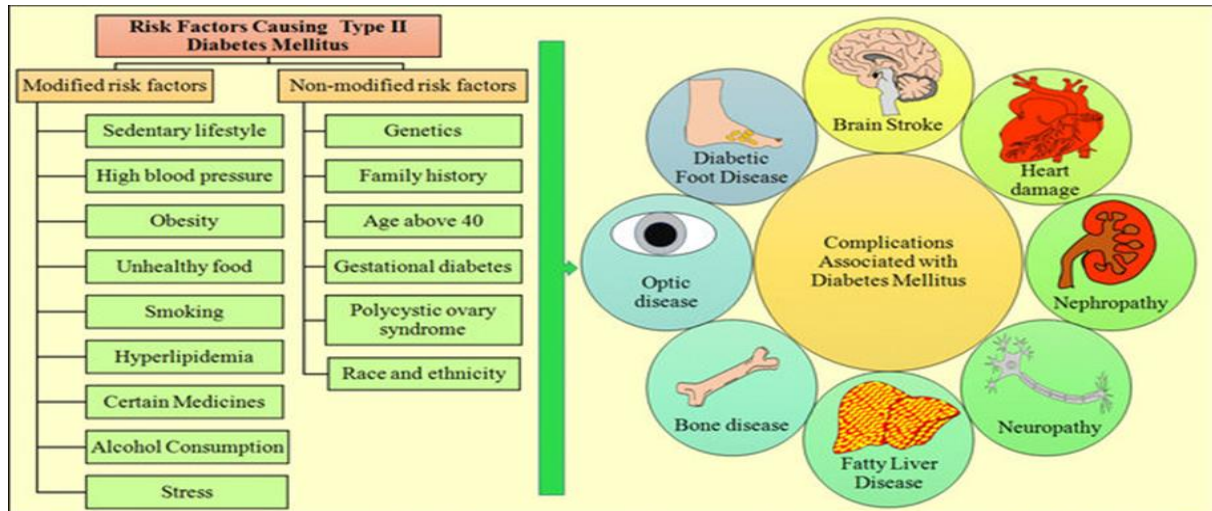


Figure 2.2 Type 2 diabetes mellitus risk factors and complications (Tiwary et al., 2023)

The medical methods typically used to diagnose diabetes include the fasting blood sugar test, hemoglobin A1C (HbA1C) test, or oral glucose tolerance test (OGTT), each type of these tests has a different procedure and threshold for diagnosing diabetes. For the HbA1C test, the threshold is 6.5% or higher, in the case of OGTT the threshold is 200 mg/dL or above (Association, 2020). Regarding the fasting blood sugar test, the normal range is between 70 to 99 milligrams per deciliter (mg/dL), readings between 100 and 125 mg/dL are marked as impaired levels of fasting blood sugar and might be an indication of the condition of prediabetes, while when the fasting blood sugar measures 126 mg/dL or higher, it is recognized as raised blood sugar which may indicate diabetes if the results of the fasting blood sugar for two separate tests exceed this threshold (The World Health Organization, 2024a). The treatment methods for diabetes are different according to case diagnosis, and the diabetes type, either by medication which might be using one or more types, including oral medications drugs or injectable insulin, also changing the person's lifestyle to be healthier is considered a type of treatment (J. D.

Rockefeller, 2015). Maintaining a healthy lifestyle is important to treat or prevent type 2 diabetes and its consequences, and this includes maintaining a healthy weight, engaging in at least 30 minutes of moderate exercise each day, eating a healthy diet, avoiding sugar and saturated fat, and not smoking (Burns & Francis, 2023).

Decreasing the percentage of diabetic patients and controlling diabetes and its complications is one of the main mandates adopted by WHO (Gregg et al., 2023). Several effective measures are being applied to meet this mandate, which include providing scientific guidelines for preventing major NCDs including diabetes, developing standards for diabetes diagnosis and care, raising awareness about the global epidemic of diabetes, commemorating World Diabetes Day on 14 November, and monitoring diabetes risk factors and prevalence (Parija, 2023; The World Health Organization, 2006).

2.4 Using Machine Learning in Healthcare

Many studies have presented machine learning-based models as tools for the automatic detection of some diseases or as computerized diagnosis systems, intending to help healthcare workers and service providers in detecting suspected cases and making more confident clinical decisions (Cuevas-Chávez et al., 2023; Kumar et al., 2023). In light of this, several researches work and automated models that demonstrated the usefulness of utilizing machine-learning-based systems in the healthcare sector have been published, one type of these systems is disease prediction models. The disease detection and diagnosis models were proposed for different diseases, using different algorithms from classical machine learning algorithms to advanced deep learning algorithms and applied to different types of data such as tabular data, images like X-ray, MRI, CT scan, and ultrasound images, unstructured text data, or signals data like electrocardiogram (ECG) and electroencephalography (EEG) signals, which are collected from different sources either from surveys, cohort studies, wearables devices, medical devices and monitors, electronic health records, or disease electronic surveillance systems.

The authors of [150] developed a computerized application that serves as a clinical aid tool to identify patients in a hospital who are recognized to have an expected risk of thrombophlebitis and are not prescribed the needed prophylactic treatment. The developed application supported notifying the hospital doctor of that potential risk, by processing historical data of patients collected in the electronic health records of the

hospital. The machine learning-based application has assured high performance in optimizing the prevention of new thrombophlebitis cases. The study in (Hug et al., 2009) discussed the advantages of integrating a software application in the process of prescribing drugs electronically through the system, to check and validate the side effects of prescribed medications on patients with renal failure automatically. The study utilized historical data collected by a certain cohort study that targeted chronic kidney patients attending Boston hospitals, which included data related to pharmacy drug orders, lab test results, and kidney medication doses. Likewise, another study was conducted to examine the impact of ordering drugs for kidney failure patients electronically via a computer system, using data from an observational study conducted on resident bed patients at the intensive care unit in a hospital in Belgium. The study results concluded that prescribing medication for those patients with chronic kidney failure in the hospital using the electronic system helped significantly in reducing mistakes in drug prescriptions, at an estimated rate three times lower than using the paper approach in prescribing medications (Colpaert et al., 2006). Along with the same goal of engaging the prescription of drugs into electronic systems for patients with renal failure to improve health outcomes, another study investigated the capability of adjusting the prescribed dose and frequency of the ordered medications through that system, the concluded results from that study proven that the engagement of computer systems in the medication prescription process is an effective method, that lead to enhanced health outcomes on patients (Chertow et al., 2001). Moreover, this study concluded that incorporating computer-based systems in the intervention strategies of disease management is a useful method to gain optimal health outcomes.

In the study of (Pereboom et al., 2019), the use of an alert system during treatment with antibiotics gentamicin or vancomycin was examined, to improve adherence to the clinical guidelines of testing plasma concentrations. Implementation of the proposed clinical software data mining-based system, along with the daily review of the patient list, resulted in high compliance with guideline recommendations and optimal therapy for inpatients treated with antibiotics of vancomycin and gentamicin.

Using a series of papers and previous studies conducted in this field between 2012 and 2021, the authors of (Ahsan et al., 2022) provided a comprehensive review of machine learning-based disease diagnosis models. An overview of developed disease diagnosis

systems using machine learning was presented in the study, which analyzed key factors that contribute to these models, including the algorithms used, the types of diseases, the data types, the applications, and the evaluation criteria. The study highlights key findings relating to future opportunities and trends in the field of machine learning-based disease diagnosis. According to the study, previous works that used machine learning to build disease detection models had clearly focused on certain diseases, including diabetes, heart disease, kidney disease, Alzheimer's, Parkinson's, and breast cancer, and the majority of them used deep learning algorithms.

By reviewing all preceding studies, it can be inferred that integrating machine learning techniques into the healthcare sector has proven its efficiency and accuracy in enhancing healthcare services and aiding the health workforce.

2.5 Using Machine Learning in Detecting Blood Pressure and Glucose Metabolism Disorders

Several studies have been conducted with the aim of early detection of metabolic disorders linked to high blood pressure and glucose metabolism disorders, based on machine-learning techniques including supervised learning algorithms, neural networks, or image processing techniques, using different approaches and datasets, which had resulted in good outcomes (Afsaneh et al., 2022; Montagna et al., 2023; Ogunpola et al., 2024; Silva et al., 2022).

Similarly, the authors of (Fitriyani et al., 2019) developed a machine learning-based model for predicting type 2 diabetes and hypertension diseases by integrating a set of machine learning techniques to achieve the purpose of early prediction. The proposed prediction model consists of three main stages including the detection and handling of outliers using the isolation Forest technique, then handling the class imbalance issue using the SMOTETomek oversampling technique, and finally training the disease prediction model using two levels ensemble technique by combining the decision tree, support vector machines, multilayer perceptron, support vector machines in the first level, and logistic regression for the second level. Additionally, the proposed prediction model was applied to four different datasets. Two of these datasets include a subset of features extracted from Dr. Golino's public dataset (H. Golino, 2013b, 2013a) for training the prehypertension and hypertension models, and another one extracted from Dr John

Schorling's dataset (Schorling et al., 2002) for training the diabetes model. The results of the proposed model outperformed the results of previous studies that were implemented on the same tested datasets with the highest accuracy of 100% for one of the datasets that were constructed with the aim of finding the relationship between hypertension, diabetes, and age variables. The lowest accuracy of 75.78% was obtained for the pre-hypertension prediction model. The diabetes prediction models achieved an accuracy of 96.74%, and 85.73% for the hypertension model. Furthermore, the experimental results of this study emphasized that individual age, waist, and hip circumferences, weight, and BMI are considered attributes contributing to hypertension disease.

Based on the study in (Ijaz et al., 2018), a variety of machine learning techniques were used to detect hypertension and type 2 diabetes mellitus in a timely manner. The DBSCAN clustering algorithm was used to detect and handle outliers, the SMOTE algorithm was used to handle class imbalance, and the Random Forest was used to classify detected cases. The detection models were applied to three separate sets of data comprising several variables that are marked as related risk factors in the preceding research works that were conducted. The first dataset consists of diabetic subjects to predict diabetes including the following variables as risk factors stabilized glucose, age, cholesterol to HDL ratio, waist, total cholesterol, systolic blood pressure, gender of subject, HDL, height, hip, diastolic blood pressure, and weight. The second dataset that was used comprised hypertensive male subjects as (yes/ no) binary variable which was derived based on the systolic blood pressure whether it is higher than 140 mmHg then the subject is classified as hypertensive. Also, the following set of features used in the second dataset as risk factors, is obese (yes/no) binary feature, waist-hip ratio, hip circumference, BMI, waist circumference, and age. The third dataset was used to analyze the relationship between diabetes and hypertension, which is demonstrated in a dataset that was collected originally for the purpose of predicting chronic kidney diseases. As the purpose of the developed model is different in that study, the authors used a subset of features consisting of age, blood pressure, is hypertensive (yes/ no) variable, while the used outcome feature indicates whether the subject is diabetic (yes/ no) variable. The best accuracy was achieved when implementing the diabetes prediction model on the first dataset, which achieved an accuracy of 92.56%, while the hypertension prediction model that was applied using the second dataset yielded an accuracy of 76.42%. For the third model that

was constructed to foresee the relationship between diabetes and hypertension, the achieved prediction accuracy was 83.64%. Additionally, the authors concluded that age and blood pressure are significant variables for predicting diabetes.

The authors in the study of (Farran et al., 2013) initiated a prediction framework based on machine learning algorithms for anticipating the risks of high blood pressure and diabetes mellitus (type 2), using a huge dataset collected from EHR for a hospital in Kuwait, containing about 13 million records for diabetes and/or hypertension patients of different ethnicities living in Kuwait. The proposed models in this study were based on using SVM, KNN, logistic regression, and Multi-factor Dimensionality Reduction algorithms while integrating the cross-validation technique using 5 folds. The used set of features for building the prediction models included BMI, age, ethnicity, diagnosis of diabetes, diagnosis of hypertension, and family history of diabetes or hypertension. The proposed model achieved promising results, the SVM-based models outperforms the other used classification models with a value of 81.3%, and 82.4% in terms of accuracy for the diabetes, and hypertension prediction models. Another remarkable finding in the study is that ethnicity was identified as a risk factor for predicting hypertension and diabetes.

As another type of demonstration of the utilization of machine learning algorithms in the NCDs domain, the authors of (Pittoli et al., 2018) proposed an intelligent system that can contribute to the care of NCD patients, by sending strategic messages to the chronic patients on their mobiles. The proposed system was based on using the Bayesian Networks model for processing the NCDs-related risk factors including systolic blood pressure, diastolic blood pressure, total cholesterol, blood sugar, weight, waist measurement, age, and HDL variables. The authors of the study used a simple approach by determining whether the predicted value of a target variable belongs to the normal or high class based on the known normality ranges or cutoff values for the processed variables. Then, accordingly, feedback is sent to the patient that might include a warning or medical advice.

2.5.1 Using Machine Learning in Detecting Glucose Metabolism Disorders

This section discusses various studies related to the prediction of glucose metabolism disorders that involved machine learning-based models for predicting either diabetes, prediabetes, hyperglycemia, or raised blood sugar.

The study in (Patil et al., 2010) presented a model for predicting type 2 diabetes using the K-means algorithm, C4.5 tree-based classification algorithm, and k-fold cross-validation technique, the model was applied to a dataset collected from native American Indian females (PIDD dataset) (UCI Machine Learning and Kaggle, 2016), who were diabetic, the proposed model achieved a high classification accuracy of 92.38%. Using the same dataset, the author of (Mahboob Alam et al., 2019) applied the diabetes prediction model by utilizing Artificial Neural Networks, Random Forest, and K-means clustering algorithms, the ANN-based model achieved the best results among the three implementations with an accuracy of 75.7%, also this study aimed to characterize the relationship between the significant risk factors responsible for causing diabetes using Apriori association rules, which showed that diabetes is strongly associated with body mass index and glucose level. Another comparative study in (Khanam & Foo, 2021) was demonstrated on the PIDD dataset for predicting diabetes using a set of machine-learning algorithms including SVM, Decision Trees, KNN, Random Forest, Naïve Bayes, AdaBoost, Logistic Regression, and Artificial Neural Networks for building the models, each algorithm was tested using two methods, once by splitting into training and testing datasets and another time using k-fold cross validation techniques. The model built using ANN had the highest accuracy score of 88.6%.

The authors of (Kandhasamy & Balamurali, 2015) presented a comparative study for building a diabetes classification model using different machine learning classification algorithms including random forest, SVM, KNN, and Decision Tree J48, the study was demonstrated to the UCI diabetes public dataset consisting of variables of diabetes symptoms (AITBAYEV, 2021), although the experimental results of the proposed model were excellent in terms of accuracy it still needs more testing and investigation by demonstrating a more challenging dataset since it was applied to a dataset of a very small set of features having binary classification only.

The study in (Yahyaoui et al., 2019) presented a machine learning-based diabetes detection model that can be used as a clinical decision support system for assisting healthcare workers in diagnosing diabetes cases, the proposed model was implemented using a set of machine learning algorithms including random forest, SVM, and deep convolutional neural networks (CNN), the diabetes detection model performed the best results by utilizing Random Forest, achieving an accuracy of 83.6% model, also the accuracy was 76.81% using deep CNN, and 65.38% using SVM.

The authors in (Naz & Ahuja, 2020) proposed a deep learning model for diabetes prediction using artificial neural networks, Naïve Bayes, decision tree, and deep learning algorithms, the model has been applied to the American Indian females' diabetes public dataset and resulted in a high accuracy of 98.07%, however, the study didn't discuss the model performance in case of using a low-quality dataset that might contain missing values or unbalanced data.

The study in (H. Wu et al., 2018) proposed a hybrid prediction model for predicting type 2 diabetes, that is based on using K-means and logistic regression techniques. Using the K-means algorithm in this model was helpful in eliminating incorrect clustered data in the preprocessing stage, before applying the logistic regression algorithm for performing the classification. The demonstrated diabetes prediction model achieved an excellent prediction accuracy of 93.9% in comparison with previous works applied to the same dataset. The authors in (X. H. Meng et al., 2013) presented a comparative study for a different set of machine learning algorithms that were used in building models for predicting diabetes based on the related risk factors including age, gender, BMI, family history of diabetes, marital status, education level, stress, sleep, physical activity, diet, in-salt taking, and drinking coffee, the study focused on comparing the performance of ANNs, decision trees, and logistic regression machine learning algorithms in building the diabetes prediction models. Using the decision tree in the diabetes prediction model achieved the best performance results with an accuracy of 77.87% while using the ANN resulted in the lowest accuracy. The proposed model was applied to a balanced dataset collected from Chinese adults including observations from both diabetes patients and normal adults not diagnosed with diabetes.

The authors of (Dinh et al., 2019) developed a machine learning-based model to predict diabetes and coronary heart disease by utilizing a subset of data granted from the National Health and Nutrition Examination Survey (NHANES) (Centers for Disease Control and Prevention, 2023). The NHANES dataset was gathered by ongoing research that had conducted peer interviews, medical checkups, and blood tests for participants. The authors divided the collected NHANES dataset into two subsets one having the diagnosis of diabetes reported by participants. The second subset included subjects having FBS greater or equal to 126 mg/dL and diabetes pre-diagnosis not reported by participants, to be considered diabetic patients. The main features used in this study are physical characteristics (age, waist size, leg length, etc.), dietary intake (sodium, fiber, caffeine intake), demographics (ethnicity and income), and laboratory test results (HDL, LDL, cholesterol, urine). The proposed diabetes prediction model in that study tested several machine learning classifiers, including Random Forest, SVM, Gradient Boosting Trees, and Logistic Regression on the two subsets one time without including laboratory variables as input for the models and another time using laboratory tests. In the next stage, these algorithms were integrated as a combination to develop an ensemble model to enhance the achieved performances. The testing results using the five models of the 4 different cases of the dataset, achieved the best performance using the XGBoost algorithm when applied to the subset that includes the lab test results and reported diabetes diagnosis as the target feature. Another study was presented and utilized the NHANES dataset for predicting prediabetes (Vangeepuram et al., 2021). The proposed prediabetes models employed several machine-learning algorithms including Random Forest, AdaBoost, LogitBoost, Logistic Regression, J48, Naïve Bayes, PART, sequential minimal optimization algorithm (SMO) SVM, and instance-based learner (IBk). The used features for training the prediabetes prediction models were BMI, family history of diabetes, race, hypertension patient, and total cholesterol, while the target categorical yes/no feature was derived from any one of the available tests of either FBS, HbA1c, or 2-hour postprandial glucose test (2hrPG). The Naïve Bayes prediabetes prediction model achieved the best positive predictive value of 74.5%.

The work in (Maeta et al., 2018) presented machine learning-based models for predicting the risk of glucose metabolism disorders using a private dataset collected from employees working in a certain Japanese corporation who underwent annual medical examinations.

The proposed glucose metabolism disorders prediction models were based mainly on using XGBoost and Logistic Regression classifiers, which were trained on a dataset that includes age, sex, BMI, systolic and diastolic blood pressure, triglyceride, HDL, LDL, creatinine, immunoreactive insulin (IRI), two plasma glucose (PG) testing results, post one hour and another one post two hours, and total cholesterol. The OGTT, FBS, and HbA1C test results were used to determine the class of the target feature that represents whether an observation belongs to the glucose metabolism disorders risky group or not. The model yielded the highest performance when the XGBoost algorithm was utilized.

Another study proposed a prediction model for diabetes using a dataset collected from a hospital patient in Thailand, the collected dataset consists of related risk factors such as gender, age, BMI, weight, height, blood pressure measurements, history of diabetes, and hypertension family history, alcohol and tobacco consumption. The proposed prediction model was built using several classification algorithms that are based on supervised learning, such as Decision Tree, Artificial Neural Networks, Logistic Regression, and Naive Bayes. The performance of the model utilized using the Random Forest classifier was the best which outperformed other models with an accuracy of 85.56% (Nai-Arun & Moungrmai, 2015).

Table 2.1 presents an overview of the previous works that were conducted to predict metabolic disorders associated with raised blood sugar. The table summarizes the used dataset, predictor features, target feature, algorithms employed, and the best-achieved accuracy.

Table 2.1 Research findings related to machine learning prediction model for glucose metabolism disorders.

Reference	Dataset	Features	Target Feature	Algorithms	Best Model	Outcome%
(Fitriyani et al., 2019)	Dr. John Schorling dataset	total cholesterol, HDL, cholesterol to HDL ratio, age, gender, weight, systolic and diastolic blood pressure, waist and hip circumferences	HbA1c ≥ 7 (yes/ no)	2 levels Ensemble approach	Ensemble approach	Accuracy 96.74
(Ijaz et al., 2018)	Dr. John Schorling dataset	age, cholesterol to HDL ratio, waist, total cholesterol, systolic blood pressure, gender of subject, HDL, height, hip, diastolic blood pressure, weight	HbA1c ≥ 7 (yes/ no)	RF	RF	Accuracy 92.56
(Farran et al., 2013)	Privately collected EHR dataset	BMI, age, ethnicity, presence of hypertension, diabetes/ hypertension family history	Reported diabetes diagnosis (yes/ no)	SVM, KNN, LR	SVM	Accuracy 81.3
(Patil et al., 2010)	PIDD	number of pregnancies, glucose, blood pressure, skin thickness,	Diabetic class (yes/ no)	K-means, C4.5 tree-based, k-	Hybrid model (K-	Accuracy 92.38 Sensitivity 90.38

		insulin, BMI, diabetes pedigree function, age		fold cross-validation	means, C4.5 tree-based, k-fold cross-validation)	Specificity 93.29
(Mahboob Alam et al., 2019)	PIDD	age, number of pregnancies, glucose, diabetes pedigree function, blood pressure, skin thickness, insulin, BMI	Diabetic class (yes/ no)	ANN, RF, K-means	ANN	Accuracy 75.7
(Khanam & Foo, 2021)	PIDD	number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age	Diabetic class (yes/ no)	SVM, DT, KNN, RF, AdaBoost, NB, LR, ANN	ANN	Accuracy 88.6
(Kandhasamy & Balamurali, 2015)	UCI Diabetes Dataset	age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity	Diabetes patient (yes/ no)	RF, SVM, KNN, and DT J48	DT J48 (with noisy data) RF (without noisy data)	Accuracy 73.82 Accuracy 100.0
(Yahyaoui et al., 2019)	PIDD	number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age	Diabetic class (yes/ no)	RF, SVM, CNN	RF	Accuracy 83.6
(Naz & Ahuja, 2020)	PIDD	number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age	Diabetic class (yes/ no)	NB, DT, ANN, DL	DL	Accuracy 98.07
(H. Wu et al., 2018)	PIDD	number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age	Diabetic class (yes/ no)	K-means and LR	Hybrid model (K-means and LR)	Accuracy 93.9
(X. H. Meng et al., 2013)	Privately collected dataset	age, gender, BMI, family history of diabetes, marital status, education level, stress, sleep, physical activity, diet, in-salt taking, and drinking coffee	Reported diabetes diagnosis (yes/ no)	ANN, DT, LR	DT	Accuracy 77.87 Sensitivity 80.68 Specificity 75.13
(X. H. Meng et al., 2013)	NHANES	age, waist size, leg length, sodium, fiber, caffeine intake, ethnicity and income	Reported diabetes diagnosis (yes/ no)	LR, SVM, RF, XGBoost, Ensemble	XGBoost	ROC AUC 86.2 Precision, Recall, F1-score 78.0
(X. H. Meng et al., 2013)	NHANES	age, waist size, leg length, sodium, fiber, caffeine intake, ethnicity and income, HDL, LDL, cholesterol, urine	Reported diabetes diagnosis (yes/ no)	LR, SVM, RF, XGBoost, Ensemble	XGBoost	ROC AUC 95.7 Precision, Recall, F1-score 89.0
(X. H. Meng et al., 2013)	NHANES	age, waist size, leg length, sodium, fiber, caffeine intake, ethnicity, and income	FBS \geq 126 (yes/ no)	LR, SVM, RF, XGBoost, Ensemble	Ensemble	ROC AUC 73.7 Precision, Recall, F1-score 68.0
(Dinh et al., 2019)	NHANES	age, waist size, leg length, sodium, fiber, caffeine intake, ethnicity and income, HDL, LDL, cholesterol, urine	FBS \geq 126 (yes/ no)	LR, SVM, RF, XGBoost, Ensemble	XGBoost	ROC AUC 80.2 Precision, Recall, F1-score 68.0
(Vangeepuram et al., 2021)	NHANES	BMI, family history of diabetes, race, hypertension, cholesterol	FBS \geq 100, or 2hrPG \geq 140, or HbA1C \geq 5.7% (yes/ no)	RF, AdaBoost, LR, J48, NB, PART, SMO, IBk, LogitBoost	NB	Accuracy 74.5

(Maeta et al., 2018)	Privately collected dataset	age, sex, BMI, blood pressure, triglyceride, HDL, LDL, creatinine, total cholesterol, FBS, HbA1C, IRI, PG	FBS \geq 100, or 2hrPG \geq 140, or HbA1C \geq 5.7% (yes/ no)	LR, XGBoost	XGBoost	ROC AUC 78.0
(Nai-Arun & Mounngmai, 2015)	Privately collected HER dataset	gender, age, BMI, weight, height, blood pressure measurements, presence of diabetes, and hypertension family history, Alcohol and tobacco consumption	Diabetic class (yes/ no)	DT, ANN, LR, NB	RF	Accuracy 85.56

2.5.2 Using Machine Learning in Detecting Blood Pressure Disorders

Researchers had proposed several studies that aimed to predict blood pressure, using various types of datasets including either clinical variables, physiological variables, or both. The prediction models proposed in these studies are developed through several machine learning algorithms and techniques, which are constructed as classification or regression problems.

The work in (López-Martínez et al., 2018) presented a machine-learning model for detecting hypertension patients and evaluating the association between the related risk factors. The model was based mainly on the logistic regression classification algorithm and demonstrated a dataset including a set of the related clinical variables that are considered risk factors for developing hypertension, the used dataset in the proposed model was a public dataset (NHANES) that is collected by a population survey conducted by the National Center for Health Statistics in the united states for the purpose of assessing the health, and nutritional status of the population in the USA (Johnson et al., 2013). The evaluation of the proposed hypertension detection model using different performance metrics yielded good results, achieving 77%, 68%, and 73% for sensitivity, specificity, and AUC metrics respectively. In this study also, the authors concluded that individuals' weight, age, and gender are considered determinant factors for the detection of high blood pressure. The authors in (Patnaik et al., 2018) proposed a model for predicting hypertension by utilizing a set of different machine-learning algorithms and then evaluating the outcomes and performance of each algorithm. The used machine learning classification techniques in that study were logistic regression, Random Forest, Support Vector Machine, Naïve Bayes, and Multilayer Perceptron. The model was implemented using a dataset collected by the Korean National Health Insurance

Corporation that consists of electronic health records of general medical annual checkups for a group of patients (Seong et al., 2017). The processed variables in the used dataset were age, gender, BMI, systolic blood pressure, diastolic blood pressure, cholesterol, triglycerides, LDL, HDL, microalbumin, and urine analysis. The performance evaluation results indicate that the model performed the best results using the Support Vector Machine classification algorithm, achieving an accuracy of 80.23% and 89.77% for AUC. The study in (Ye et al., 2018) presented a model for predicting the 1-year risk of hypertension using a dataset of more than 1 million electronic health records that contain patients' clinical information, health conditions, patients' history, clinical measures, and demographic factors, that were collected from hospitals and health centers in a state in the USA between 2013 and 2015. The model utilized a set of powerful machine-learning techniques for preparing the dataset and training the model. In the first stage of data preprocessing, the KNN imputation method was used for handling missing values in the dataset, and then for feature selection, the univariate correlation between the independent variables and the target variable was used to eliminate irrelevant features. For training the prediction model the XGBOOST classification algorithm was used. The AUC of the hypertension risk prediction model was 0.917 in the retrospective cohort and 0.87 in the prospective cohort.

Along the same lines, hypertension prediction models were implemented using deep learning algorithms as a demonstration of using advanced techniques in machine learning. For instance, the authors of (López-Martínez et al., 2020) presented a study about a prediction model they had developed for the detection of hypertension that was implemented by artificial neural networks deep learning algorithm as an advanced version of their previous work that was based on using the Logistic Regression classification supervised learning algorithm (López-Martínez et al., 2018), both proposed models were tested on the same NHANES dataset. Using the ANN shows a slight improvement in the model performance, obtaining 40%, 87%, 57.8%, and 77% for sensitivity, specificity, precision, and AUC metrics respectively. Similarly, the study in (Lafreniere et al., 2017) proposed a deep learning-based model for predicting hypertension using the feedforward neural network (FNN) algorithm. The model training and testing were done on a big dataset that included over 200,000 observations of clinical data related to patients, one group of data was collected from hypertension patients, and the second one

related to referred patients who are not hypertensive and not having another type of NCDs. The collected dataset comprised of age, gender, weight, BMI, measurements of systolic and diastolic blood pressure, cholesterol, triglyceride blood lipids, HDL, and LDL, which are marked as associated risk factors, in addition to other blood test results like urine albumin-creatinine ratio, and micro-albumin. The hypertension prediction model was developed using Artificial Neural Networks, and achieved good performance results up to 82% in terms of accuracy. Other studies also utilized deep learning techniques in developing applications for providing healthcare services remotely. Likewise, the study proposed in (Kwong et al., 2018) aims to predict systolic blood pressure and send an alarm to the application users if the predicted value is recognized as alarming according to an identified criterion that determines the normal systolic blood pressure according to age. The approach followed in the proposed study is highly useful for individuals who do not measure their blood pressure regularly. There are two models proposed by the authors in this study, the first model is implemented using perceptron neural networks with four hidden nodes, and the other one uses radial basis neural networks with five hidden nodes. The proposed models were trained using a dataset that includes variables related to systolic blood pressure, gender, age, BMI, smoking status, physical activity level, alcohol consumption level, stress level, and salt intake, the dataset consisted of records collected from 498 individuals. The model implemented using the radial basis function neural networks showed 91.06% accuracy, while the Multilayer Perceptron-based model was able to reach 94.28% accuracy. Some studies utilized complex deep-learning algorithms in combination with signals datasets for presenting more advanced hypertension prediction models, such as the study in (Tjahjadi et al., 2020) that aims to classify blood pressure into three classes norm tension, prehypertension, and hypertension. The authors in this study used the bidirectional LSTMs technique of recurrent neural networks (RNN) to train the blood pressure classification model, by taking a dataset of photoplethysmography (PPG) signals as an input for training the model. The achieved performance of their model was promising with an Accuracy of 97.33%, sensitivity of 100%, and 94.87% for specificity. Another work in (Tjahjadi & Ramli, 2020) was presented as an enhanced version of the previous study in (Tjahjadi et al., 2020) for the purposes of improving the model accuracy of the blood pressure classification model and reducing the time for training the model. In this study, the KNN

algorithm was used for training the model instead of RNN and using the same dataset of PPG signals. The performance of the enhanced model was improved, resulting in 100%, 100%, and 90.8% for the F1-score metric for the three identified classes of normal tension, prehypertension, and hypertension, in comparison with the previous proposed work that resulted in 97.29%, 97.39%, and 93.93% for F1-score.

Another study was proposed to predict high blood pressure by utilizing the Classification Trees machine learning technique. The predictor variables that were used in that study were physical measurements of participants encompassing height, weight, hip circumference, and waist circumference, which belong to two sets of participants, men, and women (H. F. Golino et al., 2014). The outcomes of the model were good in comparison to the previous models developed using usual classification algorithms, which is an advantage of using the Classification Trees approach that contributed to determining the best settings of cut-offs for the predictor input features. Employing the approach played a part in enhancing the performance of the model predicting the outcome variable. As mentioned earlier, the model outcomes are good but not high enough, resulting in 45.65% as a true positive rate, and 65.15% for the true negative rate from testing the model on the dataset of female subjects, and a true positive rate of 58.38%, and true negative rate of 69.70% when the same testing is applied using the dataset consisting of male subjects. Nevertheless, the researchers involved in the study announced that the implemented model needs further review and development to resolve a set of limitations that exist in their model, due to problems in the methodology of sampling and data collection in the used dataset. The authors suggested that resolving these issues can support model generalization.

The authors in (Teimouri et al., 2016) presented a comparative study for building a diagnosis model for high blood pressure among diabetes patients. The proposed hypertension diagnosis model was designed using several machine learning algorithms, Decision Tree, ANN, KNN, Support Vector Machine, and Logistic Regression, and demonstrated to a dataset collected from diabetic patients in Iran. The used dataset included age, gender, BMI, family history of diabetes, systolic and diastolic blood pressure, FBS, triglyceride, total cholesterol, HDL, creatinine, urea, and albumin. The model performance was evaluated using Youden's index which obtained a value of 68% as the best result for the model, which was accomplished using the Decision Tree that

utilized without misclassification cost scenario. The researchers in (T. H. Wu et al., 2014) introduced a regression model that was developed using machine learning techniques to estimate systolic blood pressure. The Artificial Neural Networks technique was chosen to demonstrate the regression model, which was applied to a privately collected dataset consisting of 498 observations and a set of variables known as risk factors related to hypertension. The used dataset included BMI, age, physical exercise level, alcohol consumption level, stress level, salt intake level, smoking status, and the measurement of systolic blood pressure as the target feature in the prediction model. The used dataset was divided into two subsets one including subjects belonging to male participants, the other to female participants. The proposed model may contribute to the identification of future risks of hypertension and cardiovascular disease promptly and provide healthcare practitioners in obtaining another measurement as an additional reference for a more accurate evaluation of blood pressure. The model's best performance was obtained when utilizing the back-propagation neural network, which resulted in 51.9%, and 52.5% in terms of mean absolute error that is less than 10mmHg when applying the model to the men, and women groups respectively. The experimental results were not outstanding but within the acceptable rate, and were higher than the same model utilized using Radial Basis Neural Networks.

Within the same scope, the authors in (Tayefi et al., 2017) presented a study to identify the determinant risk variables for hypertension, their study was based on building two machine learning models, both based on using the CART decision tree algorithm and applied to two datasets, each dataset comprises a different set of risk factors. The used dataset in this study was obtained from a cross-sectional study. The variables set used in the first implemented model were age, gender, BMI, marital status, education level, occupation status, physical activity level, smoking status, depression and anxiety status, LDL, triglyceride, total cholesterol, FBS, uric acid, and Hs-CRP as a new feature that has not been used previously in the related preceding studies. For the second model, the variables set comprises age, gender, and a set of blood count parameters involving WBC, RBC, HGB, HCT MCV, MCH, PLT, RDW, and PDW. Model1 in this study achieved higher performance results in terms of accuracy, sensitivity, specificity, and AUC metrics of 73%, 63%, 77%, and 0.72, respectively. The findings in this study indicate that the variables used in building the first model are more relevant and significant in predicting

hypertension. The second model obtained the following values for the same metrics 70%, 61%, 74%, and 0.68, respectively. The study of (Owess et al., 2023) proposed a decision support system to detect disorders in blood pressure among adults, involving both hypertension and hypotension cases. The model employed a set of powerful machine learning algorithms that are based on the supervised learning approach, like Random Forest, Decision Tree, and XGBoost classifiers. The model was applied to a public dataset originally collected to study the risk factors of coronary heart disease (National Heart Lung and Blood Institute, 2023), which also shares a set of factors contributing to the disease of blood pressure. The proposed prediction model for predicting disorders in blood pressure performed the optimal highest accuracy of 85.81% by employing the Random Forest algorithm in implementing the model, along with integrating the tenfold cross-validation method.

Within the same scope of predicting blood pressure, some researchers proposed their models using live data collected from monitors or wearable devices that were connected to the participants in these studies. Most of these studies aimed to develop regression models using datasets that consist of signals. The study in (Zhang et al., 2018) presented a research work that was conducted by a group of machine learning experts for estimating systolic and diastolic blood pressures. The research was done by collecting multiple biological markers, gathered regularly using a CM400 device, for training the model that is developed as a machine learning-based solution relies mainly on the Classification and Regression Trees (CART) algorithm. The processed features used for performing the blood pressure regression task were ECG signal, PPG, heart rate (HR), Pulse transit time (PTT), and oxygen saturation (SPO2). It was concluded that PTT and HR are the most significant variables in estimating blood pressure. The proposed model utilized the cross-validation method, which contributes to avoiding overfitting and yielding optimized results in contrast to other traditional regression models used for estimating blood pressure, that utilized Ridge Regression, Linear Regression, the Support Vector Machine, and Neural Network algorithms. The model achieved an accuracy of 90% in estimating blood pressure. The researchers in (Ferdousi et al., 2021) proposed a machine learning model for the early prediction of NCDs by processing dynamic data collected from wearable devices, for experimental purposes the model was first trained on a dataset collected from diabetes patients to predict diabetes only as a demonstration for predicting

other NCDs, then the model was tested on a dataset collected artificially from wearable sensors. The paper in (Sideris et al., 2016) presented a regression model to predict blood pressure, by utilizing a dataset obtained through pulse oximetry, and employed the Artificial Neural Networks for learning the prediction model. Even though the testing results were acceptable, a technical limitation was detected in the model, that is related to the processing of live data which caused slowness in the learning process within the layers of neural networks.

In the study of (X. Li et al., 2017), the authors examined the use of Recurrent Neural Networks with the contextual layers in the application of predicting blood pressure on a dataset that involved historical biological measurements such as heart rate, and blood pressure measurements, anthropometric measurements, age, sex, and other information about medical history for the participant. The biological measurements were taken from a wireless monitor connected to a mobile device for measuring blood pressure and sending the taken measurements to a server remotely for storing the data. However, obtaining an accurate estimation for blood pressure by the proposed model requires that historical measurements were collected previously on the data server, which might threaten the security, accuracy, and performance of the model.

Another domain that is related to blood pressure prediction and was found that machine learning techniques can contribute to, is hypertension control transitions. According to the authors of (Sun et al., 2014), machine learning-based models can be useful in monitoring the management of hypertension patients whether their blood pressure is controlled or not. The proposed model in that study was applied to a dataset collected from HER related to a group of hypertensive patients enrolled in a medical center comprising of age, gender, race, BMI, body mass index, systolic and diastolic blood pressure, creatinine, presence of diabetes or other chronic disease, and the used medications. The model was implemented using the Random Forest classifier which achieved an accuracy of 77.3%.

The author of (Martinez-Ríos et al., 2021) presented a review of the machine learning-based applications for estimating blood pressure and detecting hypertension, in addition to studying the types of data used within previous works including data extracted from demographic, and clinical datasets, signal data such as PPG and ECG, or datasets

encompassing multiple types of data collected from clinical and physiological sources. As it was found through this comparative study, the previous studies that presented hypertension detection models were proposed as classification problems using clinical data, while the proposed models for estimating blood pressure were demonstrated on physiological signals datasets for training the regression models. The authors provided a recommendation for researchers to construct blood pressure prediction models using datasets that involve both clinical data and PPG waveforms (photoplethysmography), as there is a lack of such models, despite the known correlation between these types of data.

Table 2.2 presents an overview of the previous works that were conducted to predict blood pressure, specifically focusing on those relevant to the blood pressure prediction model proposed in this thesis, in terms of the used features and classification type of prediction. The table summarizes the used dataset, predictor features, target feature, algorithms employed, and the best-achieved accuracy.

Table 2.2 Research findings related to machine learning prediction models for blood pressure

Reference	Dataset	Independent Features	Target Feature	Prediction Algorithms	Best Model	Outcome%
(Fitriyani et al., 2019)	Dr. Golino's public dataset	age, waist, and hip circumferences, obesity status, BMI, WHR	SBP \geq 140 hypertensive (yes/ no)	2 levels Ensemble approach	Ensemble approach	Accuracy 85.73
(Fitriyani et al., 2019)	Dr. Golino's public dataset	age, waist, and hip circumferences, obesity status, BMI, WHR	SBP \geq 120 prehypertensive (yes/ no)	2 levels Ensemble approach	Ensemble approach	Accuracy 75.78
(Ijaz et al., 2018)	Dr. Golino's public dataset (Men)	obesity status, WHR, hip circumference, BMI, waist circumference, age	SBP \geq 140 hypertensive (yes/ no)	RF	RF	Accuracy 76.42
(Farran et al., 2013)	Privately collected HER dataset	BMI, age, ethnicity, history of diabetes, diabetes/ hypertension family history	Reported hypertension diagnosis (yes/ no)	SVM, KNN, LR	SVM	Accuracy 82.4
(López-Martínez et al., 2018)	NHANES	age, gender, race, BMI, cigarette smoking, kidney disease presence, diabetes	SBP \geq 140 (yes/ no)	LR	LR	Sensitivity 77 Specificity 68 ROC AUC 73
(Patnaik et al., 2018)	NHIC	age, gender, BMI, systolic blood pressure, diastolic blood pressure, cholesterol, triglycerides, LDL, HDL, microalbumin and urine analysis	Reported hypertension diagnosis (yes/ no)	LR, RF, SVM, NB, MLP	SVM	Accuracy 80.23 ROC AUC 89.77
(Ye et al., 2018)	Privately collected dataset	clinical information, health conditions, patients' history, clinical measures, demographic factors	hypertensive (yes/ no)	XGBOOST	XGBOOST	ROC AUC 91.7

(López-Martínez et al., 2020)	NHANES	age, gender, race, BMI, cigarette smoking, kidney disease presence, diabetes	SBP \geq 140 (yes/ no)	ANN	ANN	Sensitivity 40 Specificity 87 Precision 57.8 ROC AUC 77
(Lafreniere et al., 2017)	CPCSSN	gender, BMI, SBP, DBP, LDL, HDL, triglyceride, cholesterol, urine albumin-creatinine ratio, and micro-albumin	hypertensive (yes/ no)	FNN	FNN	Accuracy 82
(Kwong et al., 2018)	Hypertension Case Study Dataset	blood pressure, gender, age, BMI, smoking status, physical activity level, alcohol consumption level, stress level, and salt intake	Continuous SBP	RBNN, MLP	MLP	Accuracy 94.28
(Tjahjadi et al., 2020)	Privately collected Dataset	PPG signals	Reported Diagnosis Class (NT, PHT, HT)	RNN	RNN	Accuracy 97.33 Sensitivity 100 Specificity 94.87
(Tjahjadi & Ramli, 2020)	Privately collected Dataset	PPG signals	Reported Diagnosis Class (NT, PHT, HT)	KNN	KNN	F1-score 90.8
(H. F. Golino et al., 2014)	Dr. Golino's public dataset (Men)	BMI, height, waist, WHR, ratio	SBP \geq 140 hypertensive (yes/ no)	CART	CART	Sensitivity 58.38 Specificity 65.15
(H. F. Golino et al., 2014)	Dr. Golino's public dataset (Women)	BMI, height, waist, WHR	SBP \geq 120 prehypertensive (yes/ no)	CART	CART	Sensitivity 45.65 Specificity 65.15
(Teimouri et al., 2016)	Privately collected dataset	age, gender, BMI, family history of diabetes, systolic and diastolic blood pressure, FBS, triglyceride, total cholesterol, HDL, creatinine, urea, and albumin	SBP \geq 130 hypertensive (yes/ no)	DT, ANN, KNN, SVM, LR	DT	Youden Index 68
(T. H. Wu et al., 2014)	Privately collected dataset (Men)	BMI, age, physical exercise level, alcohol consumption level, stress level, salt intake level, smoking status	Continuous SBP	RBNN, BPNN	BPNN	MAE 51.9
(T. H. Wu et al., 2014)	Privately collected dataset (Women)	BMI, age, physical exercise level, alcohol consumption level, stress level, salt intake level, smoking status	Continuous SBP	RBNN, BPNN	BPNN	MAE 52.5
(Tayefi et al., 2017)	MASHAD study dataset	age, gender, BMI, marital status, education level, occupation status, physical activity level, smoking status, depression and anxiety status, LDL, triglyceride, total cholesterol, FBS, uric acid, Hs-CRP	Reported hypertension diagnosis (yes/ no)	CART	CART	Accuracy 73 Sensitivity 63 Specificity 77 AUC 72
(Owess et al., 2023)	Framingham public dataset	sex, age, BMI, cholesterol level, heart rate, glucose level	SBP \geq 140 < 90 DBP \geq 90 < 60 Abnormal BP (yes/ no)	RF, DT, XGBoost	RF + 10-CV	Accuracy 85.81

2.6 List of Challenges in Adopting Machine Learning Models in Healthcare

Although machine learning and data mining techniques proved their usefulness, and importance in providing automated smart solutions for a wide range of applications, utilizing them in healthcare applications still has a set of limitations and challenges, that need to be taken into consideration when adopting machine learning-based models. These challenges can be outlined as follows: -

- One of the significant challenges of utilizing machine learning techniques in developing clinical/ medical models for risk predictions is related to data availability, as machine learning models are data-driven models that rely on the data as an input for learning the models. The availability of data was determined as one of the top challenges concerning data. Although it's the 21st century there is still an issue in collecting case-level health records either due to collecting data in aggregated forms only for reporting purposes, or the use of conventional paper forms, especially in developing countries (Mohd Faizal et al., 2021).
- Another identified challenge related to data is data completeness. In the case of data obtained from electronic health records, there is an obvious issue of incompleteness, as this data is usually collected from patients who come for treatment, and specific needed metrics are collected by clinicians at the patient visit (Goldstein et al., 2017).
- Quality of health data is considered a key concern when developing machine learning models as it can affect the model performance, which includes a range of common issues such as heterogeneity, ambiguity, noise, duplication, and missing values of processed health data (Miotto et al., 2017).
- Another related concern is data accessibility where privacy and security of health data must be considered, as it includes personal sensitive information of individuals (Mohd Faizal et al., 2021).
- It is crucial to note that machine learning-based predictions are also often not supported by reasons unless applying algorithms like decision trees that support intuition (Price et al., 2019).

Not considering all these concerns and challenges when utilizing machine learning techniques in implementing healthcare-supportive models might have drawbacks and could be harmful (Castillo & Kelemen, 2013).

2.6.1 Current Problems and Limitations in Hypertension and Diabetes Risk Prediction Models

Most of the proposed studies for predicting hypertension and diabetes risk through machine learning techniques are implemented using datasets from cohort studies or using the same dataset but applying different machine learning algorithms, such as the study proposed in (Kandhasamy & Balamurali, 2015) despite it yielded excellent experimental results, the authors concluded that proposed model still needs to be tested and investigated further by demonstrating its effectiveness on a more challenging dataset since its application was limited to only a very small set of features with binary classification.

Another issue that is related to datasets is the outliers, as well as class imbalance, which has been demonstrated in several studies from the past literature that these two issues are common problems in medical datasets, that pose challenges to classification systems and can affect the performance and the accuracy of machine learning-based models (H. F. Golino et al., 2014; Ijaz et al., 2018; Sakr et al., 2018). Besides the issue of missing values which represents a major challenge in building models for chronic disease prediction (Du et al., 2020).

Among the challenges in obtaining accurate prediction results was the high degree of overlap between classes according to the study in (Deberneh & Kim, 2021) that presented a predictive model for type 2 diabetes. An additional limitation in the proposed models is not taking features related to socio-economic or mental health variables into consideration when utilizing hypertension and diabetes risk prediction models (Ye et al., 2018).

For hypertension and diabetes prediction models that are based on deep learning, the size of the datasets is considered an important factor for successful training of the model, as it is not expected to achieve high-performance results in case of using data of small volume and works more properly in case of training models using larger datasets (Martinez-Ríos et al., 2021).

Many studies for predicting hypertension or diabetes were proposed using data from electronic health records (EHR), such as the study in (Sakr et al., 2018). Despite that EHR data is easy to obtain as it is pre-collected through patients' visits for treatment in healthcare centers or hospitals, utilizing it for building hypertension and diabetes prediction models is challenging as EHR data tends to be sparse, messy, and noisy in general.

2.7 Research Gap

After reviewing the art of literature for works proposed in the field of predicting blood pressure and glucose metabolism disorders, it can be inferred that using machine learning can contribute strongly in this domain.

Section 2.5 demonstrated that numerous studies have presented machine learning and data mining-based models for detecting disorders and health issues related to blood pressure and glucose metabolism. As reviewed earlier, some of these studies have a set of recognized limitations and drawbacks, which can be summarized as follows: -

- Most of the models and applications proposed in previous studies were presented as tools for predicting either diabetes or hypertension. These models often relied on using target variables that are not recognized from the medical side as diagnostic features for providing a definitive diagnosis. Such as using a single measurement of fasting blood sugar, or one reading of blood pressure, which contradicts the standard guidelines for diagnosis and detection. Likewise, in other studies, the target variables were based on the reported diagnosis of hypertension or diabetes, which requires studying the onset of these reported diseases over a specific period and analyzing the associated risk factors during the same period, which wasn't met. This poses a challenge to the adoption of their models in serving the healthcare sector. The proposed raised blood pressure and sugar presented in this study are based on objective criteria in identifying these cases, by taking into consideration the standard global guidelines and protocols in the definition of the health conditions related to high blood pressure and raised blood sugar. Thus, the proposed models are designed with an emphasis that these models serve as a screening tool, not an automatic diagnosis tool, to ensure accuracy, especially from a medical perspective.

- Only a few studies have presented prediction models that combine both blood pressure and glucose metabolism disorders detection, even though it is recognized that both conditions share the same risk factors, consequences, and preventive actions. Therefore, this study aims to fill this gap by developing models to detect both raised blood pressure and raised blood sugar levels.
- It was noticed through the reviewed studies that they had been often based on the frequent use of old datasets from cross-sectional or cohort studies for a group of patients, without including the recently identified risk factors. In this proposed study, the high blood pressure and hyperglycemia (raised blood sugar) detection model will be demonstrated to a local dataset collected using a household survey specifically designed to study the NCDs' risk factors and their prevalence among adults in Palestine.

Therefore, there are still some gaps in research within this field that need to be filled. In general, the raised blood pressure and sugar prediction models proposed in this study can play a key role at different levels, such as supporting the healthcare system by reducing the prevalence of hypertension and diabetes as a result of screening and early detection, and thus taking the needed procedures and follow-up for prevention or controlling the detected cases. In addition to saving costs spent on high blood pressure and diabetes medications and treating complications resulting from them.

2.8 Conclusion

This chapter presented a comprehensive literature review about the raised blood pressure and sugar prediction models using machine learning, starting from the investigation from a scientific medical perspective to employing the techniques of machine learning in predicting these health conditions. Exploring the standard medical and clinical side of raised blood pressure and sugar health conditions including their risk factors, symptoms, detection methods, and consequences, is an important preliminary step to establishing a scientific base for general conceptualizing of the proposed prediction models. Investigating the previous research and works conducted in the domain of raised blood pressure and sugar models has a significant positive impact on drafting the workflow and framework of the prediction models proposed in this study, in terms of the used variables, the utilized algorithms, and the performance evaluation criteria.

The next chapter will present the methodology that was followed to implement the raised blood pressure and sugar prediction models, while digging into the used components in details.

Chapter 3: Methodology

3.1 Introduction

This chapter illustrates the phases of the proposed raised blood sugar and pressure prediction models, including the data collection and description of the used dataset, the data cleaning and preparation, the exploratory data analysis, data preprocessing, and the utilized machine learning classifiers for building the raised blood pressure and sugar prediction models using Random Forest, Decision Tree, XGBoost, Adaboost, Bagging Decision Trees, and Multi-Layer Perceptron classification algorithms.

Figure 3.1 depicts the main phases of the implemented raised blood pressure and sugar detection models. The next sections within this chapter will provide detailed insights into these six phases and their inside components utilized in constructing the proposed models for raised blood pressure and sugar detection.

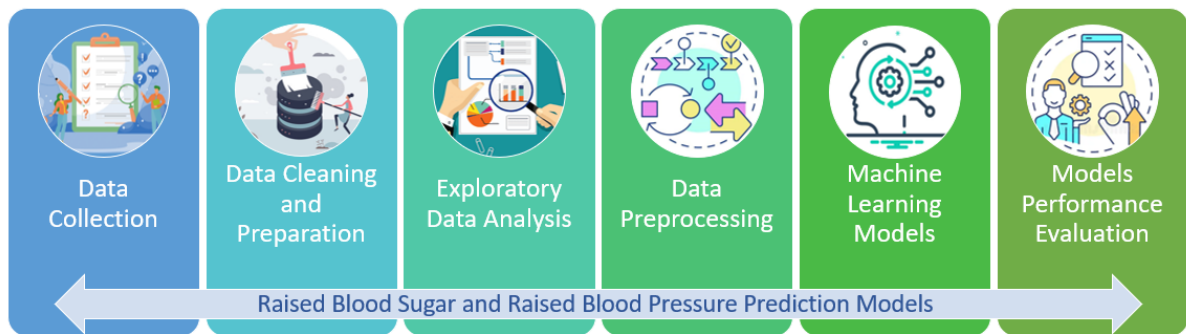


Figure 3.1 Main Phases of the Raised Blood Pressure and Sugar Detection Models

3.2 Used Technology

All the activities of data preparation, exploration, preprocessing, and modeling, used for implementing the proposed models presented in this study were accomplished using Python programming language, executed via the “Jupyter Notebook” web-based platform.

Python is an open-source programming language, that is used widely these days in developing data-driven applications and models. It is built on top of an extensive set of powerful libraries for processing and manipulating data, in addition to the modeling utilities provided by the built-in machine learning libraries, which make it an advanced, flexible, capable coding option for these purposes. The main Python libraries that were

used throughout the implementation process of the proposed models in this study are Pandas and Numpy, Seaborn, and Matplotlib, which are two of the most important core libraries in Python.

- Numpy is the fundamental package for scientific computing with Python (VanderPlas, 2016).
- Pandas is a Python package for high-level data manipulation and analysis which is built on the Numpy package (VanderPlas, 2016).
- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python (Waskom, 2021).
- Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics (Embarak, 2018).
- Scikit-learn is a Python library designed to support prediction analysis and machine learning modeling. It offers a various set of powerful machine-learning algorithms for supervised and unsupervised learning (Hao & Ho, 2019).

3.3 Data Collection and Dataset Description

The used dataset in this study for building the hypertension and diabetes prediction model is collected from STEPS study surveys (Stepwise approach to NCD risk factor surveillance) (The World Health Organization, 2022b). The STEPS is a cross-sectional population study standardized by WHO to study the NCDs' risk factors and their prevalence among countries. As the name of the study indicates it consists of three levels of risk factor assessment, the first level is a questionnaire, the second is for physical measures, and the last step is biochemical measures.

STEPS study was conducted by the Palestinian National Institute of Public Health in collaboration with the Palestinian Ministry of Health in Palestine in 2022. The STEPS dataset was collected by conducting interviews with participants, it consisted of demographic, social, economic, eating habits, health history, and mental health questions. Physical measurements were also recorded, including blood pressure, heart rate, weight, height, and BMI calculations. Additionally, the dataset includes laboratory test results for fasting blood glucose, HDL, total cholesterol, and triglyceride levels. It consists of more

than five thousand records and around 130 variables, gathered from the study population as a representative sample collected through a stratified random sampling approach.

The target population of the collected dataset is Palestinian adults aged between 18 to 69 of both sexes who have been living in Palestine for at least 12 months. The used sampling approach in collecting this dataset is three-stage stratified cluster sampling to select a random sample of adults from 525 enumeration areas in the West Bank and Gaza Strip, then selecting 11 households from each enumeration area. The first stage was selecting the enumeration areas, the second stage was selecting households from each enumeration area by using blind maps provided by the Palestinian Central Bureau of Statistics and the Kish table sampling method, and the third stage was selecting a participant from households. The data collection team consisted of two members, the first member is a field worker responsible for conducting personal interviews with participants that aim to collect demographic, socio-economic, and personal lifestyle information. The second one is a nurse who is responsible for collecting medical history, recording the physical measurements, and performing the blood tests for participants.

The standard WHO NCDS stepwise survey comprises generally from core modules which include questionnaires for the following modules: demographic information, tobacco use, alcohol consumption, diet, dietary sugar intake, dietary salt intake, physical activity, history of NCDs, cervical cancer screening for women, sleep pattern, mental health, and wellbeing, anxiety and depression, physical measurements of weight, Height, hip, waist, three measurements of heart rate, and blood pressure, and finally the biochemical measurements. The STEPS survey can be expanded with additional modules that can be optionally added, depending on the area of interest identified by the Research Director responsible for overseeing the NCD stepwise study.

For the use of building prediction models in this thesis, a selected subset of features is obtained from the original STEPS dataset. This subset included the features that were recognized in the previous studies as known risk factors for developing hypertension and diabetes and can be used as predictor variables for developing such prediction models.

It is worth pointing out that the adoption of the STEPs dataset in this study to build machine learning models represents a contemporary pattern, as there is a very limited

number of studies that utilized the STEPs dataset in training machine learning-based models since the start of conducting the NCDs stepwise approach study.

Additionally, the work proposed in this study is considered the first of its type in utilizing the STEPs dataset and building a model for predicting both raised blood sugar and high blood pressure as a screening tool for hypertension and diabetes, using all possible variables that can contribute to developing NCDS, like the variables related to the mental health and well-being, anxiety and depression, and sleep disturbances.

The STEPS Palestine 2022 dataset is owned by the Palestinian Ministry of Health. It is obtained through a formal letter for conducting this study, and not available publicly. The approval letter for using the STEPS dataset is provided in the appendix at the end of this thesis.

Table 3.1 presents the collected subset of variables from the STEPS dataset and a description of each variable.

Table 3.1 The Features of STEPS Dataset and their Description

Variable	Description	Type	Unit
QR	A QR code is provided for each participant as a unique identifier	Text	
Consent	A binary field indicates obtaining the consent to participate in the study or not. 1: Yes, 2: No	Categorical	
Age	Age of the participant at the study time	Continuous	Year
Sex	Sex of the participant 1: Male, 2: Female	Categorical	
Smoking Tobacco	If the participant smokes tobacco products 1: Yes, 2: No	Categorical	
Passive Smoker	If the participant is exposed to passive smoking 1: Yes, 2: No	Categorical	
Former Smoker	If the participant was a former smoker 1: Yes, 2: No	Categorical	
Alcohol consumption	If the participant drinks alcohol 1: Yes, 2: No	Categorical	

Fruit/ Vegetable Intake	The average number of servings of vegetables or fruits consumed per day. 1: Yes, 2: No	Continues	
Dietary Salt Intake	Level of salt intake 1: High, 2: No	Categorical	
Dietary Sugar Intake	Level of sugar intake 1: High, 2: No	Categorical	
Physical activity	Insufficient level of daily physical activity 1: Yes, 2: No	Categorical	
History of NCDs	History of being diagnosed with any of the following diseases: 1: Cancer 2: Cardiovascular 3: Asthma 4: Osteoporosis 5: Renal failure	Categorical	
Cervical Cancer	Have the female participants ever undergone cervical cancer screening 1: Yes, 2: No	Categorical	
History of raised blood pressure	History of being diagnosed with diabetes or raised blood sugar, taking medications, type of medications.	Categorical	
History of diabetes	History of being diagnosed with diabetes or raised blood sugar, taking medications, type of medications.	Categorical	
History of raised total cholesterol	History of being diagnosed with raised cholesterol (raised fat level in blood), taking medications, type of medications.	Categorical	
Is Pregnant	The pregnancy status was collected for women participants only.	Categorical	
Blood pressure	The systolic and diastolic blood pressure (3 readings with 3 minutes in between). Average blood pressure was calculated from the 3 readings.	Continues	mmHg
Heart rate	The heart rate. (3 readings with 3 minutes between)	Continues	bpm

	Average heart rate calculated from the 3 readings.		
Height	The height measurements for the participant.	Continues	cm
Weight	weight measurements of the participant.	Continues	Kg
Waist circumference	The measurements of waist circumference for the participant.	Continues	cm
Hip circumference	The measurements of hip circumference for the participant.	Continues	cm
Step3 Consent	A binary field indicates obtaining the consent to participate in step 3 or not. 1: Yes, 2: No	Categorical	
Fasting	Has the participant had anything to eat or drink, other than water during the past 12 hours? 1: Yes, 2: No	Categorical	
Fasting blood sugar	The level of blood sugar inside the participant's body.	Continues	mg/dl
Total cholesterol	Total cholesterol inside the participant's body.	Continues	mg/dl
HDL cholesterol	The level of high-density lipoprotein.	Continues	mg/dl
Triglyceride	The level of triglycerides fat inside the participant's body.	Continues	mg/dl

One of the utilized modules in the NCDs stepwise survey is mental health assessment, which is used to measure the level of psychological distress among participants. The assessment of mental health was performed using the WHO-5 Well-Being questionnaire and the Four-Item Patient Health Questionnaire.

Figure 3.2 below shows the WHO-5 Well-Being questionnaire used in the STEPS survey (The World Health Organization - Regional Office for Europe, 1998). These questions were part of the collected subset for use in this dissertation.

WHO-5 wellbeing index. The next questions ask about your mental health and wellbeing.

	Please indicate which is the closest to how you have been feeling over the last two weeks.	All the time	Most of the time	More than half of the time	Less than half of the time	Some of the time	At no time
1	In the past two weeks, how often have you felt cheerful and in good spirits?	5	4	3	2	1	0
2	In the past two weeks, how often have you felt calm and relaxed?	5	4	3	2	1	0
3	In the past two weeks, how often have you felt active and vigorous?	5	4	3	2	1	0
4	In the past two weeks, how often have you woken up feeling fresh and rested?	5	4	3	2	1	0
5	In the past two weeks, how often has your daily life been filled with things that interest you?	5	4	3	2	1	0

Figure 3.2 WHO-5 Well-Being Questionnaire

Figure 3.3 shows the Four-Item Patient Health Questionnaire for anxiety and depression that is used in the STEPS survey, which is also a part of the collected subset for use in this thesis (Kroenke et al., 2009).

Over the last two weeks, how often have you been bothered by the following problems?	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious or on edge	0	1	2	3
Not being able to stop or control worrying	0	1	2	3
Feeling down, depressed, or hopeless	0	1	2	3
Little interest or pleasure in doing things	0	1	2	3

Figure 3.3 Four-Item Patient Health Questionnaire (PHQ-4)

The sleep pattern was among the utilized modules in the NCDs stepwise survey, to measure the level of sleep disturbances among participants. Figure 3.4 shows the sleep disturbances questionnaire as part of the collected variables in the STEPS dataset using the Jenkins sleep scale (JSS) (Shahid et al., 2012). JSS is an instrument used as a standard tool in clinical research to evaluate sleep difficulties.

Jenkins Sleep Scale

How many times during the last month you?	Not at all	1–3 days/month	4–7 days/month	8–14 days/month	15–21 days/month	22–31 days/month
(1) Had trouble falling asleep	0	1	2	3	4	5
(2) Had trouble staying asleep	0	1	2	3	4	5
(3) Woke up several times per night	0	1	2	3	4	5
(4) Woke up after usual amount of sleep feeling tired and worn out	0	1	2	3	4	5

Figure 3.4 Jenkins Sleep Scale

3.4 Data Cleaning

Cleaning the data is considered a crucial stage in the data utilization process, which has a significant impact on the following phases, particularly on exploratory data analysis and the phase of modeling as well.

The STEPs dataset was collected through a population survey, though digitalized forms were used in the data collection process, it still needs further investigation for detecting and handling any anomalies. The utilized structural dataset in this study was obtained as two separate MS Excel files, the first file containing the data collected in steps 1 and 2 of the study referred to in the previous section, and the second one containing the biochemical measurements that were collected in the third step of the STEPs study.

Figure 3.5 outlines the techniques applied for cleaning the STEPS dataset.

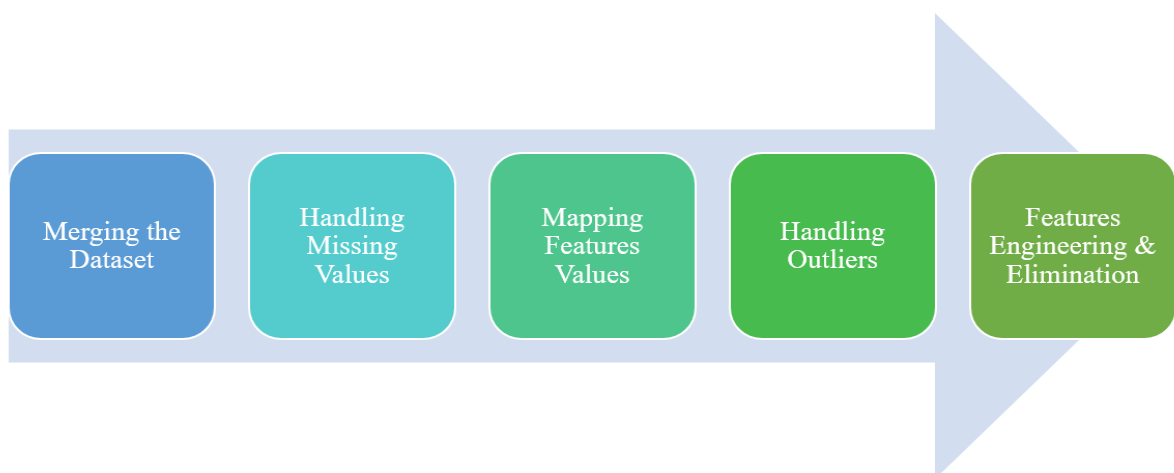


Figure 3.5 The Applied Steps for Cleaning STEPS Raw Dataset

In this phase of dataset cleaning, the primary steps involved are merging datasets, handling missing values, mapping feature values, managing outliers, and performing feature engineering and elimination. The details of the methods applied in each step are clarified in the following subsections from 3.4.1 to 3.4.5.

3.4.1 Merging the Dataset

The first step in developing this model involved extracting the dataset that was provided as two separate files into two data structures. The next step is merging the extracted data, using the provided QR code of the participant in both files as a mutual key in the two datasets. The merged dataset is ready for exploring its dimensionality, the variable names and types, the number of missing values, and the content of different variables in this dataset.

3.4.2 Missing Values

The issue of missing values is considered a common problem in any collected dataset used for building data-driven models. Missing values have to be checked and handled as a first step before proceeding to any other task within the data science process. Hence, the following steps present the methods used for checking and handling missing values in the STEPS dataset.

- On checking the missing values, it was found that there are some records having consent values equal to 2 and the remaining variables in the tuples are empty, which indicates that these records are related to selected people who refused to participate in the study. In this step, such observations were handled by excluding them from the dataset.
- Another reason for missing values in the dataset is due to records related to pregnant women, which did not have available physical measurements. According to the instructions in the guidelines of the STEPs survey, the physical measurements are not recorded in case of pregnancy. Thus, any missing values in the physical measurement were handled by dropping the associated record from the dataset to maintain data accuracy.
- Some variables in the obtained dataset contain a specially identified code used as an alternative for missing values. These special values were handled by imputing

the most frequent value in the case of categorical variables using the mode function.

- Other missing values in the dataset were in conditional variables that were either redirected or skipped depending on the answer given in the previous question, like the daily smoking based on current smoking, such a case was handled by filling the missing value with zero.
- For missing values in biochemical measurement variables, exploration of these observations showed that missing values were due to either subjects not providing consent to participate in Step 3 of the study or participants not adhering to fasting instructions provided by field workers to perform necessary blood tests.
- The raw dataset initially contained 5529 records. After applying the data cleaning steps to address missing values, the dataset was reduced to 5054 observations. This indicates that 475 records had missing values and were consequently removed during the cleaning process.

Therefore, in general, the handling criteria to deal with inconsistency issues of missing values in the dataset was applied in two approaches. For categorical features, the used criteria is imputation by the common value within the variable. For continuous features containing missing values associated with physical measurements and blood test results, in particular, the approach was to drop relevant observations entirely from the data set to ensure data consistency and accuracy.

3.4.3 Mapping Features Values

For the binary categorical variables associated with Yes/ No questions, in the initial collected dataset, the value of 1 indicates a Yes answer while 2 indicates an answer with No. Features of this type were handled by mapping a value of no answers or negative categories from 2 to 0.

3.4.4 Checking Outliers

Maintaining data consistency is considered an important step for building powerful prediction models. Outliers are one of the common issues in data inconsistency regards (Ackerman et al., 2020). In a general definition, a value is recognized as an outlier if it has a significant obvious deviation from other observations among the processed subjects (Aguinis et al., 2013). In the case of clinical data and research, this definition might not

be applicable, particularly in dealing with values obtained from clinical measurements or lab test results, since sometimes these values that are detected as outliers might be correct values indicating a serious health condition which implies keeping it instead of removing (Gress et al., 2018).

Outlier management in this study adhered to the NCDs Stepwise Manual guidelines established by the WHO (The World Health Organization, 2017). Variable boundaries were defined by domain experts based on global standards, encompassing acceptable low, normal, and high values. These ranges served to exclude implausible data points potentially caused by device calibration issues or human error. Table 3.2 outlines the specific ranges applied to the dataset.

Table 3.2 The Accepted Ranges of STEPS Dataset Variables		
Variable	Minimum Value	Maximum Value
Height (cm)	100	270
Weight (kg)	20	350
Waist circumference (cm)	30	200
Hip circumference (cm)	45	300
BMI (kg/m ²)	14	60
Systolic BP (mmHg)	40	300
Diastolic BP (mmHg)	30	200
Fasting blood glucose (mg/dl)	18	630
Total cholesterol (mg/dl)	67	773
Fasting triglycerides (mg/dl)	22	700
HDL cholesterol (mg/dl)	3.8	190

3.4.5 Features Engineering and Elimination

The process of selecting the final features list from the dataset was demonstrated through various approaches including deriving new features and eliminating useless features.

The process of extracting new measurable features from the collected variables in the initial dataset included several features: mental health assessment, presence of anxiety and depression symptoms, insufficient sleep duration, and poor sleep quality. The steps taken for defining these new features are explained in detail in the following lines.

- 1- Mental health illbeing: in the collected dataset, were 5 variables associated with the WHO-5 Well-Being questionnaire referred to in Section 3.1, which is used as a standard tool for assessing the mental health of humans. The sum of the values of the 5 given answers is used to calculate the total score. Subsequently, the total score is multiplied by 4, which results in a percentage score between 0 and 100. If the result of the calculated WHO-5 index is less than or equal to 50, then this indicates subjects of ill-being mental health status (Topp et al., 2015). A value of 50 for the WHO-5 index is identified as a cutoff value for poor well-being according to the diagnostic criteria defined using the WHO-5 instrument (Downs et al., 2017). The final derived feature is a binary variable carrying the value of 1 if the subject is recognized as ill in terms of psychological distress, or 0 in the case of a normal subject.
- 2- Anxiety and depression PHQ-4: another instrument that is used for the screening of beings' mental health, particularly anxiety and depression, is the Patient Health Questionnaire consisting of 4 questions (PHQ-4). The PHQ-4 had been incorporated into the mental health module of the NCDs Stepwise survey (Kerper et al., 2014). The PHQ-4 score is calculated as the total of the figures for the answers to the 4 questions. An obtained score of value greater or equal to 6 is determined as an indicator of symptoms of anxiety and depression (Löwe et al., 2010). Finally, at this stage of data cleaning, a new binary feature was derived based on this score, where positive values of this new variable indicate people with anxiety and depression or not.
- 3- Inadequate sleeping hours: the average daily sleeping hours variable was among the variables collected in the STEPs dataset. Insufficient sleeping duration was identified as a risk factor for increased blood pressure and sugar levels in many studies (Spiegel et al., 1999; Vierra et al., 2022). This variable was used to determine the adequacy of the sleeping duration of participants. The cutoff value of 7 hours was used for deriving the sleeping inadequacy variable, which is recommended as the minimum needed amount of sleeping hours for adults from a health perspective (Watson et al., 2015).
- 4- JSS Sleep disturbances: Another component that was used in the survey of the collected dataset is the Jenkins Sleep Scale, which was utilized for evaluating

the sleep quality of participants, which is known to have effects on the public health of people (Jenkins et al., 1988; Yoda et al., 2015). The Jenkins Sleep Scale is a clinical research instrument designed in the form of a questionnaire to evaluate the severity of sleeping difficulties. The answers provided for the questions in the JSS questionnaire were used to obtain the total JSS score. The total score of 11 is the threshold value for recognizing sleep disturbances among the evaluated subjects (Juhola et al., 2021). The newly defined JSS feature is set to 1 for subjects having a total JSS score greater than 11 indicating disorders in participant sleeping.

- 5- Insufficient fruit/vegetable intake: this variable was defined using the number of servings of fruit and vegetables on average per day, which was provided in the collected dataset. The derived feature of Insufficient fruit/vegetable intake is set to positive if a subject is consuming less than five servings of fruit and vegetables per day. The cutoff value of five daily servings as per the recommendation of the daily minimum amount of fruits and vegetables for a healthy diet according to the WHO experts (Agudo & Joint FAO, 2005).
- 6- BMI class: the BMI continuous variable was provided in the dataset as an auto-calculated variable using the physical measurement of weight and height. It is used to define the classes of participants weights to four main classes according to the standard criteria for classifying adults as follows: BMI below 18.5 is classified as underweight, BMI between 18.5 and 24.9 falls into the class of normal weight, while BMI ranged between 25 and 29.9 categorized in the overweight class, and higher or equal to 30 are in the obesity class (CDC, 2021).
- 7- Waist-hip ratio: the participants' physical measurements of the waist and hip are provided in the initial dataset. These measurements were used in defining a new continuous feature WHR, which is simply calculated as the ratio of waist circumference to hip circumference. The waist-hip ratio was determined in several studies among the anthropometric risk factors used for predicting NCDs (Nishida et al., 2010; The World Health Organization, 2008) .
- 8- Raised blood pressure (RBP): the average systolic and diastolic blood pressure are provided in the dataset. The global guidelines in defining high blood

pressure indicate that blood pressure is recognized as high among adults if the systolic blood pressure reading is greater than or equal to 140 mmHg and/ or diastolic blood pressure is greater than or equal to 90 mmHg (Chobanian et al., 2003; James et al., 2014; J. H. Lee et al., 2018). A new feature was derived based on this global standard definition, which is used as the outcome class in the raised blood pressure prediction model.

- 9- Raised blood sugar (RBS): the provided fasting blood sugar test result was used to define this feature. Per the guidelines of WHO and IDF, a value greater than or equal to 126 mg/dl for fasting blood sugar is defined as raised blood sugar (WHO, 2006). This criterion was used to define the RBS variable, which will be utilized as the outcome predicted feature in developing the prediction model of the raised blood sugar.

The next part of the feature selection stage is the elimination of unused and irrelevant features.

- 1- The feature elimination was applied also to variables that are irrelevant to the purpose of this study, such as the household's information (number of household members, number of adults, number of children in the house, relation to household), marital status of the participant, and education level. Such variables are used usually in population surveys to study the general demographics composition of a community, which is not the area of interest of this thesis work. Additionally, these variables are not identified medically as related risk factors to raised blood sugar and pressure.
- 2- Variables related to cervical cancer screening were found irrelevant to the aim of this study. On investigation of the source questionnaire, it is found that these variables were used for measuring the screening rate of cervical cancer among women, by asking female participants if they had ever undergone a test for cervical cancer, and when in case of yes answers. These features were handled by dropping from the dataset.
- 3- All the variables related to mental health, anxiety, and depression, sleep disturbances that were used in the previous steps to derive the new measurable variables based on the calculated scores, were eliminated from the dataset after the score calculation and the features-defining process was completed.

By applying all the previous steps to the dataset, it can be ensured the data is consistent to be fed into the following phases of the applying the data science cycle, in particular, for use in the exploratory data analysis and building prediction models.

3.5 Exploratory Data Analysis

In data-driven models, exploratory data analysis using either statistical or graphical methods plays an important role in investigating the dataset (Owda et al., 2023). Additionally, it can contribute to getting a brief overview, detecting general trends, and inferring helpful assumptions to build on before proceeding to the phase of machine learning modeling.

The process of analyzing data encompasses univariate analysis which is used to study a single feature within the dataset, and multivariate analysis to explore the relationship between two features or more. An important part to consider in the process of exploring the dataset variables is the selection of the exploration techniques, which needs to be determined wisely according to various considerations like the type of variable to explore, the desired information of interest that has to be evaluated, and the audience who will use or look into the results of exploration (Deming et al., 2018).

Therefore, in this study, exploratory data analysis techniques using a graphical approach were used, which provided a comprehensive understanding of every single feature in the dataset, as well as to identify the different relationships between the variables in the dataset. The illustrated figures below show some characteristics of the used STEPS dataset, in addition to different relationships between the dataset variables.

3.5.1 Main Characteristics of the Population Sample in the STEPS Dataset

Figures 3.6 – 3.8 present the main characteristics and demographics of the population sample in the collected STEPS dataset.

Figure 3.6 shows the frequency of participants' ages in the STEPS dataset.

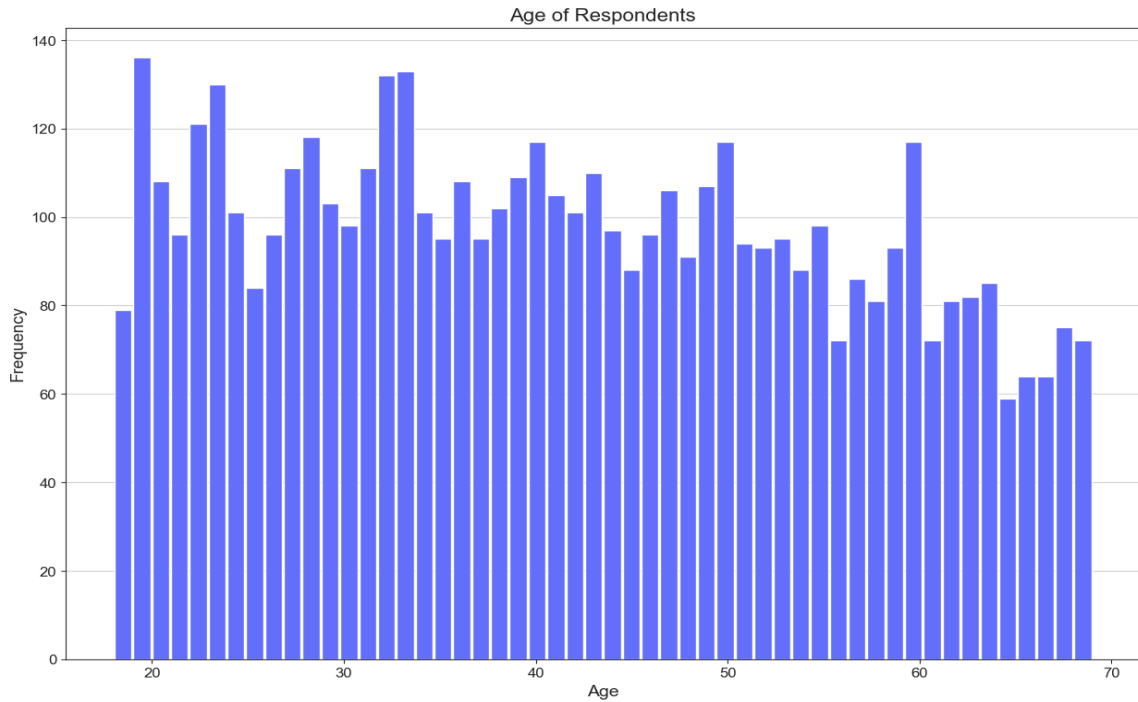


Figure 3.6 Age Distribution of Participants in the STEPS Dataset

As previously mentioned, the target population of the STEPS study comprises adults aged between 18 and 69 years, which is reflected in the histogram shown in Figure 3.6.

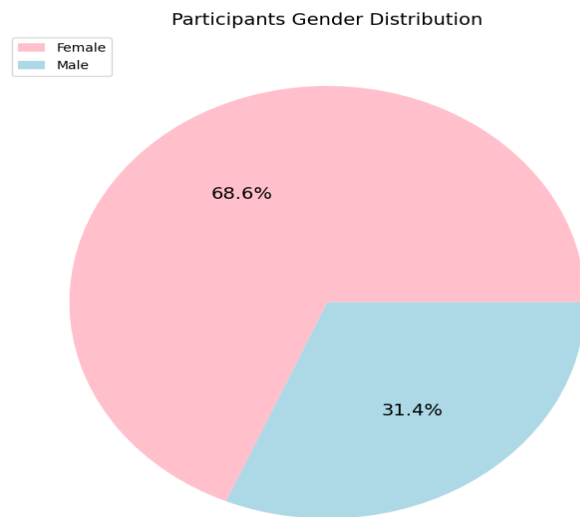


Figure 3.7 The Distribution of Participants in the STEPS Dataset by Sex

Figure 3.7 illustrates the distribution of study participants by gender, showing that the majority are females. This predominance may be attributed to the selection criteria of the STEPS study, which targets eligible residents who have lived in Palestine for at least 12

months. Consequently, more women participated as they were more likely to be available at home.

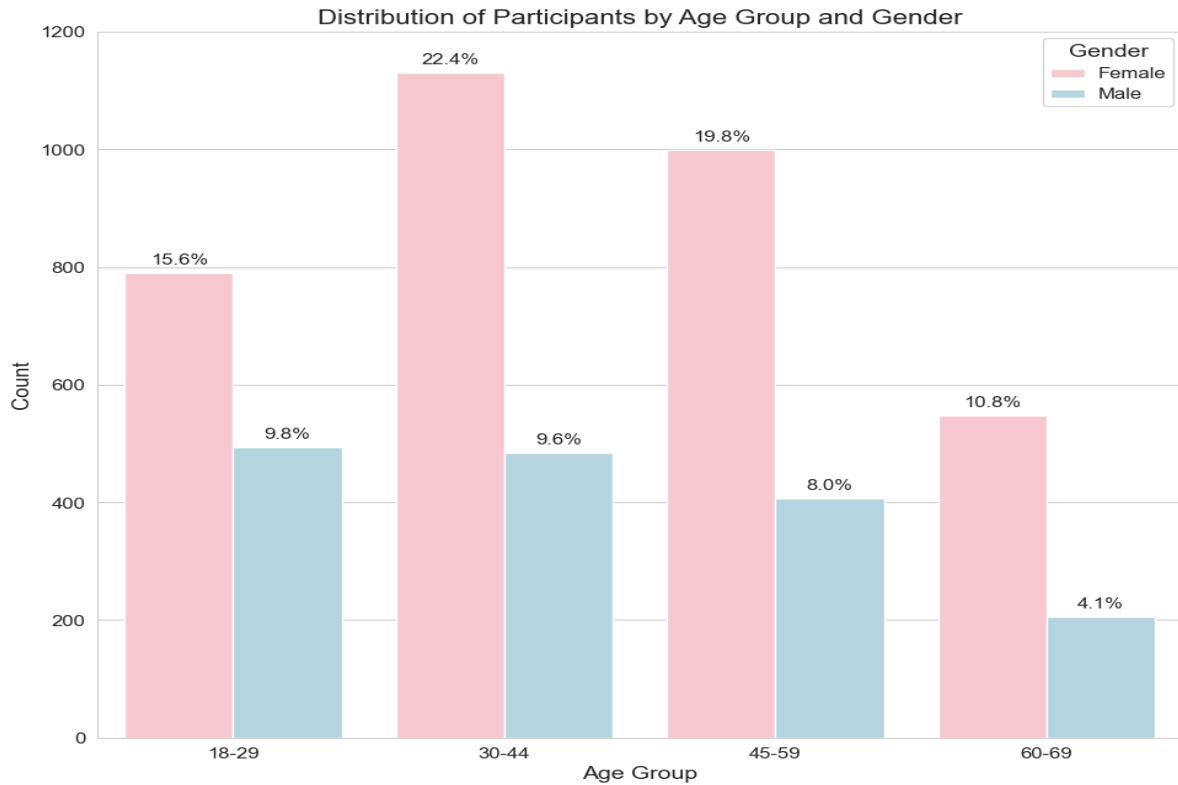


Figure 3.8 The Distribution of Participants in the STEPS Dataset by Age Group and Sex

Figure 3.8 illustrates the distribution of participants in the study categorized by age group and gender. The majority of the participants are adults between 30 and 59 years old, who fall within the age range of 30 to 44 years, and 45 to 59 years, indicating a higher representation of middle-aged adults in the sample. The criteria for age grouping employed in this analysis adhere to the standard aggregation guidelines established in the STEPS study, ensuring consistency and comparability in the analysis of participant demographics. This approach helps in maintaining the integrity of the data and provides a clear understanding of the age and gender distribution within the study population.

3.5.2 Exploring Variables Relationships Using Bivariate Analysis

Figures 3.9 to 3.23 present illustrations for the relationships between a set of dataset variables to obtain clear insights and an understanding of the correlations and dependencies between different variables.

These visual representations are crucial for identifying patterns, trends, and potential interactions within the dataset, thereby facilitating a deeper analysis and interpretation of the data. By examining these relationships, we can better understand how different factors may influence each other and contribute to the overall outcomes observed in the study.

Figure 3.9 shows a comparison between the female group and the male group of participants, in terms of their BMI.

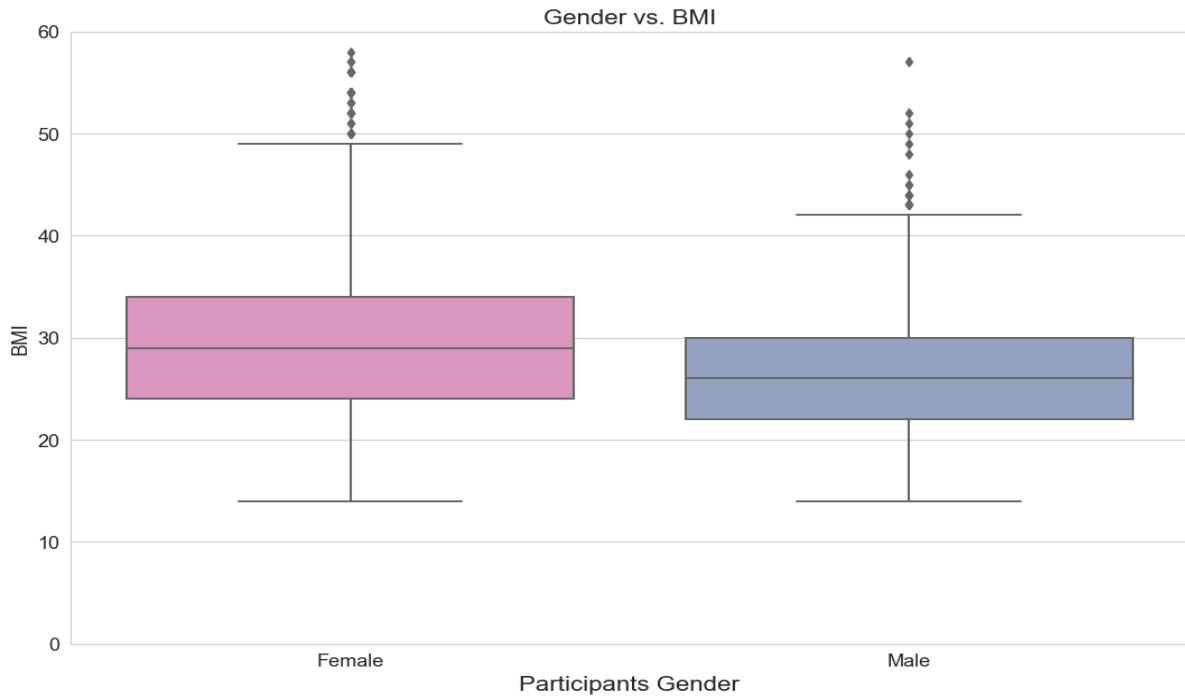


Figure 3.9 Comparison of BMI Rates by Participants' Gender

It is evident in Figure 3.9 that female participants have higher BMI rates rather than males. That might be attributed to several reasons including biological factors like body composition and metabolic rates, social and behavioral factors like dietary habits and physical activity, socioeconomic factors like the affordability of going to the gym and consuming healthy food, physiological factors including stress and mental health, and hormonal factors associated with period, pregnancy and childbirth (Kanter & Caballero, 2012).

The illustration in Figure 3.10 shows the distribution of the weight categories of participants by age and gender.

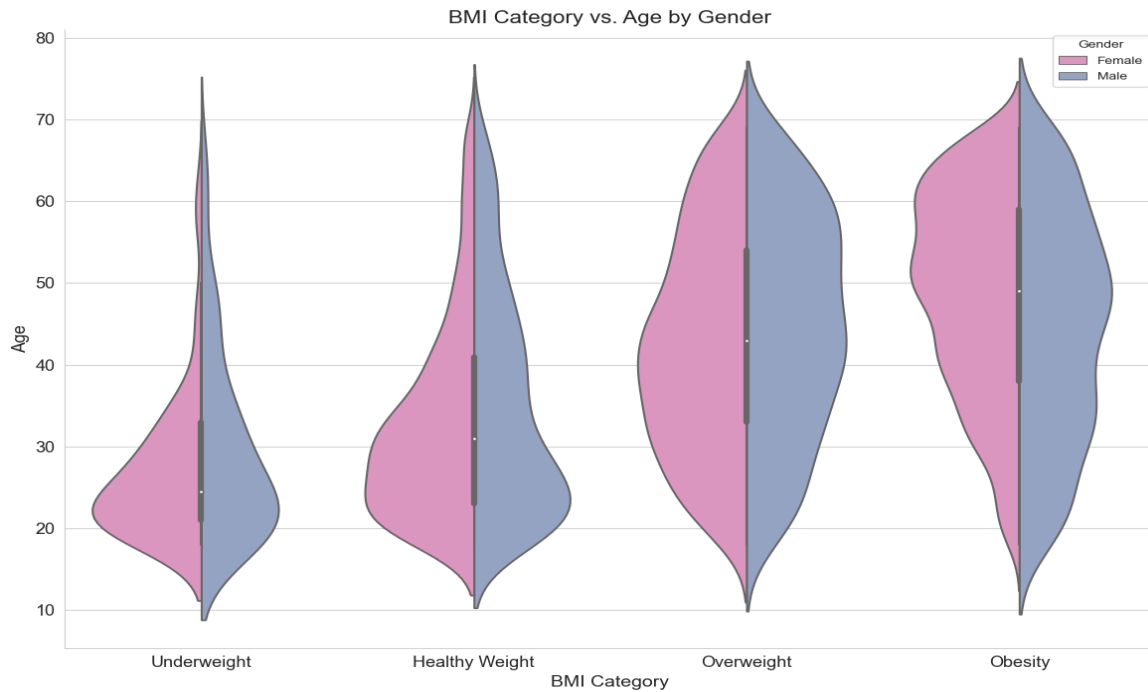


Figure 3.10 Comparison of BMI Class by Age and Sex

In Figure 3.10, it is clear that overweight and obesity are more prevalent among adult participants at older ages.

Figure 3.11 shows the relationship between the weight and systolic blood pressure variables.

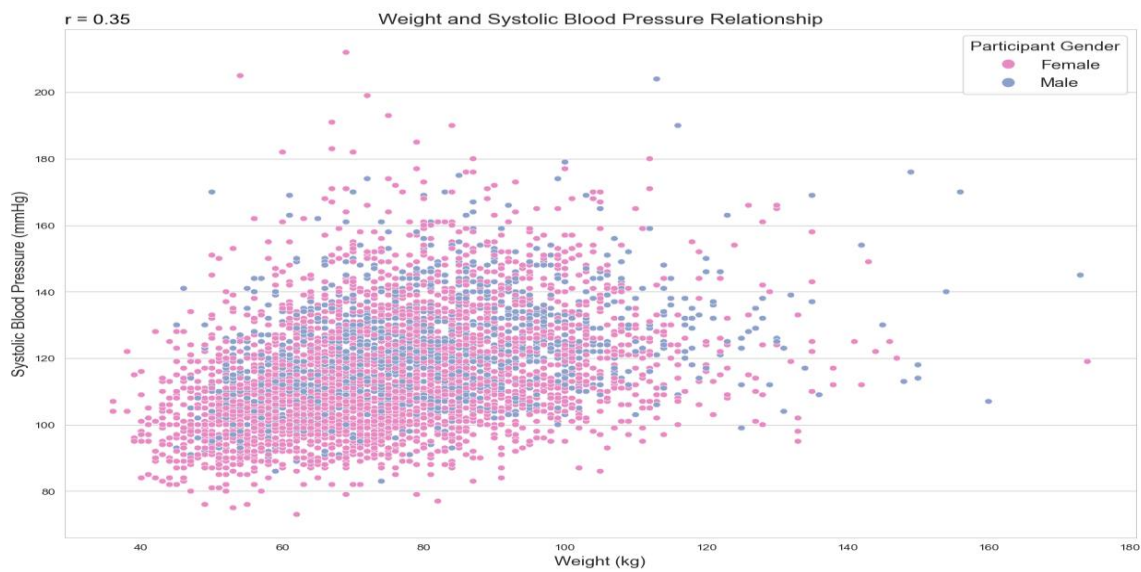


Figure 3.11 The Relationship Between Weight and Systolic Blood Pressure

The relationship between weight and systolic blood pressure, illustrated in Figure 3.11, shows a positive moderate correlation with a correlation coefficient (r) of 0.35. This indicates that as the weight of participants increases, there tends to be a corresponding increase in their systolic blood pressure.

Figure 3.12 shows the distribution of participants by smoking status and gender.

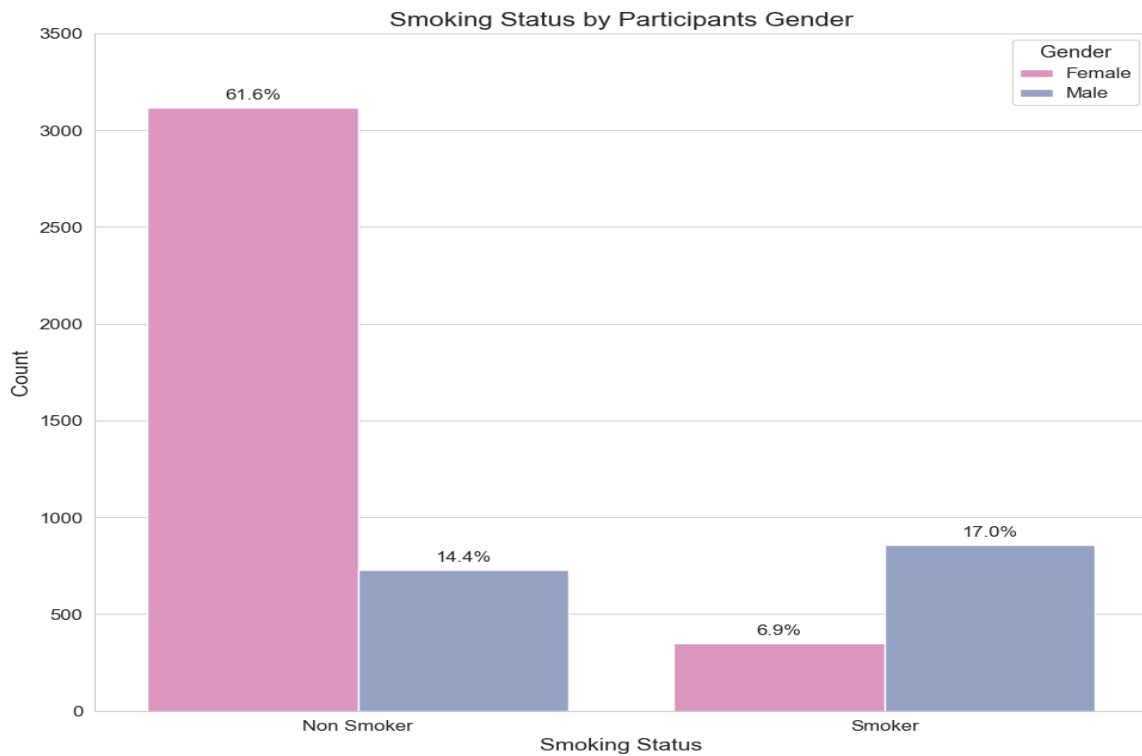


Figure 3.12 The Distribution of Participants by Smoking Status and Gender

It is noticed in Figure 3.12 that the percentage of men in the smokers' group is higher than women within the same group. In comparison to the non-smokers group, the percentage of women who are not smokers is much higher than the men group.

Figure 3.13 below illustrates the relationship between the classification of the fasting blood sugar and the BMI of the participant

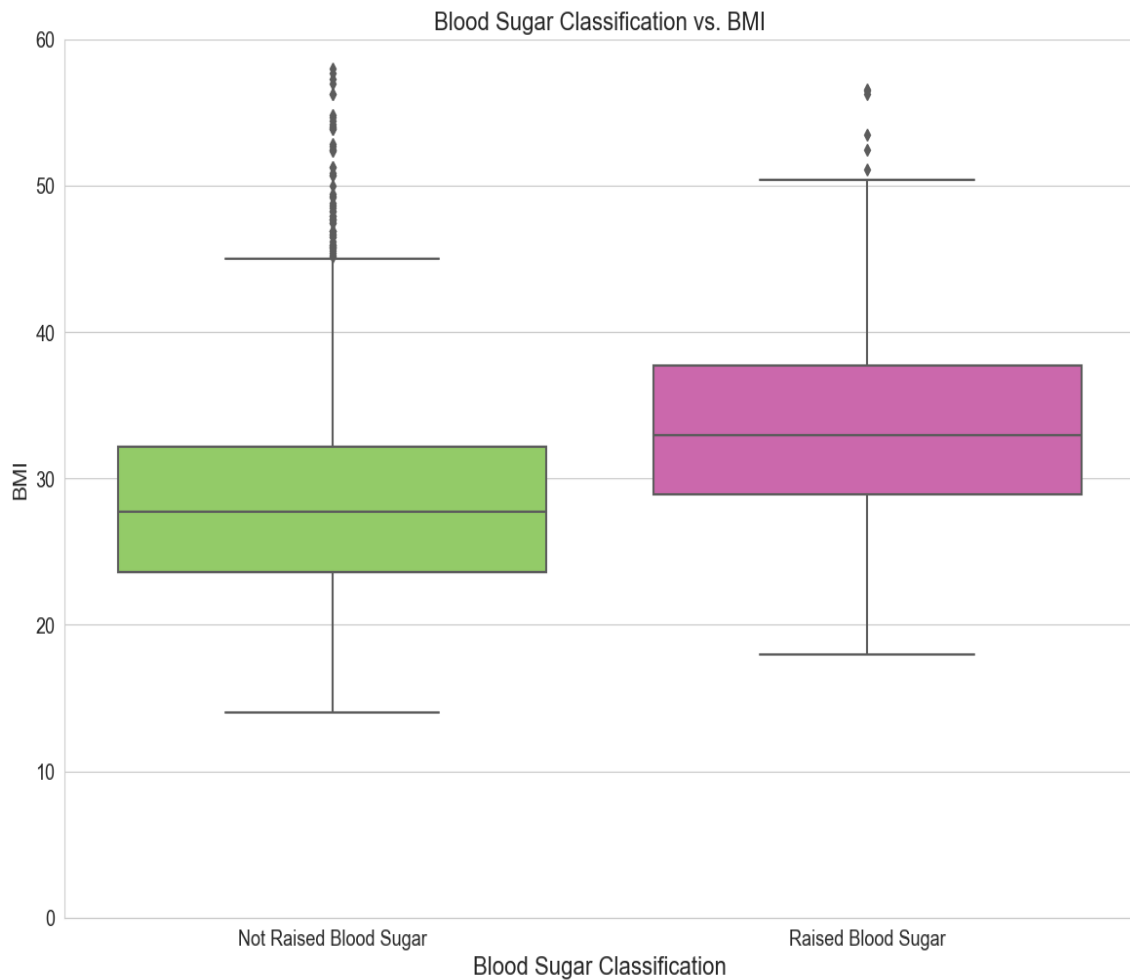


Figure 3.13 Blood Sugar Level Classifications by BMI

The relationship between fasting blood sugar classification and participants' BMI, illustrated in Figure 3.13, shows that subjects with raised blood sugar levels tend to have higher BMI rates compared to those with normal blood sugar levels. This indicates a positive relationship between raised glucose levels and higher BMI, consistent with the findings of previous studies (Amelia, 2017; Owei et al., 2019).

It should be noted that the term “Not Raised Blood Sugar” in all of the graphs below refers to the category of subjects that involves cases of normal and impaired blood sugar levels.

Figure 3.14 shows the relationship between the classification of the blood pressure and the BMI of the participant.

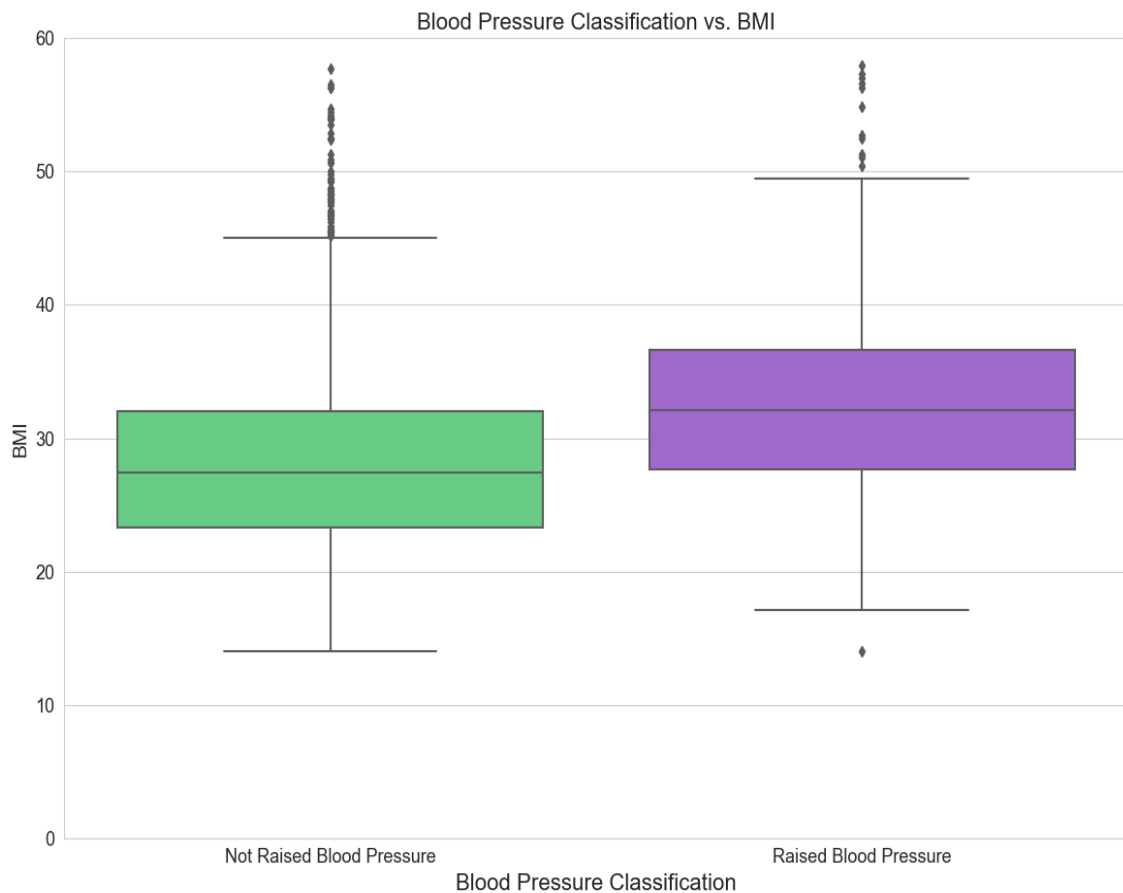


Figure 3.14 Blood Pressure Measurements' Classifications by BMI

In Figure 3.14, it can be seen that instances corresponding to raised blood pressure have higher BMI rates than those who are not recognized as raised pressure cases.

The results shown in the above Figures 3.13 and 3.14 are very near to each other. These findings underscore unhealthy weight as a prevalent risk factor for both raised blood pressure and raised blood sugar, aligning with previous research (Brown et al., 2000; Swartz et al., 2008).

The term "Not Raised Blood Pressure" in all of the figures below indicates subjects that have blood pressure measurement $< 140/90$ mmHg, in the systolic and diastolic components.

Figure 3.15 depicts the diagnosed diabetes reported by the participant through the medical history interview question "Have you ever been told by a doctor or health worker that you have diabetes".

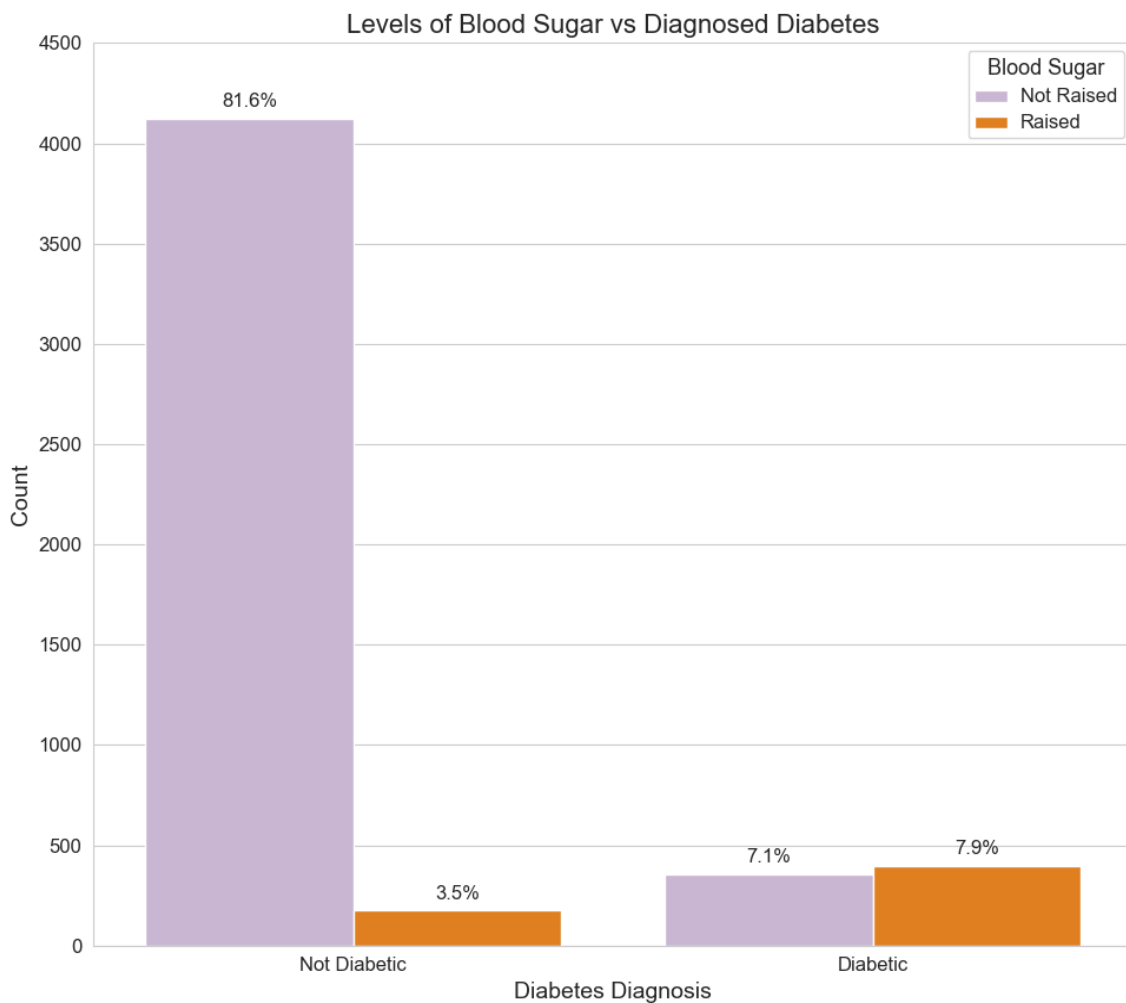


Figure 3.15 Comparison of Blood Sugar Between Diabetic and Nondiabetic Groups

As shown in Figure 3.15, the raised blood sugar cases detected based on the results of the fasting blood sugar test, who are tested with a fasting blood sugar level of 126 md/dl or

higher, represent around 8% among the diabetic group, which indicates that diabetes among this diagnosed group of participants is not controlled.

On the other hand, in the non-diabetic group of participants, the percentage of participants who have raised levels of blood sugar is 3.5%, which may indicate undiagnosed cases of diabetes.

However, this finding among the non-diabetic group of participants is not a definitive conclusion that they are diabetic, but it is a good supportive screening method for paying attention to this sign and adopting the frequent monitoring of blood sugar levels in order to mitigate this risk.

Figure 3.16 illustrates the diagnosed cases of hypertension reported by the participant through the medical history interview question “Have you ever been told by a doctor or health worker that you have hypertension”.

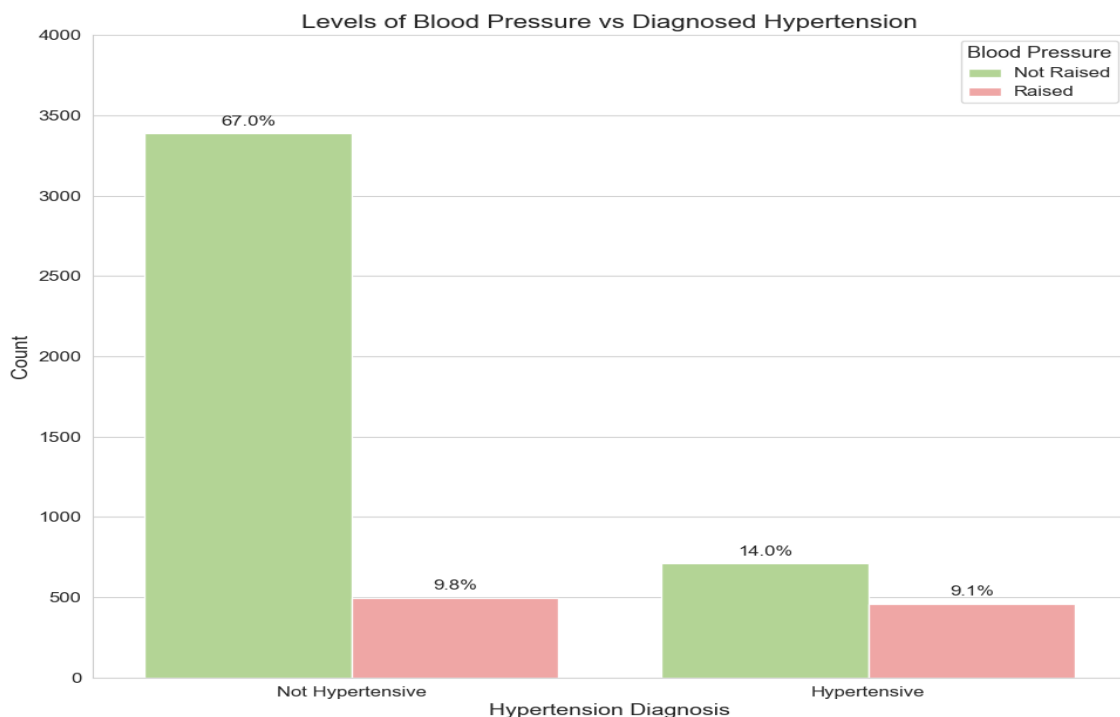


Figure 3.16 Comparison of Blood Sugar Between Hypertensive and Non-hypertensive Groups

Figure 3.16 shows that there is a number of recognized high blood pressure cases that were detected based on the average blood pressure of the three measurements taken for participants in the Stepwise NCDs study. These cases that were identified as raised blood

pressure cases were distributed between the non-hypertensive group and the hypertensive group, who reported having previously been diagnosed with hypertension or had not been diagnosed. As shown in the illustrated graph almost half of the raised blood pressure measured cases belong to participants who are not hypertensive, which means that they might not be aware of this condition. Indeed, the other group of cases that detected as raised blood pressure cases is among the hypertensive participants, this finding points out that hypertension is not managed properly among that group of patients pre-diagnosed with hypertension.

Figure 3.17 illustrates the relationship between the classification of the blood pressure levels measured by three separate readings and the fasting blood sugar levels among the participants in the STEPS dataset.

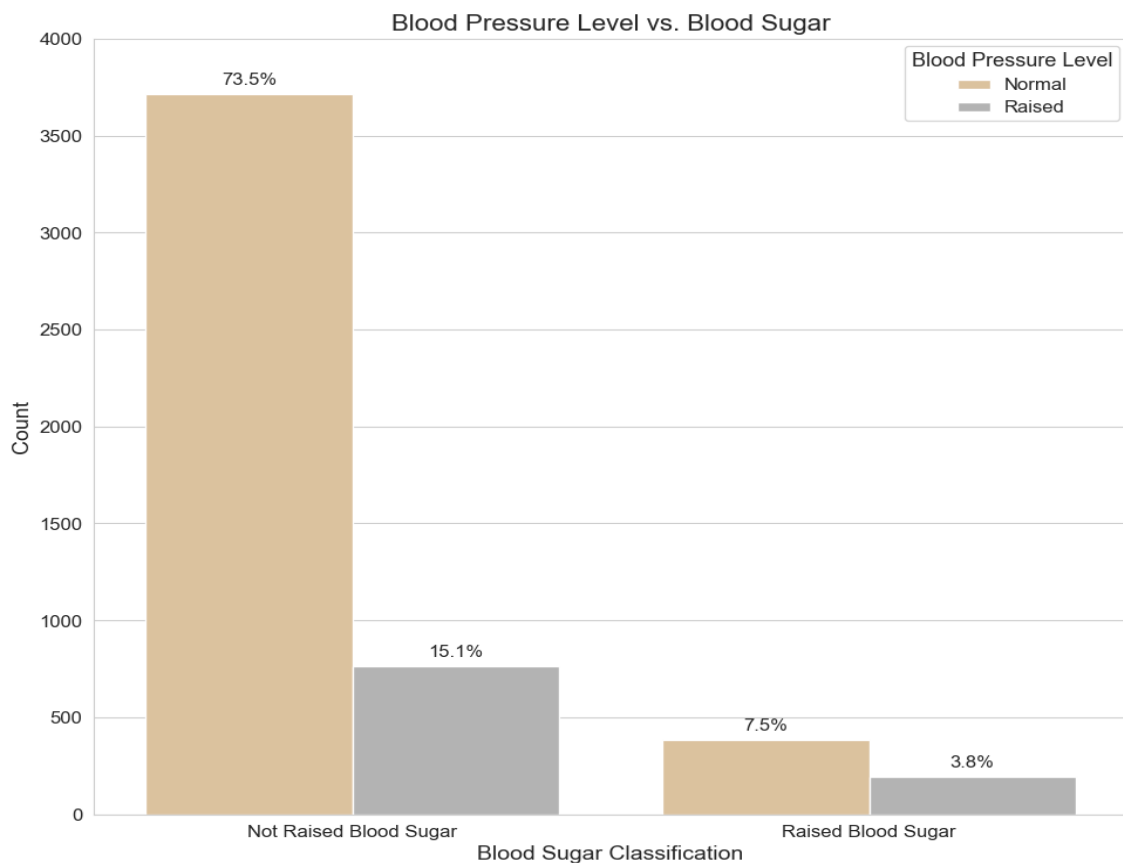


Figure 3.17 The Relationship Between Blood Pressure and Blood Sugar Levels

Focusing on the raised blood sugar group in Figure 3.17, it is evident that approximately twice as many cases in the raised blood pressure class also exhibit raised blood sugar

levels. This underscores the strong association between high blood pressure and raised blood sugar, as concluded by several previous studies (De Boer et al., 2017; Ferrannini & Cushman, 2012).

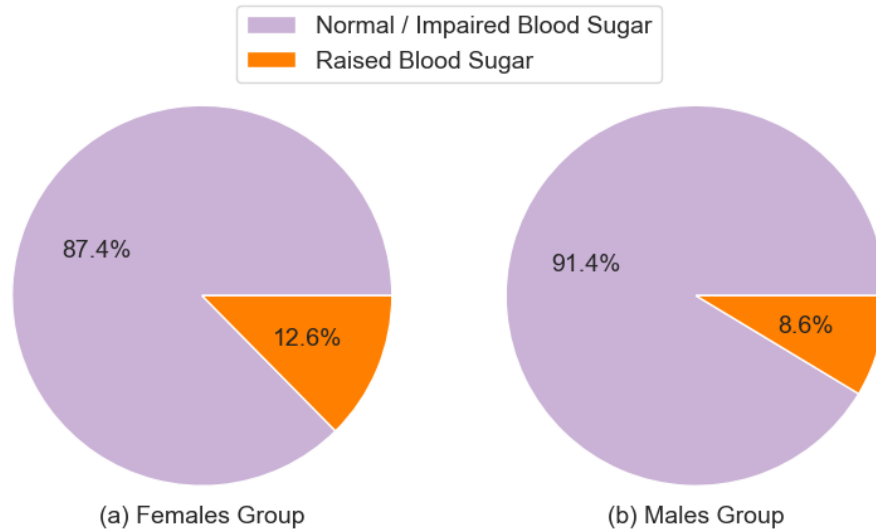


Figure 3.18 Comparison of the Raised Blood Sugar Prevalence Between Female and Male Groups

Figure 3.18 shows a comparison of the prevalence of raised blood sugar levels between the female participants group and male participants, which is higher among the women group than the men group.

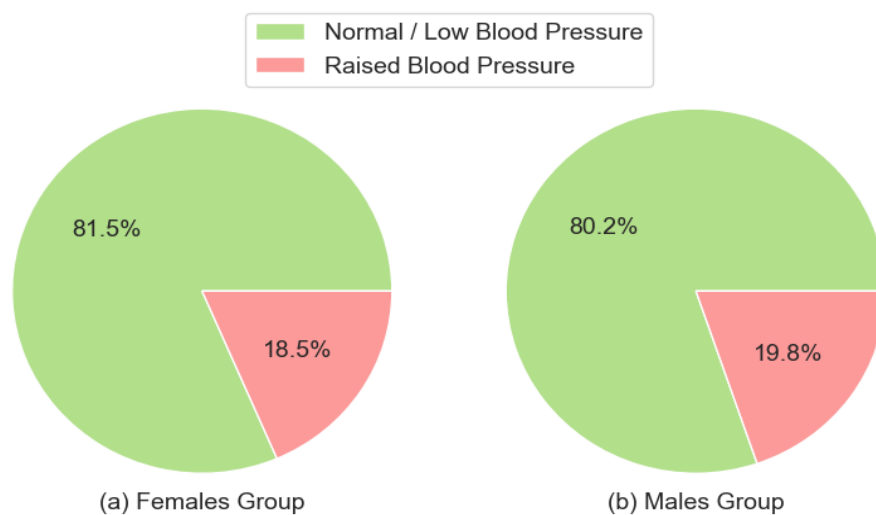


Figure 3.19 Comparison of the Raised Blood Pressure Prevalence Between Female and Male Groups

In Figure 3.19 the comparison between the women and men groups is presented in terms of the prevalence of high blood pressure. The assessment of the blood pressure was based on the readings of blood pressure measured during the study after interviewing the participants. The percentages of raised blood pressure between the male and female groups were very close.

Figure 3.20 shows the prevalence of reported diabetes among participants distributed by the BMI category according to their weight.

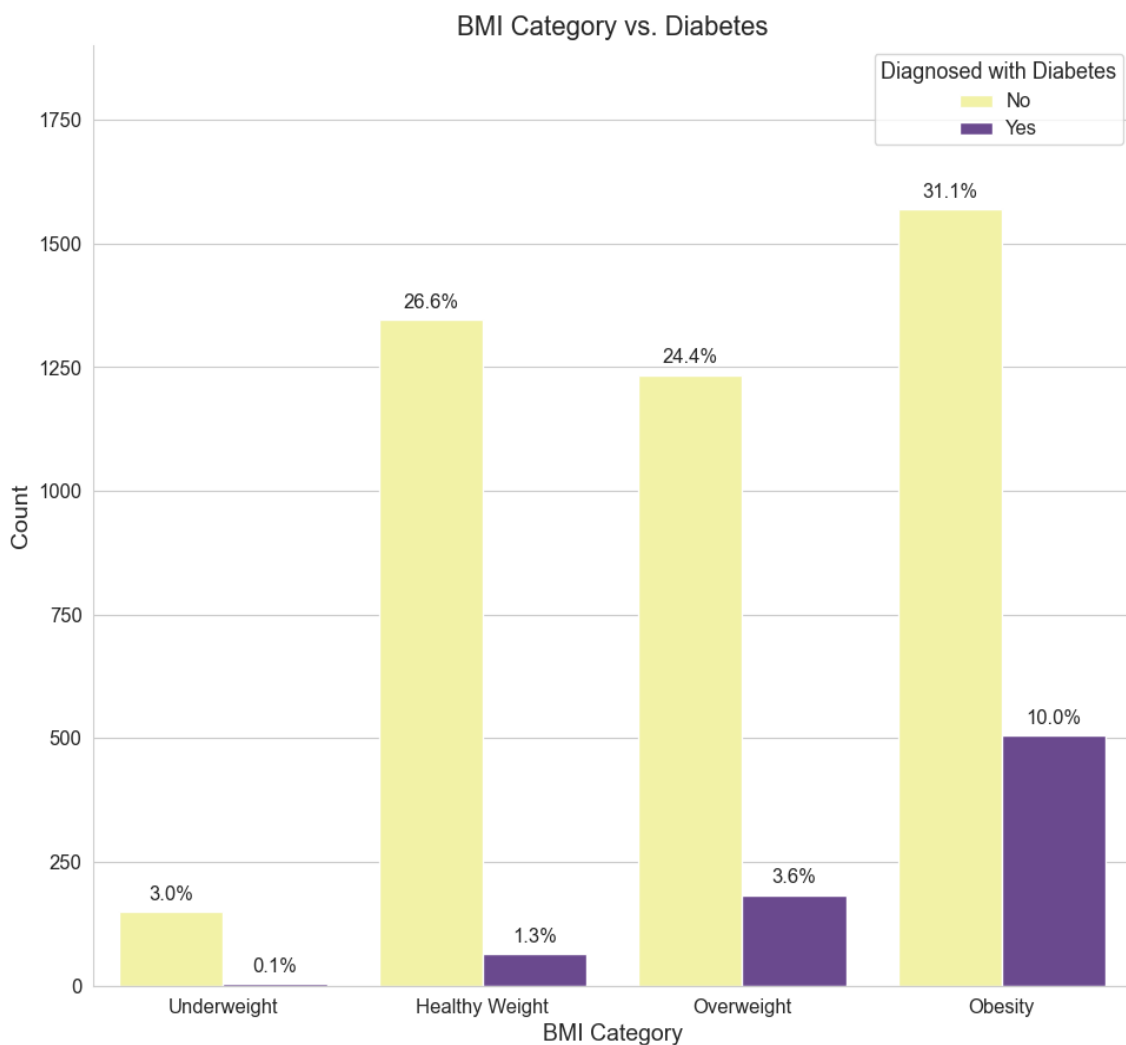


Figure 3.20 The Prevalence of Diabetes Among Participants by their BMI Class

It is observed in Figure 3.20 that in the case of participants who are classified as obese, it is obvious that around a third of them are diagnosed with diabetes, for participants in the

overweight class, around 15% of them are diabetic patients, while these percentages are significantly lower in the healthy weight and underweight categories. This indicates a relationship between obesity and overweight with diabetes, which is determined as a risk factor for developing diabetes (Abdullah et al., 2010; Jafar et al., 2006; Nguyen et al., 2008).

Figure 3.21 shows the prevalence of reported hypertension among participants distributed by the BMI category according to their weight. The majority of participants who are diagnosed with hypertension are found in the obesity and overweight categories based on BMI.

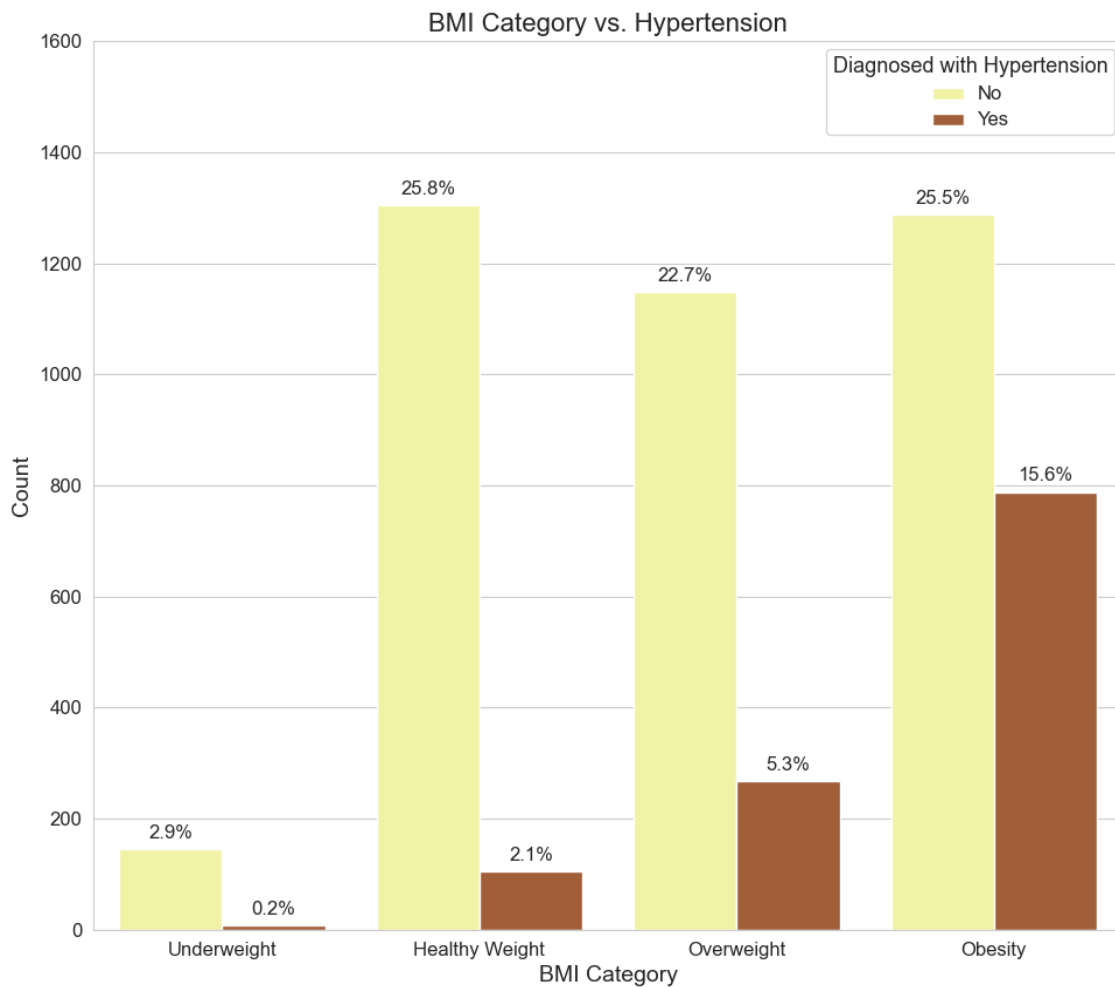


Figure 3.21 The Prevalence of Hypertension Among Participants by their BMI Class

As observed in Figure 3.21, the largest segment of participants who were diagnosed with hypertension fell into the obesity class. The second largest segment of hypertensive

participants were in the overweight BMI class. This finding emphasizes the strong correlation between unhealthy weight and hypertension, which is determined as a risk factor for developing hypertension and NCDs in general (Abdullah et al., 2010; Jafar et al., 2006; Nguyen et al., 2008).

Figure 3.22 shows a comparison of the levels of sugar intake between the two groups of raised and not-raised blood sugar.

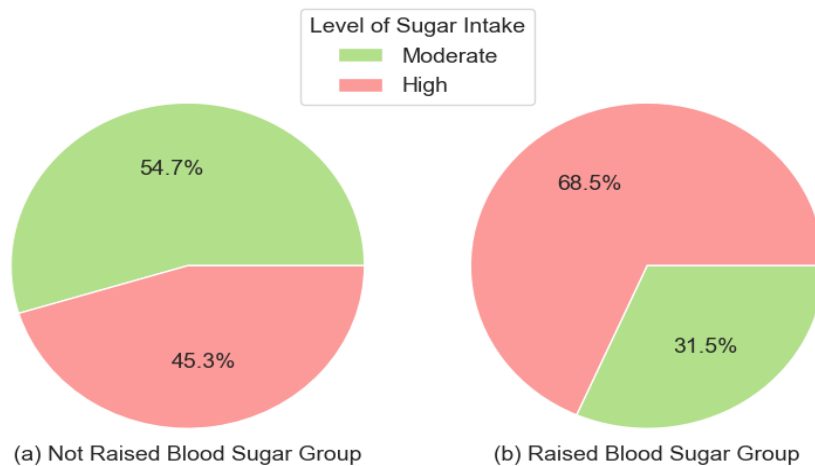


Figure 3.22 Comparison of the Sugar Intake Between the Raised and Not Raised Blood Sugar Groups

In Figure 3.22, it can be observed that in the raised blood sugar group shown in the second pie chart, the percentage of high levels of sugar intake is significantly higher than the corresponding one in the first group of “Not Raised Blood Sugar”.

Figure 3.23 shows a comparison of the levels of salt intake in relation to blood pressure.

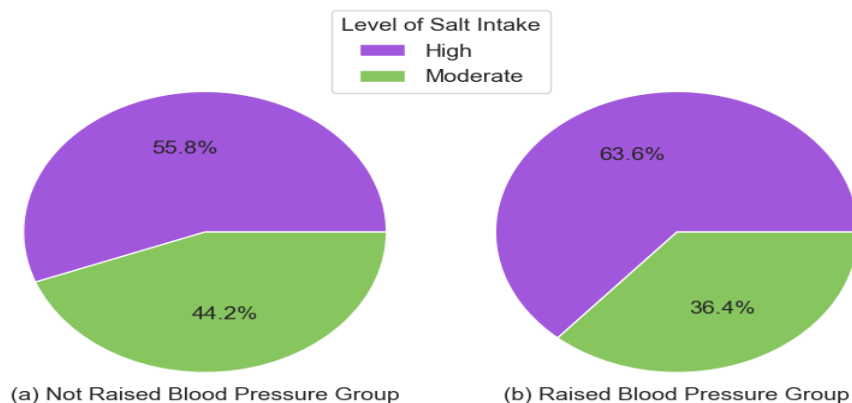


Figure 3.23 Comparison of the Salt Intake Between the Raised and Not Raised Blood Pressure Groups

As shown in Figure 3.23, a comparison of the two pie charts shows that the percentage of high levels of salt intake is greater in the raised blood pressure group on the second pie chart than it is in the “Not Raised Blood Pressure” group on the first pie chart.

3.5.3 Analysis of Key Features of Interest

In the raised blood pressure and sugar models proposed in this study, the systolic and diastolic blood pressure, and the fasting blood sugar represent the key features of interest that are used in deriving the target variables utilized in building the proposed prediction models.

Figure 3.24 shows the distribution of the observations according to the results of the fasting blood sugar test.

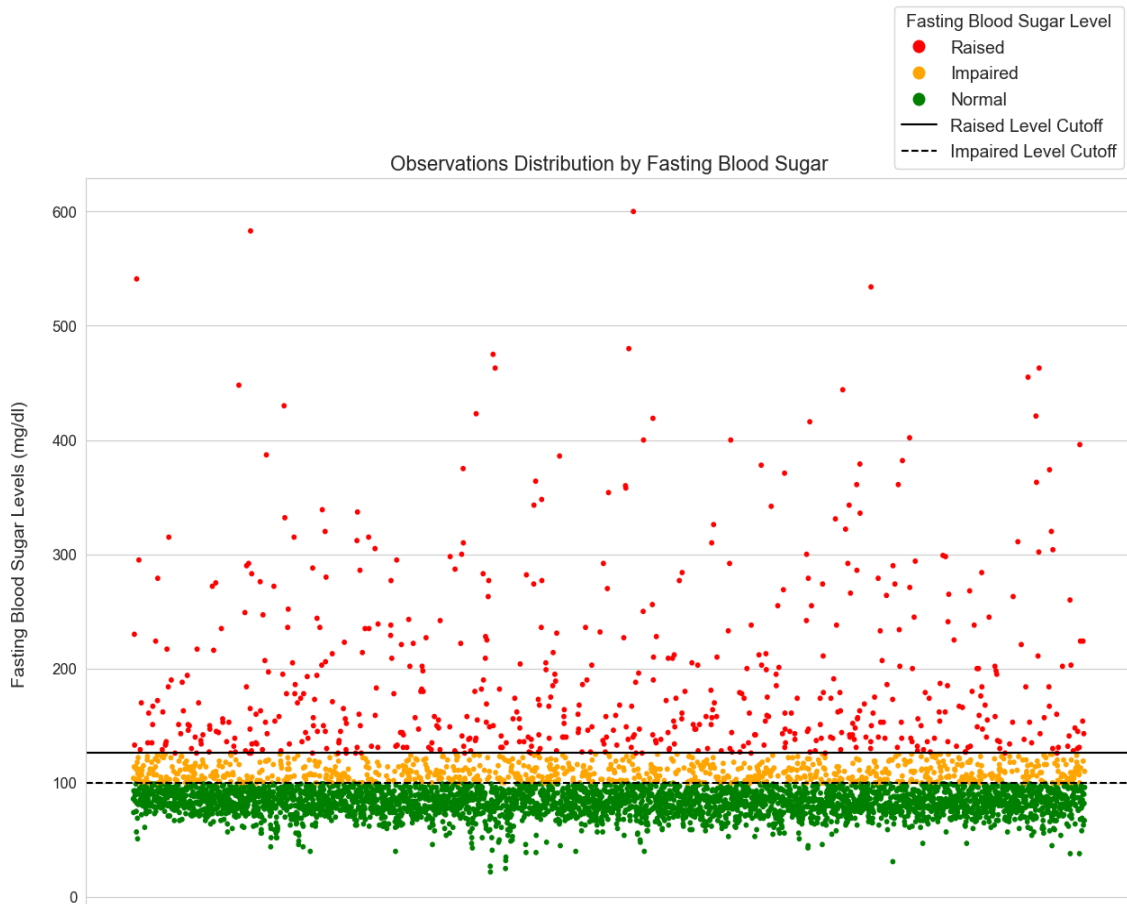


Figure 3.24 The Distribution of Participants by the Levels of Fasting Blood Sugar

In Figure 3.24 the green observations indicate subjects with fasting blood sugar levels less than 100 mg/dL, which belong to the normal class, while the observations illustrated

in orange, which are located above the dashed line represent the impaired cases who have measurements of fasting plasma glucose from 100 to 125 mg/dL. The observations of fasting blood sugar that are higher or equal to 126 mg/dL, which are depicted in green in the graph and situated above the continuous line that represents the threshold of raised levels, are classified as raised blood sugar cases. For the raised blood sugar cases, it is evident that the observations are sparse indicating irregular sugar levels in blood that might be due to a certain disorder.

Figure 3.25 shows the distribution of the observations according to the measurements of systolic blood pressure.

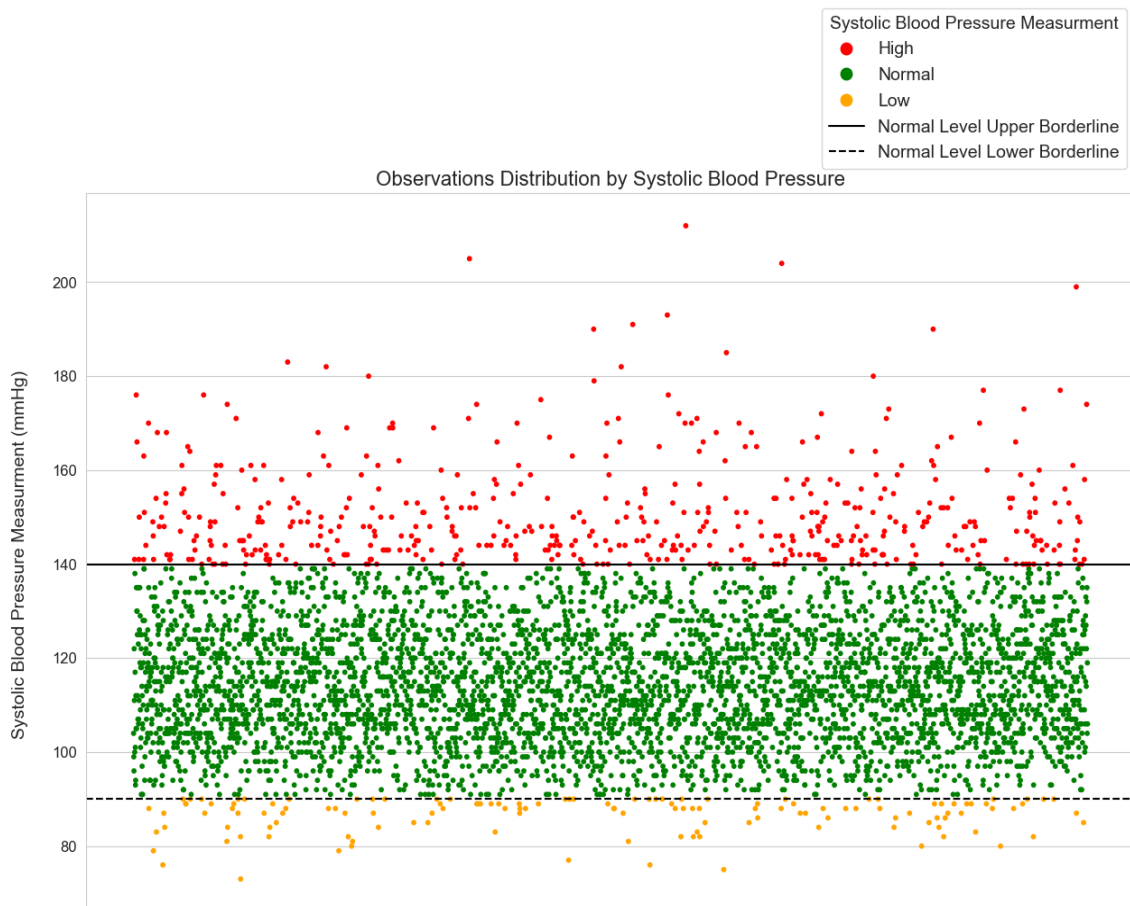


Figure 3.25 The Distribution of Participants by the readings of Systolic Blood Pressure

In Figure 3.25 the shown distribution is based on the systolic reading of the blood pressure. The observations illustrated in green indicate cases with normal systolic blood

pressure that fell between 90 and 140 mmHg. The observations of systolic blood pressure that is lower than 90 mmHg, which is depicted in yellow in the graph and situated below the dashed line that represents the lower threshold of normal blood pressure reading, are classified as low blood pressure cases. The observations illustrated in red, which are located above the continuous line represent the elevated cases, who have systolic blood pressure readings higher than 140 mmHg.

Figure 3.26 shows the distribution of the observations according to the measurements of diastolic blood pressure.



Figure 3.26 The Distribution of Participants by the readings of Diastolic Blood Pressure

Figure 3.26 represents the distribution of observation based on diastolic blood pressure. The observations illustrated in green indicate normal cases having diastolic pressure

between 60 and 90 mmHg. The yellow observations of diastolic blood pressure that are lower than 60 mmHg, are recognized as low diastolic blood pressure cases. The observations illustrated in red, which are located above the continuous line represent the elevated cases, who have diastolic blood pressure readings higher than 90 mmHg.

3.6 Data Preprocessing

The preparation of data is an imperative major component of modeling, in order to rectify any problems within the data prior to passing it into the model training phase. The impact of poor data quality is not limited to affecting the performance of the model only, it also can hinder the process of building the machine learning-based model. Thus, applying the data preparation and preprocessing techniques properly ensures optimal results by utilizing clean and uniform data. Given that data quality significantly influences the effectiveness of prediction models, the prior step applied before implementing the proposed model is preprocessing the STEPS cleaned dataset. Referring to Figure 3.1 presented in the first section of this chapter, the steps involved in the fourth phase of data preprocessing are outlined in this section and illustrated in Figure 3.27.

Figure 3.27 outlines the employed techniques in preprocessing the dataset before the implementation of the machine learning algorithms.

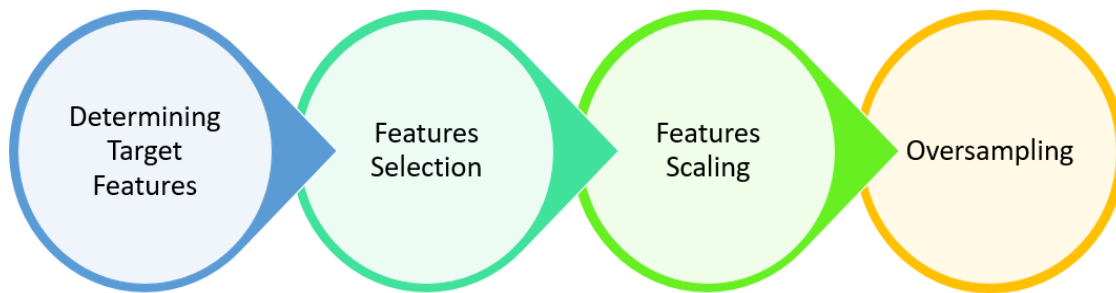


Figure 3.27 The Preprocessing Methods Applied on STEPS Dataset

Data preparation and preprocessing of the STEPS dataset for building the machine learning-based models to predict raised blood sugar and pressure involved the following activities.

3.6.1 Determining Target Feature

In the proposed “Raised Blood Pressure and Sugar Automatic Detection and Prediction Models”, the constructed models operate as supervised classification tools. This implies that the predicted target feature is a variable of categorical type, where categories within are indexed as discrete values.

In the raised blood sugar prediction model, the target feature is a binary variable derived from the fasting blood sugar reading that is available in the provided STEPS raw dataset. The newly created feature indicates whether the participant has a raised blood sugar or not, based on the threshold identified by WHO, which determines levels of 126 mg/dL or higher of fasting blood sugar as elevated blood glucose cases (The World Health Organization, 2024a). The positive class in this feature reflects the raised blood sugar cases, while the negative class means not raised cases with fasting blood sugar levels less than 126 mg/dL, which involves both normal and impaired cases.

For the second model to predict the raised blood pressure, the target feature to predict, was created using the average systolic and diastolic blood pressure measurements, that are available in the collected dataset. The positive class in this feature reflects high blood pressure cases, involving subjects who recorded a value of blood pressure reading higher than or equal to 140 mmHg for the systolic component or higher than or equal to 90 mmHg for the diastolic component [225]. The negative class in this target feature reflects not raised pressure cases, which involve both normal and low-pressure cases. Both target features were created in the earlier steps within the data cleaning phase, to be studied in the exploratory data analysis.

3.6.2 Features Selection

Various techniques were employed to identify the most significant features for constructing the Raised Blood Sugar prediction model and eliminating the irrelevant features, which was concluded by several researchers and previous studies to have an impact on improving the accuracy of classification algorithms (Amarnath & Appavu Alias Balamurugan, 2016; R. C. Chen et al., 2020). The involved feature selection techniques in this study encompassed using correlation matrix, Chi-Square, and random forest classifier feature importance.

- 1- The first method used in the feature selection process is calculating the correlation coefficients between the features of the clean dataset. The correlation matrix is a helpful method as an initial step to explore the relationships between the set of independent variables. This approach can support in determining the set of independent variables that could be eliminated due to high correlation which may affect the model performance badly (Sundus et al., 2022). The impact of multicollinearity is not an issue specific to regression models only, but it may affect classification models as well (Misra & Yadav, 2020). Its impact on classification models involves both the stability and the interpretability of the model (Drobnič et al., 2020).

In this study, the criteria that were followed in handling the multicollinearity issue between independent variables were based on eliminating one of the independent features that has a correlation coefficient value greater than 0.7 with another independent one (Dormann et al., 2013).

Figure 3.28 shows the correlation matrix for a subset of independent features related to the medical history of participants, the raised blood pressure (RBP), and raised blood sugar (RBS) target features.

In Figure 3.28, it can be seen there is a high collinearity between a set of independent features, represented in the relation between prevalent hypertension and taking high blood pressure medication features, the second collinearity is found between prevalent diabetes and taking diabetes medication features, and the third one between prevalent cholesterol and raised cholesterol medication features. All of these combinations have collinearity with a correlation coefficient value of 0.8, which is greater than the threshold of 0.7. In this step, the selected features to eliminate are those related to taking medication variables.

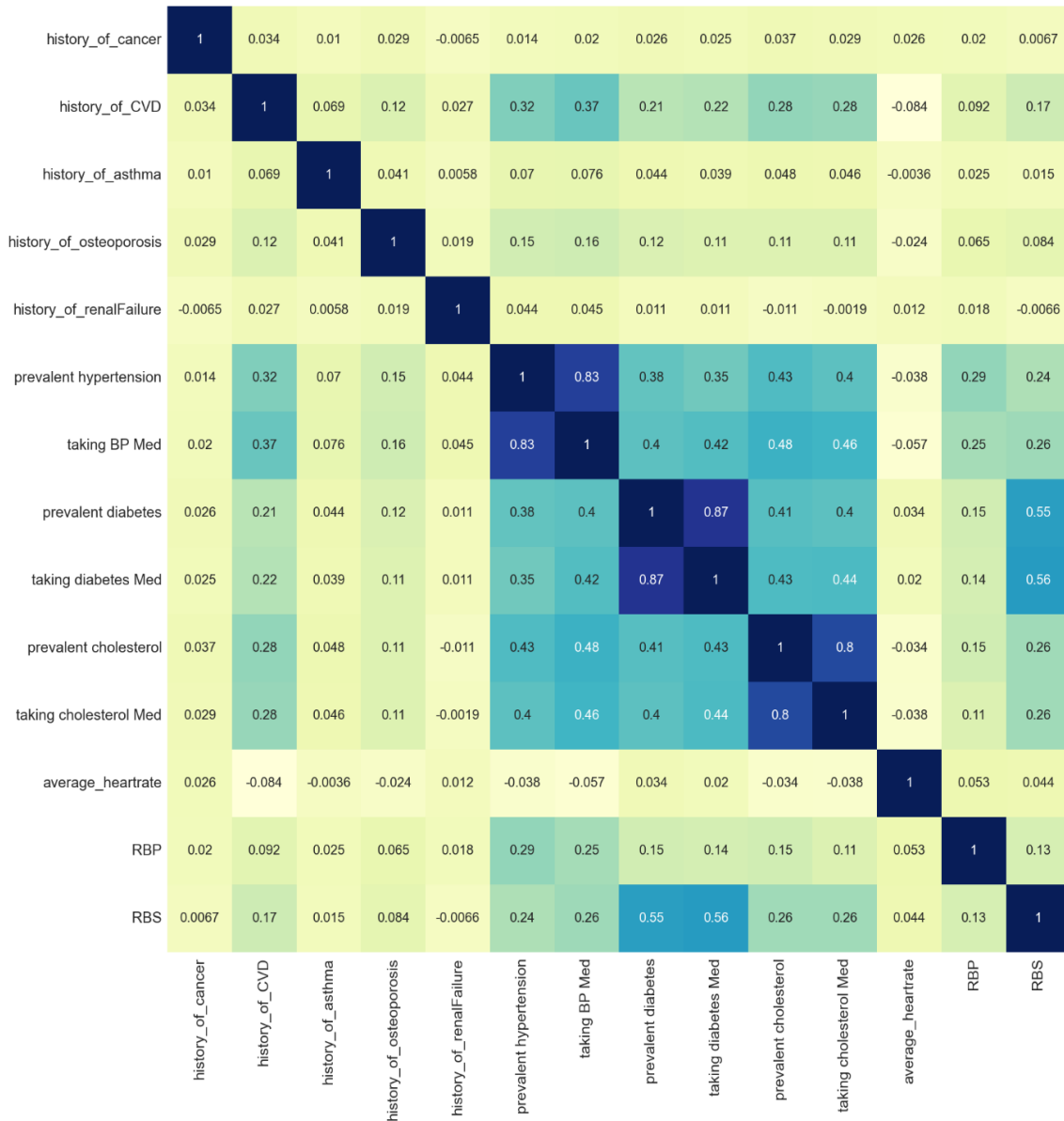


Figure 3.28 Variables Correlation Matrix

- 2- The Chi-Square test feature importance was used to identify the categorical independent set of variables that are correlated with the target variables for raised blood sugar and raised blood pressure. Incorporating the Chi-Square method in the process of feature selection can optimize the preprocessing of the dataset and thus the predictive power of the model since it helps in objectively selecting the final set of input features to be used in training the machine learning-based models (Sikri et al., 2023). Figure 3.29 below shows the list of independent categorical features identified as the top 15 important features in predicting the target feature

of raised blood sugar. This list included the history of cholesterol, history of hypertension, history of CVD, raised blood pressure, level of sugar intake, history of osteoporosis, physical inactivity, sleep disturbances, gender, former smoker, level of salt intake, current smoker, anxiety and depression (PHQ4), insufficiency of fruit and vegetable intake, and history of asthma, as the most important features identified by Chi-Square test for the prediction of raised blood sugar. The history of diabetes was not forwarded to this test of feature importance since it is correlated with the outcome variable which can affect the predictive power of the model.

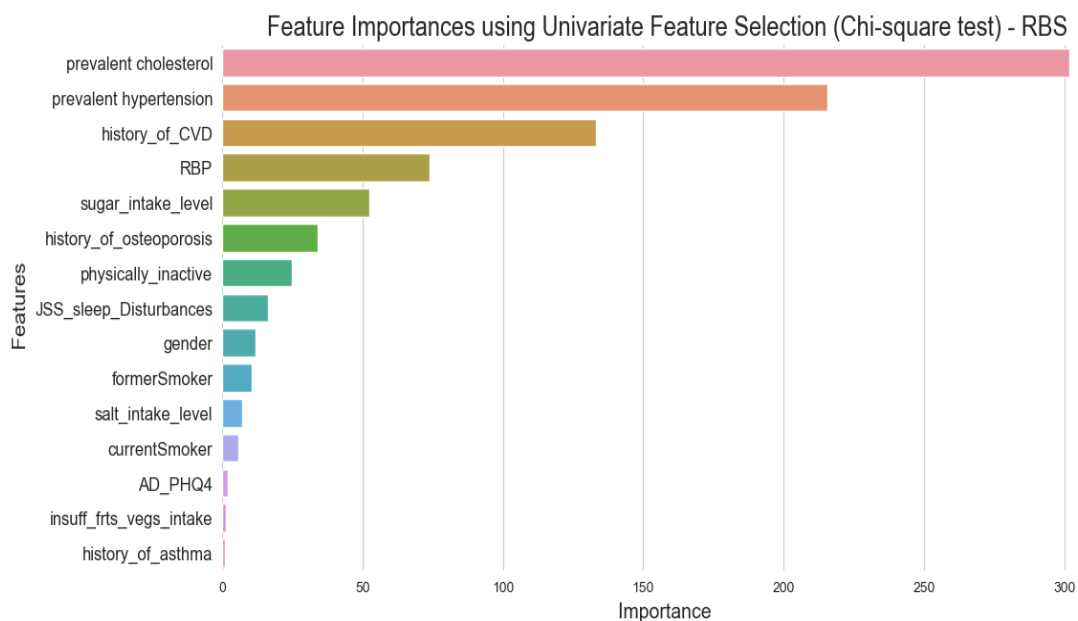


Figure 3.29 Feature Importance using Univariate Feature Selection (Chi-square test) - RBS
Target Feature

The same approach was applied to the target feature for predicting the raised blood pressure, the identified important features according to the Chi-Square test were, the history of high cholesterol, history of diabetes, raised blood sugar, history of CVD, history of osteoporosis, sugar intake level, former smoker, sleep disturbances, salt intake level, physical inactivity, inadequate sleeping hours, history of asthma, insufficient intake of fruits and vegetables, history of cancer, history of renal failure, as shown in Figure 3.30 below. Hypertension history was

not included in the test of feature importance, as it is correlated with the target variable and could impact the model's predictive power.

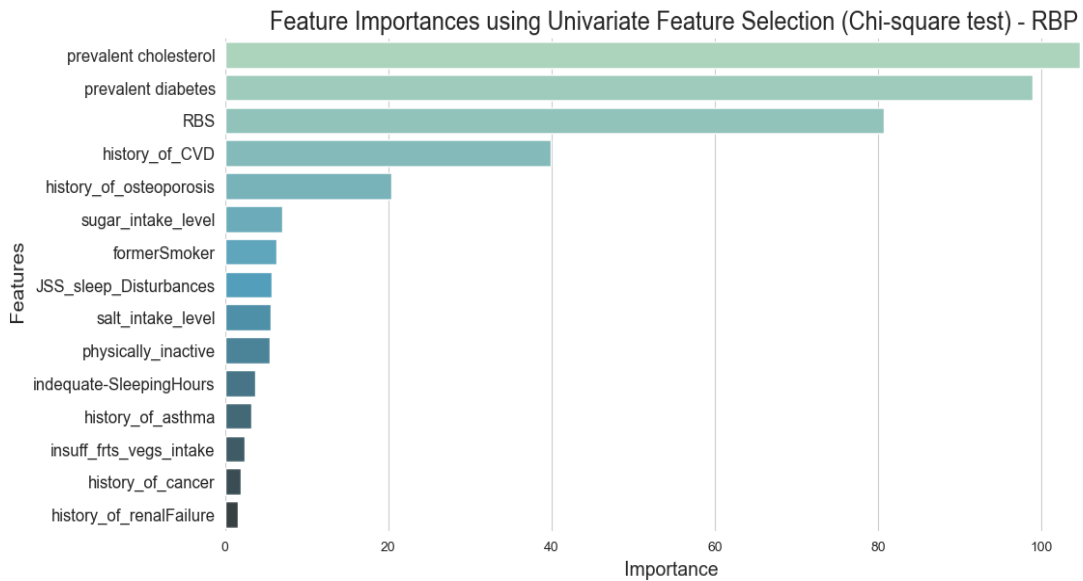


Figure 3.30 Feature Importance using Univariate Feature Selection (Chi-square test) - RBP Target Feature

This step helped in identifying the important predictor variables for the outcome features related to raised blood sugar and raised blood pressure prediction models. However, no variable was eliminated within this step, the variables set were forwarded for further phase of exploring features importance using the Random Forest classifier.

- 3- A Random Forest classifier was utilized as the final step in the feature selection process. The Random Forest classifier was used to obtain the optimal feature selection process, by forwarding all types of variables either categorical or continuous as input for identifying their performance to the outcome variables in the proposed models. Utilizing the Random Forest in the variable's selection process is considered one of the most powerful methods to determine the appropriate and significant features that contribute to the prediction of the outcome feature in machine learning models (Khan et al., 2020). The high efficiency of the Random Forest classifier in identifying features is due to its mechanism of building multiple decision trees, in which each one of these trees is

trained using a bootstrap sample of input features, which are selected on a random basis. The results of all trees are aggregated to obtain the final result and to determine the contribution of each variable in predicting the outcome feature (Reif et al., 2006).

Figure 3.31 illustrates the results of feature importance that were obtained by integrating the Random Forest classifier in selecting the top 30 features for predicting the raised blood sugar outcome variable. Based on the given results, the final list of features that will be used for the next stage of training the raised blood sugar prediction model are age, triglyceride, waist-to-hip ratio, BMI, waist circumference, heart rate, total Cholesterol, hip circumference, HDL cholesterol, history of cholesterol, sugar intake level, history of hypertension, inadequate sleeping hours, insufficient physical activity, anxiety and depression, level of salt intake, history of CVD, raised blood pressure, passive smoking, mental health ill-being, JSS sleep Disturbances, insufficient intake of fruits and vegetables, gender, smoking, history of osteoporosis, former smoker, history of asthma, history of cancer, history of renal Failure, and alcohol consumption.

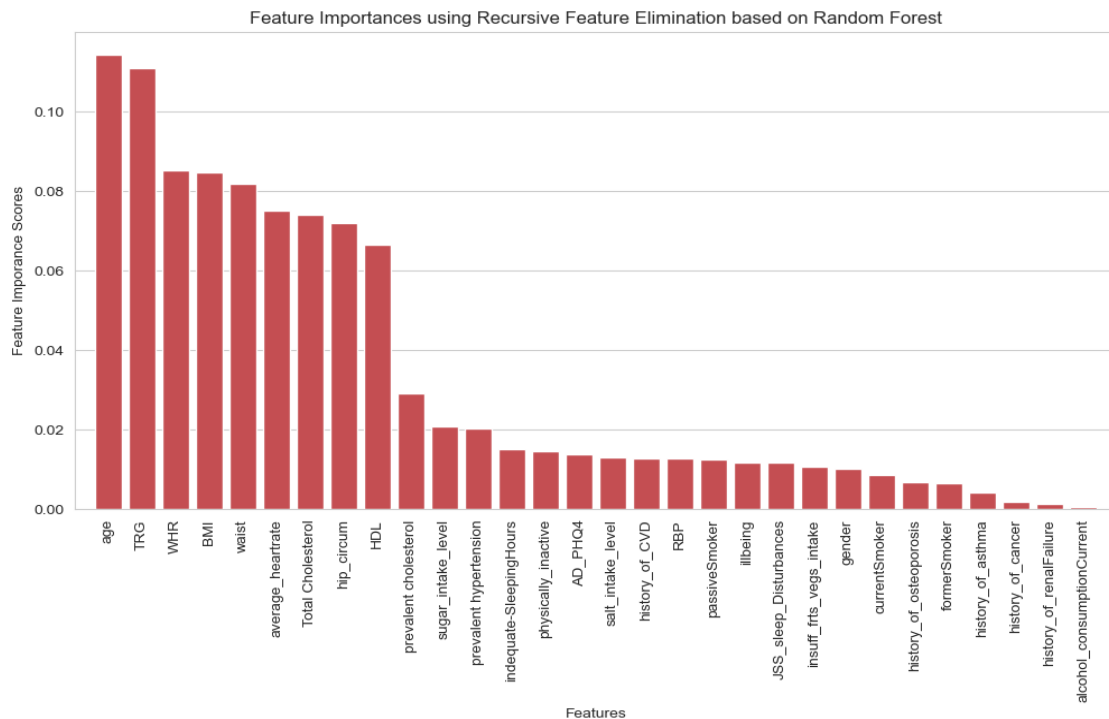


Figure 3.31 Feature Importance using Recursive Feature Elimination based on Random Forest - RBS

The same steps were applied to identify the top 30 features for predicting the raised blood pressure outcome variable, as shown in Figure 3.32. Based on the shown results, the final list of features that will be used for the next stage of training the raised blood pressure prediction model is age, BMI, triglyceride, waist-to-hip ratio, waist circumference, Fasting blood sugar, heart rate, total cholesterol, HDL, hip circumference, sugar intake level, salt intake level, inadequate sleeping Hours, ill-being, physical inactivity, gender, passive smoking, history of cholesterol, anxiety and depression, sleep disturbances, dietary fruits and vegetable intake, history of diabetes, smoking, history of CVD, RBS, former smoking, history of osteoporosis, history of asthma, history of cancer, history of renal failure.

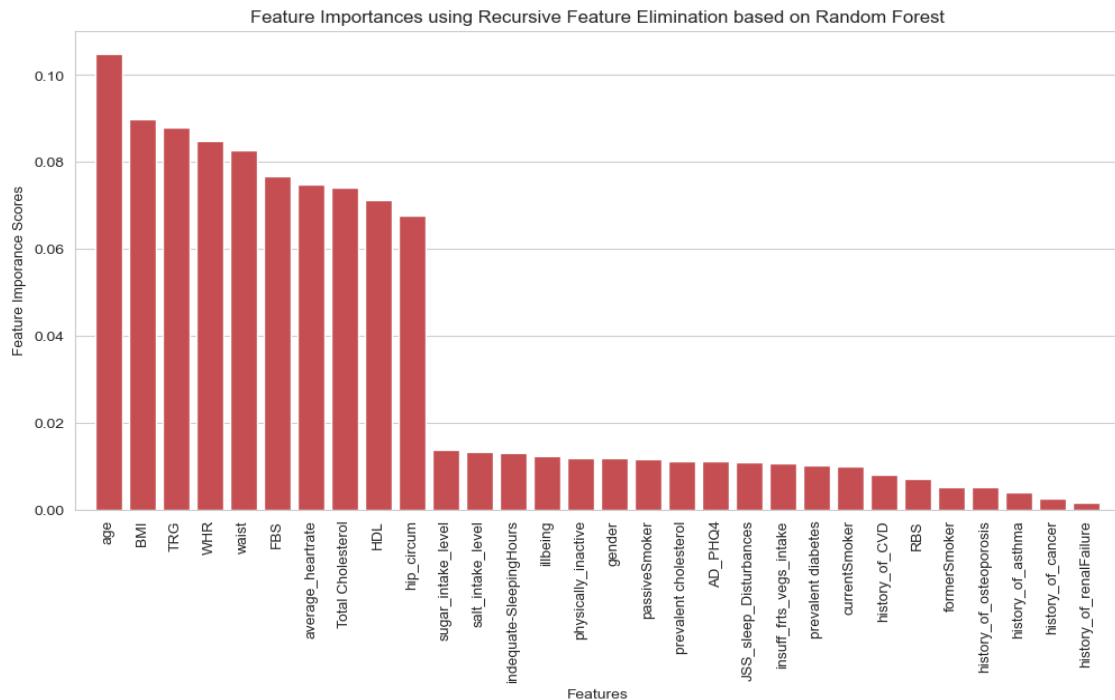


Figure 3.32 Feature Importance using Recursive Feature Elimination based on Random Forest – RBP

As observed in Figure 3.31 and Figure 3.32, the features that are identified as predictors in both models for predicting raised blood sugar and raised blood pressure are almost the same with different rankings, which is consistent with the findings of previous studies that the risk factors for raised blood pressure and raised blood sugar are common (Huang, 2009; Tarleton et al., 2014).

The feature selection process combined Chi-square and Random Forest methods to ensure a robust selection of relevant variables. Chi-square evaluated the statistical significance of categorical features, while Random Forest assessed feature importance based on their contribution to model accuracy. Combining both methods in this process provided complementary insights, Chi-square identified features with strong associations, while Random Forest captured complex, non-linear relationships that enhanced model performance.

3.6.3 Features Scaling

The STEPs dataset contains numerical features from various scales, which can pose computational challenges during predictive modeling. Therefore, the min-max transformation algorithm was applied to the data to normalize it, by scaling the continuous features into smaller magnitudes, to minimize the computational complexity. This technique normalizes the features to a range between 0 and 1, as shown in Equation (1) (Raju et al., 2020).

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}, \quad (1)$$

3.6.4 Dataset Oversampling

Through exploration of the processed dataset, it was obvious that the target features, indicating raised blood sugar level, and raised blood pressure exhibit class imbalance. Specifically, only 11.4% of observations in the dataset indicate raised blood sugar levels, while 88.6% signify normal blood sugar levels, as shown in Figure 3.33 below. This imbalance can adversely affect machine learning algorithms, as class imbalance is a significant challenge in classification algorithms (Cecchini et al., 2019). To mitigate this problem, the oversampling method was implemented (Gosain & Sardana, 2017). This involved replicating observations from the minority class, which refers to subjects recognized with fasting blood sugar levels ≥ 126 mg/dl, to match the size of the subjects in the majority class that were tested with fasting glucose levels lower than 126 mg/dl, to resolve this imbalance issue.

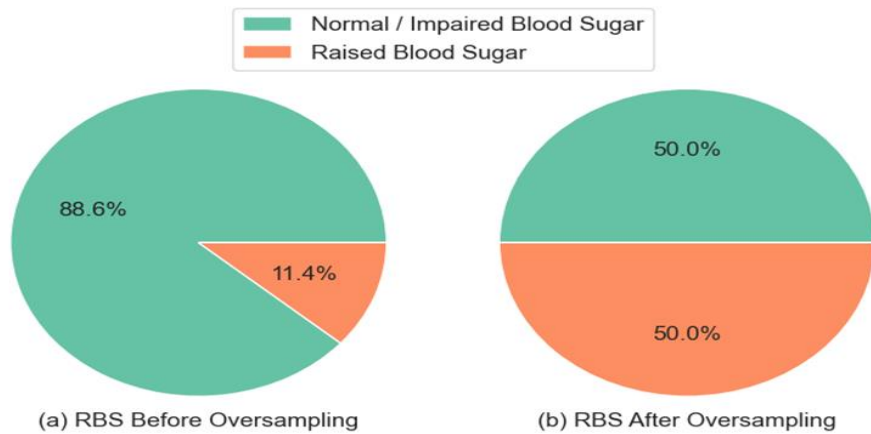


Figure 3.33 Distribution of Cases by the Level of Blood Sugar Before (A) and After (B) Applying Oversampling

Likewise, the target feature of the second model for predicting raised blood pressure, which indicates a high reading of systolic blood pressure that is higher or equal to 140 mmHg, or diastolic blood pressure of 90 mmHg or higher, has the same problem of imbalance within the positive class of the dataset. The same resampling mechanism was applied to address this issue in predicting raised blood pressure, by duplicating the instances of the smaller segment corresponding to cases with raised blood pressure, to align with the number of instances from the negative class encompassing normal or low-pressure cases. Figure 3.34 depicts the presence of the class imbalance problem in the target feature of the raised blood pressure detection model.

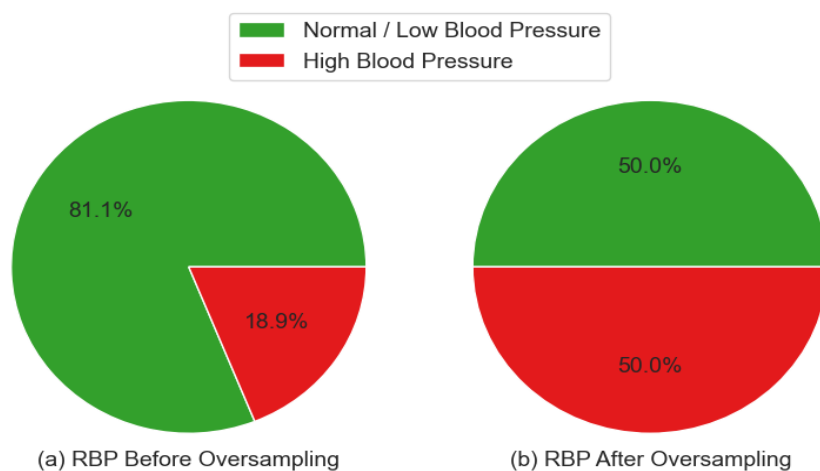


Figure 3.34 Distribution of Cases by the Blood Pressure Before (A) and After (B) Applying Oversampling

3.7 Machine Learning Models

This section outlines the workflow and machine-learning algorithms used to implement the proposed models and the algorithms utilized to predict raised blood pressure and raised blood sugar.

The techniques utilized in this phase and their results are highly dependent on the preprocessing steps presented in the previous section. Referring to Figure 3.1 in the first section of this chapter, Figure 3.35 below outlines the details of the last three phases and their components used to construct the proposed models. These three phases are closely linked, particularly for building machine learning models.

The workflow in Figure 3.35 highlights the core phases of the entire process, for constructing the proposed models from a machine learning perspective. This includes the essential data preprocessing phase before model training outlined in the previous section 3.6, the training and testing of the constructed models presented in this section, and the final phase for performance evaluation outlined in the next section.

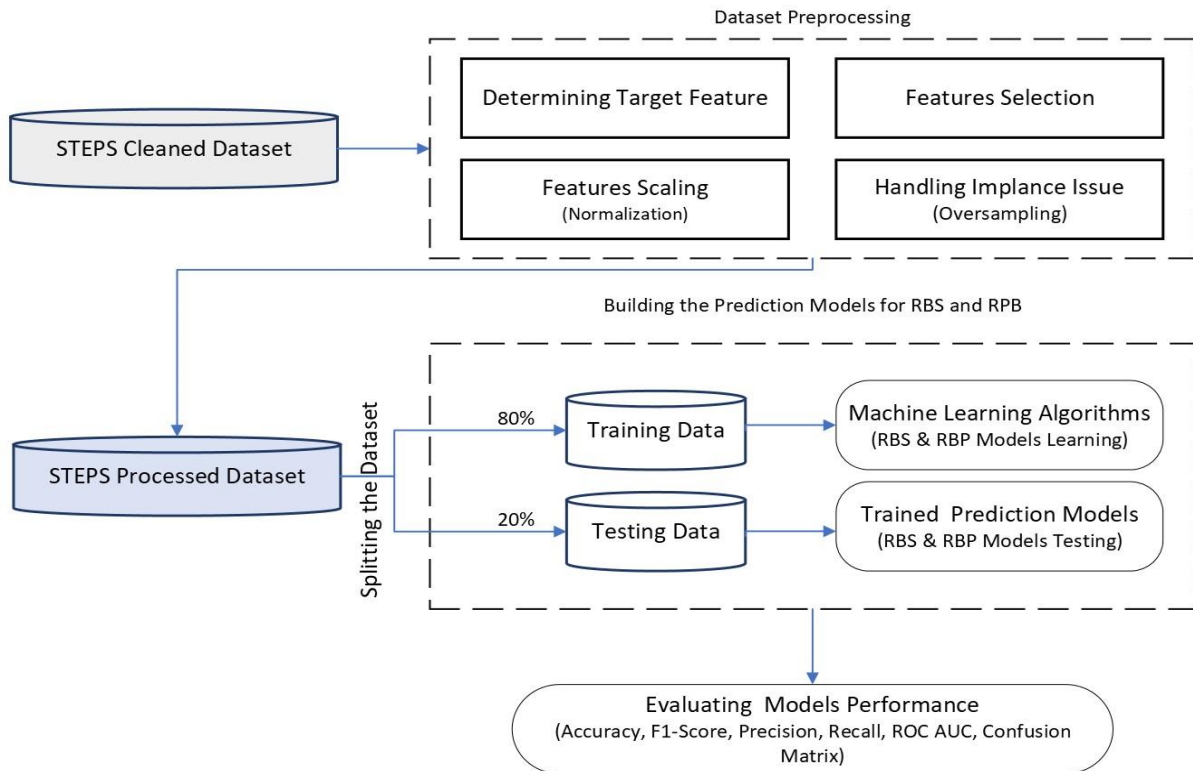


Figure 3.35 The Workflow of the Proposed Prediction Models for Raised Blood Pressure and Sugar

The two prediction models for raised blood pressure and raised blood sugar were implemented as separate models, each one used its clone of the processed dataset, with changing the target feature and the set of input features according to the results concluded in the feature selection process.

The technical framework that was used for implementing both models is the same. The framework of building the machine learning-based models for predicting raised blood pressure and raised blood sugar started with passing the cleaned data into the preprocessing phase. Although the cleaned dataset was well prepared, with proper handling of any possible inconsistencies or anomalies. However, this might be sufficient in statistical modeling. For modeling by machine learning, the situation is different, due to the higher computational complexity, which requires that data is prepared in the best representation. Therefore, the application of the previous preprocessing steps was highly necessary. The next step in the modeling process is splitting the processed dataset into two subsets, one of them for training the models using 80% of observations from the entire processed dataset, and the remaining portion of 20%, to be used later for testing the results of the trained models. Subsequently, after the phase of model learning is completed, the testing subset is passed into the trained models. Finally, using the predicted outcomes, the models' performance measures are calculated to evaluate the effectiveness of the implemented models.

3.7.1 Machine Learning Classifiers

The raised blood pressure and raised blood sugar prediction models are implemented using a selection of machine-learning algorithms that are known for their high predictive performance including Random Forest, Decision Tree, XGBoost, and Multi-Layer Perceptron. The parameter settings for each algorithm employed in these models are configured using the GridSearch approach to set up the parameters that yield the best performance results.

- Decision Tree is a machine-learning algorithm, that is based on a supervised learning approach, it can be utilized to build classification and regression models. It is marked among the easiest straightforward machine learning algorithms, that is based on arranging the features in a tree structure, and recursively splitting them

based on chosen impurity criteria, such as the Gini index, entropy, or information gain value (Sharma & Kumar, 2016).

- Random Forest is among the top robust supervised machine-learning algorithms commonly utilized in classifying subjects and predicting an outcome based on an input of labeled data. It relies on constructing a group of several independent decision trees, variables in the built trees among the group, are chosen on a random basis. The prediction outcomes of all the generated trees are aggregated to obtain the optimized accumulated result, which improves the accuracy and stability of decision trees by using the ensemble approach. Through utilizing the ensemble technique, the Random Forest achieves improved performance by mitigating the high variance issue that is known as a common issue in the Decision Tree (Ziegler & König, 2014).
- XGBOOST algorithm, short for Extreme Gradient Boosting, is an enhanced edition of its earlier Gradient Boost algorithm, based on a supervised learning approach in machine learning, and its mechanism relies on generating prediction trees. It can be applied for regression and classification problems, in particular for high-volume data due to its optimized performance in generating accurate predictions within a fast training pace. The working approach of the XGBOOST is based on passing the outcome of a processed tree into the next tree sequentially (T. Chen & Guestrin, 2016).
- The Multi-Layer Perceptron Classifier is a feedforward ANN algorithm used for classification problems, consists of multiple layers of interconnected nodes associated with weights, and uses an iterative optimization algorithm of backpropagation to minimize error and optimize classification results. The MLP classifier has the well-known advantage of handling the nonlinearity issue of relationships in the processed data (Fiesler & Beale, 2020).

3.7.2 Experimental Settings

The following table highlights the parameters hyper-tuning and settings of each machine learning model employed in developing the models throughout this thesis for building the raised blood pressure and sugar detection models, including the Decision Tree, Random Forest, XGBoost, and Multilayer Perceptron neural networks.

Table 3.3 outlines the adopted machine learning algorithms for implementing the proposed models and their parameter settings in this thesis. Hyper-tuning for the parameters in Decision Tree, Random Forest, and K-Nearest Neighbor algorithms was done using the “Grid Search” technique for specifying hyperparameter values that result in the highest accuracy values for each algorithm (Belete & Huchaiah, 2022).

Table 3.3 Configuration of RBS & RBP Models Parameters

Algorithm	Parameters Setting	
Random Forest	criterion	Gini
	max_depth	300
	n_estimators	3000
	min_samples_split	2
	min_samples_leaf	1
Decision Tree	criterion	Gini
	max_depth	22
	min_samples_split	2
	min_samples_leaf	1
	class_weight	0: 0.5, 1: 0.5
XGBoost	n_estimators	5000
	max_depth	3
	learning_rate	0.1
Multi-layer Perceptron	hidden_layer_sizes	(1000, 300, 300)
	solver	Adam
	shuffle	False
	tolerance	0.0001

3.8 Models Performance Evaluation Criteria

Various measures are employed to evaluate the models' performance in predicting and detecting raised blood sugar cases and raised blood pressure cases as well. These metrics encompass accuracy, precision, recall, F1-score, confusion matrix, and ROC AUC, which are mainly calculated from fundamental measures, true positive (TP), false positive (FP), true negative (TN), and false negative (FN), which are used commonly to evaluate classification models (Dj Novakovi et al., 2017).

- Confusion matrix: used commonly to summarize the prediction results of the machine learning classification models, by comparing the actual values versus

predicted values, by which also the different performance metrics can be calculated. The confusion matrix consists of the following four measures:

TP: the number of records from the positive class predicted correctly by the model.

FP: the number of records from the negative class predicted incorrectly as a positive class by the model.

TN: the number of records from the negative class predicted correctly by the model.

FN: the number of records from the positive class predicted incorrectly as a negative class by the model.

- Accuracy: performance measure used to evaluate the efficiency of machine learning classification models. It is computed as the ratio of correct prediction to the total number of predictions (Dj Novakovi et al., 2017).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (2)$$

- Precision is calculated as the fraction of correct prediction from the positive class to the total number of predictions as a positive class (Dj Novakovi et al., 2017).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (3)$$

- Recall metric is the fraction of correctly predicted observations out of all actual observations of the positive class (Dj Novakovi et al., 2017).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4)$$

- F1-Score is calculated as the harmonic mean of precision and recall metrics (Dj Novakovi et al., 2017).

$$\text{F1-score} = 2 * \frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}, \quad (5)$$

- The Receiver Operator Characteristic (ROC) is a visual representation that shows how well a machine-learning model can differentiate between several classes, by plotting the True Positive Rate (TPR) and the False Positive Rate. ROC curve is

a visual evaluation method for the performance of classification models, which works by calculating the area under the curve (AUC), the greater the AUC value the better the performance of the classification model (Sø, 2009).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} , \quad (6)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} , \quad (7)$$

3.9 Conclusion

This chapter presented the proposed methodology used in constructing the raised blood pressure and raised blood sugar prediction models.

The next chapter will present the results of implementing the proposed models, along with a discussion of the obtained results, to obtain the model with the highest performance among the examined models utilized using different machine learning algorithms.

Chapter 4: Results and Discussion

4.1 Introduction

This chapter discusses the experimental results of the raised blood sugar and raised blood pressure detection models. The performance metrics presented in the previous chapter were used for evaluating the performance of the raised blood pressure and raised blood sugar detection models.

4.2 Raised Blood Sugar Detection Model

Table 4.1 presents the results of the performance evaluation yielded by testing the raised blood sugar detection models, examined with a selection of robust machine learning algorithms, that help to identify the best algorithm in building the proposed raised blood sugar detection models. Evaluating the models' performance was done using various measures including accuracy, recall, precision, and F1 Score. Table 4.1 outlines the experimental results using the classification performance metrics.

Table 4.1 The results of Performance Metrics for the Raised Blood Sugar Prediction Models

Algorithm	Random Forest	XGBoost	Decision Tree	MLP
Accuracy	98.05%	96.10%	94.65%	94.26%
F1-Score	98.11%	96.30%	94.98%	94.64%
Precision	96.3%	92.86%	90.54%	89.91%
Recall	100.00%	100.00%	99.89%	99.89%

According to the results presented in Table 4.1, the Random Forest algorithm has the highest scores in accuracy, precision, and F1-Score among all examined models. The scores were 98.05%, 98.11%, 96.3%, and 100.00% for accuracy, F1-Score, precision, and recall respectively.

Figure 4.1 presents the testing results for the raised blood sugar detection and prediction models using the ROC AUC, represented by the ratio of negative observations (not raised blood sugar levels cases) that were predicted incorrectly as raised blood sugar levels out of the total number of observation from the negative class (specificity), versus the ratio of positive observations (raised blood sugar levels cases) that were predicted correctly as

raised blood sugar levels cases out of the total number of observation from the positive class (sensitivity).

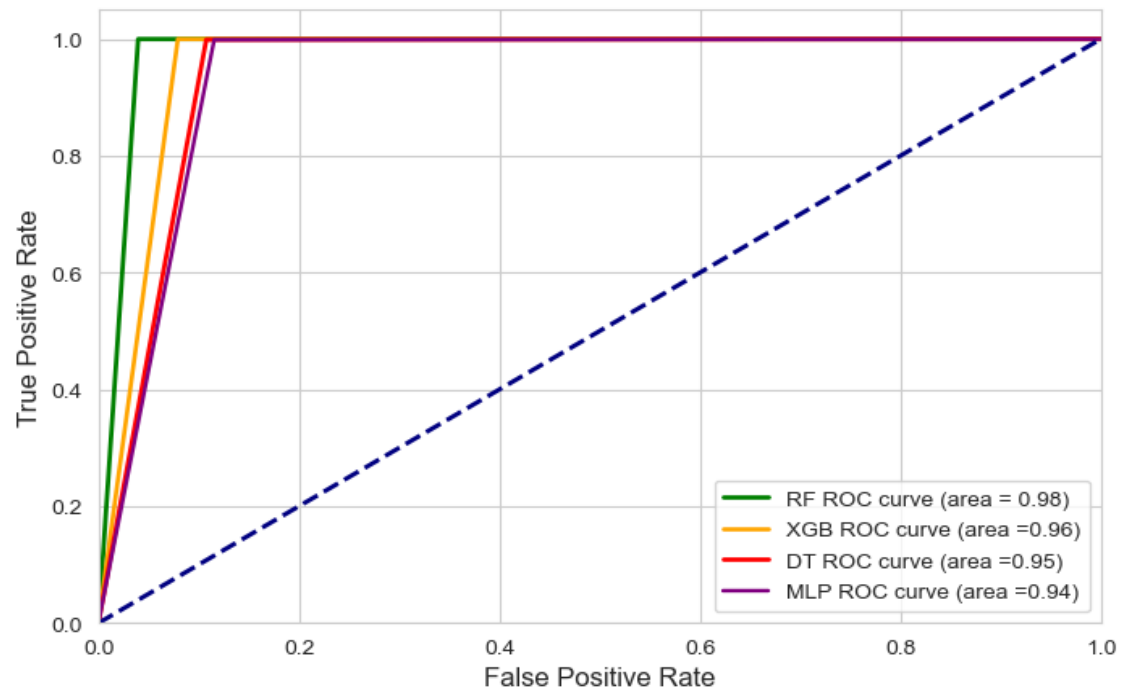
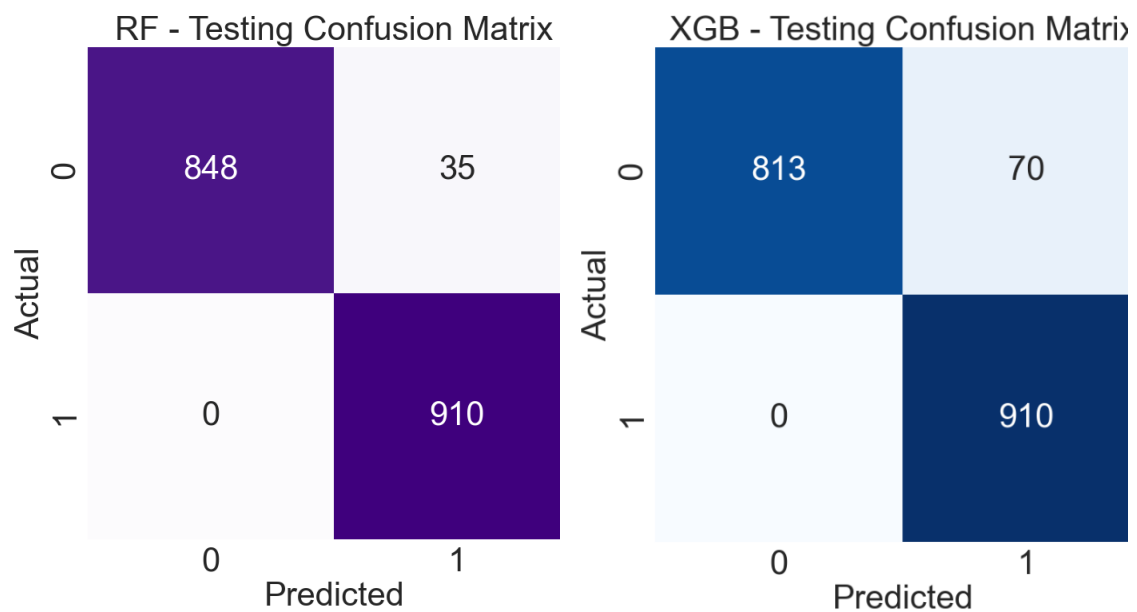


Figure 4.1 Raised Blood Sugar Prediction Models Performances Comparison using ROC AUC

The confusion matrix in Figure 4.2 summarizes the performance of the raised blood sugar prediction model implemented using the selected set of machine learning classifiers.



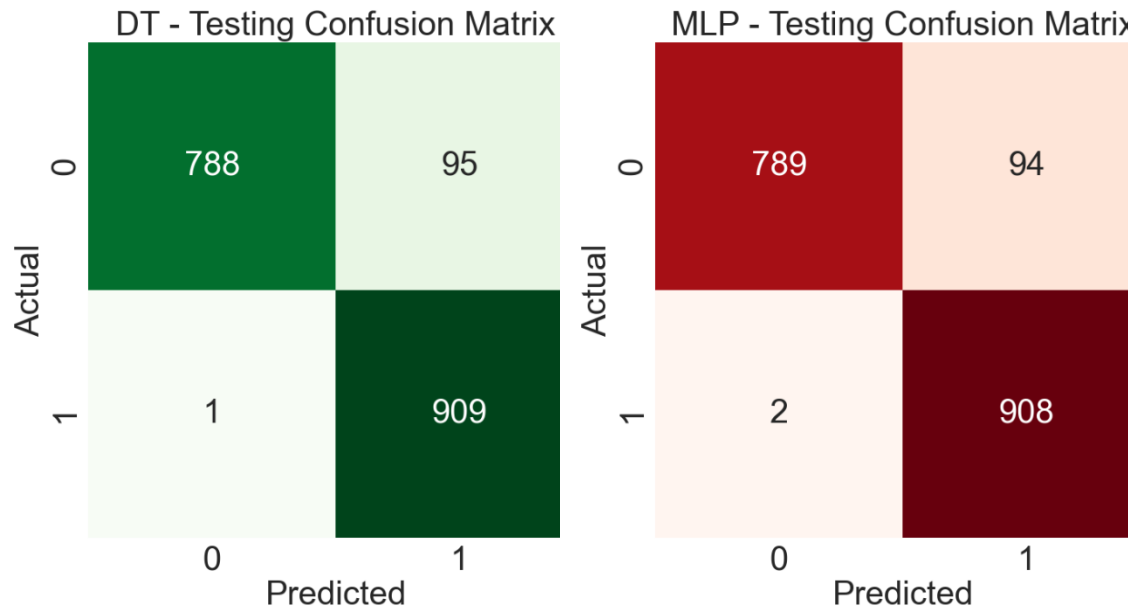


Figure 4.2 Comparison of Performance of the Raised Blood Sugar Detection Models Using Different Classifiers by Confusion Matrix

As shown in Figure 4.2, the Random Forest classifier, achieved the highest performance among the tested models, in showing the minimum number of misclassified predictions from the positive and negative classes, and a higher number of correctly classified instances, highlighted with the dark colors in the above figures.

From the presented results in Figure 4.1 and 4.2, it can be concluded that the Random Forest algorithm had the highest performance, achieving a sensitivity of 100%, a specificity of 96.04% an AUC value of 0.98, and an accuracy of 98.05%, followed by the XGBoost model with an AUC-value of 0.96, while the MLP-based model had the lowest performance with an AUC of 0.94 and accuracy of 94.26%.

4.3 Raised Blood Pressure Detection Model

This section presents the results of the performance evaluation yielded by testing the raised blood pressure detection models using Random Forest, Decision Tree, XGBoost, and Multiple Perceptron Neural Networks.

Using multiple algorithms in building the proposed model aims to ensure the construction of a model with optimized performance for detecting raised blood pressure cases.

The raised blood pressure detection models were evaluated through a set of metrics that are usually used for the assessment of the predictive power of classification models.

Table 4.2 outlines the experimental results using the classification performance metrics including accuracy, precision, recall, and F1-Score.

Table 4.2 The Results of Performance Metrics for the Raised Blood Pressure Prediction Models

Algorithm	Random Forest	XGBoost	MLP	Decision Tree
Accuracy	94.76%	91.59%	90.24%	89.45%
F1-Score	95.15%	92.46%	90.84%	90.69%
Precision	92.85%	87.67%	90.01%	84.81%
Recall	97.57%	97.80%	91.68%	97.46%

According to the results presented in Table 4.2, the Random Forest algorithm has the highest scores in accuracy, precision, and F1-Score among all examined models. The scores were 94.76%, 95.15%, and 92.85%, for accuracy, F1-Score, and precision respectively. Figure 4.3 presents the testing results for the raised blood pressure detection models using the ROC AUC, represented by the ratio of negative observations (not raised blood pressure subjects) that were predicted incorrectly as raised blood pressure cases out of the total number of observations from the negative class (specificity), versus the ratio of positive observations (raised blood pressure subjects) that were predicted correctly as raised blood pressure cases out of the total number of observation from the positive class (sensitivity).

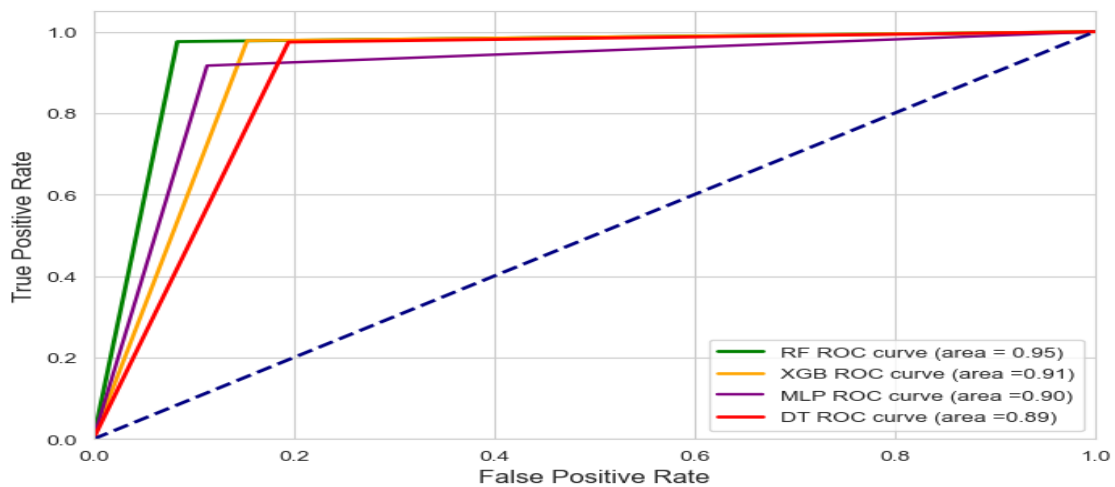


Figure 4.3 Raised Blood Pressure Prediction Models Performances Comparison using ROC AUC

Figure 4.4 presents the confusion matrix summarizing the performance of the raised blood pressure prediction models using the tested set of machine learning classifiers, including Random Forest, XGBoost, Decision Tree, and MLP.

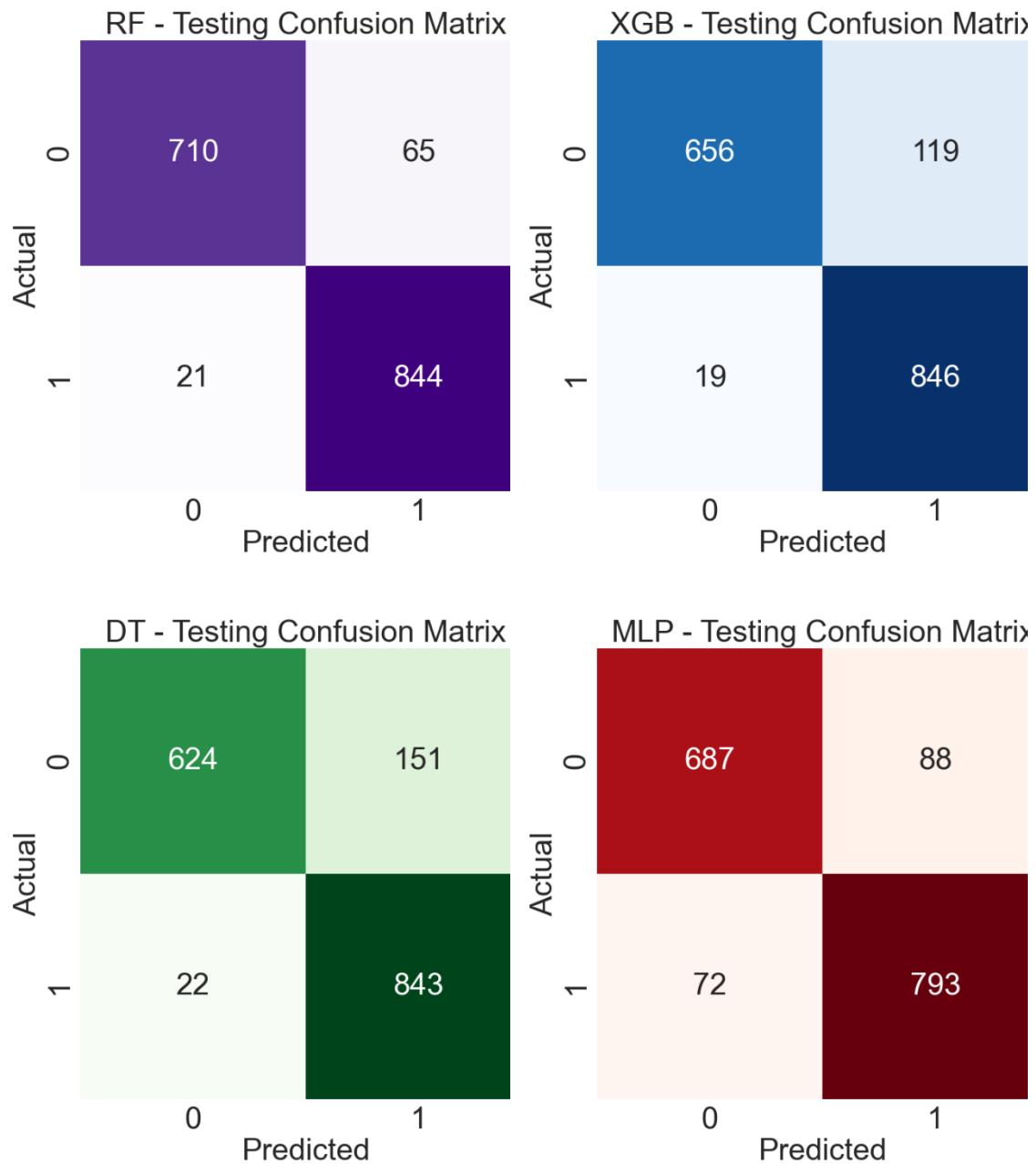


Figure 4.4 Comparison of Performance of the Raised Blood Pressure Detection Models Using Different Classifiers by Confusion Matrix

As shown in Figure 4.4, the Random Forest classifier performed the best prediction results among all the tested classifiers, showing the minimal number of misclassifications in

predicting incorrect raised blood pressure cases who are actually subjects without raised blood pressure condition. From the presented results in Figure 4.3 and 4.4, it can be concluded that the Random Forest algorithm had the highest performance, achieving a sensitivity of 97.57%, a specificity of 91.61%, an AUC value of 0.95, and an accuracy of 94.76%, followed by the XGBoost model with accuracy, and AUC values of 91.59%, and 0.91, while the Decision Tree model had the lowest performance with an AUC of 0.89 and accuracy of 89.45%.

4.4 Discussion

In light of the presented results, it is clear that the Random Forest algorithm had the highest performance, with an AUC value of 0.95 and an accuracy of 94.76% for the raised blood pressure prediction model. For the raised blood sugar prediction models the obtained testing results using Random Forest were 98.05% and 0.98 for accuracy, and ROC AUC respectively.

The highest performance result for the raised blood sugar and the raised blood pressure prediction models that were implemented using the Random Forest classifier might be attributed to the ensemble working approach in Random Forest, minimizing the high variance issues that are present in the Decision Tree, as well as to its low sensitivity to overfitting problems. On the other hand, the model that was utilized using the Multilayer Perceptron algorithm which is based on Neural Networks did not yield the best performance results, although Neural Network algorithms are well known for their high predictive power and robustness. However, this can vary depending on the characteristics and complexity of the used dataset, in addition to the fact that using MLP ANN requires very careful configuration of the algorithm parameters. For models utilized using the XGBoost algorithms it is noticed that their performance is higher than the decision tree-based model, this result aligns with the basics of this algorithm, of using the Decision Tree as a weak learner and improving its performance by aggregating multiple weak predictions to reduce bias and high variance issues.

Based on the previous results, it can be concluded that the proposed model for predicting raised blood sugar in this study outperforms the results achieved in other models implemented in previous studies, that were applied to similar datasets of risk factors related to glucose metabolism disorders and high blood pressure. In addition, the accuracy

and other performance metrics are significantly higher in the prediction models proposed through this study, despite it being applied to a larger dataset including all possible variables that are found to contribute to the development of diabetes and other NCDs.

Analysis of the features' importance in the proposed model for raised blood pressure and sugar prediction models shows that age, gender, BMI, waist-hip ratio, total cholesterol, HDL, history of cardiovascular diseases, hypertension and diabetes, physical activity, sugar intake, salt intake, anxiety and depression, and sleep disturbances variables contribute substantially towards the predictive power of the proposed models for predicting raised blood pressure and sugar.

4.5 Conclusion

This chapter presented the main findings and the experimental results of the constructed models in this study and discussed these results, the performance, and the accuracy achieved after implementing the raised blood pressure and sugar prediction models.

The next chapter will present the conclusion inferred after conducting this study, the encountered obstacles and the mitigation methods, and the future works.

Chapter 5: Conclusion and Future Work

5.1 Conclusion

This study proposes machine learning-based models to detect and predict raised blood sugar levels, and raised blood pressure. The proposed prediction models in this study are implemented using a curated selection of robust supervised machine-learning classification algorithms. These algorithms leverage various risk factors present in the dataset that is collected as secondary anonymized data from the standardized NCDs Stepwise survey conducted to study the risk factors of NCDs. The collected dataset included variables that are identified globally as common risk factors causing the development of NCDs. The collected variables within that study included age, body mass index, waist circumference, hip circumference, dietary habits, physical activity, history of other chronic diseases, and blood lipids as NCDs' common risk factors, in addition to fasting blood sugar and blood pressure measurement.

This study focuses on addressing the health issue of raised blood sugar and raised blood pressure. The interest in these two health conditions is due to multiple reasons. First, raised blood sugar and raised blood pressure are identified as the key signs of type 2 diabetes, and hypertension, which are recognized as the top common NCDs affecting adults, and the leading causes of death in the world. Secondly, these health conditions can be preventable, or at least managed if they are early detected. Hence, stopping their adverse effect, progression to other chronic diseases, and development of serious health complications. Finally, on the community level, the risks of raised blood pressure and raised blood sugar can be avoided through conducting population screening, however, this approach poses additional costs to national health systems. Integrating machine learning data-driven models can support significantly in implementing screening on a wide scale, as an inexpensive or even no-cost approach.

Therefore, this study aimed to construct machine learning-based models for detecting and predicting the risk of raised blood pressure and raised blood sugar, as metabolic disorders that represent the onset signs of hypertension and diabetes. The constructed models throughout this study used the NCDs risk factors obtained from the STEPS dataset as the set of input features for training the detection models. The prediction results of the built

models were based on using the fasting blood sugar levels, and the average systolic and diastolic blood measurements for defining the outcome features of the raised blood sugar model, and raised blood pressure model respectively.

The raised blood pressure and raised blood sugar detection models are built using four primary classification algorithms: Random Forest, Decision Tree, XGBoost, and Multilayer Perceptron from neural networks.

In the raised blood sugar detection models, the Random Forest classifier achieved the highest accuracy rate among the other classifiers, at 98.05%. There were also promising results for MLP, XGBoost, and Decision Tree, achieving accuracy rates of 94.26%, 96.10%, and 94.65%, respectively.

For the raised blood pressure detection models, the Random Forest classifier outperformed the other used classifiers, which yielded the highest accuracy rate at 94.76%. The results of the models utilized with the XGBoost, MLP, and Decision Tree algorithms were also high, achieving accuracy rates of 91.59%, 90.24%, and 89.45%, respectively.

The significance of using these models lies in their potential to support the healthcare sector and alleviate the workload of healthcare workers. This will contribute to improved public health and well-being.

5.2 Study Obstacles and Mitigation Methods

In the initial phases of conceptualizing this study, the focus was on building machine learning-based models for the detection of hypertension and diabetes. However, after conducting a thorough literature review and investigating related works in this field, the study's direction shifted slightly while remaining within the same conceptual framework. The focus was redirected towards the detection and prediction of raised blood pressure and blood sugar levels.

The literature review was one of the most challenging phases conducted throughout this study. However, it helped me highly in getting a wider look, by understanding the scientific and medical facts, determining the needed variables, techniques, and algorithms, and finally drafting the technical framework and development plan to implement this study. Indeed, the comprehensive review of related works contributed to

the identification of the research gap. Interestingly, most of the previous studies that had proposed models for the prediction models of glucose metabolism disorders or high blood pressure disorders presented their models as a diagnosis tool for hypertension and diabetes using a one-time reading of the predicted variable, or using the reported diagnosis of hypertension or diabetes, where this might be unobjective, since using the reported diagnosis needs that the development of these diagnosed health condition in terms of their risk factors over a period of time using an ongoing approach.

Through the implementation journey of the proposed models, it was noticed that activities done within the data preparation, exploring, and preprocessing phases took most of the time in the process of technical development. This is due to the sensitivity in dealing with clinical data, and the need to understand every single feature and investigate the relationships between different features. Nevertheless, as an IT professional working with colleagues in the healthcare sector, this supported me in better understanding the dataset features, the clinical protocols and guidelines, and the difference between screening and diagnosis approaches.

5.3 Future Work

For future endeavors, the raised blood pressure and raised blood sugar prediction models proposed in this study can be improved by incorporating multiple separate readings of blood pressure and fasting blood sugar that are taken over multiple days. That can contribute to strengthening the raised blood pressure and sugar detection models for advanced usability, efficiency, and accuracy to diagnose diabetes and hypertension, as well as for screening purposes.

Further, the proposed models for detecting the raised blood sugar cases can be enhanced to determine whether fasting blood sugar is normal, impaired, or high. Likewise, the raised blood pressure detection models can be improved for classifying the level of blood pressure as high, normal, or low. Additionally, the classification power of the model can be extended to determine which component of blood pressure that is causing a disorder.

Another potential opportunity for improving the study results could be achieved by collecting datasets from the STEPS NCDs studies that were conducted in different communities and countries for performing model generalization by training it on datasets collected from various ethnicities.

References

- Abdullah, A., Peeters, A., de Courten, M., & Stoelwinder, J. (2010). The magnitude of association between overweight and obesity and the risk of diabetes: A meta-analysis of prospective cohort studies. *Diabetes Research and Clinical Practice*, 89(3), 309–319. <https://doi.org/10.1016/J.DIABRES.2010.04.012>
- Abhari, S., Kalhori, S. R. N., Ebrahimi, M., Hasannejadasl, H., & Garavand, A. (2019). Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods. In *Healthcare Informatics Research* (Vol. 25, Issue 4). <https://doi.org/10.4258/hir.2019.25.4.248>
- Abiodun, O. O., Anya, T., Chukwu, J. C., & Adekanmbi, V. (2024). Prevalence, Risk Factors and Cardiovascular Comorbidities of Resistant Hypertension among Treated Hypertensives in a Nigerian Population. *Global Heart*, 19(1). <https://doi.org/10.5334/gh.1296>
- Ackerman, S., Farchi, E., Raz, O., Zalmanovici, M., & Dube, P. (2020). Detection of data drift and outliers affecting machine learning model performance over time. *ArXiv Preprint ArXiv:2012.09258*.
- Ackland, M., Choi, B., & Puska, P. (2003). Rethinking the terms non-communicable disease and chronic disease. *Journal of Epidemiology & Community Health*, 57(11), 838–839. <https://doi.org/10.1136/JECH.57.11.838>
- Adler, A., Bennett, P., Colagiuri Chair, S., Gregg, E., Venkat Narayan, K., Inês Schmidt, M., Sobngwi, E., Tajima, N., Tandon, N., Unwin, N., Wild, S., Yudkin, J., Levitt, N., Mohan, V., Montgomery, S., Nyirenda, M. J., Tuomilehto, J., Den Boon, S., & Hocking, S. (2021). REPRINT OF: CLASSIFICATION OF DIABETES MELLITUS. *Diabetes Research and Clinical Practice*. <https://doi.org/10.1016/j.diabres.2021.108972>
- Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and

management of diabetes: a comprehensive review. In *Diabetology and Metabolic Syndrome* (Vol. 14, Issue 1). <https://doi.org/10.1186/s13098-022-00969-9>

Agudo, A., & Joint FAO. (2005). Measuring intake of fruit and vegetables. *World Health Organization*.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In *Organizational Research Methods* (Vol. 16, Issue 2). <https://doi.org/10.1177/1094428112470848>

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. In *Healthcare (Switzerland)* (Vol. 10, Issue 3). <https://doi.org/10.3390/healthcare10030541>

AITBAYEV, A. (2021). *Diabetes UCI Dataset*. <https://www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset>

Akil, A. A. S., Yassin, E., Al-Maraghi, A., Aliyev, E., Al-Malki, K., & Fakhro, K. A. (2021). Diagnosis and treatment of type 1 diabetes at the dawn of the personalized medicine era. In *Journal of Translational Medicine* (Vol. 19, Issue 1). <https://doi.org/10.1186/s12967-021-02778-6>

Alberti, K. G. M. M., & Zimmet, P. Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *Diabetic Medicine*, 15(7). [https://doi.org/10.1002/\(SICI\)1096-9136\(199807\)15:7<539::AID-DIA668>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S)

Al-Hadlaq, S. M., Balto, H. A., Hassan, W. M., Marraiki, N. A., & El-Ansary, A. K. (2022). Biomarkers of non-communicable chronic disease: an update on contemporary methods. *PeerJ*, 10. <https://doi.org/10.7717/peerj.12977>

Allotey, P., Davey, T., & Reidpath, D. D. (2014). NCDs in low and middle-income countries - Assessing the capacity of health systems to respond to population needs. *BMC Public Health*, 14(SUPPL. 2). <https://doi.org/10.1186/1471-2458-14-S2-S1>

Almeida, C. S. de, Miccoli, L. S., Andhini, N. F., Aranha, S., Oliveira, L. C. de, Artigo, C. E., Em, A. A. R., Em, A. A. R., Bachman, L., Chick, K., Curtis, D., Peirce, B. N., Askey, Dale., Rubin, J., Egnatoff, Dr. W. J., Uhl Chamot, A., El-Dinary, P. B., Scott, J.; Marshall, G., Prensky, M., ... Santa, U. F. De. (2016). Clinical Guideline for the Management of Hypertension. In *Revista Brasileira de Linguística Aplicada* (Vol. 5, Issue 1).

Almomani, M. H., Rababa, M., Alzoubi, F., Alnuaimi, K., Alnatour, A., & Ali, R. A. (2021). Effects of a health education intervention on knowledge and attitudes towards chronic non-communicable diseases among undergraduate students in Jordan. *Nursing Open*, 8(1). <https://doi.org/10.1002/nop2.634>

Alnasir, F. A. L. (2008). Hypertension the silent killer. In *Journal of the Bahrain Medical Society* (Vol. 20, Issue 2). <https://doi.org/10.36106/ijsr/1231728>

Alqadi, S. F. (2024). Diabetes Mellitus and Its Influence on Oral Health: Review. *Diabetes, Metabolic Syndrome and Obesity*, 17, 107–120. <https://doi.org/10.2147/DMSO.S426671>

Amarnath, B., & Appavu Alias Balamurugan, S. (2016). Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. In *Journal of Engineering Science and Technology* (Vol. 11, Issue 11).

Amelia, R. (2017). The correlation between body mass index and self-efficacy with blood glucose level in type 2 diabetes mellitus. *Advanced Science Letters*, 23(4). <https://doi.org/10.1166/asl.2017.9192>

Amelia, R., & Luhulima, D. (2020). *Relationship Between Levels of Fasting Blood Glucose and HbA1C in Prediabetes Patients*. <https://doi.org/10.2991/ahsr.k.200523.001>

American Diabetes Association Professional Practice Committee, ElSayed, N. A., Aleppo, G., Bannuru, R. R., Bruemmer, D., Collins, B. S., Ekhlaspour, L., Gibbons, C. H., Giurini, J. M., Hilliard, M. E., Johnson, E. L., Khunti, K., Lingvay, I., Matfin, G., McCoy, R. G., Perry, M. Lou, Pilla, S. J., Polsky, S., Prahalad, P., ... Gabbay, R. A. (2024). 12. Retinopathy, Neuropathy, and Foot Care: Standards of Care in Diabetes—2024. *Diabetes Care*, 47(Supplement_1), S231–S243. <https://doi.org/10.2337/DC24-S012>

Anand, T., Kishore, J., Isaakidis, P., Gupte, H. A., Kaur, G., Kumari, S., Jha, D., & Grover, S. (2018). Integrating screening for non-communicable diseases and their risk factors in routine tuberculosis care in Delhi, India: A mixed-methods study. *PLoS ONE*, 13(8). <https://doi.org/10.1371/journal.pone.0202256>

Asif, M. (2014). The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. *Journal of Education and Health Promotion*, 3(1), 1. <https://doi.org/10.4103/2277-9531.127541>

Association, A. D. (2020). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. *Diabetes Care*, 43(Supplement_1), S14–S31. <https://doi.org/10.2337/DC20-S002>

Atun, R., Jaffar, S., Nishtar, S., Knaul, F. M., Barreto, M. L., Nyirenda, M., Banatvala, N., & Piot, P. (2013). Improving responsiveness of health systems to non-communicable diseases. In *The Lancet* (Vol. 381, Issue 9867). [https://doi.org/10.1016/S0140-6736\(13\)60063-X](https://doi.org/10.1016/S0140-6736(13)60063-X)

Balwan, W. K., & Kour, S. (2021). A Systematic Review of Hypertension and Stress - The Silent Killers. *Scholars Academic Journal of Biosciences*, 9(6). <https://doi.org/10.36347/sajb.2021.v09i06.002>

Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna Journal of Medicine*, 10(04). https://doi.org/10.4103/ajm.ajm_53_20

Bauer, U. E., Briss, P. A., Goodman, R. A., & Bowman, B. A. (2014a). Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the USA. In *The Lancet* (Vol. 384, Issue 9937). [https://doi.org/10.1016/S0140-6736\(14\)60648-6](https://doi.org/10.1016/S0140-6736(14)60648-6)

Bauer, U. E., Briss, P. A., Goodman, R. A., & Bowman, B. A. (2014b). Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the USA. *The Lancet*, 384(9937).

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 44(9). <https://doi.org/10.1080/1206212X.2021.1974663>

Bergman, M. (2013). Pathophysiology of prediabetes and treatment implications for the prevention of type 2 diabetes mellitus. In *Endocrine* (Vol. 43, Issue 3). <https://doi.org/10.1007/s12020-012-9830-9>

Bernell, S., & Howard, S. W. (2016). Use Your Words Carefully: What Is a Chronic Disease? *Frontiers in Public Health*, 4, 212747. <https://doi.org/10.3389/FPUBH.2016.00159/BIBTEX>

Brody, H., & Waters, D. B. (1980). Diagnosis is treatment. *J Fam Pract*, 10(3), 445–449.

Bromfield, S., & Muntner, P. (2013). High blood pressure: The leading global burden of disease risk factor and the need for worldwide prevention programs. *Current Hypertension Reports*, 15(3). <https://doi.org/10.1007/s11906-013-0340-9>

Brown, C. D., Higgins, M., Donato, K. A., Rohde, F. C., Garrison, R., Obarzanek, E., Ernst, N. D., & Horan, M. (2000). Body mass index and the prevalence of hypertension and dyslipidemia. *Obesity Research*, 8(9). <https://doi.org/10.1038/oby.2000.79>

Budreviciute, A., Damiani, S., Sabir, D. K., Onder, K., Schuller-Goetzburg, P., Plakys, G., Katileviciute, A., Khoja, S., & Kodzius, R. (2020). Management and Prevention

Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors. In *Frontiers in Public Health* (Vol. 8). <https://doi.org/10.3389/fpubh.2020.574111>

Burns, C., & Francis, N. (2023). Type 2 Diabetes: Etiology, Epidemiology, Pathogenesis, and Treatment. *Metabolic Syndrome*, 509–528. https://doi.org/10.1007/978-3-031-40116-9_34

Castillo, R. S., & Kelemen, A. (2013). Considerations for a successful clinical decision support system. *CIN - Computers Informatics Nursing*, 31(7), 319–326. <https://doi.org/10.1097/NXN.0B013E3182997A9C>

CDC. (2021). About Adult BMI: Healthy Weight, Nutrition, and Physical Activity. *Centers for Disease Control and Prevention*.

Cecchini, V., Nguyen, T. P., Pfau, T., Landtsheer, S. De, & Sauter, T. (2019). An efficient machine learning method to solve imbalanced data in metabolic disease prediction. *Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019*. <https://doi.org/10.1109/KSE.2019.8919337>

Centers for Disease Control and Prevention. (2017). *Hypertension Prevalence and Control Among Adults: United States, 2015–2016*. <https://www.cdc.gov/nchs/products/databriefs/db289.htm>

Centers for Disease Control and Prevention. (2023). *NHANES Questionnaires, Datasets, and Related Documentation*. <https://wwwn.cdc.gov/nchs/nhanes/Default.aspx>

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1), 1–26. <https://doi.org/10.1186/S40537-020-00327-4/FIGURES/13>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*. <https://doi.org/10.1145/2939672.2939785>

Chertow, G. M., Lee, J., Kuperman, G. J., Burdick, E., Horsky, J., Seger, D. L., Lee, R., Mekala, A., Song, J., Komaroff, A. L., & Bates, D. W. (2001). Guided medication dosing for inpatients with renal insufficiency. *JAMA*, 286(22). <https://doi.org/10.1001/jama.286.22.2839>

Chi, J. H., & Lee, B. J. (2022). Risk factors for hypertension and diabetes comorbidity in a Korean population: A cross-sectional study. *PLoS ONE*, 17(1 January). <https://doi.org/10.1371/journal.pone.0262757>

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T., & Roccella, E. J. (2003). Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. In *Hypertension* (Vol. 42, Issue 6). <https://doi.org/10.1161/01.HYP.0000107251.49515.c2>

Christe, D. M., Vijaya, S., & Tharangini, K. (2020). Screening for non-communicable diseases. *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, 9(3). <https://doi.org/10.18203/2320-1770.ijrcog20200881>

Clark, N. G., Fox, K. M., & Grandy, S. (2007a). Symptoms of diabetes and their association with the risk and presence of diabetes: Findings from the study to help improve early evaluation and management of risk factors leading to diabetes (SHIELD). *Diabetes Care*, 30(11). <https://doi.org/10.2337/dc07-0816>

Clark, N. G., Fox, K. M., & Grandy, S. (2007b). Symptoms of diabetes and their association with the risk and presence of diabetes: Findings from the study to help improve early evaluation and management of risk factors leading to diabetes (SHIELD). *Diabetes Care*, 30(11). <https://doi.org/10.2337/dc07-0816>

Colpaert, K., Claus, B., Somers, A., Vandewoude, K., Robays, H., & Decruyenaere, J. (2006). Impact of computerized physician order entry on medication prescription errors

in the intensive care unit: A controlled cross-sectional trial. *Critical Care*, 10(1).
<https://doi.org/10.1186/cc3983>

Cuevas-Chávez, A., Hernández, Y., Ortiz-Hernandez, J., Sánchez-Jiménez, E., Ochoa-Ruiz, G., Pérez, J., & González-Serna, G. (2023). A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases. In *Healthcare (Switzerland)* (Vol. 11, Issue 16).
<https://doi.org/10.3390/healthcare11162240>

Dabelea, D., Mayer-Davis, E. J., Saydah, S., Imperatore, G., Linder, B., Divers, J., Bell, R., Badaru, A., Talton, J. W., Crume, T., Liese, A. D., Merchant, A. T., Lawrence, J. M., Reynolds, K., Dolan, L., Liu, L. L., & Hamman, R. F. (2014). Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. *JAMA*, 311(17).
<https://doi.org/10.1001/jama.2014.3201>

Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health* 2021, Vol. 18, Page 3317, 18(6), 3317. <https://doi.org/10.3390/IJERPH18063317>

De Boer, I. H., Bangalore, S., Benetos, A., Davis, A. M., Michos, E. D., Muntner, P., Rossing, P., Zoungas, S., & Bakris, G. (2017). Diabetes and hypertension: A position statement by the American diabetes association. *Diabetes Care*, 40(9).
<https://doi.org/10.2337/dci17-0026>

Demaio, A. R., Nielsen, K. K., Tersbøl, B. P., Kallestrup, P., & Meyrowitsch, D. W. (2014). Primary Health Care: A strategic framework for the prevention and control of chronic non-communicable disease. *Global Health Action*, 7(1).
<https://doi.org/10.3402/gha.v7.24504>

Deming, C., Dekkati, S., & Desamsetti, H. (2018). Exploratory Data Analysis and Visualization for Business Analytics. *Asian Journal of Applied Science and Engineering*, 7(1). <https://doi.org/10.18034/ajase.v7i1.53>

Desai, N., Unni, G., Agarwala, R., Salagre, S., Godbole, S., Dengra, A., Abhyankar, M. V., & Revankar, S. (2021). Risk factors and comorbidities in young indian patients with hypertension: Real young (hypertension) study. *Integrated Blood Pressure Control*, 14. <https://doi.org/10.2147/IBPC.S272548>

Dhimal, M., Neupane, T., & Lamichhane Dhimal, M. (2021). Understanding linkages between environmental risk factors and noncommunicable diseases—A review. In *FASEB BioAdvances* (Vol. 3, Issue 5). <https://doi.org/10.1096/fba.2020-00119>

Diabetes Quick Facts / Basics / Diabetes / CDC. (2024). <https://www.cdc.gov/diabetes/about/>

Di Cesare, M. (2019). Global trends of chronic non-communicable diseases risk factors. *European Journal of Public Health*, 29(Supplement_4). <https://doi.org/10.1093/eurpub/ckz185.196>

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1), 1–15. <https://doi.org/10.1186/S12911-019-0918-5/FIGURES/8>

Dj Novakovi, J., Veljovi, A., Ili, S. S., Zeljko Papi, ˇ, & Tomovi, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, 7(1), 39–46.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1). <https://doi.org/10.1111/j.1600-0587.2012.07348.x>

Downs, A., Boucher, L. A., Campbell, D. G., & Polyakov, A. (2017). Using the WHO–5 well-being index to identify college students at risk for mental health problems. *Journal of College Student Development*, 58(1). <https://doi.org/10.1353/csd.2017.0008>

Drobnič, F., Kos, A., & Pustišek, M. (2020). On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics (Switzerland)*, 9(5). <https://doi.org/10.3390/electronics9050761>

Du, Z., Yang, Y., Zheng, J., Li, Q., Lin, D., Li, Y., Fan, J., Cheng, W., Chen, X. H., & Cai, Y. (2020). Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation. *JMIR Medical Informatics*, 8(7). <https://doi.org/10.2196/17257>

Embarak, O. (2018). Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems. *Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems*, 1–374. <https://doi.org/10.1007/978-1-4842-4109-7/COVER>

Esmailnasab, N., Moradi, G., & Delaveri, A. (2012). Risk factors of non-communicable diseases and metabolic syndrome. *Iranian Journal of Public Health*, 41(7).

Ezzati, M., & Riboli, E. (2013). Behavioral and Dietary Risk Factors for Noncommunicable Diseases. *New England Journal of Medicine*, 369(10). <https://doi.org/10.1056/nejmra1203528>

Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open*, 3(5). <https://doi.org/10.1136/bmjopen-2012-002457>

Ferdousi, R., Hossain, M. A., & Saddik, A. El. (2021). Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS. *IEEE Access*, 9. <https://doi.org/10.1109/ACCESS.2021.3094063>

Ferrannini, E., & Cushman, W. C. (2012). Diabetes and hypertension: the bad companions. *The Lancet*, 380(9841), 601–610. [https://doi.org/10.1016/S0140-6736\(12\)60987-8](https://doi.org/10.1016/S0140-6736(12)60987-8)

Ferrannini, E., Gastaldelli, A., & Iozzo, P. (2011). Pathophysiology of Prediabetes. In *Medical Clinics of North America* (Vol. 95, Issue 2). <https://doi.org/10.1016/j.mcna.2010.11.005>

Fiesler, E., & Beale, R. (Russell). (2020). *Multilayer perceptrons*. C1.2:1-C1.2:30. <https://doi.org/10.1201/9780429142772-60>

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. *IEEE Access*, 7. <https://doi.org/10.1109/ACCESS.2019.2945129>

Flack, J. M., & Adekola, B. (2020). Blood pressure and the new ACC/AHA hypertension guidelines. In *Trends in Cardiovascular Medicine* (Vol. 30, Issue 3). <https://doi.org/10.1016/j.tcm.2019.05.003>

Flack, J. M., Peters, R., Shafi, T., Alrefai, H., Nasser, S. A., & Crook, E. (2003). Prevention of hypertension and its complications: Theoretical basis and guidelines for treatment. *Journal of the American Society of Nephrology*, 14(SUPPL. 2).

Franco, O. H., Peeters, A., Bonneux, L., & De Laet, C. (2005). Blood Pressure in Adulthood and Life Expectancy With Cardiovascular Disease in Men and Women. *Hypertension*, 46(2), 280–286. <https://doi.org/10.1161/01.HYP.0000173433.67426.9B>

Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K. B., Ostolaza, H., & Martín, C. (2020). Pathophysiology of type 2 diabetes mellitus. In

International Journal of Molecular Sciences (Vol. 21, Issue 17).
<https://doi.org/10.3390/ijms21176275>

Gheorghe, A., Griffiths, U., Legido-Quigley, H., Murphy, A., Perel, Pa., & Lamprey
 Citation Adrian Gheorghe, P. (2014). The economic burden of non-communicable
 diseases (NCDs) in low-and middle- income countries (LMICs): a systematic review.
National Institute for Health Research.

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities
 and challenges in developing risk prediction models with electronic health records data:
 A systematic review. *Journal of the American Medical Informatics Association*, 24(1).
<https://doi.org/10.1093/jamia/ocw042>

Golino, H. (2013a). *Men's dataset from the "Predicting increased blood pressure using
 Machine Learning" paper.*
https://figshare.com/articles/dataset/Men_s_dataset_from_the_Predicting_increased_blood_pressure_using_Machine_Learning_paper/845665/1

Golino, H. (2013b). *Women's dataset from the "Predicting increased blood pressure
 using Machine Learning" paper.*
https://figshare.com/articles/dataset/Women_s_dataset_from_the_Predicting_increased_blood_pressure_using_Machine_Learning_paper/845664/1

Golino, H. F., Amaral, L. S. D. B., Duarte, S. F. P., Gomes, C. M. A., Soares, T. D. J.,
 Reis, L. A. Dos, & Santos, J. (2014). Predicting increased blood pressure using machine
 learning. *Journal of Obesity*, 2014. <https://doi.org/10.1155/2014/637635>

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling
 techniques: A review. *2017 International Conference on Advances in Computing,
 Communications and Informatics, ICACCI 2017, 2017-January*, 79–85.
<https://doi.org/10.1109/ICACCI.2017.8125820>

Gregg, E. W., Buckley, J., Ali, M. K., Davies, J., Flood, D., Mehta, R., Griffiths, B., Lim, L. L., Manne-Goehler, J., Pearson-Stuttard, J., Tandon, N., Roglic, G., Slama, S., Shaw, J. E., Agoudavi, K., Aryal, K. K., Atun, R., Bahendeka, S., Bicaba, B. W., ... Zhumadilov, Z. (2023). Improving health outcomes of people with diabetes: target setting for the WHO Global Diabetes Compact. In *The Lancet* (Vol. 401, Issue 10384). [https://doi.org/10.1016/S0140-6736\(23\)00001-6](https://doi.org/10.1016/S0140-6736(23)00001-6)

Gress, T. W., Denvir, J., & Shapiro, J. I. (2018). Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine*, 4(2). <https://doi.org/10.18590/MJM.2018.VOL4.ISS2.9>

Gupta, N., Mocumbi, A., Arwal, S. H., Jain, Y., Haileamlak, A. M., Memirie, S. T., Larco, N. C., Kwan, G. F., Amuyunzu-Nyamongo, M., Gathecha, G., Amegashie, F., Rakotoarison, V., Masiye, J., Wroe, E., Koirala, B., Karmacharya, B., Condo, J., Pierre Nyemazi, J., Sesay, S., ... Bukhman, G. (2021). Prioritizing health-sector interventions for noncommunicable diseases and injuries in low- And lower-middle income countries: national ncdi poverty commissions. *Global Health Science and Practice*, 9(3). <https://doi.org/10.9745/GHSP-D-21-00035>

Gupta, R., & Xavier, D. (2018). Hypertension: The most important non communicable disease risk factor in India. In *Indian Heart Journal* (Vol. 70, Issue 4). <https://doi.org/10.1016/j.ihj.2018.02.003>

Haffner, S. M. (2006a). Relationship of Metabolic Risk Factors and Development of Cardiovascular Disease and Diabetes. *Obesity*, 14(S6), 121S-127S. <https://doi.org/10.1038/OBY.2006.291>

Haffner, S. M. (2006b). Relationship of Metabolic Risk Factors and Development of Cardiovascular Disease and Diabetes. *Obesity*, 14(S6), 121S-127S. <https://doi.org/10.1038/OBY.2006.291>

Haldar, R. N. (2013). Global Brief on Hypertension: Silent Killer, Global Public Health Crisis. *Indian Journal of Physical Medicine and Rehabilitation*, 24(1). <https://doi.org/10.5005/ijopmr-24-1-2>

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. In *Journal of Educational and Behavioral Statistics* (Vol. 44, Issue 3). <https://doi.org/10.3102/1076998619832248>

Hewson, T., Minchin, M., Lee, K., Liu, S., Wong, E., Edge, C., Hard, J., Forsyth, K., Senior, J., & Shaw, J. (2024). Interventions for the detection, monitoring, and management of chronic non-communicable diseases in the prison population: an international systematic review. *BMC Public Health*, 24(1). <https://doi.org/10.1186/s12889-024-17715-7>

Hodgkinson, J., Mant, J., Martin, U., Guo, B., Hobbs, F. D. R., Deeks, J. J., Heneghan, C., Roberts, N., & McManus, R. J. (2011). Relative effectiveness of clinic and home blood pressure monitoring compared with ambulatory blood pressure monitoring in diagnosis of hypertension: Systematic review. *BMJ*, 343(7814). <https://doi.org/10.1136/bmj.d3621>

Huang, P. L. (2009). A comprehensive definition for metabolic syndrome. In *DMM Disease Models and Mechanisms* (Vol. 2, Issues 5–6). <https://doi.org/10.1242/dmm.001180>

Huck, W. (2023). Transforming our world: the 2030 Agenda for Sustainable Development. In *Sustainable Development Goals*. <https://doi.org/10.5040/9781509934058.0025>

Hug, B. L., Witkowski, D. J., Sox, C. M., Keohane, C. A., Seger, D. L., Yoon, C., Matheny, M. E., & Bates, D. W. (2009). Occurrence of adverse, often preventable, events in community hospitals involving nephrotoxic drugs or those excreted by the kidney. *Kidney International*, 76(11). <https://doi.org/10.1038/ki.2009.353>

Ijaz, M. F., Alfian, G., Syafrudin, M., & Rhee, J. (2018). Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. *Applied Sciences (Switzerland)*, 8(8). <https://doi.org/10.3390/app8081325>

Institute of Health Metrics and Evaluation. (2022). *The Global Burden of Disease Study (GBD)*. Global Health Data Exchange. <https://vizhub.healthdata.org/gbd-results>

Jafar, T. H., Chaturvedi, N., & Pappas, G. (2006). Prevalence of overweight and obesity and their association with hypertension and diabetes mellitus in an Indo-Asian population. *CMAJ*, 175(9), 1071–1077. <https://doi.org/10.1503/CMAJ.060464>

Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, 15(3), 435–443. <https://doi.org/10.1016/J.PCD.2021.02.005>

James, P. A., Oparil, S., Carter, B. L., Cushman, W. C., Dennison-Himmelfarb, C., Handler, J., Lackland, D. T., LeFevre, M. L., MacKenzie, T. D., Ogedegbe, O., Smith, S. C., Svetkey, L. P., Taler, S. J., Townsend, R. R., Wright, J. T., Narva, A. S., & Ortiz, E. (2014). 2014 Evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the Eighth Joint National Committee (JNC 8). In *JAMA* (Vol. 311, Issue 5). <https://doi.org/10.1001/jama.2013.284427>

J. D. Rockefeller. (2015). *Diabetes: Symptoms, Causes, Treatment and Prevention*. J.D. Rockefelle.

[https://books.google.ps/books?hl=en&lr=&id=WpbVCwAAQBAJ&oi=fnd&pg=PP4&dq=Good+to+Know:+Diabetes+Symptoms+and+Tests&ots=fOLxfck7gv&sig=tiqwYMUmaXNv3-](https://books.google.ps/books?hl=en&lr=&id=WpbVCwAAQBAJ&oi=fnd&pg=PP4&dq=Good+to+Know:+Diabetes+Symptoms+and+Tests&ots=fOLxfck7gv&sig=tiqwYMUmaXNv3-OVBiQ1oRCzh1w&redir_esc=y#v=onepage&q=Good%20to%20Know%3A%20Diabetes%20Symptoms%20and%20Tests&f=false)

[OVBiQ1oRCzh1w&redir_esc=y#v=onepage&q=Good%20to%20Know%3A%20Diabetes%20Symptoms%20and%20Tests&f=false](https://books.google.ps/books?hl=en&lr=&id=WpbVCwAAQBAJ&oi=fnd&pg=PP4&dq=Good+to+Know:+Diabetes+Symptoms+and+Tests&ots=fOLxfck7gv&sig=tiqwYMUmaXNv3-OVBiQ1oRCzh1w&redir_esc=y#v=onepage&q=Good%20to%20Know%3A%20Diabetes%20Symptoms%20and%20Tests&f=false)

Jenkins, C. D., Stanton, B. A., Niemcryk, S. J., & Rose, R. M. (1988). A scale for the estimation of sleep problems in clinical research. *Journal of Clinical Epidemiology*, 41(4). [https://doi.org/10.1016/0895-4356\(88\)90138-2](https://doi.org/10.1016/0895-4356(88)90138-2)

Jeong, I. K., Yoon, K. H., & Lee, M. K. (2020). Diabetes and COVID-19: Global and regional perspectives. *Diabetes Research and Clinical Practice*, 166. <https://doi.org/10.1016/j.diabres.2020.108303>

Jobe, M., Mactaggart, I., Bell, S., Kim, M. J., Hydera, A., Bascaran, C., Njai, M., Badjie, O., Perel, P., Prentice, A. M., & Burton, M. J. (2024). Prevalence of hypertension, diabetes, obesity, multimorbidity, and related risk factors among adult Gambians: a cross-sectional nationwide study. *The Lancet Global Health*, 12(1), e55–e65. [https://doi.org/10.1016/S2214-109X\(23\)00508-9](https://doi.org/10.1016/S2214-109X(23)00508-9)

Johnson, C. L., Paulose-Ram, R., Ogden, C. L., Carroll, M. D., Kruszan-Moran, D., Dohrmann, S. M., & Curtin, L. R. (2013). National health and nutrition examination survey: Analytic guidelines, 1999-2010. *Vital and Health Statistics, Series 2: Data Evaluation and Methods Research*, 161.

J, P., L, A., N, T., J, W., C, F., N, R., M, R., B, M., F, B., K, W., Pullar, J., Allen, L., Townsend, N., Williams, J., Foster, C., Roberts, N., Rayner, M., Mikkelsen, B., Branca, F., & Wickramasinghe, K. (2018). The impact of poverty reduction and development interventions on non-communicable diseases and their behavioural risk factors in low and lower-middle income countries: A systematic review. *PloS One*, 13(2).

Juhola, J., Arokoski, J. P. A., Ervasti, J., Kivimäki, M., Vahtera, J., Myllyntausta, S., & Saltychev, M. (2021). Internal consistency and factor structure of Jenkins Sleep Scale: Cross-sectional cohort study among 80 000 adults. *BMJ Open*, 11(1). <https://doi.org/10.1136/bmjopen-2020-043276>

Kabir, A., Karim, M. N., Islam, R. M., Romero, L., & Billah, B. (2022). Health system readiness for non-communicable diseases at the primary care level: A systematic review. In *BMJ Open* (Vol. 12, Issue 2). <https://doi.org/10.1136/bmjopen-2021-060387>

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, 47(C), 45–51. <https://doi.org/10.1016/J.PROCS.2015.03.182>

Kang, Y. M., Cho, Y. K., Lee, S. E., Park, J. Y., Lee, W. J., Kim, Y. J., & Jung, C. H. (2017). Cardiovascular diseases and life expectancy in adults with type 2 diabetes: A Korean national sample cohort study. *Journal of Clinical Endocrinology and Metabolism*, 102(9). <https://doi.org/10.1210/jc.2017-00643>

Kankeu, H. T., Saksena, P., Xu, K., & Evans, D. B. (2013). The financial burden from non-communicable diseases in low- and middle-income countries: A literature review. In *Health Research Policy and Systems* (Vol. 11, Issue 1). <https://doi.org/10.1186/1478-4505-11-31>

Kanter, R., & Caballero, B. (2012). Global gender disparities in obesity: A review. In *Advances in Nutrition* (Vol. 3, Issue 4). <https://doi.org/10.3945/an.112.002063>

Kaur, S., Bansal, K., Kumar, Y., & Changela, A. (2023). A Comprehensive Analysis of Hypertension Disease Risk-Factors, Diagnostics, and Detections Using Deep Learning-Based Approaches. In *Archives of Computational Methods in Engineering*. <https://doi.org/10.1007/s11831-023-10035-w>

Kavishe, B., Biraro, S., Baisley, K., Vanobberghen, F., Kapiga, S., Munderi, P., Smeeth, L., Peck, R., Mghamba, J., Mutungi, G., Ikoona, E., Levin, J., Bou Monclús, M. A., Katende, D., Kisanga, E., Hayes, R., & Grosskurth, H. (2015). High prevalence of hypertension and of risk factors for non-communicable diseases (NCDs): A population based cross-sectional survey of NCDS and HIV infection in Northwestern Tanzania and Southern Uganda. *BMC Medicine*, 13(1). <https://doi.org/10.1186/s12916-015-0357-9>

Kerper, L. F., Spies, C. D., Tillinger, J., Wegscheider, K., Salz, A.-L., Weiss-Gerlach, E., Neumann, T., & Krampe, H. (2014). Screening for Depression, Anxiety, and General Psychological Distress in Pre-operative Surgical Patients: A Psychometric Analysis of the Patient Health Questionnaire 4 (Phq-4). *Clinical Health Promotion - Research and Best Practice for Patients, Staff and Community*, 4(1). <https://doi.org/10.29102/clinhp.14002>

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://doi.org/10.1016/J.ICTE.2021.02.004>

Khan, N. M., Madhav C, N., Negi, A., & Thaseen, I. S. (2020). Analysis on Improving the Performance of Machine Learning Models Using Feature Selection Technique. *Advances in Intelligent Systems and Computing*, 941. https://doi.org/10.1007/978-3-030-16660-1_7

Kiani, M. M., Takian, A., Farzadfar, F., Rezaei, S., & Zandian, H. (2023). The Relationships between Social Capital, Metabolic, and Behavioral Risk Factors of Non-Communicable Diseases: A Systematic Review. In *Iranian Journal of Public Health* (Vol. 52, Issue 9). <https://doi.org/10.18502/ijph.v52i9.13563>

Kifle, Z. D., Adugna, M., Chanie, G. S., & Mohammed, A. (2022). Prevalence and associated factors of hypertension complications among hypertensive patients at University of Gondar Comprehensive Specialized Referral Hospital. *Clinical Epidemiology and Global Health*, 13. <https://doi.org/10.1016/j.cegh.2021.100951>

Kjeldsen, S. E. (2018). Hypertension and cardiovascular risk: General aspects. In *Pharmacological Research* (Vol. 129). <https://doi.org/10.1016/j.phrs.2017.11.003>

Kobashi, Y., Haque, S. E., Sakisaka, K., Amir, I., Kaneko, M., Mutahara, M., Mubassara, S., Kashem, A., & Tsubokura, M. (2024). Community-based intervention for managing

hypertension and diabetes in rural Bangladesh. *Tropical Medicine and Health*, 52(1), 1–10. <https://doi.org/10.1186/S41182-023-00574-0/FIGURES/2>

Koye, D. N., Magliano, D. J., Nelson, R. G., & Pavkov, M. E. (2018). The Global Epidemiology of Diabetes and Kidney Disease. In *Advances in Chronic Kidney Disease* (Vol. 25, Issue 2). <https://doi.org/10.1053/j.ackd.2017.10.011>

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, 50(6). <https://doi.org/10.1176/appi.psy.50.6.613>

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, 14(7). <https://doi.org/10.1007/s12652-021-03612-z>

Kwong, E. W. Y., Wu, H., & Pang, G. K. H. (2018). A prediction model of blood pressure for telemedicine. *Health Informatics Journal*, 24(3). <https://doi.org/10.1177/1460458216663025>

Kyung, W. P., Ki, H. C., & Kwan, W. L. (2006). Relationship Between Metabolic Syndrome and Familial History of Hypertension/Stroke, Diabetes, and Cardiovascular Disease. *Journal of Korean Medical Science*, 21(4), 701. <https://doi.org/10.3346/JKMS.2006.21.4.701>

Laditka, S. B., & Laditka, J. N. (2015). Active life expectancy of Americans with diabetes: Risks of heart disease, obesity, and inactivity. *Diabetes Research and Clinical Practice*, 107(1). <https://doi.org/10.1016/j.diabres.2014.10.008>

Lafreniere, D., Zulkernine, F., Barber, D., & Martin, K. (2017). Using machine learning to predict hypertension from a clinical dataset. *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*. <https://doi.org/10.1109/SSCI.2016.7849886>

Lee, J. H., Kim, S. H., Kang, S. H., Cho, J. H., Cho, Y., Oh, I. Y., Yoon, C. H., Lee, H. Y., Youn, T. J., Chae, I. H., & Kim, C. H. (2018). Blood Pressure Control and Cardiovascular Outcomes: Real-world Implications of the 2017 ACC/AHA Hypertension Guideline. *Scientific Reports*, 8(1). <https://doi.org/10.1038/s41598-018-31549-5>

Lee, M.-R., & et al. (2017). Diabetes mellitus and latent tuberculosis infection: A systemic review and metaanalysis. In *Clinical Infectious Diseases* (Vol. 64, Issue 6, pp. 719–727). <https://doi.org/10.1093/cid/cix300>

Legesse, E., Nigussie, T., Girma, D., Adugna Geleta, L., Dejene, H., Senbeta Deriba, B., Abeya Geleta, T., Hailu, D., Teshome, F., Midaksa, G., Worku, A., Tessema, M., & Negash, D. (2022). Does knowledge of non-communicable diseases risk factors influence screening service utilization? The finding from North Shewa Zone, Central Ethiopia. *Preventive Medicine Reports*, 30. <https://doi.org/10.1016/j.pmedr.2022.102040>

Lenfant, C. (2001). Can we prevent cardiovascular diseases in low- and middle-income countries? *Bulletin of the World Health Organization*, 79(10).

Ling, W., Huang, Y., Huang, Y. M., Fan, R. R., Sui, Y., & Zhao, H. L. (2020). Global trend of diabetes mortality attributed to vascular complications, 2000–2016. *Cardiovascular Diabetology*, 19(1). <https://doi.org/10.1186/s12933-020-01159-5>

Lin, H., Li, Q., Hu, Y., Zhu, C., Ma, H., Gao, J., Wu, J., Shen, H., Jiang, W., Zhao, N., Yin, Y., Pan, B., Jeekel, J., Hofman, A., & Gao, X. (2017). The prevalence of multiple non-communicable diseases among middle-aged and elderly people: the Shanghai Changfeng Study. In *European Journal of Epidemiology* (Vol. 32, Issue 2). <https://doi.org/10.1007/s10654-016-0219-6>

Li, Q., Li, N., Liang, X., Liu, Y., Chen, L., Lao, H., Wei, S., Xiao, J., & Qi, X. (2023). The study of antecedent clinical manifestations of hypertensive heart disease in cohort of hypertension. *Aging*, 15(6). <https://doi.org/10.18632/aging.204510>

Li, X., Wu, S., & Wang, L. (2017). Blood pressure prediction via recurrent models with contextual layer. *26th International World Wide Web Conference, WWW 2017*. <https://doi.org/10.1145/3038912.3052604>

Lombard, L. (2013). Non-communicable diseases: an ongoing challenge. *SA Journal of Diabetes & Vascular Disease Editorial* , 10(2).

Long, A. N., & Dagogo-Jack, S. (2011). Comorbidities of Diabetes and Hypertension: Mechanisms and Approach to Target Organ Protection. In *Journal of Clinical Hypertension* (Vol. 13, Issue 4). <https://doi.org/10.1111/j.1751-7176.2011.00434.x>

López-Martínez, F., Núñez-Valdez, E. R., Crespo, R. G., & García-Díaz, V. (2020). An artificial neural network approach for predicting hypertension using NHANES data. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-67640-z>

López-Martínez, F., Schwarcz.MD, A., Núñez-Valdez, E. R., & García-Díaz, V. (2018). Machine learning classification analysis for a hypertensive population as a function of several risk factors. *Expert Systems with Applications*, 110. <https://doi.org/10.1016/j.eswa.2018.06.006>

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., & Brähler, E. (2010). A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders*, 122(1–2). <https://doi.org/10.1016/j.jad.2009.06.019>

L. Tuck, M., & B. Corry, D. (2010). Prevalence of Obesity, Hypertension, Diabetes, and Metabolic Syndrome and Its Cardiovascular Complications. *Current Hypertension Reviews*, 6(2), 73–82. <https://doi.org/10.2174/157340210791171010>

Maeta, K., Nishiyama, Y., Fujibayashi, K., Gunji, T., Sasabe, N., Iijima, K., & Naito, T. (2018). Prediction of Glucose Metabolism Disorder Risk Using a Machine Learning Algorithm: Pilot Study. *JMIR Diabetes*, 3(4), e10212. <https://doi.org/10.2196/10212>

Magder, S. (2018). The meaning of blood pressure Luigi Forni. In *Critical Care* (Vol. 22, Issue 1). <https://doi.org/10.1186/s13054-018-2171-1>

Mahboob Alam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204. <https://doi.org/10.1016/J.IMU.2019.100204>

Marques da Silva, P., Lima, M. J., Neves, P. M., & Espiga de Macedo, M. (2019). Prevalence of cardiovascular risk factors and other comorbidities in patients with hypertension in Portuguese primary health care populations: The PRECISE study. *Revista Portuguesa de Cardiologia*, 38(6). <https://doi.org/10.1016/j.repc.2018.09.011>

Martinez-Ríos, E., Montesinos, L., Alfaro-Ponce, M., & Pecchia, L. (2021). A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data. *Biomedical Signal Processing and Control*, 68, 102813. <https://doi.org/10.1016/J.BSPC.2021.102813>

Meher, M., Pradhan, S., & Pradhan, S. R. (2023). Risk Factors Associated With Hypertension in Young Adults: A Systematic Review. *Cureus*. <https://doi.org/10.7759/cureus.37467>

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, 29(2), 93–99. <https://doi.org/10.1016/J.KJMS.2012.08.016>

Meng, X., Song, M., Vilaj, M., Štambuk, J., Dolikun, M., Zhang, J., Liu, D., Wang, H., Zhang, X., Zhang, J., Cao, W., Momčilović, A., Trbojević-Akmačić, I., Li, X., Zheng, D., Wu, L., Guo, X., Wang, Y., Lauc, G., & Wang, W. (2021). Glycosylation of igg associates with hypertension and type 2 diabetes mellitus comorbidity in the chinese muslim ethnic minorities and the han chinese. *Journal of Personalized Medicine*, 11(7). <https://doi.org/10.3390/jpm11070614>

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6). <https://doi.org/10.1093/bib/bbx044>

Miranda, J. J., Kinra, S., Casas, J. P., Davey Smith, G., & Ebrahim, S. (2008). Non-communicable diseases in low- and middle-income countries: Context, determinants and health policy. In *Tropical Medicine and International Health* (Vol. 13, Issue 10). <https://doi.org/10.1111/j.1365-3156.2008.02116.x>

Misra, P., & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal on Emerging Technologies*, 11(3).

Mocumbi, A. O. (2024). Cardiovascular Health Care in Low- and Middle-Income Countries. *Circulation*, 149(8). <https://doi.org/10.1161/CIRCULATIONAHA.123.065717>

Mohd Faizal, A. S., Thevarajah, T. M., Khor, S. M., & Chang, S. W. (2021). A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Computer Methods and Programs in Biomedicine*, 207. <https://doi.org/10.1016/j.cmpb.2021.106190>

Monaghan, M., Helgeson, V., & Wiebe, D. (2015). Type 1 Diabetes in Young Adulthood. *Current Diabetes Reviews*, 11(4). <https://doi.org/10.2174/1573399811666150421114957>

Montagna, S., Pengo, M. F., Ferretti, S., Borghi, C., Ferri, C., Grassi, G., Muiesan, M. L., & Parati, G. (2023). Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data. *Journal of Medical Systems*, 47(1). <https://doi.org/10.1007/s10916-022-01900-5>

Nai-Arun, N., & Moungrmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science*, 69. <https://doi.org/10.1016/j.procs.2015.10.014>

National Heart Lung and Blood Institute. (2023). *BioLINCC: Framingham Heart Study-Cohort (FHS-Cohort)*. <https://biolincc.nhlbi.nih.gov/studies/framcohort/>

Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, 19(1), 391–403. <https://doi.org/10.1007/S40200-020-00520-5/METRICS>

Ndubuisi, N. E. (2021). Noncommunicable Diseases Prevention In Low- and Middle-Income Countries: An Overview of Health in All Policies (HiAP). In *Inquiry (United States)* (Vol. 58). <https://doi.org/10.1177/0046958020927885>

Neira, M., & Prüss-Ustün, A. (2016). Preventing disease through healthy environments: A global assessment of the environmental burden of disease. *Toxicology Letters*, 259, S1.

Nguyen, N. T., Magno, C. P., Lane, K. T., Hinojosa, M. W., & Lane, J. S. (2008). Association of Hypertension, Diabetes, Dyslipidemia, and Metabolic Syndrome with Obesity: Findings from the National Health and Nutrition Examination Survey, 1999 to 2004. *Journal of the American College of Surgeons*, 207(6), 928–934. <https://doi.org/10.1016/J.JAMCOLLSURG.2008.08.022>

Nishida, C., Ko, G. T., & Kumanyika, S. (2010). Body fat distribution and noncommunicable diseases in populations: Overview of the 2008 WHO Expert Consultation on Waist Circumference and Waist-Hip Ratio. In *European Journal of Clinical Nutrition* (Vol. 64, Issue 1). <https://doi.org/10.1038/ejcn.2009.139>

Nojilana, B., Bradshaw, D., Pillay-Van Wyk, V., Msemburi, W., Laubscher, R., & Somdyala, N. I. (2014). Strategic Plan for the Prevention and Control of Non-Communicable Diseases. In *South African Medical Journal* (Vol. 104, Issue 10).

Ogedegbe, G., & Pickering, T. (2010). Principles and Techniques of Blood Pressure Measurement. In *Cardiology Clinics* (Vol. 28, Issue 4). <https://doi.org/10.1016/j.ccl.2010.07.006>

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics*, 14(2). <https://doi.org/10.3390/diagnostics14020144>

Okosun, I. S., Chandra, K. M. D., Choi, S., Christman, J., Dever, G. E. A., & Prewitt, T. E. (2001). Hypertension and type 2 diabetes comorbidity in adults in the United States: Risk of overall and regional adiposity. *Obesity Research*, 9(1). <https://doi.org/10.1038/oby.2001.1>

Owda, M., Owda, A. Y., & Fasli, M. (2023). An Exploratory Data Analysis and Visualizations of Underprivileged Communities Diabetes Dataset for Public Good. *Proceedings - 2023 22nd IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2023*, 581–585. <https://doi.org/10.1109/WI-IAT59888.2023.00096>

Owei, I., Umekwe, N., Ceesay, F., & Dagogo-Jack, S. (2019). Awareness of prediabetes status and subsequent health behavior, body weight, and blood glucose levels. *Journal of the American Board of Family Medicine*, 32(1). <https://doi.org/10.3122/jabfm.2019.01.180242>

Owess, M. M., Owda, A. Y., & Owda, M. (2023). Decision Support System in Healthcare for Predicting Blood Pressure Disorders. *2023 International Conference on Information Technology: Cybersecurity Challenges for Sustainable Cities, ICIT 2023 - Proceeding*, 62–67. <https://doi.org/10.1109/ICIT58056.2023.10226098>

Pan American Health Organization. (2022). *Economics of NCDs*. Pan American Health Organization. <https://www.paho.org/en/topics/economics-ncds>

Pan American Health Organization, & The World Health Organization. (2021). Noncommunicable Diseases. In *World Health Organization*. <https://www.paho.org/en/topics/noncommunicable-diseases>

Parija, S. C. (2023). World Diabetes Day 2022: Access to Diabetes Care. *Pondicherry Journal of Nursing*, 15(4). <https://doi.org/10.5005/jp-journals-10084-13160>

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

Patnaik, R., Chandran, M., Lee, S. C., Gupta, A., & Kim, C. (2018). Predicting the occurrence of essential hypertension using annual health records. *Proceedings of 2018 2nd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2018*. <https://doi.org/10.1109/ICAIECC.2018.8479458>

Pearce, K. A., & Furberg, C. D. (1994). The primary prevention of hypertension. In *Cardiovascular Risk Factors* (Vol. 4, Issue 3).

Peer, N., de Villiers, A., Jonathan, D., Kalombo, C., & Kengne, A. P. (2020). Care and management of a double burden of chronic diseases: Experiences of patients and perceptions of their healthcare providers. *PLoS ONE*, 15(7 July). <https://doi.org/10.1371/journal.pone.0235710>

Peng, W., Chen, S., Chen, X., Ma, Y., Wang, T., Sun, X., Wang, Y., Ding, G., & Wang, Y. (2024). Trends in major non-communicable diseases and related risk factors in China 2002–2019: an analysis of nationally representative survey data. *The Lancet Regional Health - Western Pacific*, 43. <https://doi.org/10.1016/j.lanwpc.2023.100809>

Pereboom, M., Mulder, I. J., Verweij, S. L., van der Hoeven, R. T. M., & Becker, M. L. (2019). A clinical decision support system to improve adequate dosing of gentamicin and vancomycin. *International Journal of Medical Informatics*, 124. <https://doi.org/10.1016/j.ijmedinf.2019.01.002>

Perng, W., Conway, R., Mayer-Davis, E., & Dabelea, D. (2023). Youth-Onset Type 2 Diabetes: The Epidemiology of an Awakening Epidemic. In *Diabetes Care* (Vol. 46, Issue 3). <https://doi.org/10.2337/dci22-0046>

Pinto, I. C., & Martins, D. (2017). Prevalence and risk factors of arterial hypertension: A literature review. *J Cardiovasc Med Ther*, 1(2).

Pittoli, F., Vianna, H. D., Victória Barbosa, J. L., Butzen, E., Gaedke, M. Â., Dias da Costa, J. S., & Scherer dos Santos, R. B. (2018). An intelligent system for prognosis of noncommunicable diseases' risk factors. *Telematics and Informatics*, 35(5). <https://doi.org/10.1016/j.tele.2018.02.005>

Plows, J. F., Stanley, J. L., Baker, P. N., Reynolds, C. M., & Vickers, M. H. (2018). The pathophysiology of gestational diabetes mellitus. In *International Journal of Molecular Sciences* (Vol. 19, Issue 11). <https://doi.org/10.3390/ijms19113342>

Prakash, S. (2017). Non-Communicable Diseases (NCDs): A Global Challenge. *Biomedical Journal of Scientific & Technical Research*, 1(2). <https://doi.org/10.26717/bjstr.2017.01.000185>

Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential Liability for Physicians Using Artificial Intelligence. In *JAMA - Journal of the American Medical Association* (Vol. 322, Issue 18). <https://doi.org/10.1001/jama.2019.15064>

Prihanti, G. S., Wilyani, D., Isnaini, F., Ailani, A. F., Humaira, A. A., Kurniawan, I. K., Rahayu, K. P., Hadiyanti, L. A., & Santoso, S. R. (2022). Community Awareness for Screening Non-Communicable Diseases. *Kemas*, 18(2). <https://doi.org/10.15294/kemas.v18i2.35039>

Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., Hannan, M. A., Uddin, M. J., & Pang, M. G. (2021). Role of insulin in health and disease: An update. In *International Journal of Molecular Sciences* (Vol. 22, Issue 12). <https://doi.org/10.3390/ijms22126403>

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *Proceedings of the 3rd International Conference on Smart Systems and*

Inventive Technology, ICSSIT 2020, 729–735.
<https://doi.org/10.1109/ICSSIT48917.2020.9214160>

Rampal, L. (2017). Prevention of non communicable diseases: A paradigm shift required. *Medical Journal of Malaysia*, 72.

Reddy, K. S. (2002). Cardiovascular diseases in the developing countries: dimensions, determinants, dynamics and directions for public health action. *Public Health Nutrition*, 5(1a). <https://doi.org/10.1079/phn2001298>

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., & Moore, J. H. (2006). Feature selection using a random forests classifier for the integrated analysis of multiple data types. *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB'06*, 171–178.
<https://doi.org/10.1109/CIBCB.2006.330987>

Rohwer, A., Uwimana Nicol, J., Toews, I., Young, T., Bavuma, C. M., & Meerpohl, J. (2021). Effects of integrated models of care for diabetes and hypertension in low-income and middle-income countries: A systematic review and meta-analysis. In *BMJ Open* (Vol. 11, Issue 7). <https://doi.org/10.1136/bmjopen-2020-043705>

Sakhaee, K., Maalouf, N. M., & Sinnott, B. (2017). Diagnosis and management for patients with hypertension, A noncommunicable disease education manual for primary health care professionals and patients. *World Health Organization CC BY-NC-SA 3.0 IGO Licence.*, 25(May).

Sakr, S., Elshaw, R., Ahmed, A., Qureshi, W. T., Brawner, C., Keteyian, S., Blaha, M. J., & Al-Mallah, M. H. (2018). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) Project. *PLoS ONE*, 13(4). <https://doi.org/10.1371/journal.pone.0195344>

Samadian, F., Dalili, N., & Jamalian, A. (2016). Lifestyle modifications to prevent and control hypertension. *Article in Iranian Journal of Kidney Diseases*, 10(5). www.ijkd.org

Schiffrin, E. L. (2020). How Structure, Mechanics, and Function of the Vasculature Contribute to Blood Pressure Elevation in Hypertension. In *Canadian Journal of Cardiology* (Vol. 36, Issue 5). <https://doi.org/10.1016/j.cjca.2020.02.003>

Schorling, J., Department of Medicine, & University of Virginia School of Medicine. (2002). *Diabetes Dataset*. <https://hbiostat.org/data/repo/diabetes>

Seong, S. C., Kim, Y. Y., Khang, Y. H., Park, J. H., Kang, H. J., Lee, H., Do, C. H., Song, J. S., Bang, J. H., Ha, S., Lee, E. J., & Shin, S. A. (2017). Data resource profile: The national health information database of the national health insurance service in South Korea. *International Journal of Epidemiology*, 46(3). <https://doi.org/10.1093/ije/dyw253>

Shahid, A., Wilkinson, K., Marcu, S., & Shapiro, C. M. (2012). Jenkins Sleep Scale. In A. Shahid, K. Wilkinson, S. Marcu, & C. M. Shapiro (Eds.), *STOP, THAT and One Hundred Other Sleep Scales* (pp. 203–204). Springer New York. https://doi.org/10.1007/978-1-4419-9893-4_45

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR)*, 5(4). <https://doi.org/10.21275/v5i4.nov162954>

Shrestha, R. (2023). *Hypertension: Causes, Symptoms & Treatment*. Public Health Notes. <https://www.publichealthnotes.com/hypertension/>

Sideris, C., Kalantarian, H., Nemati, E., & Sarrafzadeh, M. (2016). Building Continuous Arterial Blood Pressure Prediction Models Using Recurrent Networks. *2016 IEEE International Conference on Smart Computing, SMARTCOMP 2016*. <https://doi.org/10.1109/SMARTCOMP.2016.7501681>

Sikri, A., Singh, N. P., & Dalal, S. (2023). Chi-Square Method of Feature Selection: Impact of Pre-Processing of Data. *International Journal of Intelligent Systems and Applications in Engineering*, 11(3s).

Silva, G. F. S., Fagundes, T. P., Teixeira, B. C., & Chiavegatto Filho, A. D. P. (2022). Machine Learning for Hypertension Prediction: a Systematic Review. In *Current Hypertension Reports* (Vol. 24, Issue 11). <https://doi.org/10.1007/s11906-022-01212-6>

Sø, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. In *Journal of Clinical Pathology* (Vol. 62, Issue 1). <https://doi.org/10.1136/jcp.2008.061010>

Solomons, N., Kruger, H. S., & Puoane, T. R. (2017). Adherence challenges encountered in an intervention programme to combat chronic non-communicable diseases in an urban black community, Cape Town. *Health SA Gesondheid*, 22. <https://doi.org/10.1016/j.hsag.2016.11.006>

Soomro, M. H., & Jabbar, A. (2024). Diabetes etiopathology, classification, diagnosis, and epidemiology. *BIDE's Diabetes Desk Book*, 19–42. <https://doi.org/10.1016/B978-0-443-22106-4.00022-X>

Spiegel, K., Leproult, R., & Van Cauter, E. (1999). Impact of sleep debt on metabolic and endocrine function. *Lancet*, 354(9188). [https://doi.org/10.1016/S0140-6736\(99\)01376-8](https://doi.org/10.1016/S0140-6736(99)01376-8)

Stamler, J. (1991). Blood pressure and high blood pressure aspects of risk. *Hypertension*, 18(3). https://doi.org/10.1161/01.hyp.18.3_suppl.i95

Stumvoll, M., Goldstein, B. J., & Van Haeften, T. W. (2005). Type 2 diabetes: Principles of pathogenesis and therapy. *Lancet*, 365(9467). [https://doi.org/10.1016/S0140-6736\(05\)61032-X](https://doi.org/10.1016/S0140-6736(05)61032-X)

Subramanian, S., Gakunga, R., Kibachio, J., Gathecha, G., Edwards, P., Ogola, E., Yonga, G., Busakhala, N., Munyoro, E., Chakaya, J., Ngugi, N., Mwangi, N., Rege, D. Von, Wangari, L. M., Wata, D., Makori, R., Mwangi, J., & Mwanda, W. (2018). Cost and affordability of non-communicable disease screening, diagnosis and treatment in Kenya: Patient payments in the private and public sectors. *PLoS ONE*, 13(1). <https://doi.org/10.1371/journal.pone.0190113>

Sundus, K. I., Hammo, B. H., Al-Zoubi, M. B., & Al-Omari, A. (2022). Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. *Informatics in Medicine Unlocked*, 33, 101088. <https://doi.org/10.1016/J.IMU.2022.101088>

Sun, J., Mcnaughton, C. D., Zhang, P., Perer, A., Gkoulalas-Divanis, A., Denny, J. C., Kirby, J., Lasko, T., Saip, A., & Malin, B. A. (2014). Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association*, 21(2). <https://doi.org/10.1136/amiajnl-2013-002033>

Swartz, A. M., Strath, S. J., Parker, S. J., & Miller, N. E. (2008). The impact of body-mass index and steps per day on blood pressure and fasting glucose in older adults. *Journal of Aging and Physical Activity*, 16(2). <https://doi.org/10.1123/japa.16.2.188>

Tarleton, H. P., Smith, L. V., Zhang, Z. F., & Kuo, T. (2014). Utility of anthropometric measures in a multiethnic population: Their association with prevalent diabetes, hypertension and other chronic disease comorbidities. *Journal of Community Health*, 39(3). <https://doi.org/10.1007/s10900-013-9780-z>

Tayefi, M., Esmaeili, H., Saberi Karimian, M., Amirabadi Zadeh, A., Ebrahimi, M., Safarian, M., Nematy, M., Parizadeh, S. M. R., Ferns, G. A., & Ghayour-Mobarhan, M. (2017). The application of a decision tree to establish the parameters associated with hypertension. *Computer Methods and Programs in Biomedicine*, 139. <https://doi.org/10.1016/j.cmpb.2016.10.020>

Taylor, C., & Ward, A. (2003). Patients' views of high blood pressure, its treatment and risks. *Australian Family Physician*, 32(4).

Teimouri, M., Ebrahimi, E., & Alavinia, S. M. (2016). Comparison of various machine learning methods in diagnosis of hypertension in diabetics with/without consideration of costs. *Iranian Journal of Epidemiology*, 11(4).

The World Health Organization. (2006). Guidelines for the Prevention, Management and Care of Diabetes Mellitus. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

The World Health Organization. (2008). Waist circumference and waist-hip ratio: report of a WHO Expert Consultation. *Report of a WHO Expert Consultation. Geneva: World Health Organization, 2008*, 8–11.

The World Health Organization. (2016). *An estimated 12.6 million deaths each year are attributable to unhealthy environments*. <https://www.who.int/news/item/15-03-2016-an-estimated-12-6-million-deaths-each-year-are-attributable-to-unhealthy-environments>

The World Health Organization. (2017). WHO STEPS surveillance manual : the WHO STEPwise approach to chronic disease risk factor surveillance / Noncommunicable Diseases and Mental Health, World Health Organization. *World Health Organization*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/<https://www.who.int/docs/default-source/ncds/ncd-surveillance/steps/steps-manual.pdf>

The World Health Organization. (2019). *Noncommunicable diseases: Mortality*. <https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-mortality>

The World Health Organization. (2022a). *Global NCD Compact 2020-2030* . <https://www.who.int/initiatives/global-noncommunicable-diseases-compact-2020-2030>

The World Health Organization. (2022b). *Noncommunicable Disease Surveillance, Monitoring and Reporting*. <https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps>

The World Health Organization. (2023a). *Advancing the global agenda on prevention and control of noncommunicable diseases 2000 to 2020: looking forwards to 2030*. World Health Organization.

The World Health Organization. (2023b). *Blood pressure/hypertension*. <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3155>

The World Health Organization. (2023c). *Diabetes*. <https://www.who.int/news-room/fact-sheets/detail/diabetes>

The World Health Organization. (2023d). First WHO report details devastating impact of hypertension and ways to stop it. In *WHO*. <https://www.who.int/news/item/19-09-2023-first-who-report-details-devastating-impact-of-hypertension-and-ways-to-stop-it>

The World Health Organization. (2023e). *Noncommunicable diseases*. <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>

The World Health Organization. (2023f). *Noncommunicable diseases: Risk factors*. <https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-risk-factors>

The World Health Organization. (2023g). World health statistics 2023: monitoring health for the sdgs, sustainable development goals. In *The Milbank Memorial Fund quarterly* (Vol. 27, Issue 2).

The World Health Organization. (2024a). *Mean fasting blood glucose*. <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380>

The World Health Organization. (2024b). *Noncommunicable Diseases, Rehabilitation and Disability*. <https://www.who.int/teams/noncommunicable-diseases/integrated-support/environmental-risk-factors-and-ncds>

The World Health Organization. (2024c). *The top 10 causes of death*. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

The World Health Organization and others. (2013). Global Action Plan for The Prevention and Control of Noncommunicable Diseases 2013-2020. In *World Health Organization* (p. 55).

www.who.int/about/licensing/copyright_form/en/index.html

The World Health Organization - Regional Office for Europe. (1998). *Wellbeing Measures in Primary Health Care/The Depcare Project*. World Health Organization. Regional Office for Europe. <https://iris.who.int/bitstream/handle/10665/349766/WHO-EURO-1998-4234-43993-62027-eng.pdf?sequence=1&isAllowed=y>

Thomas, C. C., & Philipson, L. H. (2015). Update on Diabetes Classification. In *Medical Clinics of North America* (Vol. 99, Issue 1). <https://doi.org/10.1016/j.mcna.2014.08.015>

Tiwary, N., Sharma, N., Singh, S., Behl, T., & Zahoor, I. (2023). Understanding the Pharmacological and Nanotechnological Facets of Dipeptidyl Peptidase-4 Inhibitors in Type II Diabetes Mellitus: a Paradigm in Therapeutics. *BioNanoScience*, 1–19. <https://doi.org/10.1007/S12668-023-01234-7/METRICS>

Tjahjadi, H., & Ramli, K. (2020). Noninvasive blood pressure classification based on photoplethysmography using K-nearest neighbors algorithm: A feasibility study. *Information (Switzerland)*, 11(2). <https://doi.org/10.3390/info11020093>

Tjahjadi, H., Ramli, K., & Murfi, H. (2020). Noninvasive Classification of Blood Pressure Based on Photoplethysmography Signals Using Bidirectional Long Short-Term Memory and Time-Frequency Analysis. *IEEE Access*, 8. <https://doi.org/10.1109/ACCESS.2020.2968967>

Tomic, D., Shaw, J. E., & Magliano, D. J. (2022). The burden and risks of emerging complications of diabetes mellitus. In *Nature Reviews Endocrinology* (Vol. 18, Issue 9). <https://doi.org/10.1038/s41574-022-00690-7>

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 well-being index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3). <https://doi.org/10.1159/000376585>

UCI Machine Learning and Kaggle. (2016). *Pima Indians Diabetes Database*.
<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data>

Uthman, O. A., Ayorinde, A., Oyeboode, O., Sartori, J., Gill, P., & Lilford, R. J. (2022). Global prevalence and trends in hypertension and type 2 diabetes mellitus among slum residents: A systematic review and meta-analysis. In *BMJ Open* (Vol. 12, Issue 2).
<https://doi.org/10.1136/bmjopen-2021-052393>

Uwimana Nicol, J., Rohwer, A., Young, T., Bavuma, C. M., & Meerphol, J. J. (2018). Integrated models of care for diabetes and hypertension in low- and middle-income countries (LMICs): Protocol for a systematic review. *Systematic Reviews*, 7(1).
<https://doi.org/10.1186/s13643-018-0865-8>

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc.

Vangeepuram, N., Liu, B., Chiu, P. hsiang, Wang, L., & Pandey, G. (2021). Predicting youth diabetes risk using NHANES data and machine learning. *Scientific Reports*, 11(1).
<https://doi.org/10.1038/s41598-021-90406-0>

Vaughan, C. J., & Delanty, N. (2000). Hypertensive emergencies. *Lancet (London, England)*, 356(9227), 411–417. [https://doi.org/10.1016/S0140-6736\(00\)02539-3](https://doi.org/10.1016/S0140-6736(00)02539-3)

Vierra, J., Boonla, O., & Prasertsri, P. (2022). Effects of sleep deprivation and 4-7-8 breathing control on heart rate variability, blood pressure, blood glucose, and endothelial function in healthy young adults. *Physiological Reports*, 10(13).
<https://doi.org/10.14814/phy2.15389>

Volkova, N. I., Davidenko, I. Y., & Degtyareva, Y. S. (2021). Gestational diabetes mellitus. *Akusherstvo i Ginekologiya (Russian Federation)*, 2021(9).
<https://doi.org/10.54393/df.v2i2.18>

Wang, X., Ma, H., Li, X., Heianza, Y., Manson, J. E., Franco, O. H., & Qi, L. (2023a). Association of Cardiovascular Health With Life Expectancy Free of Cardiovascular Disease, Diabetes, Cancer, and Dementia in UK Adults. *JAMA Internal Medicine*, 183(4), 340–349. <https://doi.org/10.1001/JAMAINTERNMED.2023.0015>

Wang, X., Ma, H., Li, X., Heianza, Y., Manson, J. E., Franco, O. H., & Qi, L. (2023b). Association of Cardiovascular Health With Life Expectancy Free of Cardiovascular Disease, Diabetes, Cancer, and Dementia in UK Adults. *JAMA Internal Medicine*, 183(4), 340–349. <https://doi.org/10.1001/JAMAINTERNMED.2023.0015>

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/JOSS.03021>

Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F., Tasali, E., Twery, M., Croft, J. B., Maher, E., Barrett, J. A., ... Heald, J. L. (2015). Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society. *Journal of Clinical Sleep Medicine*, 11(6), 591–592. <https://doi.org/10.5664/JCSM.4758>

Weiss, R., Santoro, N., Giannini, C., Galderisi, A., Umamo, G. R., & Caprio, S. (2017). Prediabetes in youths: mechanisms and biomarkers. In *The Lancet Child and Adolescent Health* (Vol. 1, Issue 3). [https://doi.org/10.1016/S2352-4642\(17\)30044-5](https://doi.org/10.1016/S2352-4642(17)30044-5)

WHO. (2006). Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: report of a WHO/IDF consultation. In *World Health Organization*.

Williams, R., & Airey, M. (2002). Epidemiology and Public Health Consequences of Diabetes. *Current Medical Research and Opinion*, 18 Suppl 1. <https://doi.org/10.1185/030079902125000183>

Wilson, J. M. G., Jungner, G., Cochrane, A. L., Holland, W. W., & Whitehead, T. G. (1968). *Principles and Practice of Screening for Disease*. Geneva: World Health Organisation. *Chronic Illness in the United States*, 27.

World Health Organization. (2024). *Priorities for research on hypertension care delivery*. <https://www.who.int/publications/i/item/9789240088184>

World Health Organization (WHO). (2023). *Hypertension*. WHO. <https://www.who.int/news-room/fact-sheets/detail/hypertension>

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10, 100–107. <https://doi.org/10.1016/J.IMU.2017.12.006>

Wu, T. H., Pang, G. K. H., & Kwong, E. W. Y. (2014). Predicting systolic blood pressure using machine learning. *2014 7th International Conference on Information and Automation for Sustainability: “Sharpening the Future with Sustainable Technology”, ICIAfS 2014*. <https://doi.org/10.1109/ICIAFS.2014.7069529>

Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings*. <https://doi.org/10.1109/UBMYK48245.2019.8965556>

Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., Xia, M., Liu, M., Zhou, X., Wu, Q., Guo, Y., Zhu, C., Li, Y. M., Culver, D. S., Alfreds, S. T., Stearns, F., Sylvester, K. G., Widen, E., McElhinney, D., & Ling, X. (2018). Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *Journal of Medical Internet Research*, 20(1). <https://doi.org/10.2196/jmir.9268>

Yoda, K., Inaba, M., Hamamoto, K., Yoda, M., Tsuda, A., Mori, K., Imanishi, Y., Emoto, M., & Yamada, S. (2015). Association between poor glycemic control, impaired sleep

quality, and increased arterial thickening in type 2 diabetic patients. *PLoS ONE*, 10(4).
<https://doi.org/10.1371/journal.pone.0122521>

Zarocostas, J. (2006). Millions of deaths from environmental causes are preventable, says WHO. *BMJ* (Clinical Research Ed.), 332(7555).
<https://doi.org/10.1136/bmj.332.7555.1412-b>

Zhang, B., Wei, Z., Ren, J., Cheng, Y., & Zheng, Z. (2018). An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees. *IEEE Access*, 6.
<https://doi.org/10.1109/ACCESS.2017.2787980>

Zheng, Y., Ley, S. H., & Hu, F. B. (2017). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology* 2017 14:2, 14(2), 88–98. <https://doi.org/10.1038/nrendo.2017.151>

Zhou, B., Carrillo-Larco, R. M., Danaei, G., Riley, L. M., Paciorek, C. J., Stevens, G. A., Gregg, E. W., Bennett, J. E., Solomon, B., Singleton, R. K., Sophiea, M. K., Iurilli, M. L. C., Lhoste, V. P. F., Cowan, M. J., Savin, S., Woodward, M., Balanova, Y., Cifkova, R., Damasceno, A., ... Zuñiga Cisneros, J. (2021). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *The Lancet*, 398(10304). [https://doi.org/10.1016/S0140-6736\(21\)01330-1](https://doi.org/10.1016/S0140-6736(21)01330-1)

Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2021). Deep Learning for Diabetes: A Systematic Review. *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2744–2757. <https://doi.org/10.1109/JBHI.2020.3040225>

Ziegler, A., & König, I. R. (2014). Mining data with random forests: Current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1). <https://doi.org/10.1002/widm.1114>

Appendices

Appendix (1) Use of STEPS dataset authorization letter



Date: May 1st, 2023.

RE: Permission to Use STEPS 2022 data

This letter serves as formal permission from the World Health Organization (WHO) to access and utilize the STEPS 2022 dataset for the purposes of developing machine learning-based models for predicting and detecting raised blood sugar.

As the STEPS 2022 coordinator, I am hereby granting the study team at Arab American University; Dr. Amani Owda and Dr. Majdi Owda, and Ms. Marwa Owess, a WHO Staff to use the STEPS 2022 anonymized dataset. The data provided will be used solely for the aforementioned project and its related activities, such as data analysis, statistical modeling, paper generation, and journal paper publication. Access to data is limited to authorized personnel only.

Data Sharing:

The authorized coauthors are not permitted to share the data with any third party.

Signatures:

This agreement is effective as of the date indicated.

Sincerely,

Salwa Massad

STEPS 2022 coordinator

World Health Organization



الملخص

يعتبر ارتفاع ضغط الدم وارتفاع نسبة السكر في الدم من عوامل الخطر الرئيسية للعديد من أنواع الأمراض المزمنة غير المعدية. يعد ارتفاع نسبة السكر في الدم مؤشراً قوياً على الإصابة بمقدمات السكري أو داء السكري. وكذلك أيضاً يعتبر ضغط الدم المرتفع علامة على الإصابة بأمراض ارتفاع ضغط الدم، والذي يعتبر عامل خطر رئيسي للإصابة بأمراض القلب والأوعية الدموية. ومن المثير للاهتمام أن ارتفاع ضغط الدم والسكري هما من أهم الأمراض غير المعدية العالمية الشائعة التي تؤثر على الأشخاص البالغين وليس فقط كبار السن. في الآونة الأخيرة، تزايد انتشار أمراض السكري وارتفاع ضغط الدم بمعدل أسرع، وخاصة في البلدان النامية. إن المصدر الرئيسي للقلق المرتبط بهذه الأمراض هو احتمال حدوث مضاعفات صحية خطيرة إذا لم يتم تشخيصها مبكراً أو لم تتم إدارتها بشكل صحيح، مثل الفشل الكلوي والسكتة الدماغية والنوبات القلبية وتلف الأعصاب وأمراض العيون التي قد تتطور بشكل سيئ وتؤدي إلى الإعاقات. ولذلك، فإن الكشف والفحص في الوقت المناسب لأمراض السكري وارتفاع ضغط الدم يعتبر عاملاً حاسماً في علاج هذه الأمراض ومكافحتها، وتجنب تطورها إلى عواقب صحية وخيمة. يهدف المسح السكاني لارتفاع ضغط الدم وارتفاع نسبة السكر في الدم إلى تحديد الأفراد المعرضين للخطر قبل ظهور الأعراض، مما يتيح التدخل في الوقت المناسب وتحسين النتائج الصحية المحتملة. ومع ذلك، فإن تنفيذ برامج الفحص واسعة النطاق يمكن أن يكون مكلفاً، حيث يتطلب تخصيص موارد مالية وبشرية لاختبار ومتابعة وإدارة هذه البرامج، مما قد يجهد أنظمة الرعاية الصحية. استناداً إلى هذه الحقائق، تقدم هذه الدراسة نماذج تعلم آلي خاضعة للإشراف لاكتشاف حالات ارتفاع ضغط الدم وسكر الدم والتنبؤ بها، كأداة للتقصي عن أمراض السكري والضغط. تستخدم نماذج ارتفاع ضغط الدم وارتفاع نسبة السكر في الدم المقترحة في هذه الدراسة عوامل الخطر ذات الصلة، المشتركة بين ارتفاع ضغط الدم وارتفاع سكر الدم. تشمل عوامل الخطر المشتركة هذه العمر، ومؤشر كتلة الجسم، والعادات الغذائية، والنشاط البدني، والتاريخ المرضي، ومستوى سكر الدم بعد الصيام، التي تم الحصول عليها من مجموعة بيانات النهج التدريجي لدراسة عوامل خطر الأمراض غير السارية، والتي تم جمعها من البالغين في المجتمع الفلسطيني. تم استخدام عوامل خطر الأمراض غير المعدية التي تم جمعها بواسطة مجموعة بيانات STEPS كمتغيرات لتدريب نماذج التنبؤ التي تم بنائها باستخدام أنواع مختلفة من خوارزميات التصنيف ذات التعلم الخاضع للإشراف بما في ذلك Random Forest و Decision Tree و XGBoost و Multilayer Perceptron. بناءً على النتائج

التجريبية، أظهرت النماذج المقترحة أداءً مثاليًا عند تنفيذها باستخدام مصنف Random Forest، الذي حقق دقة قدرها 98.05% و94.76% لنماذج الكشف عن ارتفاع نسبة السكر في الدم وارتفاع ضغط الدم على التوالي. بالإضافة إلى ذلك، فإن النتائج التجريبية للنماذج المطبقة باستخدام المصنفات الأخرى كانت واعدة. يمكن تحسين النماذج المقترحة في هذه الدراسة من خلال إدخال عدة قراءات منفصلة لضغط الدم وسكر الدم للصائم والتي يتم أخذها على عدة أيام، بحيث يمكن استخدامها كأداة تشخيص عالية الكفاءة ودقيقة لمرض السكري وارتفاع ضغط الدم، ليس فقط لأغراض التقصي. بالإضافة إلى ذلك، يمكن التوسع في تحديد مستوى نسبة السكر في الدم للصائم ما إذا كان طبيعي أو مضطرب أو مرتفع، وتصنيف نوع اضطرابات ضغط الدم.