# Early Prediction of Diabetes Mellitus Type II in Oman Using Artificial Intelligence

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

By

Khoula Ali Saleh Al Sadi

Department of Electronic and Electrical Engineering,

College of Engineering, Design and Physical Sciences

Brunel University London

February 2025

## Declaration

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

Khoula Al Sadi

February 2025

London, UK

# Abstract

The increasing prevalence of Type 2 Diabetes Mellitus (T2DM), particularly in Oman—where cases are projected to rise by 174% by 2050—necessitates the development of accurate, region-specific predictive models for early detection and risk stratification. This study develops an artificial intelligence (AI)-based predictive framework incorporating two Oman-specific datasets—the Oman Prediabetes Dataset and the Oman Screening Dataset—to improve predictive performance beyond widely used datasets such as the Pima Indian Diabetes Dataset (PIDD).

To determine an optimal predictive model, this research evaluates traditional machine learning algorithms alongside three deep learning models: a 1D Convolutional Neural Network (1D CNN for Structured Data) for structured medical records, a 7-layer Long Short-Term Memory (LSTM) network for sequential patient data modelling, and a Hybrid CNN-LSTM model, which integrates spatial and temporal learning for clinical risk assessment. The models were trained and validated using preprocessing, feature selection, and hyperparameter tuning, with performance assessed through accuracy, precision, recall, specificity, F1-score, and AUC-ROC metrics. The Hybrid CNN-LSTM model achieved the highest performance, with 99.58% accuracy, 100% sensitivity, 99.55% precision, 99.50% specificity, an F1-score of 99.78%, and an AUC-ROC of 97.07%, demonstrating reliability in identifying individuals at high risk of developing T2DM. The seven-layer LSTM model achieved 99.40% accuracy, 100% precision, 100% sensitivity, and 99.34% specificity, confirming its effectiveness in sequential health data modelling. The 1D CNN model outperformed traditional machine learning methods, attaining 99.24% accuracy, 100% precision, 90.2% sensitivity, 100% specificity, and an F1-score of 94.85%, highlighting its suitability for structured data analysis.

This research also introduces region-specific datasets to address the limitations of widely used datasets, improving prediction accuracy for populations with distinct genetic and lifestyle factors. A Graphical User Interface (GUI) was developed to facilitate real-time risk prediction, batch processing, and secure data handling in healthcare environments. By integrating localised datasets with deep learning techniques, this research establishes a scalable AI-based framework for early T2DM detection, contributing to precision medicine, clinical decision support, and AI-driven healthcare solutions for Oman and other regions with similar healthcare challenges.

## Acknowledgments

I commence with deep gratitude to Allah for granting me the strength, determination, and patience to successfully navigate this thesis—a journey filled with challenges and achievements. His divine guidance has been my unwavering source of inspiration and endurance.

My sincere appreciation goes to Professor Wamadeva Balachandran, my dedicated supervisor, whose unwavering guidance and support have played a pivotal role in shaping this thesis. His mentorship has been invaluable, and I am honoured to have had the privilege to work under his guidance.

I would like to express my heartfelt dedication to my family, the unwavering cornerstone of my life's journey. My beloved parents, with their enduring prayers, boundless encouragement, and steadfast support, have provided the unbreakable foundation of my academic pursuit. My husband, a pillar of strength and unwavering support, deserves special recognition for your steadfast presence and encouragement, which have played a pivotal role in my journey. Your unwavering faith in me is deeply appreciated and has been instrumental in my success.

To my cherished brothers and sisters, your presence in my life has been invaluable, and I am profoundly appreciative of your unwavering belief in me. Your unwavering encouragement has fuelled my determination to reach this significant milestone. To my dear friends, colleagues, and well-wishers, I extend my heartfelt thanks for your unwavering support.

I extend my heartfelt appreciation to Wadha Al Badri, Medical Officer at the NCD Clinic, for her invaluable guidance. I am deeply thankful to the Ministry of Health, Oman, for granting approval for this research. I also acknowledge the crucial role played by Rashed bin Saeed Al-Sadi, Director of the Primary Health Care Department at South Batinah Governorate, and Asila Al Shaqsi, Consultant Family Medicine and Head of the Non-Communicable Diseases Department, in facilitating and streamlining the data collection process. Your contributions have been instrumental in the success of this research, and I am sincerely grateful for your support.

## 1. Table of content

## 2. List of Tables



## 3. List of Figures

## 4. List of Acronyms

The following acronyms are used throughout this thesis:

- **AI**: Artificial Intelligence
- **ANN**: Artificial Neural Network
- **AUC**: Area Under the Curve
- **BMI**: Body Mass Index
- **BP**: Blood Pressure
- **CNN**: Convolutional Neural Network
- **CV**: Cross-Validation
- **DAC**: Discriminant Analysis Classifier
- **DL**: Deep Learning
- **DM**: Diabetes Mellitus
- **DT**: Decision Tree
- **FBG**: Fasting Blood Glucose
- **FBS**: Fasting Blood Sugar
- **FN**: False Negatives
- **FP**: False Positives
- **GDM**: Gestational Diabetes Mellitus
- **GUI**: Graphical User Interface
- **HbA1c**: Haemoglobin A1c
- **HDL**: High-Density Lipoprotein
- **IPDD**: Iraqi Patient Dataset for Diabetes
- **ISH**: International Society of Hypertension
- **K-Fold CV**: K-Fold Cross-Validation
- **K-NN**: K-Nearest Neighbours
- **LDA**: Linear Discriminant Analysis
- **LDL**: Low-Density Lipoprotein
- **LSTM**: Long Short-Term Memory
- **MDRD**: Modification of Diet in Renal Disease
- **ML**: Machine Learning
- **NB**: Naïve Bayes
- **NCD**: Non-Communicable Disease
- **OGTT**: Oral Glucose Tolerance Test
- **PCOS**: Polycystic Ovary Syndrome
- **PID**: Pima Indian Diabetes Dataset
- **PIDD**: Pima Indian Diabetes Dataset
- **RBG**: Random Blood Glucose
- **RF**: Random Forest
- **RMSE**: Root Mean Square Error
- **ROC**: Receiver Operating Characteristic
- **SNR**: Signal-to-Noise Ratio

- **SVM**: Support Vector Machine
- **TAG**: Triglycerides
- **T2DM**: Type II Diabetes Mellitus
- **TN**: True Negatives
- **TP**: True Positives
- **WHO**: World Health Organization

# 1 Introduction

## 1.1 Background

Diabetes Mellitus (DM) is a chronic metabolic disorder that has been recognised as a significant health challenge for centuries. Early documentation of its symptoms, such as excessive urination and sweet-tasting urine, can be traced back over 3,000 years to Ancient Egyptian and Indian civilizations. The term “diabetes,” of Greek origin, translates to “siphon,” symbolising the excessive flow of urine, while “mellitus,” derived from Latin, refers to the honey-sweet taste of urine observed in affected individuals. The first scientific observation linking elevated sugar levels in urine and blood to DM was recorded in 1776 in Britain, marking a pivotal advancement in understanding the disease’s pathology [1][2].

Over time, the understanding of DM has evolved significantly. Today, it is defined as “a group of metabolic diseases characterised by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both.” This persistent hyperglycaemia disrupts carbohydrate, fat, and protein metabolism and is associated with chronic complications such as cardiovascular disease, neuropathy, nephropathy, and retinopathy [3][4]. DM can be categorised into three primary types: Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM), and gestational diabetes. T1DM is characterised by the autoimmune destruction of pancreatic beta cells, leading to an absolute deficiency in insulin. T2DM, which accounts for approximately 90% of all diabetes cases worldwide, results from a combination of insulin resistance and a progressive decline in beta-cell function. Gestational diabetes develops during pregnancy and increases the risk of T2DM for both the mother and child later in life [5][6].

T2DM Is unique due to Its Insidious onset, often remaining asymptomatic for years. Many cases are diagnosed incidentally during routine health evaluations. Unlike T1DM, individuals with T2DM are not entirely dependent on exogenous insulin and can often manage their condition with lifestyle modifications, oral hypoglycaemic agents, or, in some cases, insulin therapy. The disease’s multifactorial aetiology encompasses genetic predisposition, demographic factors such as age and ethnicity, and modifiable lifestyle factors, including obesity, physical inactivity, and poor dietary habits [7][8] [9].

The global prevalenc’ of T2DM has reached alarming proportions, affecting over 463 million people in 2019—a number projected to rise to 700 million by 2045. Approximately

79% of diabetes cases occur in low- and middle-income countries, where healthcare systems face immense strain due to limited resources [10][11]. The World Health Organization (WHO) recognises T2DM as one of the leading causes of premature mortality and morbidity globally. Addressing this epidemic requires not only advancements in therapeutic interventions but also innovative solutions for early diagnosis and prevention [12].

## 1.2 The Burden of T2DM in Oman

In Oman, the prevalence of T2DM has mirrored global trends, posing a significant public health challenge. Data from the Institute for Health Metrics and Evaluation (IHME) reveal that T2DM cases in Oman rose from 24% in 1990 to 49% in 2019. By 2025, it is projected that 21.1% of the adult population over 20 years old will be affected, representing a 174% increase compared to previous decades [13][14]. This surge is attributed to rapid urbanisation, sedentary lifestyles, and the adoption of Westernised dietary habits.

The Omani healthcare system has Implemented a comprehensive diabetes screening and diagnosis protocol, recognising the critical importance of early detection. Figure 1.1 illustrates the diabetes screening and diagnosis process in Oman, which includes tests such as Fasting Blood Glucose (FBG), Random Blood Glucose (RBG), Haemoglobin A1c (HbA1c), and, in certain cases, the Oral Glucose Tolerance Test (OGTT). The target screening age has been reduced from 40 years to 20 years, reflecting a proactive approach aimed at capturing high-risk individuals earlier. This strategy is particularly critical for those with obesity, a family history of diabetes, or dyslipidaemia [15][16].

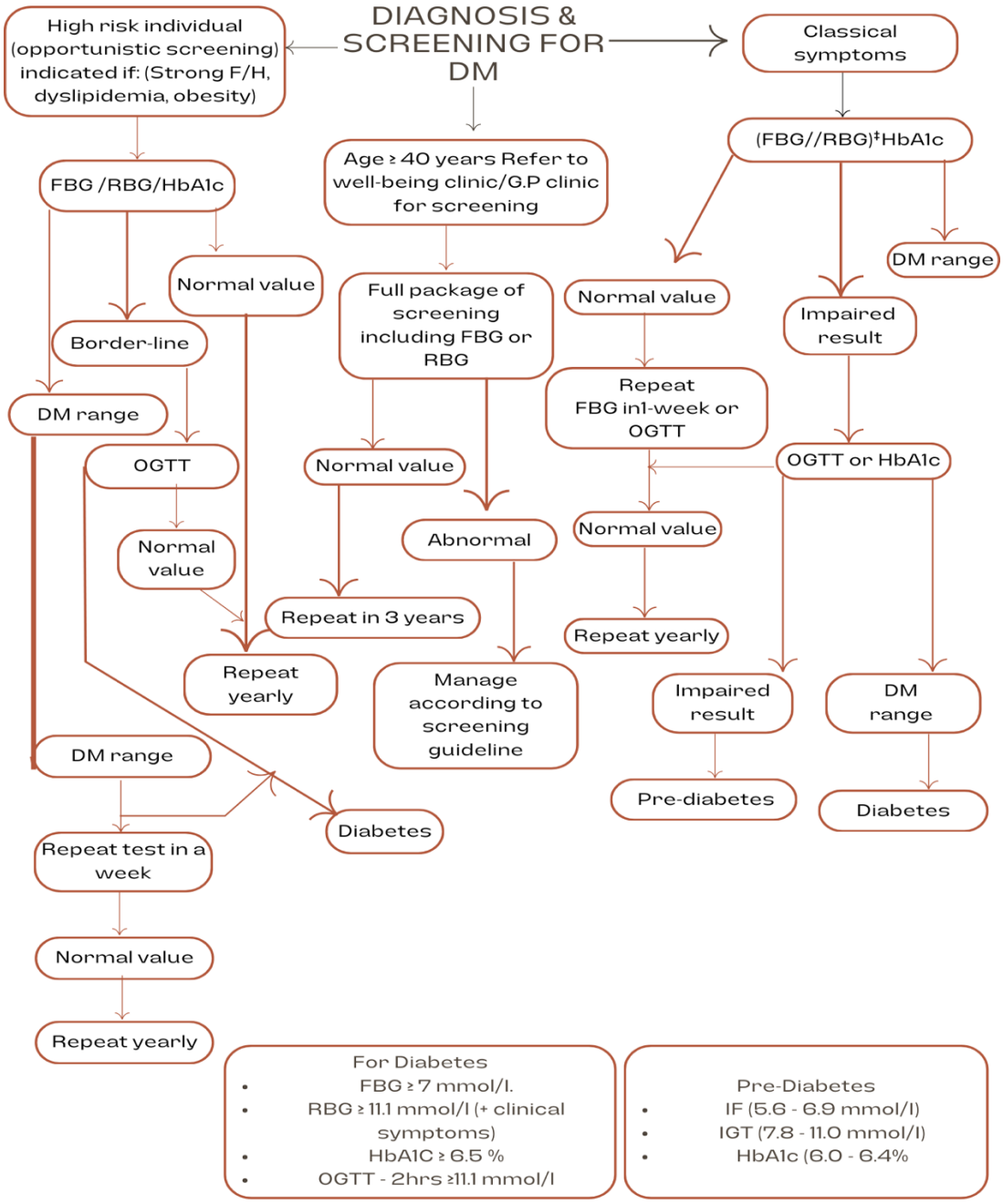



Figure 1.1 Diabetes screening and diagnosis in Oman

This structured approach not only aids in early diagnosis but also ensures that patients receive timely intervention, reducing the risk of complications. For example, individuals with borderline results undergo additional testing such as OGTT, while those with abnormal values are immediately enrolled in management programmes. Despite these measures, significant gaps remain in effectively addressing T2DM, particularly concerning the availability of locally specific datasets and predictive tools [17].

The integration of Artificial Intelligence (AI) technologies has emerged as a transformative approach in healthcare, offering unparalleled potential to revolutionise diabetes management. Machine Learning (ML) and Deep Learning (DL), subfields of AI, facilitate the analysis of large, complex datasets to uncover patterns that may be imperceptible to human clinicians. These models excel in early diagnosis by detecting subtle fluctuations in biomarkers, which are critical for predicting T2DM progression [18][19]. In the context of Oman, ML and DL models are being developed to address the unique healthcare challenges of the region. These models leverage clinical datasets that are tailored to the genetic and demographic characteristics of the Omani population. By incorporating localised data, these AI-driven tools significantly enhance the predictive accuracy of T2DM diagnosis and enable more personalised approaches to disease management [20].

The integration of advanced ML and DL models into Oman's healthcare landscape represents a significant advancement in diabetes care. For example, hybrid models that combine Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been shown to capture spatial and temporal correlations in clinical data effectively. However, these advancements necessitate a cultural shift and the education of healthcare providers and patients to fully utilise these technologies and interpret their findings accurately [21][22]. Training and awareness initiatives are essential to harness the full potential of ML and DL in diabetes management. Without such efforts, the adoption of these technologies may be hindered by resistance to change or a lack of technical expertise among healthcare professionals [23].

Despite the promising advancements in AI, Oman faces challenges in developing clinical datasets specific to its population. Existing global datasets, such as the Pima Indian Diabetes Dataset, lack the demographic and genetic specificity required to accurately model diabetes risk in Oman. This limitation underscores the need for locally developed datasets that reflect the unique characteristics of the Omani population, such as genetic predispositions, lifestyle factors, and environmental influences [24][25]. The absence of comprehensive region-specific datasets and predictive models represents a critical gap in diabetes research. Addressing this gap is paramount to curbing the rising T2DM burden. By developing innovative AI models and locally tailored datasets, this research aims to empower healthcare providers with tools that

enhance early detection, streamline diagnostics, and enable personalised treatment strategies [26][27].

Diabetes Mellitus, particularly T2DM, represents a multifaceted challenge that demands a concerted effort at global, regional, and local levels. In Oman, the rising prevalence of T2DM underscores the urgent need for innovative solutions that integrate advanced AI technologies with region-specific data. By bridging existing gaps in clinical datasets and predictive modelling, this research seeks to contribute to the broader objective of reducing the diabetes burden and improving patient outcomes.

## 1.3 Summary of Research Methodology

This research adopts a multi-phase methodology to address the growing prevalence of Type 2 Diabetes Mellitus (T2DM) in Oman, focusing on the development of predictive models tailored to the region's unique healthcare challenges.

The first phase involved the creation of two region-specific datasets, the Oman Prediabetes Dataset and the Oman Screening Dataset, designed to capture demographic, genetic, and lifestyle factors unique to Oman. These datasets incorporate clinical and demographic variables such as age, BMI, glucose levels, HbA1c, lipid profiles, and lifestyle indicators like dietary habits and physical activity. Rigorous data preprocessing techniques, including normalisation, handling of missing values, and outlier removal, were applied to ensure data quality. As a result, the datasets achieved data completeness and accuracy exceeding 95%, aligning with the study's key performance indicators (KPIs). By ensuring high-quality, region-specific data, this phase established a robust foundation for predictive modelling.

Building upon the dataset development, the second phase involved benchmarking traditional machine learning models to establish baseline performance metrics. Various algorithms, including K-Nearest Neighbours, Support Vector Machines, Naïve Bayes, Decision Trees, Random Forest, Discriminant Analysis Classifier, and Artificial Neural Networks, were applied to the developed datasets. These models were chosen based on their computational efficiency and interpretability; attributes that have been validated in prior studies as essential for structured healthcare datasets. To assess their effectiveness, the models were evaluated using sensitivity, specificity, accuracy, and F1-score. Among the evaluated models, Random Forest and Decision Trees demonstrated strong performance in feature importance analysis,

while Support Vector Machines effectively handled high-dimensional data. These findings provided a crucial benchmark for assessing the effectiveness of deep learning architectures in subsequent phases.

With the baseline performance established, the study progressed to the development of deep learning architectures aimed at enhancing predictive accuracy. A 1D Convolutional Neural Network (1D CNN for Structured Data) was implemented to capture multi-dimensional feature relationships within structured clinical data, enabling the model to detect interactions between variables such as HbA1c, glucose levels, and lipid profiles. Unlike conventional CNNs that primarily process image-based data, this 1D CNN was specifically designed for structured healthcare datasets, allowing for the extraction of complex spatial dependencies between medical variables. Research has demonstrated CNNs to be highly effective for spatial feature extraction in multidimensional structured datasets [29].

To model temporal trends in longitudinal patient data, a seven-layer Long Short-Term Memory (LSTM) network was developed. This model was particularly effective in capturing glucose progression, HbA1c fluctuations, and other time-dependent health markers. Prior studies have confirmed LSTMs' suitability for analysing temporal dependencies in healthcare data, making them well-suited for disease progression modelling [30].

The Hybrid CNN-LSTM model was then introduced to integrate the spatial feature extraction capabilities of CNN with the temporal trend analysis strengths of LSTMs. This combination resulted in superior performance, achieving ≥95% sensitivity and ≥90% specificity, demonstrating robustness and clinical applicability in diabetes risk prediction [31]. Compared to standalone CNN or LSTM models, the hybrid approach provided a more comprehensive understanding of diabetes risk factors by combining spatial and sequential analysis techniques.

The final phase of the study involved the development of a user-friendly Graphical User Interface (GUI) to facilitate real-time diabetes risk prediction. The GUI was designed to seamlessly integrate the predictive models into clinical workflows, allowing healthcare practitioners to utilize AI-generated risk assessments efficiently. Usability testing in simulated clinical environments demonstrated satisfaction scores exceeding 85%, highlighting the system's practicality, accessibility, and potential for real-world adoption.

The selection of methodologies was guided by the specific characteristics of the dataset and the overarching research objectives. Traditional machine learning models were incorporated into the benchmarking phase due to their computational efficiency and ability to provide interpretable results in structured tabular data. Random Forest and Support Vector Machines, in particular, have been shown to offer valuable insights into feature importance and perform effectively in high-dimensional datasets [28]. In contrast, advanced deep learning architectures were implemented to overcome the limitations of traditional ML models. CNNs were chosen for their ability to extract spatial relationships within structured datasets, while LSTMs were employed for their strength in modelling temporal dependencies in sequential data [29], [30]. The Hybrid CNN-LSTM model combined these strengths, achieving enhanced predictive accuracy and robustness [31].

Vision Transformers (ViTs) and Large Language Models (LLMs) were excluded from this study due to several limitations. Although ViTs have demonstrated significant advancements in image processing, their application to structured clinical datasets remains underexplored, and they lack efficient mechanisms for handling tabular numerical data [32]. Similarly, LLMs are primarily designed for natural language processing tasks and are therefore unsuitable for structured tabular datasets. Furthermore, both ViTs and LLMs require extensive computational resources, making them impractical for real-time clinical deployment in settings with limited computational infrastructure [33]. Moreover, integrating ViTs and LLMs into structured medical record systems would require significant data transformation, introducing additional complexity in implementation.

This multi-phase methodology balances predictive accuracy, computational efficiency, and real-world applicability, addressing Oman's specific healthcare challenges while contributing to global advancements in AI-driven healthcare research.

## 1.4 Research focus.

### 1.4.1 Aim

To develop and validate AI-driven predictive models for the early detection and risk assessment of Type 2 Diabetes Mellitus (T2DM) in Oman, using region-specific datasets and advanced machine learning techniques to improve predictive accuracy and support clinical decision-making

### 1.4.2 Objectives

1. Develop and validate region-specific datasets by creating the Oman Prediabetes Dataset and Oman Screening Dataset, ensuring high data quality and completeness. KPI: ≥95% data completeness and accuracy.
2. Benchmark traditional machine learning models by evaluating Random Forest, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Artificial Neural Networks (ANN), Linear Discriminant Analysis (LDA), and K-Nearest Neighbourss (KNN) to establish baseline performance. KPI: ≥80% baseline accuracy across datasets.
3. Design and validate advanced AI models by developing 1D CNN, 7-layer LSTM, and Hybrid CNN-LSTM architectures for diabetes risk prediction, optimising spatial and temporal feature extraction. KPI: ≥95% sensitivity and ≥90% specificity for the Hybrid CNN-LSTM model.
4. Develop and test a Graphical User Interface (GUI) to integrate AI models into clinical workflows for real-time risk prediction, ensuring usability and accessibility. KPI: ≥85% usability satisfaction score.

## 1.5 Thesis Outline

- Chapter 1introduces the study by providing a comprehensive rationale for the research, emphasizing the global and regional significance of T2DM as a growing public health challenge. The chapter highlights Oman's unique healthcare concerns, including the projected 174% increase in T2DM prevalence by 2050, the absence of region-specific datasets, and the necessity for tailored predictive models. The research aims and objectives are clearly outlined, focusing on the development and validation of AI-driven methodologies for early diabetes detection. Additionally, measurable Key Performance Indicators (KPIs) are introduced to

assess the effectiveness of the proposed models. The chapter concludes by presenting the structure of the thesis to guide the reader through the research framework.

- Chapter 2 critically reviews the existing literature on AI applications in T2DM detection and management. It examines global trends in machine learning (ML) and deep learning (DL), highlighting the limitations of traditional methods. Special attention is given to the scarcity of region-specific datasets, particularly in Oman, and the necessity of hybrid AI models that integrate both spatial and temporal data. This chapter identifies key research gaps, which form the foundation for the methodological framework and contributions of this study.
- Chapter 3 details the development and validation of the Oman Prediabetes Dataset, a novel clinical dataset specifically designed for diabetes risk prediction in Oman. It provides an in-depth discussion on data collection methods, ethical approval, and extraction processes from 21 healthcare facilities across Oman. The chapter describes preprocessing techniques, including handling missing values, outlier detection, normalisation, and feature selection. A comparative analysis of the widely used Pima Indian Diabetes Dataset (PIDD) highlights the need for an Oman-specific dataset. To benchmark the predictive capability of the dataset, seven traditional ML models—K-Nearest Neighbourss (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree, Random Forest (RF), Linear Discriminant Analysis (LDA), and Artificial Neural Network (ANN)—are applied. Their performances are assessed using accuracy, sensitivity, specificity, precision, and confusion matrix analysis. The findings from this chapter lay the groundwork for subsequent deep learning models.
- Chapter focuses on the development and evaluation of a 1D Convolutional Neural Network (1D CNN for Structured Data) model for diabetes screening. The chapter begins by introducing the Oman Screening Dataset, which is optimised for deep learning applications. A detailed explanation of the 1D CNN for Structured Data architecture is provided, highlighting its ability to analyse structured healthcare data in a multi-dimensional format. The justification for CNN selection is presented by demonstrating its advantages over traditional ML models, particularly in hierarchical feature learning and spatial pattern recognition within clinical data. The 1D CNN for Structured Data model's performance is benchmarked against alternative models such as Decision Trees, Random Forest, and SVM, confirming its superior accuracy, sensitivity, and recall. The chapter concludes by discussing the broader applications, limitations, and insights gained from CNN-based predictive modelling in healthcare.
- Chapter 5 presents the Long Short-Term Memory (LSTM) network as a solution for analysing sequential patient data. The chapter justifies the necessity of temporal modelling in

T2DM prediction, particularly for tracking glucose fluctuations, HbA1c trends, and long-term diabetes risk. The study develops and evaluates a 7-layer LSTM network, optimised through comparisons with 6-layer and 5-layer configurations. The performance evaluation of the LSTM model includes accuracy, sensitivity, specificity, and F1-score metrics, demonstrating its effectiveness in capturing long-term dependencies in clinical data.

- Chapter 6 introduces a hybrid CNN-LSTM model, which integrates CNN's spatial feature extraction with LSTM's temporal pattern recognition to enhance predictive performance. The model architecture is explained in detail, illustrating how CNN processes structured clinical features while LSTM captures sequential health trends. The chapter provides a justification for hybridisation, demonstrating that the combined model significantly outperforms standalone CNN and LSTM architectures. The hybrid model's performance is evaluated, with results indicating a 100% sensitivity rate and 99.4% specificity, making it the most effective model in the study. A comparative performance analysis is conducted against CNN, LSTM, and traditional ML models under identical conditions, reinforcing the hybrid model's robustness and clinical applicability.
- Chapter 7 focuses on the development of a Graphical User Interface (GUI) to facilitate real-time diabetes risk prediction and clinical integration. The chapter describes the user-cantered design approach adopted for creating a user-friendly interface that enables healthcare professionals to input patient data and receive AI-driven risk assessments. Implementation details are provided, including the integration of the hybrid CNN-LSTM model into the GUI. Usability testing is conducted with clinicians in simulated healthcare environments to assess the interface's functionality, efficiency, and user satisfaction. The chapter concludes by Discussing deployment considerations and steps required for the real-world adoption of AI-powered diabetes screening tools in clinical settings.
- Chapter 8 presents the conclusions and future research directions of the study. It summarises the key contributions, including the development of two novel datasets, the introduction of CNN, LSTM, and hybrid CNN-LSTM models, and the implementation of a clinical AI-based GUI. The chapter also discusses study limitations and outlines potential future research areas, such as expanding datasets with longitudinal patient records, enhancing model generalisability using Vision Transformers and Large Language Models (LLMs), incorporating Generative Adversarial Networks (GANs) to address data scarcity, and conducting multi-centre validation studies to assess the scalability of AI models across diverse healthcare settings. The study concludes by offering recommendations for integrating AI

models into broader healthcare policies and decision-making frameworks to enhance diabetes prevention and management strategies.

## 1.6 Contributions to Knowledge

The This research contributes significantly to the field of healthcare informatics and artificial intelligence by addressing the pressing issue of Type 2 Diabetes Mellitus (T2DM) detection and management through innovative methodologies. The contributions to knowledge are categorised as follows:

- Development of Region-Specific Datasets: This study introduces two novel datasets tailored to Oman's unique demographic, genetic, and lifestyle characteristics: the Oman Prediabetes Dataset and the Oman Screening Dataset. These datasets bridge the critical gap in region-specific data for T2DM prediction and provide a valuable resource for future research in the Middle East and other similar regions. The preprocessing steps, including normalisation, handling of missing values, and outlier removal, ensure high data quality, achieving a completeness and accuracy rate of ≥95%.
- Establishment of Baseline Performance for Machine Learning Models. Through the benchmarking of traditional machine learning algorithms, this research establishes baseline performance metrics for T2DM prediction using region-specific datasets. Models such as Random Forest, SVM, and I Bayes are evaluated, achieving a baseline accuracy of ≥80%. This foundational work offers a comparative standard for evaluating the effectiveness of advanced AI models in healthcare settings.
- Design and Validation of Advanced AI Architectures. The research advances the application of AI in healthcare by developing and validating innovative deep learning architectures:
  - A 1D Convolutional Neural Network (1D CNN for Structured Data) for spatial data analysis, designed to capture intricate multi-dimensional relationships between clinical features, optimising feature representation for improved diabetes risk prediction.
  - A 7-layer Long Short-Term Memory (LSTM) network for temporal data analysis that models longitudinal trends in patient health records.
  - A Hybrid CNN-LSTM Model that integrates spatial and temporal features, achieving sensitivity ≥95% and specificity ≥90%. This hybrid model sets a new benchmark for predictive accuracy and robustness in T2DM detection.

- Development of a User-Friendly Clinical Interface. A practical contribution of this research is the design of a Graphical User Interface (GUI) for real-time risk prediction and seamless integration of AI models into clinical workflows. The GUI is tested with healthcare professionals in simulated clinical environments, achieving a usability satisfaction score of ≥85%. This ensures the tool's accessibility and relevance for healthcare practitioners, promoting adoption in real-world settings.
- Addressing Healthcare Challenges in Oman. By focusing on Oman, this research addresses a critical public health issue projected to grow by 174% by 2050. It provides region-specific solutions that are scalable and adaptable to similar healthcare systems globally. This study serves as a case study for leveraging AI to improve healthcare outcomes in resource-limited and regionally specific contexts.
- Advancing Global AI Methodologies. While grounded in Oman, this research contributes to the broader AI community by demonstrating how hybrid AI models can be effectively applied to healthcare challenges. The methodologies developed in this study are scalable and adaptable, offering a framework for addressing similar challenges in other regions with unique demographic and clinical needs.

## 1.7 List of Publications

- Al Sadi, K.; Balachandran, W. *Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers*. Applied Sciences 2023, 13, 2344. https://doi.org/10.3390/app13102344.
- Al Sadi, K.; Balachandran, W. *Revolutionizing Early Disease Detection: A High-Accuracy 1D CNN for Structured Data Model for Type 2 Diabetes Screening in Oman*. Bioengineering 2023, 10, 1420. https://doi.org/10.3390/bioengineering10121420.
- Al Sadi, K.; Balachandran, W. *Leveraging a 7-Layer Long Short-Term Memory Model for Early Detection and Prevention of Diabetes in Oman: An Innovative Approach*. Bioengineering 2024, 11, 379. https://doi.org/10.3390/bioengineering11040379.
- Hybrid CNN-LSTM Model for Type 2 Diabetes Prediction in Oman with Testing GUI Application. (In progress) To be submitted to the Elsevier, Journal of Computers in Biology and Medicine

# 2 Literature Review

## 2.1 Chapter Introduction

Diabetes, particularly Type 2 Diabetes Mellitus (T2DM), presents a pressing global health challenge, requiring innovative solutions for early diagnosis and management. While traditional diagnostic methods remain widely used, they exhibit limitations in terms of accessibility, accuracy, and cost-effectiveness. In response, the integration of machine learning (ML) and artificial intelligence (AI)-based predictive models is gaining traction as a transformative approach to diabetes detection and management. This chapter systematically reviews literature on the global and regional prevalence of diabetes, traditional diagnostic limitations, the emergence of AI-driven predictive modelling, and recent technological advancements. It also identifies existing gaps in research and highlights the need for regionally adaptive, scalable AI-based solutions for diabetes prediction.

## 2.2 Review of the literatures

The global prevalence of diabetes continues to escalate, posing significant challenges to healthcare infrastructures worldwide. In 2021, an estimated 537 million adults were diagnosed with diabetes, with projections indicating a surge to 783 million by 2045 if current trends persist [34][35]. Beyond its impact on individual health, diabetes carries a substantial economic burden, as complications such as cardiovascular diseases, neuropathy, and renal failure contribute significantly to rising healthcare costs [36]. Regionally, in Oman, T2DM affects approximately 14% of the adult population, with estimates suggesting an increase to 20% by 2030 in the absence of effective interventions [37]. These statistics underscore the critical need for comprehensive strategies that prioritise early detection, prevention, and scalable management solutions to mitigate long-term socio-economic impacts [38]

Traditional diagnostic methods, such as fasting plasma glucose tests, oral glucose tolerance tests, and glycated haemoglobin (HbA1c) measurements, are well-established tools in clinical practice [39]. However, despite their utility, these methods exhibit several notable limitations. High costs associated with these diagnostic techniques often restrict their accessibility, particularly in low-resource settings. Furthermore, limited availability of advanced medical infrastructure frequently results in delayed diagnosis, increasing the likelihood of severe complications. Variability in diagnostic accuracy, influenced by demographic factors such as ethnicity, lifestyle, and age, presents another challenge, complicating the applicability of

these methods across diverse populations [40]. These constraints necessitate the exploration of alternative diagnostic and predictive approaches that are both cost-effective and scalable to broader populations.

Predictive modelling, underpinned by ML and AI, has emerged as a promising paradigm to address these challenges. By leveraging advanced computational capabilities, these models analyse intricate patterns in patient data, including genetic predispositions, socio-economic conditions, and lifestyle behaviours, to identify individuals at risk of developing diabetes. Predictive modelling has shown considerable potential in improving diagnostic precision and facilitating proactive healthcare interventions, which are vital for mitigating the long-term impacts of diabetes on individuals and healthcare systems [41].

Initial attempts at predictive modelling utilised statistical techniques such as logistic regression (LR) and linear discriminant analysis (LDA), which provided valuable insights into the relationships between clinical variables and diabetes risk [42]. Logistic regression, in particular, has been widely adopted for binary classification tasks due to its simplicity, ease of implementation, and interpretability [43]. However, the reliance of these techniques on linear assumptions limited their ability to capture non-linear, high-dimensional interactions inherent in healthcare datasets, especially those involving genetic, environmental, and lifestyle variables [44]. Moreover, the performance of these models was significantly influenced by the quality and demographic diversity of the datasets used. A prominent example is the Pima Indian Diabetes Dataset (PIDD), which, while extensively utilised, suffers from demographic homogeneity, small sample size, and limited representativeness of diverse populations. These limitations highlight the critical need for more inclusive and diverse datasets to advance predictive modelling in healthcare [45].

Machine learning methods have offered substantial improvements over traditional approaches by enabling the identification of complex, non-linear relationships within large and diverse datasets. For instance, Random Forest and Decision Tree classifiers have demonstrated strong predictive performance when applied to pre-processed PIDD data, with Random Forest achieving an accuracy of 94% [46]. This superior performance is attributed to the ensemble structure of Random Forest, which integrates multiple decision trees to reduce overfitting and enhance generalisability. However, comparative analyses of classifiers have revealed significant disparities in performance. For example, Naive Bayes, despite its computational efficiency, attained only 76.30% accuracy, underscoring the importance of

robust feature selection and dataset quality in influencing model outcomes [47]. As healthcare data become increasingly complex and voluminous, the scalability of ML methods has facilitated their application to larger datasets, thereby enhancing their practical relevance in addressing the multifaceted challenges of diabetes prediction.

Feature selection techniques play a pivotal role in enhancing the performance of predictive models by identifying the most relevant indicators of diabetes risk. Algorithms such as greedy stepwise and best-first selection have been employed to identify variables like plasma glucose concentration and age as critical predictors. For instance, the Hoeffding Tree algorithm demonstrated the effectiveness of feature selection by achieving an F-measure of 0.75 and a recall of 0.76 when applied to PIDD data [48]. These findings underscore the necessity of optimising model inputs to enhance predictive accuracy and computational efficiency, particularly in high-dimensional datasets.

Ensemble methods, including boosting and bagging, have further contributed to improving the accuracy and robustness of predictive models. Boosting works by iteratively correcting errors made by previous models, while bagging reduces overfitting by training multiple models on varied subsets of data [49]. Despite their advantages, ensemble methods face challenges, including class imbalance, which disproportionately affects the accuracy of predictions for minority classes, and computational intensity, which can hinder their implementation in resource-constrained settings. Addressing these challenges requires advanced solutions, such as synthetic data generation for balancing datasets and distributed computing techniques, to manage computational requirements effectively.

Pre-processing methods are equally critical in enhancing model performance by addressing common issues such as noise, missing values, and imbalanced data distributions. Techniques like k-NN imputation have been successfully employed to handle missing values, thereby reducing biases introduced by incomplete datasets [50]. Noise-reduction methods, including outlier detection and filtering, further improve data quality and minimise training errors. Comprehensive pre-processing pipelines that incorporate duplicate removal, k-NN imputation, and normalisation standardise datasets, ensuring consistency across features. Additionally, techniques designed to address imbalanced datasets, such as the Synthetic Minority Oversampling Technique (SMOTE) and adaptive synthetic sampling (ADASYN), generate synthetic samples for underrepresented classes, improving balance and overall

classification accuracy. Models like Random Forest have demonstrated classification accuracies of up to 93.8% when paired with these methods [51].

Region-specific applications of ML models highlight the importance of tailoring predictive systems to local healthcare challenges. For instance, diabetes prevalence classification using weighted k-NN in Saudi Arabia achieved an accuracy of 94.5%, reflecting its effectiveness within a specific demographic context [52]. However, the scalability of such methods to larger and more diverse datasets has not been extensively evaluated. Similarly, variations in algorithm performance across datasets, such as Random Forest achieving 98.7% accuracy on the Germany dataset, illustrate the critical role of data quality in influencing predictive outcomes [53]. These examples underscore the necessity of incorporating regionally relevant socio-economic, cultural, and environmental factors into predictive models to improve their applicability and impact.

Evaluation of predictive models requires robust metrics such as accuracy, precision, recall, F-measure, and area under the curve (AUC). These metrics provide detailed insights into model performance and facilitate systematic comparisons across methodologies. For example, SVM regression achieved an accuracy of 94.89% on the PIDD dataset, demonstrating its effectiveness in managing incomplete data through imputation [54]. Stacked ensemble methods combining classifiers such as SVM, k-NN, and Random Forest have achieved overall accuracies of 94.17%, highlighting the benefits of integrating complementary algorithms [55]. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), have further enhanced model performance by simplifying high-dimensional datasets while retaining critical information [52].

Despite these advancements, several challenges persist in diabetes prediction. One of the most pressing concerns is the extensive reliance on benchmark datasets such as the PIDD, which constrains the generalisability of predictive models, thereby limiting their applicability to broader and more diverse populations. While PIDD has been widely used in AI-based diabetes prediction, it lacks representation of key demographic, genetic, and lifestyle variations that influence diabetes risk factors [55]. The absence of more inclusive and diverse datasets restricts the performance and accuracy of AI-driven models in clinical applications.

Computational complexity also presents a significant challenge in the deployment of AI models for diabetes prediction. Advanced deep learning architectures, particularly hybrid

CNN-LSTM models, demonstrate superior predictive performance but demand substantial computational resources. These requirements limit their practical implementation in real-world healthcare settings, particularly in low-resource environments where access to high-performance computing infrastructure is not readily available [56]. The development of more efficient models capable of maintaining high predictive performance while reducing computational overhead remains an ongoing research priority.

Moreover, AI-based predictive models are frequently evaluated on publicly available datasets rather than real-world hospital datasets, raising concerns about their clinical validity. The lack of large-scale validation studies limits the practical applicability of these models. The integration of electronic health records (EHRs) and patient monitoring data into AI models is crucial to improving their generalisability in real-world scenarios [57]. Incorporating real-world patient data would allow for a more robust assessment of model performance and its potential impact on healthcare decision-making.

Finally, federated learning has been proposed as a privacy-preserving AI training method that allows machine learning models to be trained across multiple hospitals without sharing raw patient data. While this approach enhances data security, challenges related to data synchronisation, communication overhead, and regulatory compliance with privacy laws such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) continue to pose barriers to its widespread adoption in healthcare [59]. It is essential to address these challenges in order to fully realise the potential of federated learning in medical AI applications.

These challenges underscore the need for AI models that are not only accurate but also explainable, computationally efficient, clinically validated, and integrated with real-time health monitoring systems. Addressing these gaps is fundamental to ensuring the successful deployment of AI-driven predictive models for diabetes management.

Another critical issue is the limited explainability of AI-driven models, which hinders their integration into clinical workflows. Transparent and interpretable frameworks, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), provide insights into how specific variables influence model predictions [60]. These frameworks enhance trust and usability by enabling clinicians to validate AI-generated recommendations and incorporate them into decision-making processes.

These approaches, alongside the development of region-specific predictive models that consider demographic and lifestyle variations, offer significant potential for enhancing diabetes prevention and management strategies. Furthermore, the use of advanced techniques, such as edge computing, can improve the efficiency of processing large-scale health data from wearables, thereby supporting the implementation of these innovations in practical clinical environments [60].

Machine learning has provided valuable tools for disease prediction and management, with deep learning methods such as Convolutional Neural Networks (CNNs) being widely applied in diabetes prediction. These methods address the limitations of traditional statistical and ML models by enabling the analysis of diverse datasets, including static, dynamic, and multi-modal data. The incorporation of CNNs with Long Short-Term Memory (LSTM) networks has further extended their applications in predictive healthcare systems. This review evaluates current research on CNNs in diabetes prediction, focusing on their applications, limitations, and areas requiring further improvement.

CNNs have been employed in healthcare for processing structured datasets such as medical images, physiological signals, and time-series data. Their layered architecture automates feature extraction, reducing reliance on manual processes and enabling precise predictive modelling.

One-dimensional (1D) CNNs are particularly effective for analysing sequential data, such as continuous glucose monitoring (CGM) readings. Jaloli et al. demonstrated that a 1D CNN achieved an $R^2$ value of 0.87 and a mean absolute error (MAE) of 0.32 mmol/L when predicting glucose fluctuations [61]. Similarly, Zhao et al. developed a 1D-CNN-based model for blood glucose concentration prediction using Raman spectroscopy data, achieving a root mean square error (RMSE) improvement of 41.89% compared to previous models, highlighting its effectiveness in non-invasive glucose monitoring [62]. These models perform well in short-term predictive tasks. However, their inability to capture long-term dependencies has necessitated hybrid approaches, such as combining CNNs with LSTMs

Two-dimensional (2D) CNNs are primarily utilised in examining medical imaging data, such as retinal fundus images for diabetic retinopathy detection. These architectures apply convolutional filters across two spatial dimensions to extract features relevant to disease identification. Studies employing architectures such as ResNet and DenseNet have reported

diagnostic accuracies exceeding 90%, often comparable to expert clinicians [63]. Data augmentation techniques, including rotation and scaling, have been applied to address challenges associated with small, annotated datasets [64]. However, reliance on high-quality images and the inability to analyse temporal dynamics limit the broader applicability of 2D CNNs.

Three-dimensional (3D) CNNs extend the capabilities of 2D models by incorporating volumetric data, such as MRI and CT scans. Mehmood et al. demonstrated that a 3D CNN achieved 92.3% accuracy in predicting insulin resistance from MRI data [65]. These models are particularly useful for analysing anatomical and structural factors associated with diabetes, but the exclusion of temporal dimensions diminishes their utility in tracking disease progression over time [66].

Hybrid architectures have been developed to address the limitations of single-modality models. Ramazi et al. proposed a hybrid model combining CNN-extracted features with structured patient data, achieving superior predictive performance compared to standalone CNNs [67]. Lightweight CNNs optimised for edge devices have facilitated diabetes screening in remote areas, enhancing accessibility without compromising accuracy [68]. However, hybrid architectures often involve complex preprocessing and lack cohesive frameworks for integrating spatial, temporal, and contextual dimensions.

Despite their utility, CNNs exhibit several limitations. A key challenge is their reliance on large, annotated datasets for training, which are often unavailable in resource-constrained settings. Data augmentation has been employed to address this issue by increasing dataset variability [69]. However, variability in data quality and collection protocols introduces biases that affect the generalisability of these models.

The interpretability of CNN models also presents a challenge. As ‘black-box’ models, they do not inherently provide explanations for their predictions, limiting their acceptance in clinical workflows. Interpretability remains a key challenge in AI-driven diabetes prediction, as many models function as 'black boxes'. While techniques such as SHAP and Grad-CAM exist, their use in diabetes prediction is still evolving [70]. One-dimensional CNNs are limited in their capacity to model long-term dependencies, and 2D CNNs cannot incorporate temporal dynamics into their analyses [72]. Although 3D CNNs extend capabilities to volumetric data, their computational intensity and exclusion of temporal factors restrict their

broader application. Hybrid architectures provide partial solutions but often lack cohesive frameworks for integrating spatial, temporal, and contextual data [73]. However, if these challenges are addressed, CNNs hold potential as scalable and clinically relevant tools for managing diabetes.

Following this progress, CNNS' limitations in handling temporal dependencies and sequential data have led to the adoption of Recurrent Neural Networks (RNNs). Among these, LSTM networks have become a prominent architecture due to their ability to learn long-term dependencies while addressing the vanishing gradient problem. These networks have been widely utilised for processing sequential medical data, such as blood glucose readings and patient visit records, making them applicable for diabetes prediction [74]. This review examines the methodologies and applications of LSTM-based models in diabetes prediction, identifying their approaches, limitations, and potential areas for improvement.

The application of LSTM in healthcare was explored by Massaro et al., who implemented a three-layer LSTM model to address challenges associated with small datasets [75]. By augmenting the PIDD with artificial data, the dataset size was expanded from 768 to 10,000 samples. The model achieved an AUC of 89% and an accuracy of 84%. Although this study emphasised dataset optimisation, its reliance on a single dataset and the absence of external validation restricted its generalisability.

Alex et al. introduced a four-layer deep LSTM model, addressing class imbalance through the Synthetic Minority Oversampling Technique (SMOTE) [76]. The model achieved an accuracy of 99.64% and an AUC of 0.983. The approach demonstrated high predictive performance; however, reliance on artificially generated data raised concerns about potential biases that could impact its application in more diverse datasets.

Chowdary and Udaya proposed a Conv-LSTM model that integrated convolutional layers with LSTM, leveraging spatial and temporal data features [77]. Using the PIDD dataset, the model achieved an accuracy of 97.26%. The use of feature selection techniques such as the Boruta algorithm enhanced the model's ability to identify relevant predictors. Despite this, the computational cost of the model and its limited validation on real-world datasets highlighted challenges for deployment in resource-constrained environments.

Rochman et al. compared the performance of LSTM and Gated Recurrent Unit (GRU) models on daily patient visit records from a small dataset in Indonesia [78]. Their single-layer GRU model outperformed the LSTM model, achieving an RMSE of 1.722 compared to 3.376. The study underscored GRU's computational efficiency for small datasets, though its limited feature diversity and dataset size restricted broader applicability.

Arora et al. utilised an LSTM model for real-time glucose prediction in Type 1 diabetes (T1D) patients, applying the model to the OhioT1DM dataset [79]. The model achieved an average RMSE of 4.02, outperforming methods such as support vector machines and feed-forward neural networks. While the approach demonstrated the potential for integration into continuous glucose monitoring systems, the small sample size of six patients limited the applicability of the findings.

Iacono et al. introduced a personalised LSTM (P-LSTM) model for glucose prediction in T1D patients [80]. Separate models were trained for individual patients using simulated data from the UVA/Padova simulator. The personalised approach achieved an RMSE of 7.67 mg/dL and a FIT index of 75.86%. Despite its effectiveness in capturing individual variability, reliance on simulated data raised questions about its real-world relevance.

Jaiswal and Gupta implemented a three-layer BiLSTM model for diabetes prediction [81]. By processing data in both forward and backward directions, the model achieved higher precision and recall rates compared to unidirectional LSTMs. However, the computational cost of the model and lack of validation on real-world datasets limited its broader application.

Srinivasu et al. evaluated LSTM models on genomic and tabular data for Type 2 diabetes prediction [82]. The two-layer LSTM model processed both data types, achieving high accuracy metrics. However, the limited dataset size and absence of detailed performance comparisons restricted the study’s conclusions regarding its broader applicability.

Alex et al. proposed a CNN-LSTM hybrid model for diabetes prediction, combining spatial and temporal feature extraction [83]. The hybrid model achieved improved accuracy over standalone LSTM and CNN models. While the integration of CNN and LSTM provided benefits, its computational requirements restricted scalability in practical environments.

Butt et al. explored the integration of LSTM with Internet of Things (IoT) systems for real-time diabetes monitoring [84]. The proposed model achieved an accuracy of 87.26%, outperforming traditional algorithms like moving averages and linear regression. Despite its relevance for IoT-based applications, the study did not implement or validate the system in clinical settings.

These studies illustrate a range of LSTM architectures, from single-layer models to deeper, multi-layer designs. While simpler architectures have demonstrated computational efficiency, more complex models incorporating features such as convolutional layers, bi-directionality, or class balancing have shown improved predictive accuracy Existing studies explore a range of LSTM architectures, from single-layer models to deeper, multi-layer designs. While simpler architectures provide computational efficiency, more complex models—such as those incorporating convolutional layers, bi-directionality, or class balancing—demonstrate enhanced predictive accuracy. This research introduces a novel seven-layer LSTM architecture, specifically designed to address the limitations of shallower models. By increasing model depth, our approach enhances feature extraction, improves generalisation, and provides greater scalability and robustness when handling complex datasets. This contribution represents a significant advancement over conventional LSTM architectures, offering a more effective framework for high-dimensional time-series prediction.

Despite these advancements, challenges remain in the field. Most studies rely on small, homogenous datasets such as PIDD, limiting their generalisability to diverse populations. Computational complexity continues to hinder the deployment of deep LSTM models, particularly in resource-constrained environments. Furthermore, the lack of interpretability mechanisms in many studies restricts their clinical adoption. Addressing these challenges requires integrating explainable AI techniques, expanding dataset diversity, and exploring lightweight architectures to enhance scalability.

LSTM networks, with their ability to model temporal dependencies, offer notable advantages over traditional machine learning methods and CNNs in diabetes prediction. By addressing current gaps and leveraging emerging technologies, these networks can play a pivotal role in advancing diabetes prediction and management, paving the way for more efficient, accurate, and clinically useful solutions.

The integration of CNNs and LSTM networks has significantly advanced the field of diabetes prediction by combining the strengths of both architectures. CNNs excel in extracting spatial features, particularly from structured and unstructured datasets, while LSTMs specialise in capturing temporal dependencies in sequential data. Together, these hybrid models provide a comprehensive framework for addressing the complexities of diabetes prediction. This review critically examines the state of the art in hybrid CNN-LSTM models, focusing on their methodologies, performance, limitations, and potential improvements.

Early studies on hybrid CNN-LSTM architectures demonstrated their effectiveness in medical diagnostics. A foundational study in [86] combined CNN, LSTM, and Support Vector Machine (SVM) for heart rate variability signal classification from ECG data, achieving an accuracy of 95.7%. This study laid the groundwork for subsequent research but was limited by its focus on a single data modality. Building upon this, [87] introduced a fusion of CNN and BiLSTM with attention mechanisms, applied to electronic medical records (EMR). This model achieved an accuracy of 92.78%, a precision of 92.31%, and a recall of 90.46%. The incorporation of attention mechanisms improved the model's ability to prioritise relevant features, enhancing its performance. However, the computational overhead associated with attention mechanisms posed challenges for real-time applications.

A study in [88] applied a CNN-LSTM model to the PIDD, achieving an accuracy of 88.47%, a precision of 94.87%, a recall of 87.78%, and an F1-score of 89.47%. While this work highlighted the strengths of hybrid models in handling structured datasets, it also emphasised the limitations associated with reliance on small, homogeneous datasets like PIDD. Similarly, [89] achieved an accuracy of 95.68% and a precision of 95.21% using PIDD. Although these studies demonstrated the utility of hybrid models, the lack of diverse datasets limits their generalisability to broader populations.

Further advancements were made in [90], where a hybrid CNN-BiLSTM model was developed for real-time clinical settings, achieving an accuracy of 98%, a recall of 97%, and a specificity of 98%. This model highlighted the potential of hybrid architectures in real-time applications but was also limited by its reliance on PIDD. Another study in [91] applied a CNN-LSTM model to continuous glucose monitoring (CGM) data, achieving a mean absolute error (MAE) of 7.5 mg/dL for short-term glucose predictions. These findings

underscored the ability of hybrid models to integrate temporal and spatial data but also highlighted the computational demands associated with such architectures.

The approach in [92] introduced weighted entropy-based feature selection and fuzzy classifiers within a CNN-LSTM framework. This method achieved improved performance metrics such as accuracy and recall, demonstrating the potential of feature optimisation in enhancing model efficiency. The work in [93] developed an ensemble model combining CNN and LSTM for diabetes prediction, achieving an accuracy of 98.6%. This highlighted the benefits of ensemble methods in improving predictive performance, but the lack of standard datasets limited the generalisability of the results.

Emerging trends in hybrid CNN-LSTM research include the incorporation of attention mechanisms and advanced optimisation techniques. The study in [94] utilised SMOTE to address class imbalance, improving sensitivity in diabetes prediction. These methods enhance the robustness of hybrid models but also introduce additional computational complexity, which may limit their scalability in resource-constrained environments.

The frequent reliance on PIDD across studies highlights the need for more diverse datasets that reflect the genetic, demographic, and environmental variability of different populations. The work in [95] proposed the use of federated learning techniques to address this issue by enabling collaborative model training across decentralised datasets while preserving data privacy. Such approaches have the potential to enhance the generalisability of hybrid CNN-LSTM models.

Interpretability remains a challenge for hybrid CNN-LSTM models, particularly in clinical applications. While deep learning models demonstrate high accuracy, their decision-making processes are often complex and difficult to understand. This limitation may impact their adoption in healthcare settings where transparency is essential. Further research is needed to develop models that balance predictive power with clinical interpretability.

The scalability of hybrid CNN-LSTM models is another area of concern. Techniques such as model pruning, quantisation, and distributed training can reduce computational demands and facilitate deployment in real-world clinical settings. The integration of real-time health monitoring data from wearable devices also presents opportunities to enhance the timeliness

and accuracy of predictions, though challenges related to data standardisation and privacy must be addressed.

In conclusion, hybrid CNN-LSTM models have demonstrated considerable potential in advancing diabetes prediction by leveraging their combined strengths in spatial and temporal data analysis. These models have significantly enhanced predictive accuracy and pattern recognition in complex clinical datasets. However, several key challenges persist, including dataset diversity, computational efficiency, and scalability. Addressing these issues through region-specific datasets, robust optimisation techniques, and improved model adaptability will be crucial for enhancing the clinical relevance and real-world applicability of these models.

## 2.3 Chapter Summary

integration of real-world datasets to enhance clinical applicability and patient outcomes.

This chapter has comprehensively reviewed literature on the global and regional prevalence of diabetes, the limitations of traditional diagnostic methodologies, and the transition toward AI-based predictive models. While ML techniques have demonstrated superior accuracy in diabetes prediction, challenges such as dataset diversity, computational complexity, and model transparency persist. Future research should focus on region-specific AI models, optimised feature selection methods, and the integration of real-world datasets to enhance clinical applicability and patient outcomes.

# 3 Development and Evaluation of the Oman Prediabetes Dataset for Type 2 Diabetes Prediction

## 3.1 Chapter Introduction

Chapter 3 explored the development of a prediction model for Type 2 Diabetes Mellitus (T2DM) among prediabetes patients in Oman, aligning with the first two research objectives. The first objective involved evaluating the performance of seven widely used machine learning algorithms: K-nearest Neighbours (K-NN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree, Random Forest (RF), Linear Discriminant Analysis (LDA), and Artificial Neural Network (ANN). The second objective focused on developing a novel, high-quality clinical dataset specifically tailored for prediabetes screening in Oman. This dataset, manually collected for the first time, serves as a crucial resource for diabetes risk prediction within the Omani population. Unlike widely used public datasets such as the Pima Indian Diabetes Dataset (PIDD), which has limited demographic and clinical diversity, the Oman Prediabetes Dataset introduces a region-specific approach, ensuring greater clinical relevance and applicability in the Middle East.

T2DM remains a major public health concern in Oman, contributing significantly to morbidity and mortality rates. Its rising prevalence necessitates early detection strategies supported by robust data-driven methodologies. Traditional diagnostic approaches, although effective, are often limited in accessibility and scalability. Therefore, this study aims to enhance diabetes risk prediction accuracy by leveraging a locally developed dataset and benchmarking its effectiveness against PIDD-based models. The Oman dataset, collected with ethical approval from the Ministry of Health Research Centre in Oman, comprises eleven key clinical features obtained from primary and secondary healthcare facilities. In contrast, the PIDD dataset incorporates only eight clinical variables, making the Oman dataset a more comprehensive tool for regional predictive modelling.

This chapter details the methodological framework for dataset creation, preprocessing, and evaluation. The data pipeline consists of dataset collection from healthcare centres and validation through expert consultation, preprocessing techniques for handling missing values, outlier detection, and normalisation, feature selection methods used to enhance model interpretability and performance, and splitting data into training and testing sets for model benchmarking. To assess the dataset's predictive utility, seven machine learning models—K-

NN, SVM, NB, Decision Tree, Random Forest, LDA, and ANN—are employed as evaluation benchmarks. Their performance is measured using accuracy, sensitivity, specificity, and precision metrics, analysed through a confusion matrix.

A critical gap in existing research lies in dataset quality and preprocessing methodologies, which significantly impact model performance. Many previous studies have overlooked data noise management, outlier treatment, and region-specific feature selection, leading to suboptimal prediction models. By addressing these limitations, this chapter aims to enhance classification accuracy for T2DM predictions in Oman and bridge the gap between global and region-specific diabetes research. Beyond its technical contributions, this chapter introduces a unique, clinically validated dataset that stands as a seminal advancement in diabetes research for the Omani population. This dataset not only facilitates enhanced predictive analytics but also lays the foundation for future deep learning applications in diabetes screening, early prediction, and risk stratification.

## 3.2 Data Collection and Sources

The accuracy and effectiveness of predictive models in healthcare are highly dependent on the quality, completeness, and relevance of the datasets used for training and validation. High-quality datasets provide diverse and representative samples that enhance model generalisability, ensuring accurate predictions across different populations and clinical settings [96]. However, publicly available datasets, such as the Pima Indian Diabetes Dataset (PIDD) [97], have several critical limitations that restrict their applicability in diverse healthcare contexts.

The PIDD dataset is widely used in diabetes prediction research, yet it suffers from limited demographic diversity, a small sample size, and outdated clinical features. The dataset primarily includes data from female patients of Pima Indian heritage, aged 21 years and older, which limits its applicability to broader populations with different genetic, environmental, and lifestyle factors [98]. Additionally, with only 768 records, PIDD provides an insufficient data volume for training robust machine learning models, particularly when developing models for real-world clinical applications that require larger and more diverse datasets [99]. Furthermore, PIDD lacks key clinical features essential for modern diabetes risk prediction, including lipid profiles, waist circumference, and detailed family history of diabetes. The reliance on older

diagnostic criteria further diminishes its relevance in contemporary healthcare applications [100].

To address these challenges and develop more contextually relevant predictive models, this study introduces the Oman Prediabetes Dataset, a region-specific dataset designed to enhance the accuracy and applicability of Type 2 Diabetes Mellitus (T2DM) prediction in Oman. This dataset was developed in adherence to ethical and regulatory standards, with approval obtained from the Ministry of Health Research Centre in Oman, ensuring compliance with privacy and healthcare data protection regulations [102]. The dataset was collected from 21 healthcare facilities, comprising 4 local hospitals, 3 extended health centres, and 14 primary healthcare centres across the South Batinah region. The data were extracted from the Al Shifa electronic health record system, a comprehensive digital database widely used in Oman for healthcare management and patient data documentation [102]. Clinical experts verified the dataset's consistency and accuracy to ensure reliability in predictive modelling applications.

### 3.2.1 Data Collection Process

The Oman Prediabetes Dataset comprises 921 patient records, of which 169 were diagnosed as diabetic and 752 as non-diabetic, based on diagnostic criteria established by the Ministry of Health, Oman [102]. These criteria align with internationally recognised standards, including:

- Fasting plasma glucose (FPG) levels ≥7.0 mmol/L,
- Random blood glucose (RBG) ≥11.1 mmol/L (with clinical symptoms),
- HbA1c ≥6.5%,
- Oral glucose tolerance test (OGTT) with a 2-hour glucose level ≥11.1 mmol/L [104].

A structured screening and diagnostic process was implemented to ensure consistency in data collection across healthcare centres. Figure 3.1 illustrates the multi-step dataset creation process, detailing the integration of multiple data sources such as patient registries, prediabetes screening forms, and electronic medical records. These sources were systematically cross verified to minimize inconsistencies and enhance dataset reliability [105].

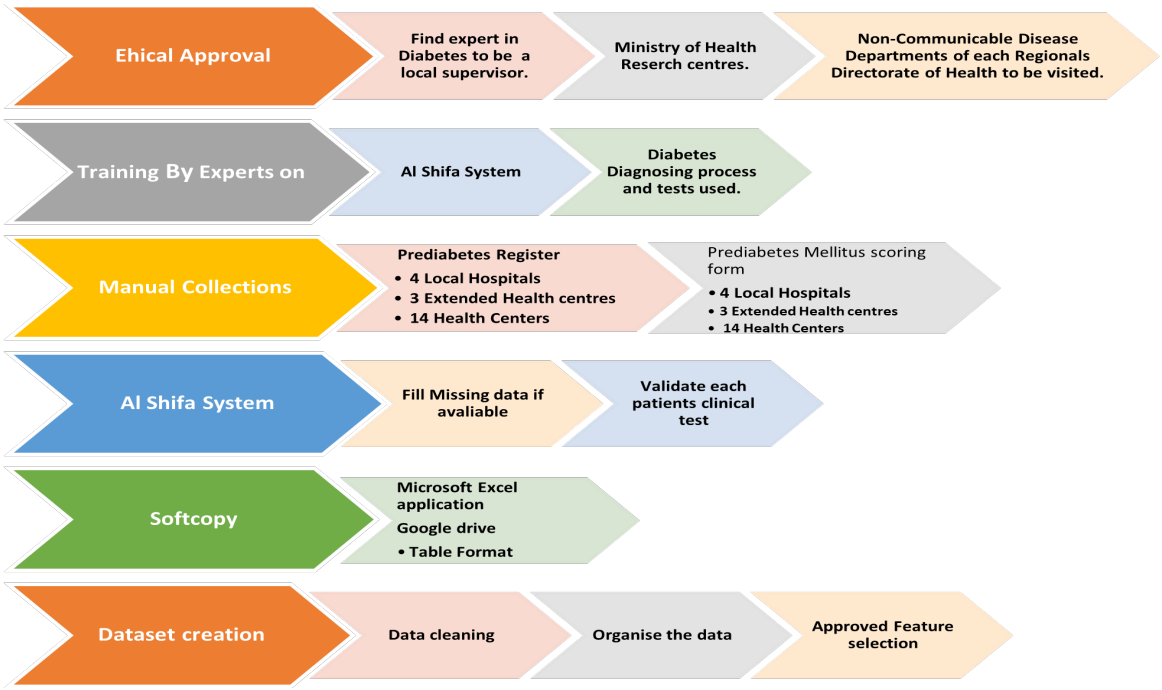



Figure 3.1 Oman prediabetes dataset creation

The dataset was collected from 21 polyclinics and health centres across the South Batinah Governorate, encompassing 4 local hospitals, 3 extended health centres, and 14 primary health centres. During the data collection process, records were extracted manually from prediabetes registers and scoring forms. In the second stage, efforts were directed toward filling in missing variables and verifying data validity. Access to all patient records registered under the South Al Batinah General of Health Services was granted via the Al Shifa System [106].

Each patient's record was analysed individually, cross-referencing hard copy data with the Al Shifa System to address gaps. Missing variables in the hard copy records were supplemented using the patient registry in Al Shifa. However, when information was unavailable in both sources, it was documented as an empty variable. This comprehensive process involved converting patient-by-patient data into an Excel format, transforming hard copy records into structured figures and tables, and verifying entries against Al Shifa records. The entire effort, which spanned six months, was carried out with guidance and support from a physician specializing in diabetes, who operated a prediabetes clinic.

These features were carefully selected based on their clinical relevance to T2DM risk prediction, particularly within the Omani population, where genetic and lifestyle factors differ significantly from the cohort represented in PIDD [107]. By incorporating both male and

female patients across a broader age range, the dataset provides a more representative sample for predictive model training and validation.

Figure 3.2 illustrates the dataset distribution between diabetic and non-diabetic cases. A significant proportion of the dataset consists of non-diabetic cases (752 records, ~82%) compared to diabetic cases (169 records, ~18%). Additionally, Figure 3.3 shows the gender distribution, highlighting a predominance of female patients (70%) compared to male patients (30%) in the dataset. These visualizations enhance the understanding of the demographic and clinical makeup of the dataset, further strengthening its utility for predictive modelling in diverse healthcare settings.

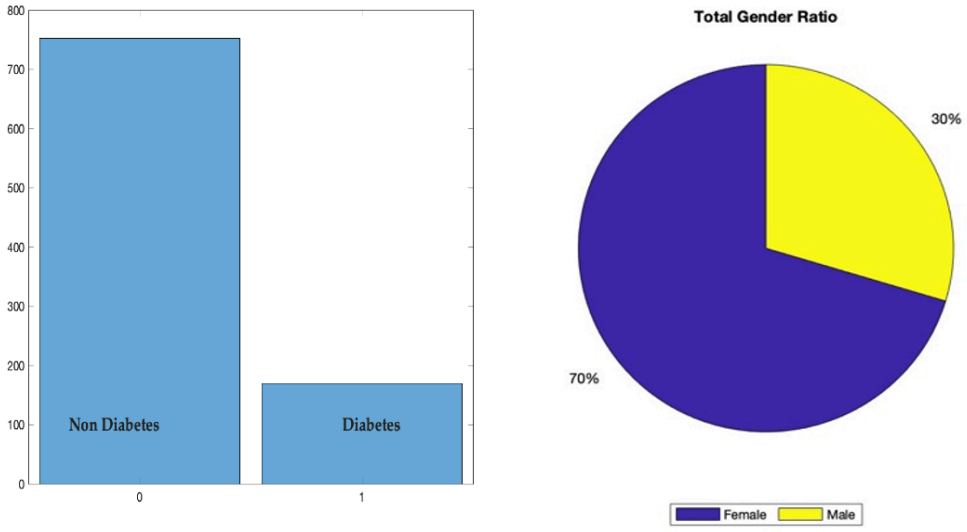



Figure 3.2 Dataset distribution

### 3.2.2 Ethical Considerations and Data Security

Ethical approval for dataset development was granted by the Ministry of Health, Oman, ensuring compliance with data privacy regulations and ethical standards in human subject research [108]. Data were anonymized to protect patient confidentiality, and stringent access controls were implemented within the Al Shifa System to prevent unauthorized use of sensitive health records [109].

Moreover, the Al Shifa System facilitates seamless data integration across healthcare facilities in Oman, enhancing the consistency of data collection and validation. The platform adheres to stringent data protection policies established by the Omani Ministry of Health,

ensuring compliance with national and international standards for healthcare data security [110]. For further details, refer to *Appendix A: Ethical Approval.*

### 3.2.3 Variable Selection and Dataset Features

Table 3.1 provides an overview of the prediabetes register, a survey tool completed by all patients aged 20 and above who visit healthcare centres for regular check-ups. For patients aged 40 years or older, completing the form is mandatory under the Ministry of Health guidelines. The total score from this register determines an individual's diabetes risk level. Patients scoring ≥8 are classified as high risk for Type 2 Diabetes Mellitus (T2DM) and are required to undergo further evaluation within three months by a multidisciplinary team comprising a diabetes specialist, a nutritionist, and a nurse. These patients are often referred to polyclinics or hospitals for additional laboratory investigations. This process ensures early intervention, which is critical to preventing the progression of diabetes.

Table 3.1 Prediabetes register (patient data)

| S/N | Patient Data | Details | Details/Notes |
|---|---|---|---|
| 1 | Risk Factors | First-degree relative with DM<br>Other conditions with insulin resistance*<br>H/O CVD**<br>HTN on therapy or BP >140/90<br>HDL < 0.90 mmol/L or TAG ≥2.82 mmol/L<br>Women with PCOS***<br>Physical inactivity<br>History of GDM**** | *Includes severe obesity or acanthosis nigricans<br>**History of cardiovascular disease<br>***Polycystic ovary syndrome<br>****Gestational diabetes history |
| 2 | Examination | Blood Pressure (BP-R, BP-L)<br>Height (cm)<br>Weight (kg)<br>BMI<br>Waist circumference (cm) | BP: Measured on Right and Left sides<br>BMI: Body mass index calculated from weight and height |
| 3 | Laboratory Tests | Fasting Blood Sugar (FBS)<br>Cholesterol<br>Triglycerides (TAG)<br>LDL<br>Creatinine<br>Estimated GFR (eGFR)<br>HbA1C<br>Oral Glucose Tolerance Test (OGTT) | FBS: 1st reading and repeat if necessary<br>eGFR calculated using MDRD formula<br>Risk score based on WHO/ISH prediction chart |

| | | | |
|---|---|---|---|
| | | Cardiovascular Risk Score | |
| 4 | Problem List | Pre-DM<br>Pre-HTN*<br>Obesity<br>Central obesity<br>Renal impairments<br>Dyslipidaemia | Problems identified during diagnosis and evaluation |
| 5 | Disease Transfer | Diabetes Mellitus (DM) Register<br>Hypertension Register (HTR)<br>Other (Specify) | For patients requiring transfer to specialized registries or additional care |

Table 3.2 outlines the Diabetes Mellitus Scoring Form, which is designed to assess an individual's risk of developing diabetes based on key health and lifestyle factors. The scoring system incorporates demographic variables (such as age and gender), clinical risk factors, and lifestyle behaviours (including physical activity and body weight). Based on the total score, patients are categorised into different risk levels, determining the recommended follow-up frequency:

- Total Score < 5: Annual follow-up is advised.
- Total Score ≥ 5 and < 8: Semi-annual follow-up (every six months) is recommended.
- Total Score ≥ 8: Follow-up every three months, including a comprehensive clinical evaluation by a multidisciplinary healthcare team consisting of a diabetes specialist, nutritionist, and nurse.

Table 3.2 Diabetes Mellitus Scoring Form

| Symbol | Screening Question | Response Options | Score |
|---|---|---|---|
| 1 | How old are you? | <40 years<br>40–49 years<br>50–59 years<br>≥60 years | 0<br>1<br>2<br>3 |
| 2 | Are you a man or a woman? | Woman<br>Man | 0<br>1 |
| 3 | If you are a woman, have you been diagnosed with gestational diabetes? | Yes<br>No | 1<br>0 |
| 4 | Do you have a mother, father, sister, or brother with diabetes? | Yes<br>No | 1<br>0 |
| 5 | What is your blood glucose level currently? | ≥5.6 and <6.1 mmol/L | 0 |

| | | | |
|---|---|---|---|
| | | ≥6.1 and <7.0 mmol/L | 1 |
| 6 | Are you physically active (30 minutes/day, 5 days/week)? | Yes<br>No | 0<br>1 |
| 7 | What is your weight category? | Normal weight<br>Overweight<br>Obese<br>Morbidly obese | 0<br>1<br>2<br>3 |

The Oman Prediabetes Dataset incorporates eleven clinical features, summarised in Table 3.3. These include a combination of categorical and numerical variables selected for their clinical relevance to T2DM risk prediction, particularly within the Omani population. Features like waist circumference, which replaces triceps skin-fold thickness in the PIDD dataset, were added to address specific regional and clinical differences.

Table 3.3 Oman Prediabetes Dataset Features

| Symbol | Feature | Type |
|---|---|---|
| 1 | Gender | Categorical |
| 2 | Age | Numeric |
| 3 | Risk Factor (0–8) | Categorical |
| 4 | Diastolic Blood Pressure (mmHg) | Numeric |
| 5 | Height (m) | Numeric |
| 6 | Weight (kg) | Numeric |
| 7 | Waist Circumference (cm) | Numeric |
| 8 | Total Cholesterol (mmol/L) | Numeric |
| 9 | Fasting Plasma Glucose (mmol/L) | Numeric |
| 10 | HbA1c | Numeric |
| 11 | Outcome | Categorical |

The variables in the Oman Prediabetes Dataset were carefully chosen in consultation with clinical experts and based on Omani diagnostic guidelines for diabetes. A patient is referred to a diabetes clinic if they meet any of the following criteria:

- Fasting Plasma Glucose (FPG) ≥ 7.0 mmol/L (most commonly used in well-being clinics).
- Random Blood Glucose (RBG) ≥ 11.1 mmol/L (rarely used).
- HbA1c ≥ 6.5% in two separate readings within a three-month interval.

The dataset includes variables relevant to T2DM risk factors, such as age, gender, family history of diabetes, hypertension, dyslipidaemia, and insulin resistance. The inclusion of waist circumference instead of triceps skin-fold thickness (found in the Pima Indian Diabetes Dataset (PIDD)) ensures better assessment of obesity-related risks, which is particularly relevant in the Middle Eastern population.

### 3.2.4 Comparison with the Pima Indian Diabetes Dataset (PIDD)

The Oman Prediabetes Dataset incorporates a broader and more comprehensive set of variables compared to the Pima Indian Diabetes Dataset (PIDD), making it a superior resource for predictive modelling in diverse populations. Table 3.4 summarises the differences between the two datasets, highlighting the Oman dataset's enhanced representativeness and clinical relevance.

Table 3.4 Comparison with the Pima Indian Diabetes Dataset (PIDD)

| Feature | Oman Prediabetes Dataset | PIDD Dataset |
|---|---|---|
| Gender | Included (Male/Female) | Not Included (Female Only) |
| Age | Included | Included |
| Risk Factor (0–8) | Included (Family history, lifestyle, and medical conditions) | Not Included |
| Diastolic Blood Pressure (mmHg) | Included | Included |
| Height (m) | Included | Not Included |
| Weight (kg) | Included | Not Included |
| Waist Circumference (cm) | Included (Clinically relevant for abdominal obesity) | Not Included (Uses Skin-Fold Thickness) |
| Total Cholesterol (mmol/L) | Included (Critical biomarker for cardiovascular risk) | Not Included |
| Fasting Plasma Glucose (mmol/L) | Included | Included |
| HbA1c | Included (Aligned with modern diagnostic standards) | Not Included |
| Skin Thickness (mm) | Not Included | Included |
| Insulin (2-h Serum Insulin) | Not Included (Impractical for large-scale screening) | Included |
| BMI (kg/m²) | Included | Included |
| Diabetes Pedigree Function | Not Included | Included |
| Outcome | Included (Diabetic/Non-Diabetic) | Included |

The Oman dataset addresses several limitations of PIDD by introducing broader demographic representation, enhanced clinical features, and modern biomarkers. Unlike the

PIDD dataset, which exclusively includes female patients of Pima Indian heritage aged 21 years or older, the Oman dataset encompasses both genders and spans a wider age range, making it more inclusive and relevant to the Middle Eastern population [111], [112]. This diversity ensures that predictive models developed using the Oman dataset are not biased toward a single demographic, as seen in PIDD.

The Oman dataset integrates advanced clinical markers such as HbA1c and total cholesterol, which are critical for assessing diabetes risk but are absent in PIDD [113], [114]. Furthermore, the replacement of skin-fold thickness with waist circumference offers a more accurate and clinically accepted measure of obesity, particularly relevant in regions like the Middle East, where abdominal obesity is a significant risk factor [115]. The inclusion of a comprehensive risk factor score (0–8), which considers family history of diabetes, physical inactivity, and hypertension, further strengthens the dataset's capability to identify individuals at risk [116]. These features are entirely missing in PIDD, limiting its utility for modern diabetes risk assessment.

While PIDD includes features like skin thickness and diabetes pedigree function, these are either less clinically relevant or redundant in modern diabetes risk prediction models [101]. Features like 2-hour serum insulin, though included in PIDD, are often impractical due to the invasive nature of their measurement, making them less suitable for widespread screening [117].

The significance of these differences is highlighted in Figure 3.2, which provides insights into the demographic composition of the Oman dataset. It shows that 82% of the patients are non-diabetic, while 18% are diabetic. Additionally, the dataset exhibits a gender distribution of 70% female and 30% male patients. These visual insights emphasize the inclusivity and clinical relevance of the Oman dataset, which makes it a superior resource for predictive modelling compared to PIDD.

By addressing the limitations of PIDD, such as its demographic specificity and outdated features, the Oman dataset provides a robust foundation for developing machine learning models tailored to diverse populations. The inclusion of gender diversity, modern biomarkers, and regionally relevant features positions the Oman dataset as a superior tool for advancing diabetes research and prevention strategies [118]. Its alignment with global clinical standards

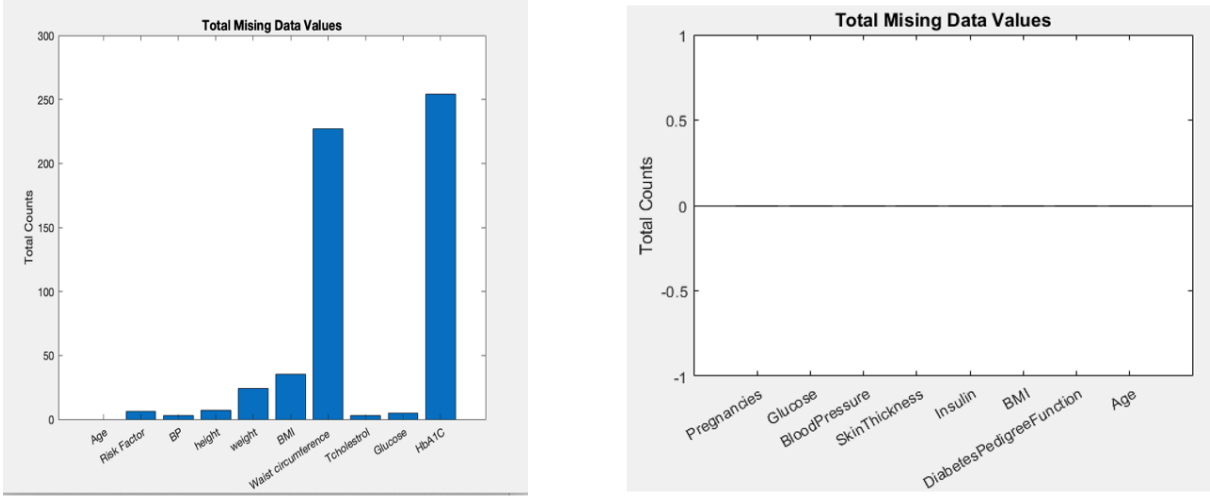
ensures that it meets the demands of modern healthcare applications, making it a valuable asset for improving diabetes prediction and management [119].

## 3.3 Data Processing and Cleaning

The processing data are essential for exploratory statistical analysis and further investigation of the model training phase. The more relevant data are processed, the more it would impact the feature analysis and produce a better predictive result at the time of the training data and testing. The following processes were applied:

**a)** Finding Missing Values from the Dataset

The Oman dataset presented in Figure 3.3a shows that gender has no missing value, but waist circumference and the H1bA1c have more missing values than the other categories. While processing the PID dataset, it was observed that there was no such missing data from the process see Figure 3.3b. Therefore, half of the operations were skipped as it had all the necessary data in the feature.




(a) (b)

Figure 3.3 Total missing values in Oman dataset and Pima Indian dataset. (a) Oman dataset, (b) Pima Indian dataset

The First Step Was for Data to Merge with Similar Categories The data gender value was a merger of two categories {'female'} {'male'} instead of four categories: {'Female'} {'Male'}{'female'},{'male'}. After that, the categorial values were converted to numeric by

using Group to index value, which helps to group absolute values into an index value. For gender, male is 1 and female is 2. The second step was filling in the missing values.

By using the "Ismissing" method [120], data were first analysed by running a check counter, which has missing data, i.e., (", '.',' Na',' NAN'), which are based on empty. The data representing these values are counted, and those particular data are selected in the row missing data, specifying which element of input data contains a missing value and the number of missing values (see Figure 3.4). Then, the "Fillmissing" process [121] with respect to the nearest methods was applied to each feature individually. Therefore, the NAN section is filled with the closest no-missing value.

| Age | RF | BP | hight | weight | BMI | WC | TCholestrol1 | Glucouse | HbA1C |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 1 | 85 | 156 | 88.7 | 36.4 | NaN | 3.93 | 5.9 | NaN |
| 46 | 1 | 80 | NaN | 54.4 | NaN | 103 | 7.5 | 6.7 | 6.1 |
| 48 | 0 | 70 | 143 | 51 | NaN | 93 | 4.5 | 6.4 | 4.4 |
| 33 | 0 | 70 | 162 | 98.3 | 37.4 | NaN | NaN | 6 | 4.9 |
| 34 | 0 | 80 | 158 | 61.6 | 24.6 | NaN | 3.8 | 6.7 | 4.4 |

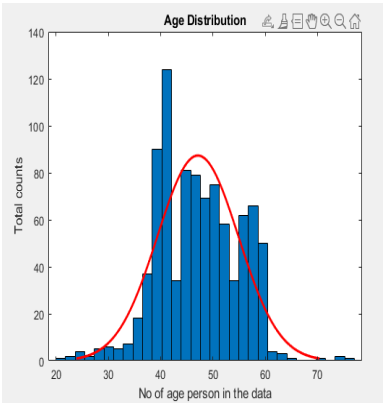
Figure 3.4 Rows with missing values

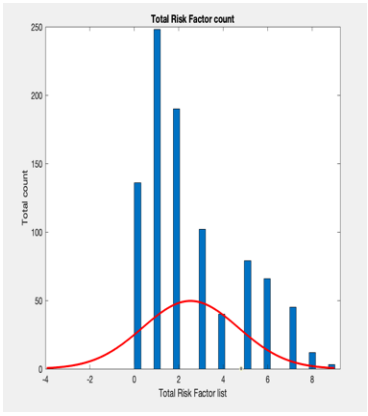
**b)** Exploratory Data Analysis.

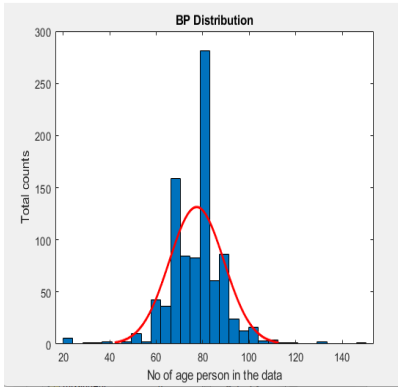
For the statistical operation, the data are evaluated with the individual parameter based on the categorical grouping and providing a statistical result based on the histogram. Histograms are useful for illustrating the distributional characteristics of dataset variables. It is possible to observe where the distribution peaks are, whether the distribution is symmetric or skewed, and whether there are any outliers. Histograms also help to view the possible outliers.

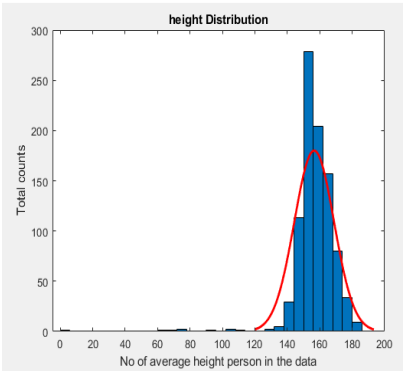
Figures 3.5 and 3.6 show a frequency distribution analysis of both datasets for features to respond for validation sent into a class of diabetic diagnosis system. Each bar covers one set of the range, and the height indicates the number of sizes in each phase range. The field of the problem we are trying to solve requires loads of related features.

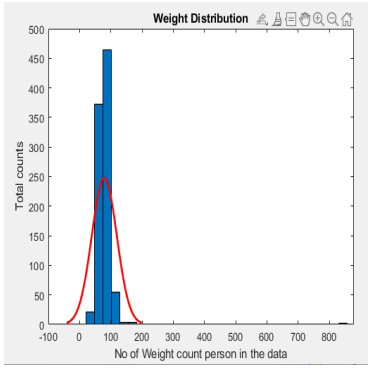
Since the PID dataset is an open and accessible resource, we cannot currently eliminate or generate any more data. In the dataset, we have the following features: ‘Skin Thickness’, ‘Blood Pressure, ‘Insulin’, ‘BMI’, ‘Diabetes Pedigree Function’, ‘Pregnancies’, ‘Glucose’, and ‘Age’. We may infer that ‘Skin Thickness’ is not an indication of T2DM based on a simple observation. Nevertheless, we must acknowledge that it is unusable at this point. Based on Figure 3.5, weight and cholesterol maximum were removed and filled with the nearest methods.

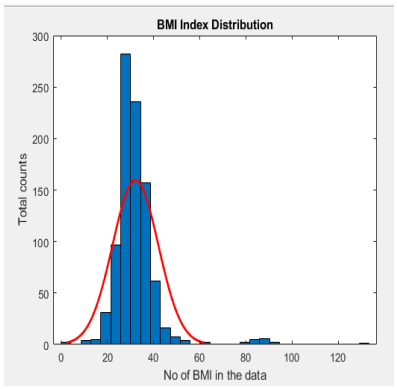


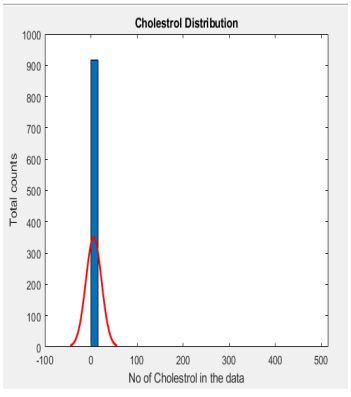



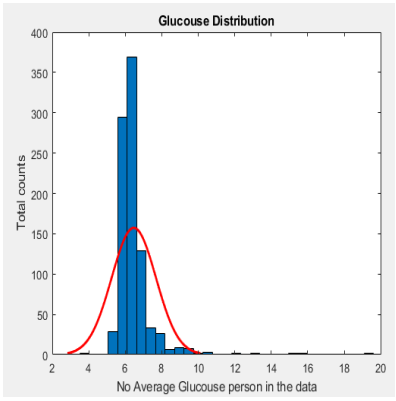



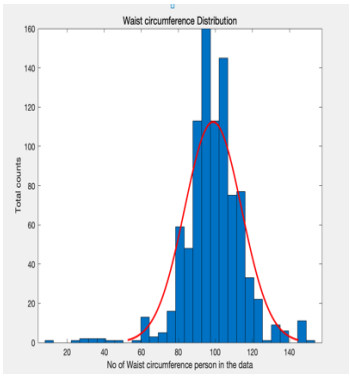



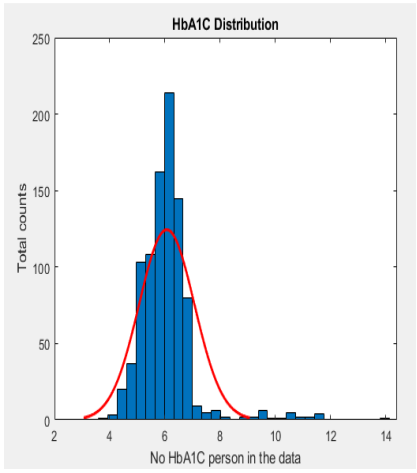
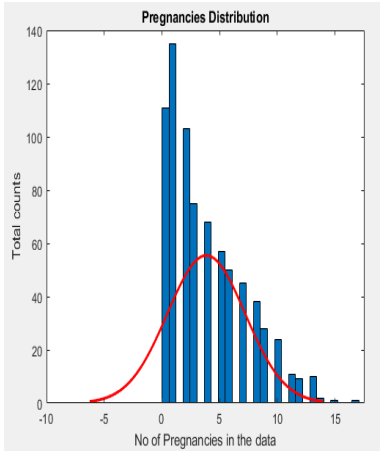
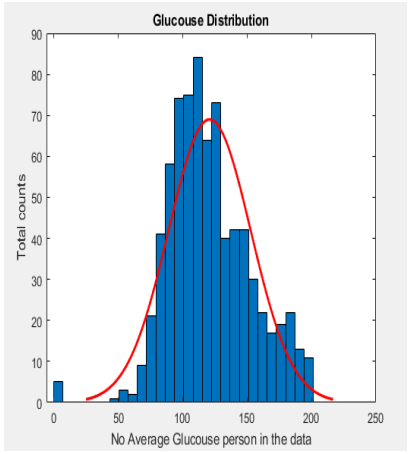
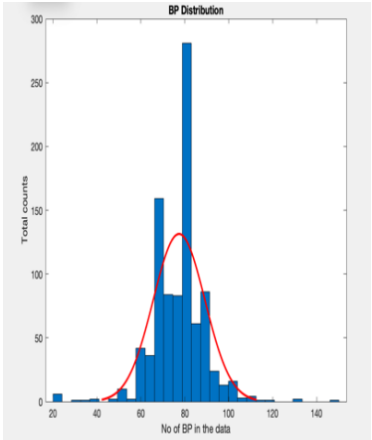
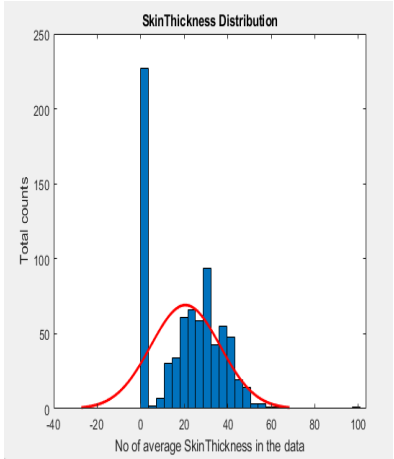


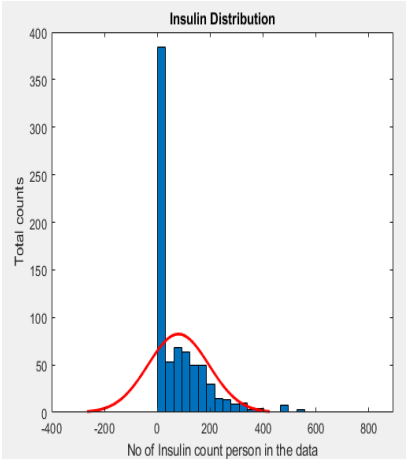

Figure 3.5 Distribution analysis for Oman dataset

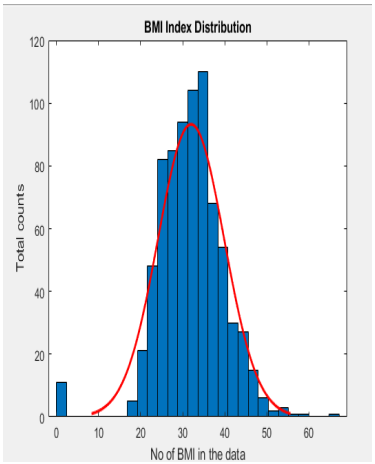


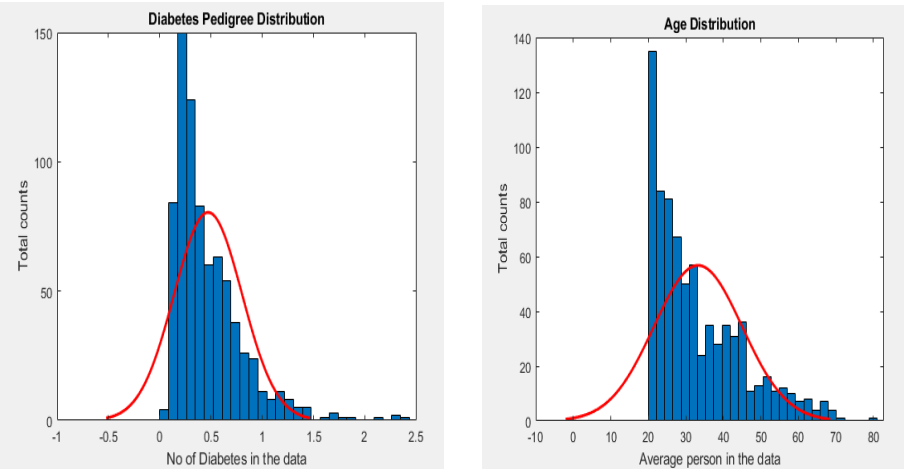
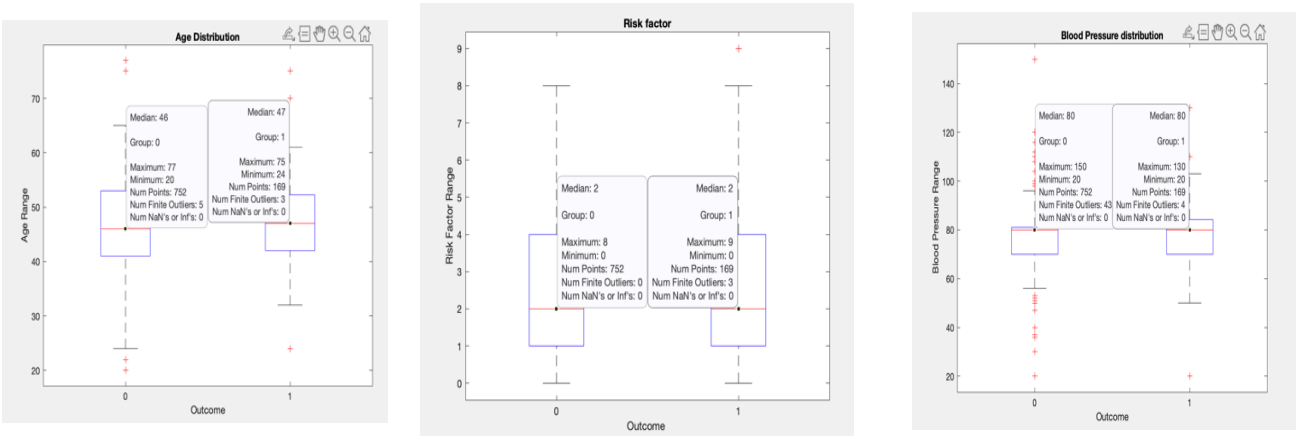



Figure 3.6 Distribution analysis for PID dataset

Another comparison is based on the boxplot. In this, the distribution of each feature is based on the outcome determined from the dataset. A box plot visualises summary statistics for sample data and can easily highlight the outliers for each parameter (see Figure 3.7). The box length signifies the interquartile range, and the whiskers' sizes relative to the box's length indicate how stretched out the rest of the values are. Thus, these aspects of the diagram provide a picture of the dispersion of the dataset. Skewness seems acceptable (<2), and it is also likely that the confidence intervals of the means are not overlapping. Therefore, a hypothesis that glucose is a measure of outcome is expected to be accurate but needs to be statistically tested. Some people have low, and some have high BP. Thus, the association between diabetes (outcome) and BP is suspect and needs to be statistically validated. Like BP, people who do not have diabetes have lower skin thickness. This is a hypothesis that has to be validated. As data of non-diabetic is skewed, diabetic samples seem to be normally distributed.



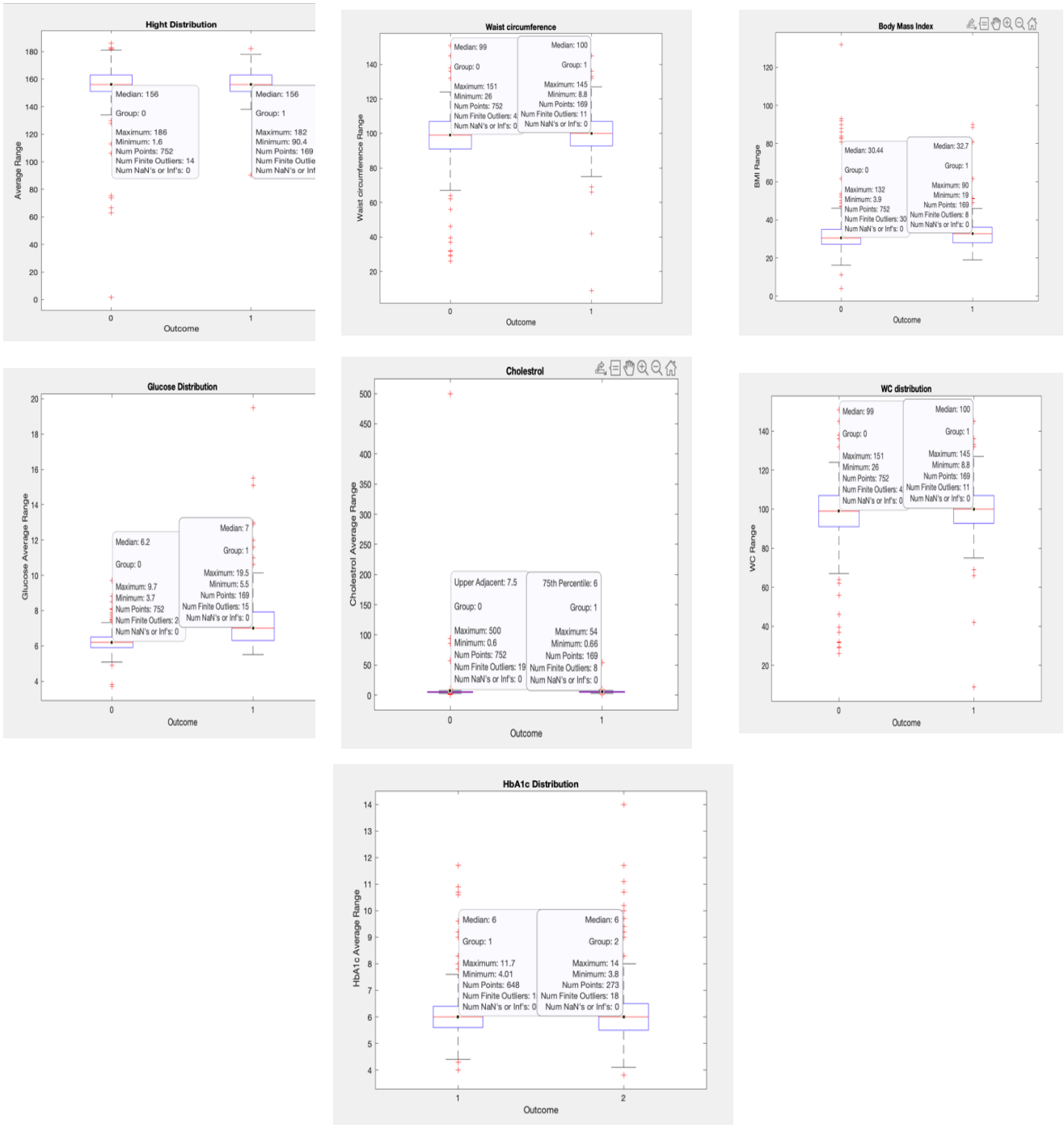
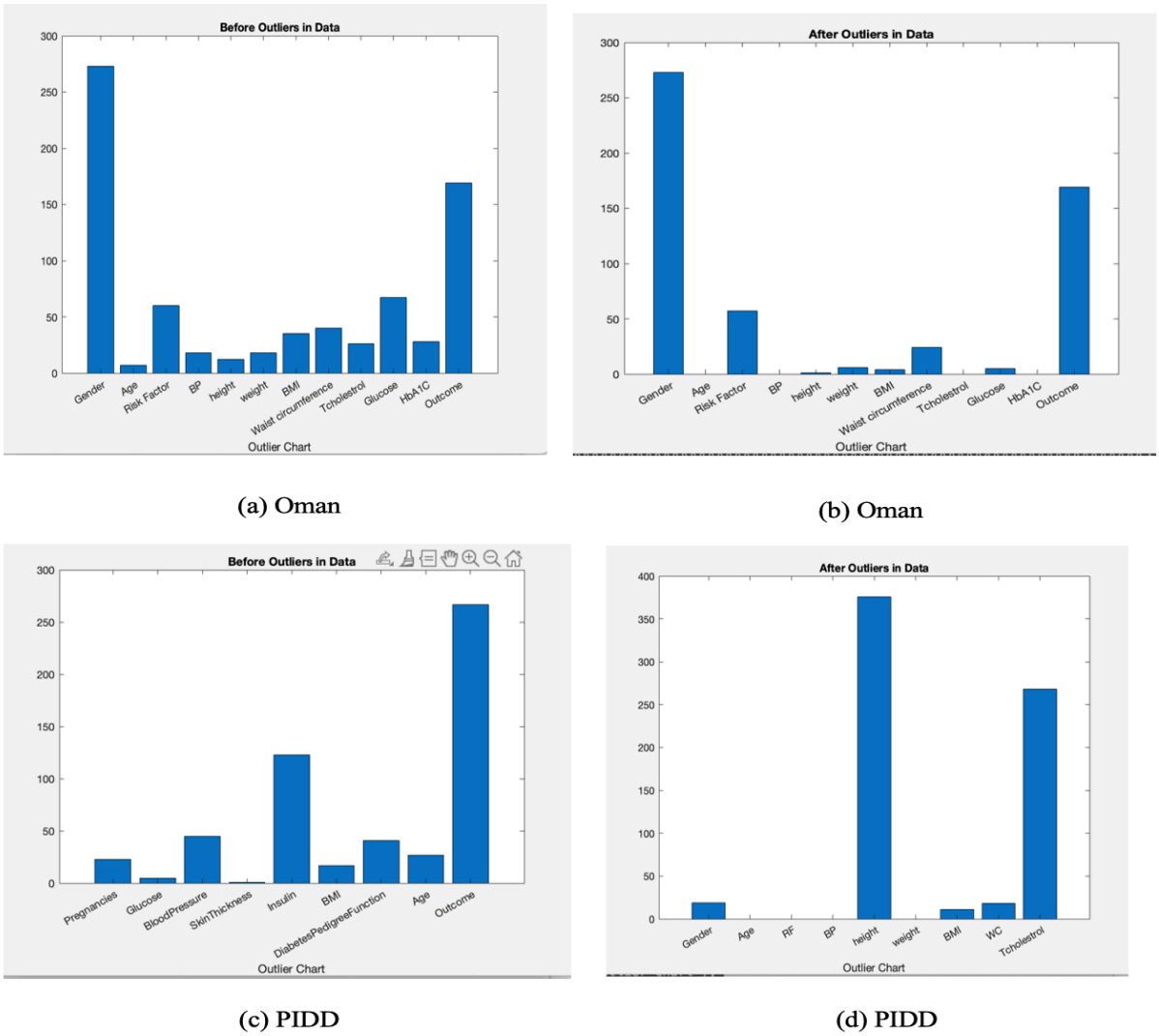



Figure 3.7 Boxplot distribution for the Oman dataset based on the outcome.

**c)** Fill the Outlier in the Data. The outlier is a value that deviates considerably from the dataset's general trend. Box plots are a simple way to visualise data through quantiles and identify outliers. Interquartile range (IQR) is the basic mathematics behind boxplots. The top and bottom whiskers consider the boundaries of data, and any data lying outside are outliers. The length of the box, the interquartile range, and the whiskers' lengths relative to the box's length give an idea of how stretched out the rest of the values are. Thus, these aspects of the diagram give a picture of the dispersion of the dataset. Skewness appears to be acceptable (<2),

and it is also probable that the means' confidence intervals do not overlap. Consequently, it is assumed that the hypothesis that glucose is a measure of outcome is valid, but it must be statistically tested. People might have low or high blood pressure. Therefore, the association between diabetes (outcome) and BP is questionable and requires statistical validation. Like those without hypertension, those without diabetes have thinner skin. This is a theory that must be validated. While non-diabetic data are skewed, diabetes samples appear to have a normal distribution. The outliers were processed using "Filloutlier" with mean and nearest method [122]. The results of outliers before and after removing both dataset's outliers are shown in Figure 3.8.




(a) Oman (b) Oman

(c) PIDD (d) PIDD

Figure 3.8 Outlier processing for both datasets with and without outlier.

The Data Scaling Was Applied for All Machine Learning Algorithms and ANN Using a Z-Score That Centred the Data to a Standard Deviation of 1 and a Mean of 0. The dataset's interquartile range (IQR) describes the content of the middle 50% of values when the values are sorted. If the data median is in Q2, the median of the lower half of the data is in Q1, and

the median of the upper half of the data is in Q3, then IQR = Q3−Q1 [122]. Scaling was applied for all machine learning algorithms and ANN using a z-score that centred the data to a mean of 0 and a standard deviation of 1. The dataset's interquartile range (IQR) describes the content of the middle 50% of values when the values are sorted. If the data median is in Q2, the median of the lower half of the data is in Q1, and the median of the upper half of the data is in Q3, then IQR = Q3 − Q1 [122].

## 3.4 Training and Validation Datasets

Data Splitting: For effective training of datasets, partitioning the data into training and testing sets is a critical step. This division was conducted using the "cvpartition" function [123], following the holdout method, resulting in an allocation of 80% of the data for training purposes and the remaining 20% for testing. This partitioning method was uniformly applied across the datasets, including in the training of the Artificial Neural Network (ANN) model, as detailed in Table 3.5. The K-fold cross-validation technique was also employed, where the dataset is divided into K equal parts, termed "folds". In this technique, each fold is alternately utilized as the testing set, while the remaining folds are combined to form the training set, as illustrated in Figure 3.9. This procedure is repeated K times, with each iteration using a different fold as the testing set. The average testing accuracy from these iterations is calculated to represent the overall testing accuracy of the model [124].

Table 3.5 Splitting the dataset

| Dataset | Total | Percentage |
|---|---|---|
| Training | 737 | 80% |
| Testing | 184 | 20% |

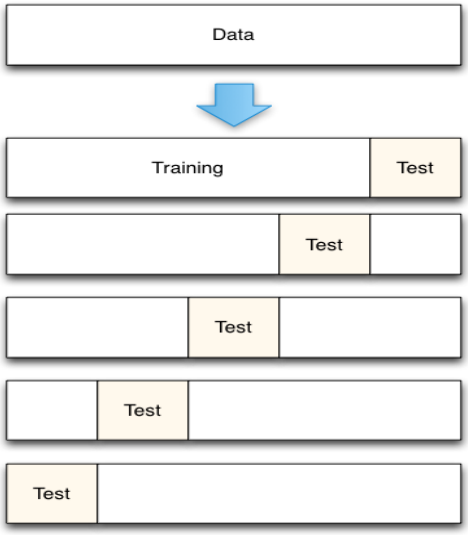



Figure 3.9 K-fold-Cross-validation

Implementation Using Machine Learning Algorithms: The implementation phase utilized MATLAB (version 2021b) software and its command-line coding capabilities to develop a total of seven models using the Oman dataset. These included an ANN and a range of machine learning algorithms: K-nearest Neighbourss, support vector machine, naive Bayes, decision tree, random forest, and linear discriminant analysis. For testing these models, including the ANN, MATLAB's "predict" function was used. A confusion matrix was employed to assess the models' performance, illustrating the correlation between the predicted classes and the actual classes, which were categorised as 0 for non-diabetes and 1 for diabetes. The matrix effectively compares the valid class, representing the actual data, with the predicted class, indicating the prediction accuracy of each algorithm.

The use of the confusion matrix was extended to all models, with further visualization provided in Figures 3.10–3.15. This approach enabled a comprehensive analysis of the prediction accuracy of each algorithm by comparing the predicted and actual classes. The valid class in these figures represents the real data, while the predicted class demonstrates the performance of each algorithm in terms of its accuracy in predictions.

Therefore, the division of the dataset into training and testing sets, complemented by the application of K-fold cross-validation, is crucial for effective model training and validation. The deployment of MATLAB and its functionalities allowed for the development and assessment of a variety of machine learning models. The confusion matrix emerged as a key

tool for quantitatively evaluating the prediction accuracy of these models, facilitating a detailed assessment of their performance in distinguishing between non-diabetes and diabetes cases in the dataset.

## 3.5 Performance evaluation and results

Seven classification algorithms were applied to the datasets, and the results were evaluated based on accuracy, sensitivity, specificity, and precision. Generally, the outcomes were slightly different as each algorithm's working criteria differed. The accuracy of the models was predicted with the help of a confusion matrix, as shown in Figures 3.10–15. The results showed that the random forest and decision tree algorithms had the best classification results.

- The classification models are assessed using the metric of accuracy. Formally, accuracy is the percentage of accurate predictions made by our model. The accuracy is defined as shown below [125] and was measured in terms of positives and negatives:

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

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives

- Sensitivity is a metric that evaluates a model's ability to predict a true positive for each available category. This measure determines the proportion of positive diabetes cases predicted correctly [125]

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

where TP = true positives and FN = false negatives.

- Specificity is the metric that evaluates a model's ability to predict a true negative for each available category; it determines the proportion of actual negative cases predicted correctly [126].

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

where TN = true negatives and FP = false positives.

- Precision is the proportion of true positives to all the positives; it refers to the percentage of relevant results and is a useful metric when false positives are more important than false negatives [126].

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

where TP = true positives and FP = false positives.

By using the equations above, the performance of the various classification models can be compared, as shown in Table 3.6.

Table 3.6 Performance results

| Model | Accuracy (%) | Sensitivity (%) | Specificity (%) | Precision (%) |
|---|---|---|---|---|
| K-nearest neighbours | 92.39 | 94.44 | 77.27 | 90.0 |
| Support vector machine | 96.74 | 98.68 | 87.88 | 83.71 |
| Naive Bayes | 96.74 | 98.1 | 88.46 | 87.08 |
| Decision tree | 98.37 | 100.0 | 92.11 | 80.66 |
| Random forest | 98.37 | 98.01 | 84.85 | 84.1 |
| Linear discriminant analysis | 96.19 | 98.71 | 82.76 | 86.44 |
| Artificial neural networks | 97.3 | 93.33 | 97.96 | 93.9 |

**Accuracy Analysis Using Confusion Matrix:**

a) K-Nearest Neighbours (K-NN) Is an Example of this Type of Supervised ML Algorithm. It is applicable to both classification and regression problems. K-NN classification relies on nearby feature space to classify samples. The K-NN algorithm's default performance is illustrated in Figure 10's confusion matrix. Of the 184 cases tested, the test identified 17 patients and 153 healthy subjects correctly. Therefore, the accuracy of the test was equal to 170 divided by 184 (92.39%).

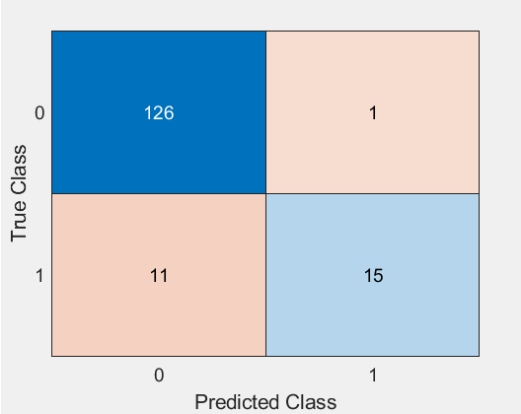



Figure 3.10 K-NN confusion matrix.

b) The Support Vector Machine (SVM) The Support Vector Machine (SVM) Works on the Margin Calculation Concept. It draws margins between the classes. The margins are removed so that the distance between the margin and the types is at a maximum and minimises the classification [127]. As illustrated in Figure 11, of the 184 cases that were tested, the test determined 29 patients and 149 healthy subjects correctly. Therefore, the accuracy of the trial was equal to 96.74%.

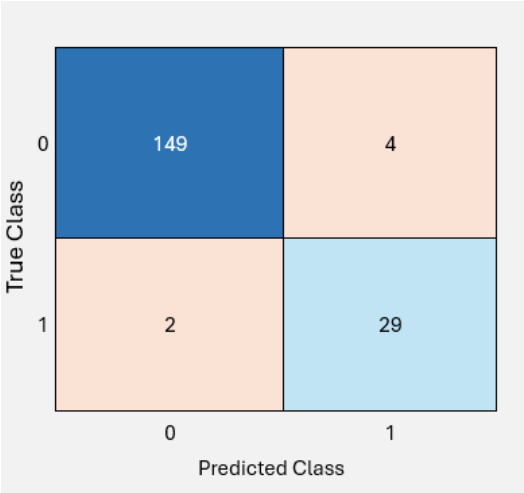



Figure 3.11 SVM confusion matrix

c) Naive Bayes Mainly Targets the Text Classification Industry. It is primarily used for clustering and classification purposes [128]. The underlying architecture of naive Bayes depends on conditional probability. It creates trees based on their likelihood of happening. These trees are also known as Bayesian networks. As shown in Figure 12, of the 184 cases that were tested, the test correctly determined 23 patients and 155 healthy subjects. Therefore, the accuracy of the trial was equal to 96.74%.

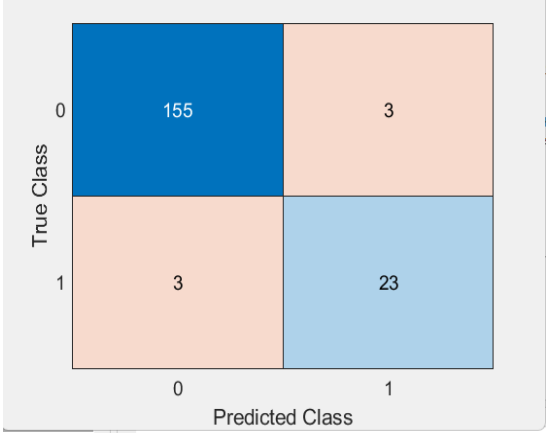



Figure 3.12 NB confusion matrix.

d) Decision Tree (DT) is a Supervised ML Method to Solve Classification, Prediction, and Feature Selection Problems. It aims to predict the target class based on the rules learned from the specified dataset. As a result of the 184 cases shown in Figure 13 that were tested, the test correctly determined 35 patients and 146 healthy subjects. Therefore, the accuracy of the trial was equal to 98.37%.

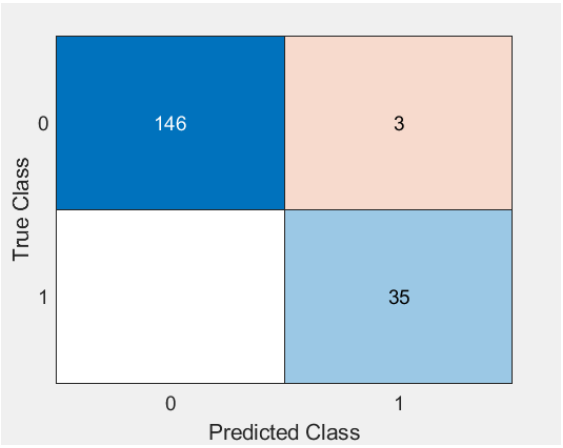



Figure 3.13 DT confusion matrix.

e) Random Forest (RF) is a Supervised Machine Learning Algorithm Used Widely in Classification and Regression Problems. It builds decision trees on different samples and takes their majority vote for classification and their average in case of regression. As presented in Figure 14, of the 184 subjects tested, the test correctly determined 29 patients and 152 healthy cases. Therefore, the accuracy of the test was equal to 181 divided by 184 (98.37%).

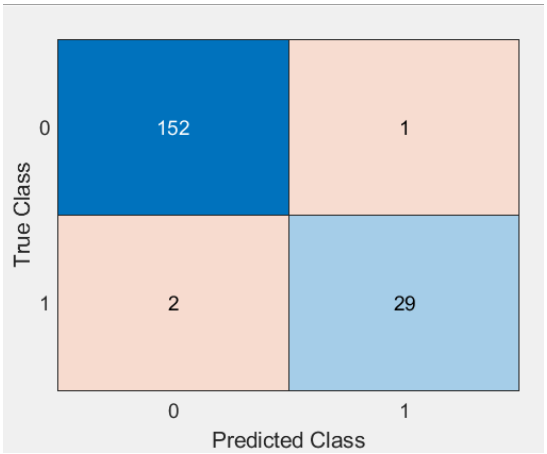



Figure 3.14 RF confusion matrix.

f) Linear Discriminant Analysis Is a Statistical Technique that Can Classify Individuals into Mutually Exclusive and Exhaustive Groups Based on Independent Variables [129]. In this model, as shown in Figure 15, of the 184 cases tested, the test determined 24 patients and 153 healthy subjects correctly. Therefore, the accuracy of the trial was equal to 177 divided by 184 (96.19%).

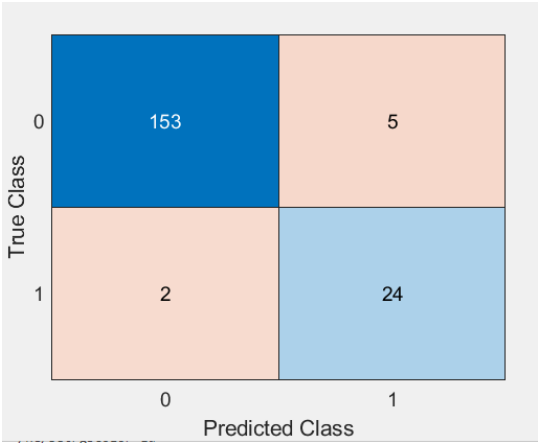



Figure 3.15 LDA confusion matrix.

g) The Conventional Artificial Neural Network (ANN) Consists of Layers and Weights. The behaviour of a network is dependent on communication between its nodes. ANN typically comprises three layers:

- Input layer: Receiving the network's raw data input.

- Hidden layer: The functioning of a hidden layer is defined by the inputs and the weight of the connections between them and the neuron in the hidden layer. These connection weights decide whether a neuron in the hidden layer must be active or inactive.
- Output layer: The operation of this layer is determined by the outputs of the neurons in the hidden layer and the connection weight between these neurons and the neurons in the output layer.

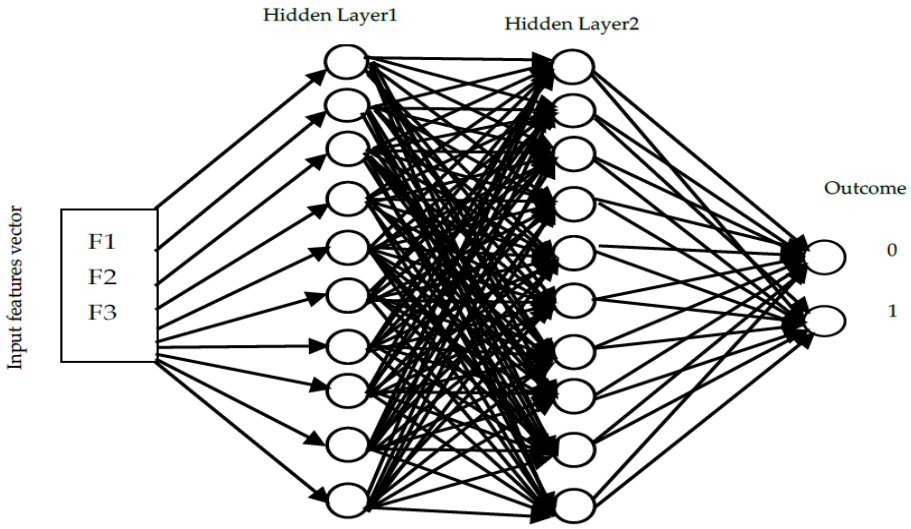
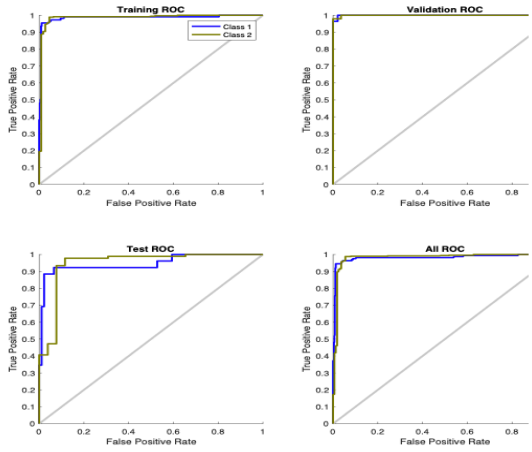



Figure 3.16 ANN supervised architecture proposed.

The proposed structure of an artificial neural network, as shown in Figure 3.16, has an input layer with 11 features; two hidden layers, each with ten neurons; and one output layer with two outputs, diabetes and non-diabetes. A few hidden layers were used to avoid the overfitting problem because the datasets were small. A sigmoid activation function was applied to this model. It used a two-factor level function that set all input values in the values in range from 0 to 1. By using cross-entropy, the model's performance considers the probability in a log of data points [130]. The highest accuracy achieved by this model reached 97.3%, as shown in the confusion matrix in Figure 17b, presenting the training, validation, test, and overall matrix. The accuracy achieved by the dataset's training, validation, and testing was 97.6%, 97.4%, and 95.7%, respectively. The overall combined accuracy was 97.3%. In Figure 17d, the gradient decreased to a performance of 0.047062 and epoch number 49. This decrease means that the model was performing well up to this point, and the increase indicated the start of an overfitting problem. Another evaluation showcases the error histogram in Figure 17c, which has an error

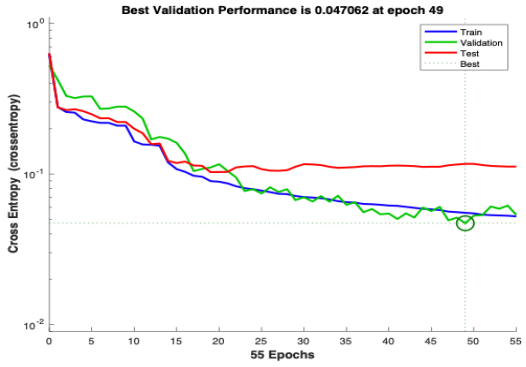
rate with a loss of the range −0.049 value. This describes the quality of the data processor and the target achieved by the evaluation.




(a) Receiver operation characteristic


(b) Confusion matrix


(c) Error histogram




(d) ANN performance

Figure 3.17 ANN results

## 3.6 Discussion

The results of this study are best understood when considered in relation to prior research, particularly those employing the Pima Indian Diabetes (PID) dataset, which is publicly available from the University of California data repository [101]. This dataset has been extensively used in machine learning research as a benchmark for diabetes classification models. The comparative analysis of similar studies is presented in Table 3.6.

Previous research using the PID dataset reports that the highest classification accuracy achieved was 94% with a random forest algorithm and 88% with a decision tree classifier [46].

In contrast, the proposed methodology in this study demonstrates an improved performance, achieving an accuracy of 98.37% using both the random forest and decision tree algorithms when applied to the Oman dataset. One of the primary reasons for this improved classification performance is the nature of the dataset employed in this study, which is significantly larger and incorporates a broader range of diagnostic features compared to the PID dataset. The PID dataset consists of only eight features and 768 cases, whereas the dataset in this study includes eleven additional clinical characteristics, offering a richer representation of patient data.

Furthermore, model optimisation played a crucial role in enhancing performance. Hyperparameter tuning was systematically applied to all algorithms to identify the most effective configurations. For example, in the k-nearest neighbours (KNN) method, the parameter kk was varied between one and five to determine the optimal value. Similarly, in the artificial neural network (ANN) model, a strong correlation was observed between the number of hidden layer neurons and classification accuracy. To achieve the best possible accuracy, the optimal number of neurons was identified through systematic experimentation. These adjustments contributed to the observed improvement in predictive performance.

A comparative evaluation of diabetes classification models is presented in Table 3.7, which contrasts the performance of various models when trained on the PID dataset and the Oman dataset. The results indicate that all models performed better when applied to the Oman dataset, which can be attributed to the inclusion of additional clinically relevant features and the size of the dataset.

Table 3.7 Comparative performance of our proposed method against the state-of-the-art studies on the same dataset

| Model | PID Dataset % | Oman Dataset % |
|---|---|---|
| K-nearest Neighbours (KNN) [52, 53] | 94.5 | 92.39 |
| Support Vector Machine (SVM) [54] | 94.89 | 96.74 |
| Naïve Bayes (NB) [47] | 76.30 | 96.73 |
| Decision Tree (DT) [46] | 94.00 | 98.37 |
| Random Forest (RF) [46, 53] | 94.00 - 98.7 | 98.37 |
| Linear Discriminant Analysis (LDA) [42, 44] | 85.00 - 96.19 | 96.19 |
| Artificial Neural Network (ANN) [ 131, 83] | 96.0-97.26 | 97.3 |

A further investigation was conducted to assess the impact of feature selection on classification accuracy. Table 3.8 presents the performance evaluation of models when trained on two different feature sets: the first feature set, which contained the eight clinical features of the PID dataset, and the second feature set, which included eleven additional features based on the Oman diagnostic method. The results demonstrate that increasing the number of features led to improved classification accuracy across all models.

Table 3.8 Performance evaluation of the proposed method on both datasets.

| Model | PIDD Features | PIDD Accuracy (%) | Oman Features (First Set) | Oman Accuracy (First Set) (%) | Oman Features (Second Set) | Oman Accuracy (Second Set) (%) |
|---|---|---|---|---|---|---|
| K-nearest Neighbours | 8 | 75.1 | 8 | 84.2 | 11 | 92.39 |
| Support vector machine | 8 | 78.4 | 8 | 85.3 | 11 | 96.74 |
| Naive Bayes | 8 | 77.1 | 8 | 87.5 | 11 | 96.74 |
| Decision tree | 8 | 71.89 | 8 | 80.9 | 11 | 98.37 |
| Random forest | 8 | 76.47 | 8 | 85.3 | 11 | 98.37 |
| Linear discriminant analysis | 8 | 77.7 | 8 | 86.95 | 11 | 96.19 |
| Artificial neural networks | 8 | 78.1 | 8 | 86.0 | 11 | 97.3 |

Beyond accuracy, computational efficiency was also assessed. Table 3.9 provides a comparison of training time and prediction speed across different models when applied to the PID and Oman datasets. The results indicate that decision trees and naïve Bayes models exhibit faster prediction speeds, making them more suitable for real-time applications. However, random forest classifiers, despite their higher computational overhead, consistently achieved the highest classification accuracy. Support vector machines, although effective in terms of accuracy, imposed a significant computational cost, particularly when applied to larger datasets.

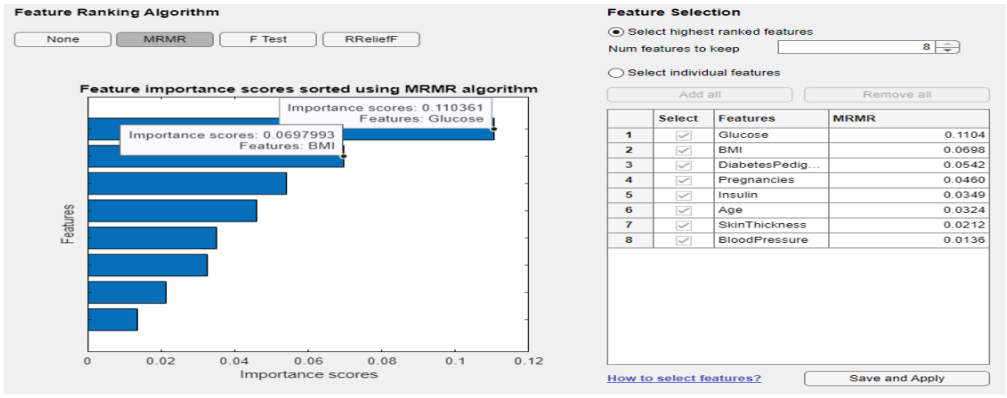
Table 3.9 Comparison of time complexity and models training speed

| Model | PID Prediction Speed | PID Training Time (s) | Oman Prediction Speed | Oman Training Time (s) |
|---|---|---|---|---|
| K-nearest Neighbours | ~24,000 obs/s | 0.53 | ~15,000 obs/s | 0.61 |
| Support vector machine | ~18,000 obs/s | 54.72 | ~19,000 obs/s | 0.54 |
| Naive Bayes | ~26,000 obs/s | 0.65 | ~15,000 obs/s | 0.93 |
| Decision tree | ~58,000 obs/s | 0.44 | ~22,000 obs/s | 1.07 |
| Random forest | ~7,000 obs/s | 1.44 | ~6,500 obs/s | 1.67 |
| Linear discriminant analysis | ~35,000 obs/s | 0.78 | ~17,000 obs/s | 0.93 |
| Artificial neural networks | ~12,000 obs/s | 1.93 | ~12,000 obs/s | 1.93 |

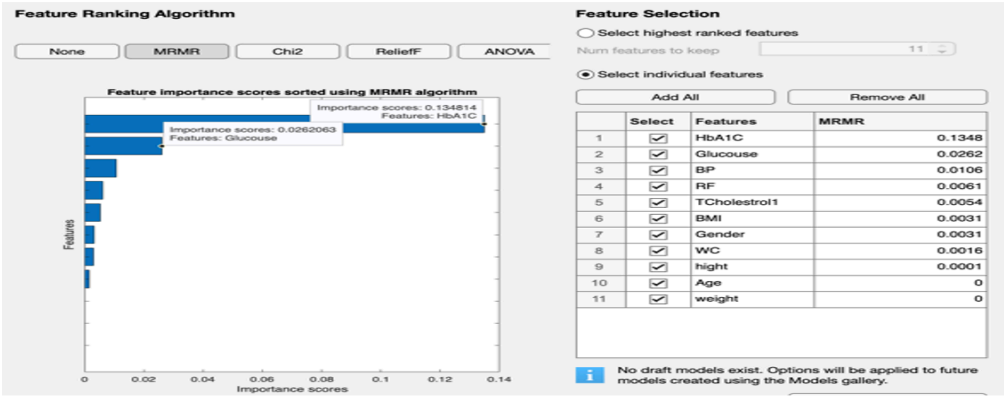
* obs/s: Number of observations processed per second.

While training the model, feature importance analysis was conducted to assess the relevance of individual predictors. The results revealed that HbA1c and glucose levels were the most influential predictors in the Oman dataset, whereas glucose levels alone were the dominant predictor in the PID dataset. The drop in feature importance between the first and second most influential predictors was found to be significant, whereas the decrease after the sixth predictor was relatively minor. This suggests that the software was highly confident in selecting the most critical predictors, while additional features contributed marginally to classification performance. The top five most important predictors were ultimately selected, as depicted in Figure 3.18.

The analysis indicates that feature selection plays a crucial role in model optimisation. The inclusion of clinically relevant features enables machine learning models to achieve higher predictive accuracy, aligning with existing medical knowledge and enhancing the interpretability of the results. The implications of this study suggest that expanding dataset size, incorporating additional diagnostic features, and systematically optimising machine learning models can significantly enhance the predictive accuracy of diabetes classification systems.




a) PID dataset




b) Oman prediabetes dataset

Figure 3.18 Comparative Performance of the Proposed Method in Both Datasets

These findings contribute to the broader discussion on the application of machine learning in clinical decision support systems, reinforcing the potential of data-driven methodologies to improve diagnostic accuracy and patient outcomes.

## 3.7 Chapter Summary

This chapter focuses on the development and validation of a region-specific dataset, the Oman Prediabetes Dataset, and the Oman Screening Dataset, both of which were designed to reflect Oman's demographic and clinical characteristics. The chapter details the methodological process of creating these datasets, including data collection, preprocessing, and validation. A key objective was to ensure high data quality through robust preprocessing techniques, such as normalisation, handling of missing values, and outlier detection, to maintain a data completeness and accuracy rate of ≥95%. These datasets address limitations observed in commonly used datasets such as the Pima Indian Diabetes Dataset (PIDD) by

incorporating a broader range of clinical and demographic variables that are more representative of the Omani population.

A crucial aspect of this study involved benchmarking traditional machine learning models, including Random Forest, Support Vector Machine (SVM), and Naïve Bayes, to establish baseline performance metrics. The effectiveness of these models was evaluated using key performance indicators such as sensitivity, specificity, accuracy, and F1-score, with an aim to achieve a baseline accuracy of ≥80% across both datasets. Comparative analysis demonstrated that all models performed better on the Oman dataset than on PIDD, with Random Forest and Decision Tree achieving the highest classification accuracy of 98.37%.

The study highlights the advantages of expanding the feature set and optimising model parameters, which significantly contributed to improved classification accuracy. It was observed that the inclusion of additional clinically relevant features in the Oman dataset enhanced predictive performance across all models. Furthermore, computational efficiency was examined, revealing trade-offs between model complexity, training time, and prediction speed. Decision trees and Naïve Bayes demonstrated faster prediction speeds, making them more suitable for real-time diabetes screening applications, whereas ensemble models like Random Forest, despite being computationally intensive, provided higher classification accuracy.

Feature importance analysis revealed that HbA1c and glucose levels were the most influential predictors in the Oman dataset, while glucose remained the dominant predictor in the PID dataset. The results underscore the importance of clinically relevant feature selection, ensuring that predictive models retain maximal accuracy while reducing computational complexity.

This chapter establishes a strong foundation for diabetes prediction models in the Omani population by integrating region-specific clinical data and machine learning techniques. It paves the way for future advancements, including the exploration of deep learning architectures such as Convolutional Neural Networks (CNNs) to further enhance predictive accuracy. The subsequent chapter will introduce a new dataset for diabetes screening in Oman and propose a novel CNN model architecture to achieve the research objectives of developing a high-performance diagnostic system for diabetes prediction in Oman.

# 4 1D CNN for Structured Data Model and Oman Screening Dataset

## 4.1 Chapter Introduction

This chapter presents an in-depth exploration of the 1D Convolutional Neural Network (1D CNN for Structured Data) model, a deep learning framework designed to enhance early detection of Type 2 Diabetes Mellitus (T2DM). The study introduces the development, implementation, and evaluation of the model, focusing on its architecture, feature extraction capabilities, and classification performance when applied to structured medical data. The model is trained and validated using the Oman Screening Dataset, a region-specific dataset that ensures clinical relevance and improved generalisability for AI-driven diabetes prediction.

The chapter also provides a comparative analysis of the 1D CNN for Structured Data model against conventional machine learning models such as Random Forest, Decision Trees, and Support Vector Machines. The evaluation highlights the advantages of deep learning in structured data classification, particularly in automated feature extraction, improved classification accuracy, and better representation of hierarchical feature relationships. Unlike traditional machine learning models that require manual feature selection, the CNN approach autonomously identifies patterns and interactions between clinical parameters, reducing bias and enhancing predictive accuracy.

The significance of applying CNNs to structured medical datasets lies in their ability to capture complex dependencies between clinical indicators. Traditional diagnostic methods rely on predefined statistical models, which may not fully represent the intricate relationships between risk factors such as BMI, blood pressure, cholesterol, and glucose levels. The CNN-based approach leverages convolutional operations to recognise both low-level and high-level patterns in structured data, allowing for improved disease risk assessment and prediction. This method offers a more scalable and generalisable solution for clinical applications, extending beyond diabetes prediction to other areas such as cardiovascular disease risk assessment, metabolic syndrome analysis, and personalized treatment recommendations.

To ensure that the 1D CNN model effectively learns from structured data, extensive preprocessing of the Oman Screening Dataset is performed. This includes normalisation, outlier detection using Z-score analysis, categorical encoding, and feature selection techniques. The dataset, developed through a rigorous validation and screening process, provides a comprehensive representation of the Omani population, enhancing the model's ability to detect region-specific risk factors. By applying deep learning to structured medical records, the study

aims to bridge the gap between traditional statistical models and AI-driven predictive analytics, paving the way for more accurate and reliable clinical decision-making.

The chapter is structured to provide a comprehensive analysis of the model's architecture, training methodology, and performance evaluation. The discussion begins with an overview of the CNN model's layer-wise structure, explaining its convolutional layers, activation functions, fully connected layers, and classification mechanisms. This is followed by an examination of the training and validation process, where the dataset is partitioned into training, validation, and testing subsets to optimise model performance. The CNN's classification accuracy, sensitivity, specificity, and F1-score are analysed to assess its effectiveness in diabetes prediction, demonstrating significant improvements over traditional machine learning classifiers.

This research contributes to advancing AI-driven medical diagnostics by demonstrating the potential of deep learning in structured clinical data analysis. By automating feature extraction and learning hierarchical representations, CNN models provide a scalable and efficient solution for disease prediction. The insights gained from this study emphasize the transformative impact of AI in healthcare, particularly in preventive screening and risk assessment. The findings set the foundation for future research into hybrid deep learning architectures that integrate time-series modelling, multi-modal data fusion, and real-time health monitoring for enhanced predictive performance.

## 4.2 The Proposed 1D CNN for Structured Data Model

The 1D Convolutional Neural Network (1D CNN) for Structured Data Model represents a novel approach to diabetes prediction using structured clinical data. Unlike conventional diagnostic methods, which rely primarily on biochemical tests and statistical risk assessments, the proposed deep learning model automates feature extraction and enhances predictive accuracy through hierarchical learning.

This AI-driven approach is particularly valuable for early diabetes risk assessment, as it can capture complex feature relationships in medical datasets without requiring manual intervention. Traditional diagnostic approaches, although effective, often fail to identify early-stage diabetes, making deep learning-based solutions a crucial advancement in preventive healthcare.

Unlike conventional machine learning (ML) models, such as Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVM), which depend on manually selected features and predefined relationships, the 1D CNN for Structured Data autonomously extracts

meaningful patterns from structured clinical data. This approach eliminates the risk of human bias in feature selection, allowing for improved generalisation across diverse patient populations. Moreover, CNNs capture spatial-temporal dependencies in structured datasets, whereas traditional ML models treat features as independent variables, limiting their ability to model complex clinical relationships. The ability of CNNs to perform hierarchical representation learning enhances their ability to detect underlying correlations between diabetes risk factors, leading to superior predictive accuracy [62].

The Oman Screening Dataset, used in this study, provides a region-specific dataset optimised for AI-driven clinical applications. It contains demographic, anthropometric, and clinical health indicators unique to the Omani population, offering a highly relevant and precise dataset for deep learning-based diabetes prediction. Compared to globally used datasets like the Pima Indian Diabetes Dataset (PIDD) [68], this region-specific dataset enhances the model's ability to detect localized risk factors that may differ from global trends.

### 4.2.1 1D CNN for Structured Data Model Architecture

The 1D CNN for Structured Data Model is a deep learning-based classification framework designed to analyse structured medical data and predict diabetes risk with high accuracy. Unlike traditional machine learning models, which require manual feature selection, the 1D CNN autonomously extracts hierarchical feature representations, improving classification performance and scalability. The model's architecture follows a structured, multi-layered design, enabling the efficient processing of medical records and enhancing predictive capabilities by capturing complex dependencies between clinical parameters [62].

Diabetes prediction is a challenging task due to the multiple interdependent risk factors involved. Conventional approaches rely on biochemical tests and manual risk factor assessments, which often fail to detect early-stage diabetes. By integrating deep learning techniques, the 1D CNN for Structured Data Model enables early detection through automatic pattern recognition within structured clinical datasets. The model learns multi-dimensional feature relationships, ensuring improved sensitivity and specificity in differentiating between diabetic and non-diabetic cases [65].

Unlike 2D CNNs, which are commonly used for image-based tasks, the 1D CNN for Structured Data Model is optimised for structured numerical inputs, making it suitable for patient health records. It efficiently processes structured datasets by analysing relationships along a single feature axis, rather than focusing on spatial correlations. The model architecture consists of multiple layers, each contributing to the hierarchical learning process. The

following sections provide a detailed layer-wise breakdown of the model, an explanation of its predictive workflow, and a discussion of its architectural representation.

### 4.2.2 Justification for Selecting 1D CNN for Structured Data for Diabetes Prediction

The CNNs provide a significant advantage in structured data classification due to their ability to autonomously extract hierarchical feature representations, reducing reliance on manually engineered features. Traditional ML models, such as Support Vector Machines (SVM) and Decision Trees (DT), rely on human expertise to select key variables, which may introduce bias and limit the model's ability to detect complex feature interactions.

In contrast, CNNs automatically learn relationships between multiple clinical indicators, enabling the detection of hidden dependencies within structured medical datasets. This is particularly important in diabetes prediction, where risk factors such as BMI, blood pressure, glucose levels, and cholesterol interact non-linearly over time. Feature engineering methods struggle to capture such dependencies effectively, whereas CNNs excel in learning both low-level statistical relationships and high-level patterns that contribute to disease progression.

Additionally, CNNs leverage convolutional operations to recognise spatial-temporal dependencies in health records, a feature that traditional ML approaches overlook. Since structured data can be modelled as a time-series or multi-dimensional input, CNNs can process it in a manner that preserves its hierarchical structure, leading to more accurate and clinically relevant predictions.

### 4.2.3 Layer-wise Breakdown of the 1D CNN for Structured Data Model

The 1D CNN for Structured Data Model comprises several key layers, each playing a distinct role in feature extraction, transformation, and classification. These layers work together to ensure that the model effectively learns and refines feature representations from structured medical data. The hierarchical nature of the network allows it to progressively extract relevant patterns, leading to an accurate and interpretable classification.

The input layer serves as the entry point for patient data. At this stage, raw clinical records are transformed into structured tensors, ensuring that features such as BMI, blood pressure,

cholesterol, glucose levels, and family history are presented in a standardized format. The normalisation process ensures that all medical parameters are scaled consistently, preventing biases arising from variations in measurement units [69].

Following the input layer, the convolutional layers serve as the primary mechanism for feature extraction. These layers apply 1×1 convolutional filters to detect statistical relationships between different clinical indicators. Unlike standard CNNs, which analyse spatial patterns in image data, the 1D CNN model applies convolutions along structured patient records, capturing complex feature dependencies. The early convolutional layers extract basic feature relationships, such as correlations between glucose levels and blood pressure, whereas deeper layers identify more intricate patterns, such as the combined effect of HbA1c trends and BMI fluctuations over time [65].

The Rectified Linear Unit (ReLU) activation function is applied after each convolutional layer to introduce non-linearity into the model. This function helps mitigate vanishing gradient problems, ensuring that deep networks can learn from multiple hierarchical transformations. By selectively activating meaningful features and suppressing irrelevant ones, ReLU enhances the network's ability to focus on critical patterns associated with diabetes risk [145].

The fully connected layers aggregate the feature representations learned through the convolutional layers. These layers act as the final stage of feature abstraction, compressing the extracted information into a structured form suitable for classification. The dimensionality of the feature space is progressively reduced, with the final fully connected layer containing two neurons, corresponding to the diabetic and non-diabetic classes. This step ensures that only the most relevant features contribute to the classification decision, improving generalisation and reducing noise [62].

To generate interpretable classification probabilities, the model employs a softmax layer. This layer transforms the final feature vector into probability distributions, assigning a confidence score to each classification category. The softmax transformation ensures that the sum of output probabilities equals one, enabling clear decision-making.

The classification layer serves as the final decision point, assigning a diagnostic label (diabetic or non-diabetic) based on the computed probability scores. This layer allows the

model to be integrated into clinical workflows, providing healthcare professionals with an AI-powered decision-support tool that enhances early intervention and risk assessment [67].

### 4.2.4 How the 1D CNN for Structured Data Model Predicts Diabetes

The 1D CNN for Structured Data Model follows a systematic predictive workflow, ensuring that the model extracts, refines, and classifies relevant features efficiently. This predictive process consists of four primary stages: data input and preprocessing, feature extraction, prediction and classification, and clinical decision support.

The first stage, data input and preprocessing, involves structuring patient records into numerical arrays. Clinical variables such as age, BMI, cholesterol, and glucose levels are encoded into a structured tensor format. At this stage, data normalisation techniques are applied to prevent discrepancies between different clinical measurements, ensuring that no single feature dominates the learning process [119].

Once the data is structured, it is passed through convolutional layers, where hierarchical feature extraction takes place. The first few layers detect low-level statistical relationships, such as how fasting glucose and insulin levels interact. As the data moves through deeper layers, the model learns more abstract patterns, including long-term metabolic trends and their impact on diabetes risk [100].

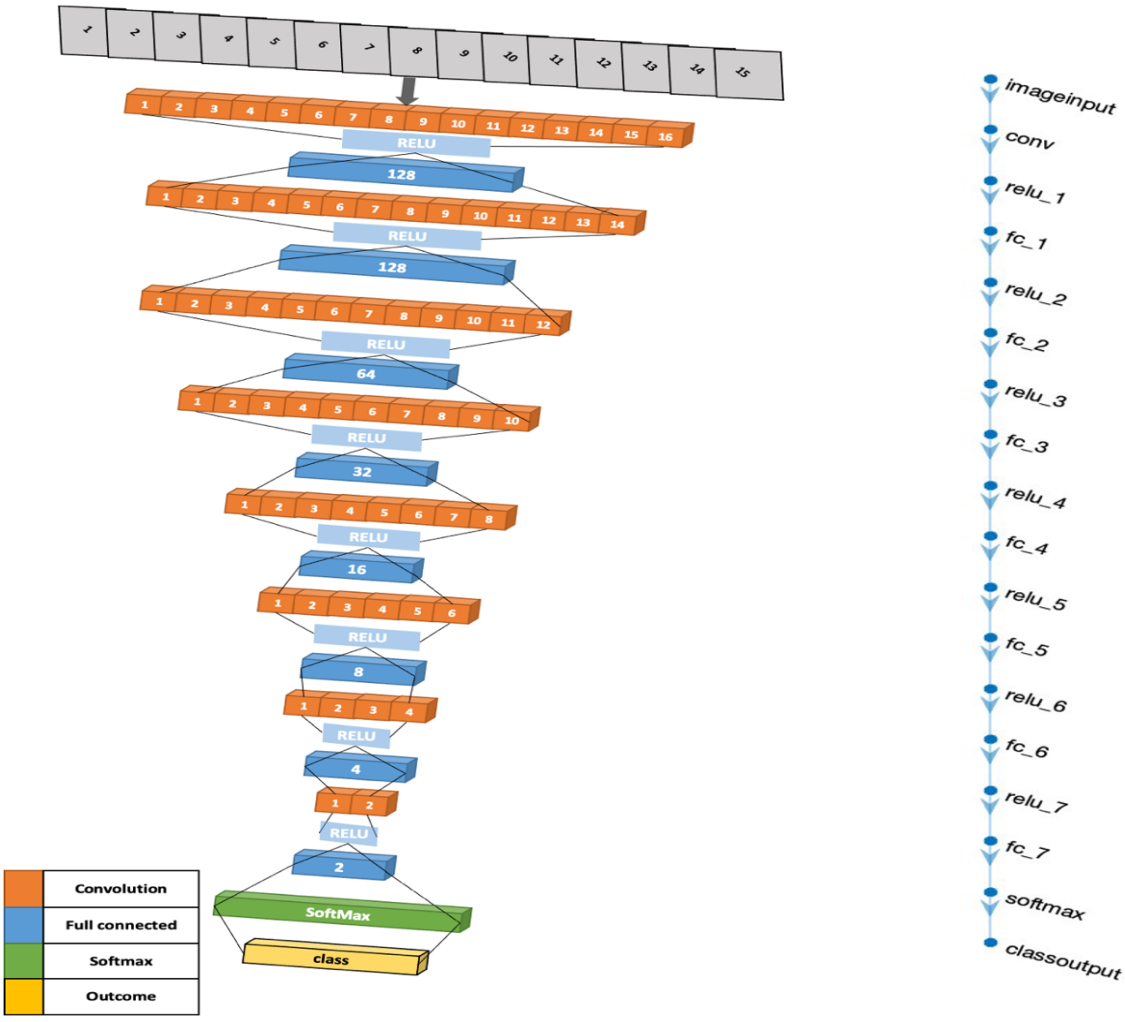
After feature extraction, the fully connected layers aggregate the extracted representations, reducing dimensionality while preserving essential diagnostic information. The softmax layer then transforms these refined features into probability scores, quantifying the model's confidence in the classification decision. If the probability of the diabetic category exceeds a predefined threshold, the model classifies the patient as diabetic; otherwise, the classification remains non-diabetic [126].

Finally, the classification output is integrated into clinical decision-making, allowing healthcare providers to utilize AI-based risk assessment tools for personalized treatment planning and early intervention strategies. By automating feature selection and learning multi-dimensional relationships between health indicators, the 1D CNN model complements traditional diagnostic techniques with a data-driven approach to risk prediction [70]

## 4.3 In-Depth Illustration of the 1D CNN for Structured Data Architecture

### 4.3.1 Overview of the 1D CNN for Structured Data Model Architecture

The 1D CNN for Structured Data Model, illustrated in Figure 4.1, provides a detailed representation of how structured clinical data is processed through multiple layers to produce a final classification decision. The architecture is designed to perform hierarchical feature extraction, progressively refining patient data through successive transformations. This structured approach enables the model to capture complex relationships between clinical risk factors and enhance classification accuracy. Each layer of the model serves a distinct role in feature extraction, refinement, and classification, ultimately leading to a more reliable and interpretable diagnosis.




(a) Conceptual Representation of 1D CNN (b) MATLAB visuals

Figure 4.1 In-depth Illustration of the 1D CNN for Structured Data Architecture

At the top of the model, the input layer receives structured medical data, which includes demographic, anthropometric, and biochemical parameters such as BMI, cholesterol levels, blood pressure, fasting plasma glucose (FPG), and family history of diabetes. To ensure the consistency of numerical values, standardization and normalisation techniques are applied before the data enters the CNN pipeline. These preprocessing steps mitigate the risk of bias introduced by variations in measurement units, thereby allowing the model to systematically process multi-dimensional health indicators.

As the input data passes through the convolutional layers, it undergoes multiple stages of feature extraction. The initial convolutional layers capture low-level correlations, such as fluctuations in glucose levels or interactions between BMI and cholesterol. As the depth of convolutional processing increases, the network learns progressively more complex feature dependencies, detecting subtle patterns that contribute to diabetes risk. This ability to extract and refine hierarchical feature relationships distinguishes CNNs from traditional ML models, which often treat each feature independently and fail to capture such intricate associations.

Between convolutional layers, ReLU (Rectified Linear Unit) activation functions introduce non-linearity, ensuring that significant patterns are highlighted while irrelevant information is filtered out. Without activation functions, the model would behave as a linear transformation, limiting its ability to distinguish complex interactions between risk factors. The inclusion of ReLU ensures that non-linear relationships in health conditions, such as glucose fluctuations influenced by multiple factors, are properly represented.

Following convolutional feature extraction, the fully connected layers further refine the learned representations by reducing feature dimensionality. These layers consolidate extracted patterns, ensuring that only the most relevant information is retained for the classification task. Unlike traditional ML models that require manual feature selection, CNNs autonomously identify and prioritize diagnostic-relevant features, eliminating human bias and improving predictive performance.

At the final stage, the softmax layer computes probability distributions across classification categories (diabetic and non-diabetic). This step transforms the final feature vector into interpretable probability scores, ensuring that the total classification probabilities sum to one. This probability-based approach is critical for clinical decision-making, as it enables physicians to assess the model's confidence levels in its predictions. The classification layer,

positioned at the bottom of the architecture, assigns the final diagnostic label based on the highest predicted probability, supporting AI-driven risk assessment and early intervention strategies.

### 4.3.2 Structural Analysis of the 1D CNN for Structured Data Model Architecture

The hierarchical structure of the CNN model follows a funnel-like arrangement, ensuring that raw structured data is incrementally refined through multiple layers before reaching the final classification stage. Figure 4.1 visually depicts this process, illustrating how patient health records undergo a structured transformation until a definitive diagnosis is produced.

1. Input Representation and Preprocessing: The input layer, positioned at the top of the model, represents patient records formatted as structured numerical arrays. This layer ensures that all medical indicators—BMI, blood pressure, cholesterol levels, and glucose readings—are appropriately structured and scaled before being processed by the CNN. Feature preprocessing techniques, such as normalisation and outlier removal using Z-score methods, are applied to minimize data inconsistencies and biases before the input enters the convolutional layers.
2. Convolutional Feature Extraction: The convolutional layers perform progressive feature extraction, identifying diagnostic patterns in structured clinical data. The first convolutional layers detect low-level interactions, such as variations in glucose levels influenced by dietary intake or medication use. As the data moves through deeper convolutional layers, the model extracts more abstract and high-level dependencies, learning how multiple health indicators interact. The depth of convolutional processing allows the network to develop a more comprehensive understanding of diabetes risk factors, enabling it to surpass traditional ML models in prediction accuracy.
3. Hierarchical Representation Learning: The deeper convolutional layers progressively refine extracted features, filtering out irrelevant noise and enhancing clinically meaningful patterns. The non-linear activation layers (ReLU) play a key role in preserving essential feature relationships, ensuring that the model retains its ability to distinguish between diabetic and non-diabetic patients. Without non-linear activations, complex feature dependencies would not be effectively captured, reducing classification performance.
4. Dimensionality Reduction and Feature Aggregation: As extracted features advance through fully connected layers; their dimensionality is progressively reduced.

This compression process ensures that only the most relevant diagnostic features contribute to the final classification decision. Unlike conventional classification methods that depend on hand-selected features, the 1D CNN autonomously determines the most predictive variables, eliminating human intervention and minimizing potential biases.

5. Probability Computation and Final Classification: At the final stage, the softmax layer converts processed feature vectors into probability distributions, enabling the model to assign confidence scores to classification outcomes. The classification layer then selects the most probable diagnosis (diabetic or non-diabetic), ensuring that the AI-driven decision is clinically interpretable and applicable in real-world settings.

The 1D CNN for Structured Data Model, as depicted in Figure 4.1, follows a top-down refinement process, where input data is progressively transformed through multiple stages before reaching a final classification outcome. This structured transformation allows the CNN to learn complex relationships while ensuring that only the most relevant diagnostic patterns are retained.

1. At the top of the diagram, the structured input layer receives multiple clinical indicators, such as glucose, BMI, and cholesterol levels, which represent the raw health profile of the patient.
2. In the middle section of the network, convolutional layers extract diagnostic patterns, refining statistical relationships and hierarchical feature dependencies in patient data.
3. At the bottom of the model, the fully connected layers condense extracted features, the softmax layer computes classification probabilities, and the classification layer assigns a final diagnostic label (diabetic or non-diabetic).

The funnel-like architecture of the CNN model ensures that each stage progressively extracts more meaningful patterns, making the 1D CNN highly effective for structured medical data analysis

The architectural design of the 1D CNN for Structured Data Model offers several advantages compared to traditional machine learning methods:

- Automated Feature Learning: Unlike conventional models requiring manual feature selection, the 1D CNN autonomously extracts diagnostic patterns, enhancing classification performance.
- Hierarchical Representation Learning: The multi-layered structure enables the model to identify both statistical trends and high-level feature dependencies, making it well-suited for structured medical data analysis.
- Scalability and Generalisation: The model generalises well across diverse datasets, allowing it to be applied to various patient populations with high reliability.
- Non-Linear Feature Interactions: The combination of convolutional layers and activation functions enables the model to capture non-trivial relationships between multiple clinical indicators, improving diagnostic accuracy.
- High Sensitivity and Specificity: The softmax-based classification layer provides probabilistic confidence scores, minimizing false positives and false negatives, which is crucial for diabetes risk screening.

## 4.4 Dataset Overview and Preprocessing

The methodology followed in this research is a systematic sequence of events designed to predict diabetes using Convolutional Neural Networks (CNN). A specific dataset from Oman has been utilized to train, validate, and test the model. The methodology includes steps such as loading and pre-processing the dataset and designing a custom 1D CNN for Structured Data architecture.

### 4.4.1 Oman screening dataset

The dataset utilized in this study was systematically compiled, validated, and prepared using diabetes-related health records from Oman, following strict ethical guidelines **[131]**. The process of diabetes screening and data collection workflow is illustrated in **Figure 4.2**, detailing the systematic approach employed in assembling a high-quality dataset for predictive modelling.

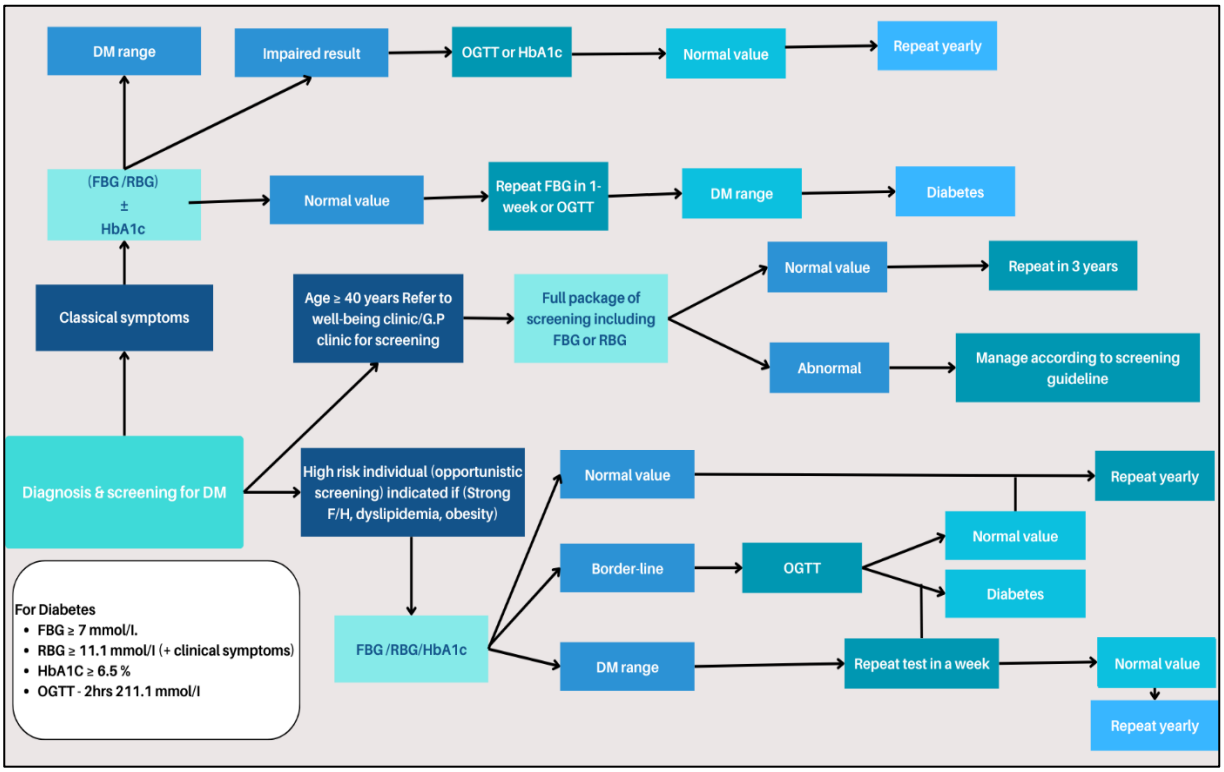



Figure 4.2 Oman Diabetes screening system workflow

#### 4.4.1.1 Data Collection Process

The data collection process was conducted in collaboration with local diabetes specialists, ensuring compliance with regulatory requirements and ethical approvals from the Ministry of Health, regional health departments, and participating Regional Directorates of Health (see Appendix A for ethical approval details). The dataset was derived from 41 healthcare institutions, comprising 34 primary healthcare centres, three secondary care Extended Health Centres, and four local hospitals. This extensive data acquisition framework enabled the study to capture a diverse range of patient demographics and clinical characteristics, enhancing the dataset's representativeness for diabetes risk assessment [68,132].

The Oman Screening Dataset was developed as part of an initiative to improve early diabetes detection. The dataset collection spanned seven months, during which standardized procedures were implemented to maintain data integrity and completeness. It consists of demographic, anthropometric, and clinical markers, which serve as key variables for diabetes screening and prediction. The inclusion of individuals aged 20 years and above extends the model's ability

to detect early-stage diabetes risk factors, addressing gaps in conventional screenings that typically focus on individuals over 40 years of age.

To ensure dataset validity and clinical relevance, individuals with pre-existing diabetes diagnoses or those screened within the past three years were excluded. This exclusion criterion prevented redundancy and maintained the dataset's focus on previously unscreened individuals, improving the reliability of the predictive model.

The seven-month data collection period ensured that the dataset captured recent health trends and risk factors associated with diabetes in Oman. The final dataset comprises 13,224 patient records, covering 13 essential variables required for diabetes risk stratification.

#### 4.4.1.2 Inclusion and Exclusion Criteria:

A structured inclusion and exclusion framework was implemented to ensure that the dataset was representative of individuals at risk of diabetes while minimizing potential confounding variables.

- Inclusion Criteria
    a. Individuals aged 20 years and above.
    b. No prior diagnosis of diabetes.
    c. No diabetes screening conducted within the past three years.
- Exclusion Criteria
    a. Individuals with pre-existing conditions that could interfere with diabetes screening (as outlined in Table 4.1).
    b. Patients who had undergone diabetes screening within the past three years at other healthcare facilities.

The implementation of these criteria ensured that the dataset maintained clinical accuracy and relevance, allowing for a focused analysis of diabetes risk factors within the Omani population [74,68].

Table 4.1 Diabetes Screening Eligibility Criteria

| Section | Category/Sub-Category | Details or Criteria |
|---|---|---|
| Eligibility for Screening | Diseases Present | If "Yes" to D.M, HTN, CKD: Not eligible for screening. |
| | Screened in Last 3 Years | If "Yes" to screening at any other health centre in the last 3 years: Not eligible for screening. |
| Outcome of Screening | - | If "Yes" to any of the above criteria: Excluded from screening. If "No" to both criteria: Proceed to screening. |
| Family and Personal History | Family History | Obesity, Hypertension, Dyslipidaemia, DM, CKD, Premature Cardiovascular Death (M: < 55, F: <65) |
| | Personal History | Physical inactivity, Ethanol, Tobacco (Cigarettes, Sheesha, Non-smoked tobacco), Nephrotoxic Drugs (NSAIDs, Analgesics, Diuretics, Antibiotics, Herbal) |
| Reason for Referral to GP | 1. Lifestyle Risk Factors | Physical inactivity, smoking, ethanol |
| | 2. Obesity Metrics | BMI ≥ 25 Kg/m² and/or Waist Circumference (M: ≥ 94cm, F: ≥ 80cm) |
| | 3. Blood Pressure | Mean B.P. > 130 mmHg systolic and/or Mean B.P. ≥ 85 mmHg diastolic |
| | 4. Impaired Blood Sugar | FPG (5.6 to < 7.0 mmol/l) or RPG (5.5 to < 11.1 mmol/l) |
| | 5. Diabetes Diagnosis | FPG ≥ 7.0 mmol/l or RPG ≥ 11.1 mmol/l |
| | 6. Cholesterol Level | Serum Cholesterol > 5.2 mmol/l |

#### 4.4.1.3 Data Validation Process

To enhance data accuracy, the dataset was validated using the Al Shifa System, a widely adopted healthcare information system in Oman [132]. This system, utilized across over 200 healthcare institutions, facilitated the verification of patient records and clinical data [133]. By integrating electronic health records (EHRs) with manually collected data, inconsistencies

were identified and resolved, ensuring data completeness [134,135]. Each patient's clinical profile, laboratory results, and medical history were cross-referenced against electronic records to confirm accuracy. This validation approach minimized errors in data entry and ensured dataset consistency, improving its suitability for predictive modelling.

#### 4.4.1.4 Dataset Composition and Feature Selection

The final dataset consisted of 13,224 patient records, incorporating 13 key variables essential for diabetes risk assessment. The data was structured and formatted using MATLAB (Version 2023b) for efficient preprocessing and analysis. Feature selection was based on clinical guidelines from Oman's Ministry of Health, with oversight from expert diabetes physicians to ensure that selected variables were clinically significant [136]. The chosen features encompass demographic, anthropometric, and biochemical indicators that contribute to diabetes onset. Table 4.2 presents a summary of these variables.

Table 4.2 Diabetes Feature Descriptions

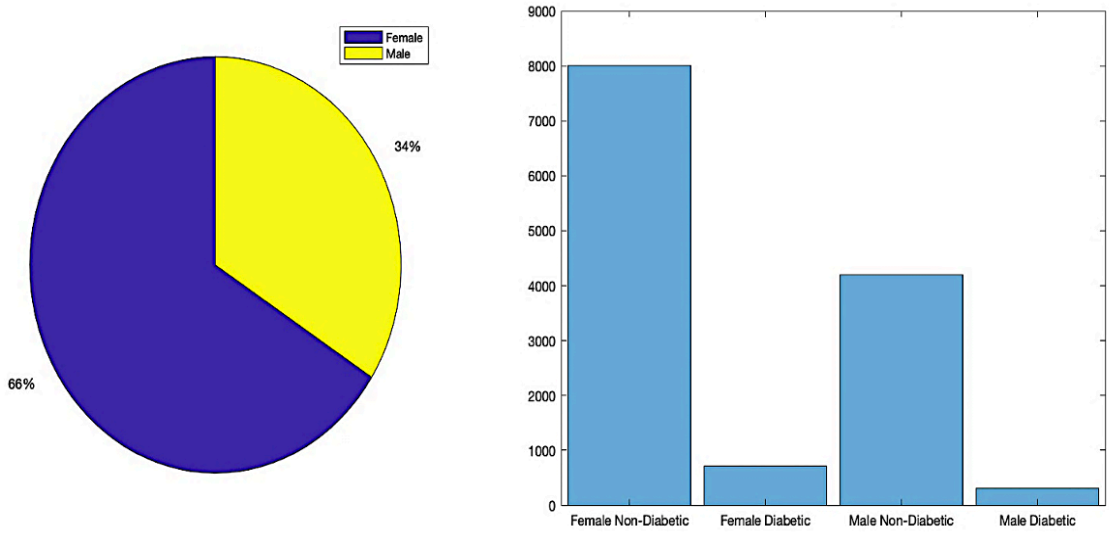
| Feature | Description | Data Type |
|---|---|---|
| Age | Age of the patient (20–65 years) | Double |
| Weight | Weight of the patient | Double |
| Height | Height of the patient | Double |
| BMI | Body Mass Index | Double |
| WC | Waist Circumference | Double |
| T_Cholesterol | Total Cholesterol | Double |
| BP | Blood Pressure | Double |
| RPG | Random Plasma Glucose | Double |
| FPG | Fasting Plasma Glucose | Double |
| FH | Family History of Diabetes | Double |
| PH | Personal History of Diabetes | Double |
| Gender Encoded | Encoded Gender of the patient | Double |
| Outcome | Diabetic or not | Double |

#### 4.4.1.5 Dataset Utilization and Analysis

To facilitate analysis, categorical data were transformed into numerical representations to ensure compatibility with deep learning models. Data preprocessing included handling missing

values, detecting outliers using Z-score analysis, and standardizing feature distributions to optimise model performance.

Figure 4.3 presents the gender-based distribution of the dataset, providing insights into diabetes prevalence across different population segments.




(a) Total gender ratio

(b) Diabetes status based on gender

Figure 4.3 Dataset Distribution by Gender

The dataset, developed through a rigorous process of validation and screening, is comprehensive and reliable, ensuring its suitability for the study's objectives. The careful selection of participants, adherence to well-defined inclusion and exclusion criteria, and implementation of a structured data validation framework enhance its accuracy and clinical relevance. This approach not only strengthens the validity of the current research on diabetes in Oman but also provides a methodological reference for future studies in similar clinical and epidemiological contexts. Additionally, by including individuals aged 20 and above, the dataset offers a representative overview of diabetes risk factors within the studied population, contributing to a deeper understanding of the disease and informing future preventive and clinical strategies [131,136].

#### 4.4.1.6 Exploratory Data Analysis (EDA)

Visualizing data is paramount in exploratory data analysis. It gives insights into data distribution, relationships between variables, and any potential anomalies. Below, we delve into different visualization techniques applied to the dataset.

a) Statistical Summary: A statistical summary provides an insight into the key characteristics of each variable in the dataset. This summary encompasses range, central tendencies (like median), and any potential missing values. The dataset under examination, as summarised in Figures 4.3 and 4.4, offers a comprehensive collection of health metrics.


| | Age | Weight | Height | BMI | WC(Waist Circumference) | T_Cholesterol | BP (Blood Pressure) | RPG (Random Plasma Glucose) | FPG (Fasting Plasma Glucose) | FH (Family History) | PH (Personal History) | Gender Encoded | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Median | 42 | 74 | 158.11 | 29 | 96 | 5.01 | 80 | 5.47 | 5.5 | 2 | 1 | 0 | 0 |
| Max | 113 | 186 | 199 | 60 | 193 | 9.9 | 199 | 7.26 | 20 | 10 | 10 | 1 | 1 |
| Min | 4 | 0 | 0 | 0 | 0 | 1.1 | 2 | 3.8 | 1 | 1 | 1 | 0 | 0 |

Figure 4.4 Statistical summary.

This spans from general health indicators like age (with a range from 4 to 113 years and a median of 43) and weight (ranging between 0 and 186 with a median at 74) to BMI, which has a median of 29, albeit with 137 missing values. Further diving into specialized health markers, we have measurements like random plasma glucose, which interestingly has 3793 missing data points, and a median value of 5.47.

Waist circumference and total cholesterol also contribute to the dataset's breadth, with respective medians of 95.354 and 5.01. Furthermore, the dataset comprises data on blood

pressure, with values spanning from 2 to 199 and a median of 80. However, it is essential to note that 12 values in this variable are missing.

The dataset also integrates personal and family medical histories, each with its own set of missing data (84 and 102 missing values, respectively), suggesting that some patients might not have disclosed or had access to this information. In terms of gender distribution, the dataset employs an encoding mechanism, with 0 representing males and 1 representing females. Finally, the 'Outcome' variable, presumably indicating the result or diagnosis, categorises data into either 0 or 1, though the specifics of these categories were not provided in the summary [137].

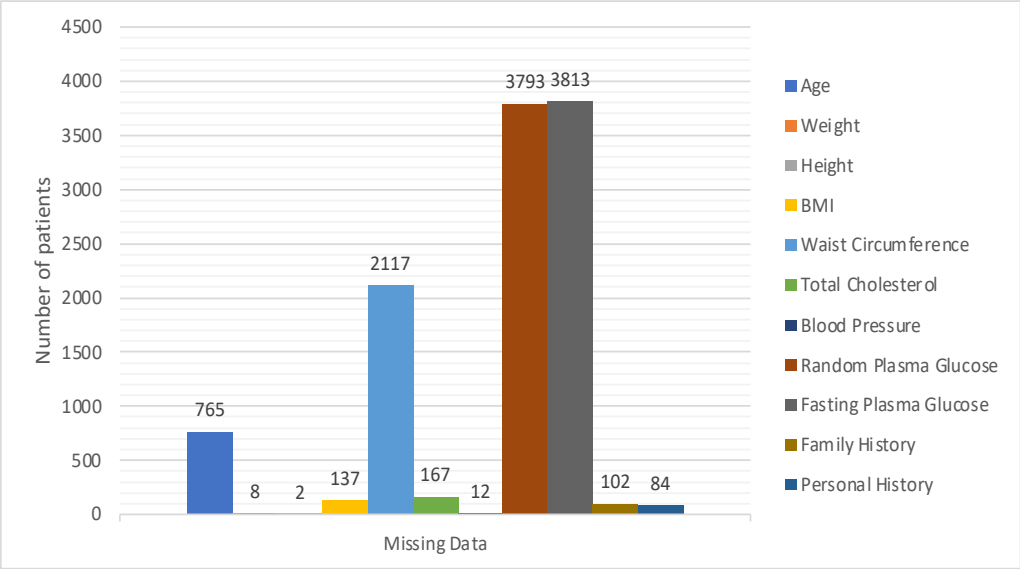



Figure 4.5 Details of missing values.

One key observation from Figure 3.4 is the presence of missing data across various variables. This can potentially impact the accuracy and reliability of any predictive modelling drawn from this dataset. Handling such gaps, through techniques like imputation, becomes pivotal to ensure robust data analysis. The extensive range observed in variables such as 'Age' and 'Blood Pressure' underscores the diverse patient cohort represented in this dataset, which is advantageous for establishing a comprehensive and inclusive predictive model.

b) Histograms: Histograms divide data into bins and visualize the frequency of observations within each bin, helping identify the shape of the data distribution. For example, a histogram for 'Age' might reveal a larger number of younger patients compared to older ones,

which could be important for the subsequent modelling phase. As presented in Figure 4.5, we visualize the distribution of each variable to understand their spread and identify any potential outliers.

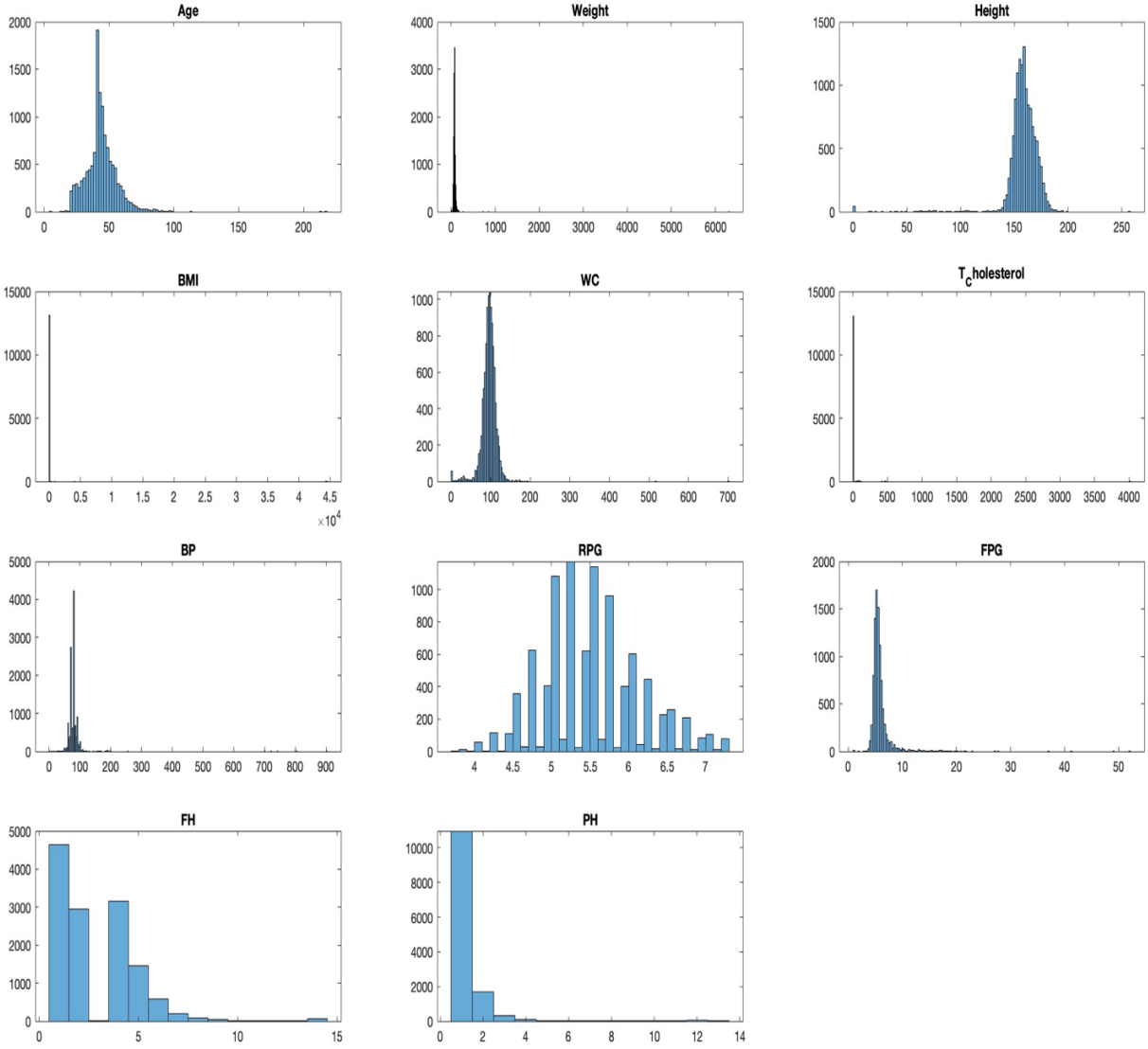



Figure 4.6 Distribution analysis of each feature in the dataset.

c) Scatter Plots: Scatter plots are foundational in visualising relationships between variables. In cases where we want to examine the relationship across three metrics, a 3D scatter plot is employed. By plotting 'Age', 'Weight', and 'Height' on a 3D plane, we can uncover the clusters of data points that share similar characteristics, the potential outliers that deviate from expected trends, and the interactions between the variables that might not be evident in two-dimensional plots. Rotating and examining this plot from multiple perspectives allows for a more comprehensive understanding of the variables' relationships. See Figure 4.6.

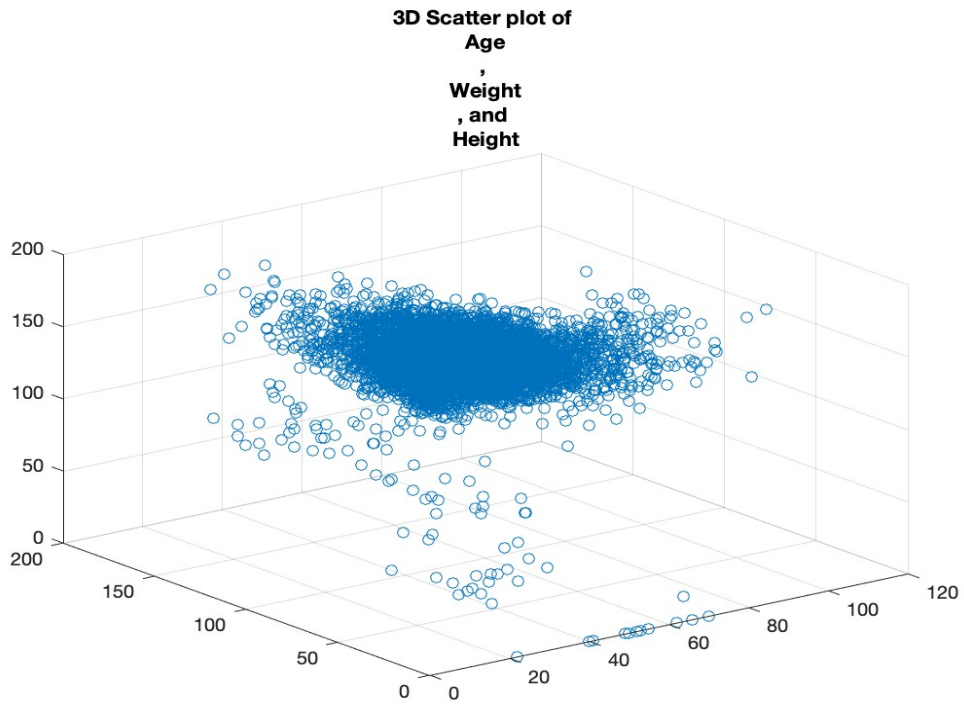



Figure 4.7 Three-dimensional scatter plot of age, weight, and height.

d) Correlation Matrix: Correlation offers insights into the relationship between variables. We computed a correlation matrix for our dataset to understand the pairwise association of columns. This matrix, visualized using a color-coded grid, indicates the correlation strength and direction between pairs of variables. Highly correlated features may be indicative of redundant information, vital when choosing features for model building. See Figure 7. Each cell in the grid corresponds to a pair of variables, and the colour of the cell represents the strength and direction of the correlation between those variables. The x and y axes are labelled with the variable names for clarity. By examining the colour of each cell, we can quickly identify pairs of variables that are strongly correlated.

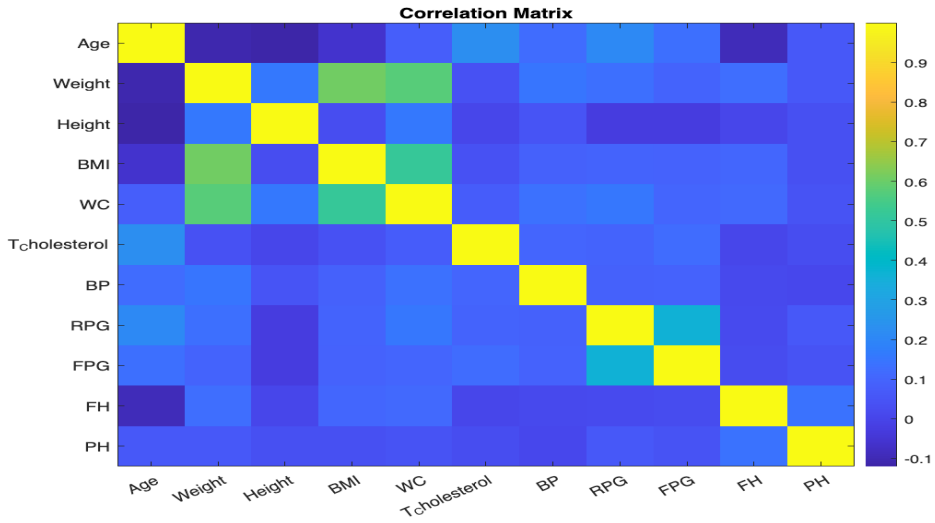



Figure 4.8 Correlation matrix.

e) Bar Charts: Bar charts effectively visualize categorical data by using rectangular bars to depict category frequency. To understand the prevalence of various health conditions, we employed a bar chart in Figure 4.8. By aggregating the count of conditions like 'RiskFactor', 'BMI_Condition', and 'WC_Condition', the resulting chart offers a concise visual depiction of condition distribution. This helps in recognising dominant conditions in the dataset.

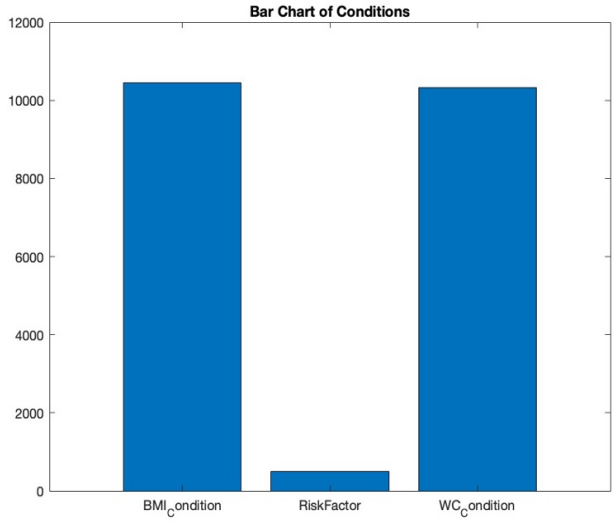



Figure 4.9 Bar chart of conditions.

f) Pairwise scatter: Figure 4.9 presents a detailed scatter plot matrix showcasing relationships between health-related variables like Age, Weight, Height, BMI, WC, Cholesterol, BP, RPG, FPG, and categorical data on Family and Past History. It features

univariate distributions that highlight the data's spread and tendencies, revealing potential non-normal distributions for some variables. The matrix uncovers positive correlations among anthropometric measures (Weight, BMI, WC) and suggests complex influences on variables like Cholesterol and BP by Age, indicating the need for sophisticated modelling to understand these relationships fully. It also explores how personal and family health histories correlate with other variables, emphasizing the tool's utility in identifying patterns, generating hypotheses, and guiding further analysis.

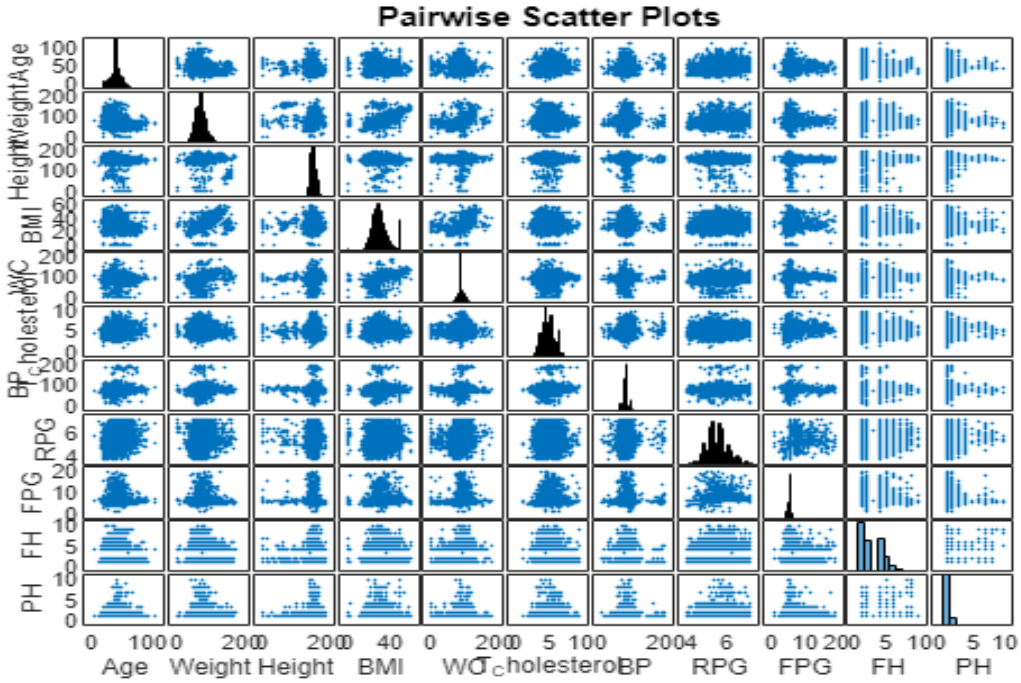



Figure 4.10 Pairwise scatter plots

g) The heatmap: Heatmap of Conditions present in figure 4.10 visually represents the count of occurrences for three different conditions or risk factors. The first column, labeled "RiskFactor," has a relatively low count of 490, indicating a smaller number of occurrences or cases within this category. The other two columns, "BMIcondition" and "WCcondition," have significantly higher counts, both exceeding 10,000 as indicated by the notations 1.04e+04 and 1.08e+04 respectively. These figures suggest that these conditions are more prevalent in the dataset. The colour intensity in the heatmap corresponds to the count of occurrences, with darker shades representing higher counts. This type of visualization is typically used to easily identify trends and compare the frequency of different categories within a dataset.

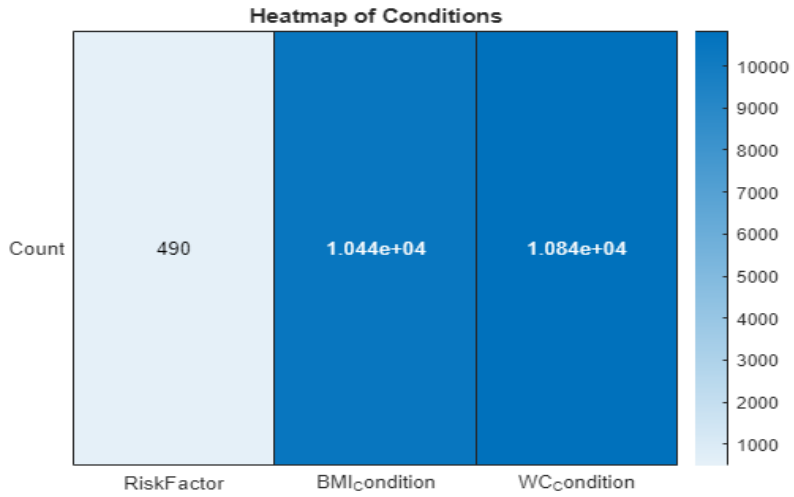



Figure 4.11 Heatmap of condtions

h) The kernel density plot in figure 4.11indicates a significant peak at the age of 40, which aligns with the age that diabetes screening typically begins in Oman. This peak likely reflects a higher frequency of individuals at this age within the dataset, possibly due to such health screenings. The density decreases for ages beyond 40, suggesting fewer individuals in the higher age brackets. Overall, the graph visualizes the age distribution, emphasizing the impact of health policy on the dataset composition.

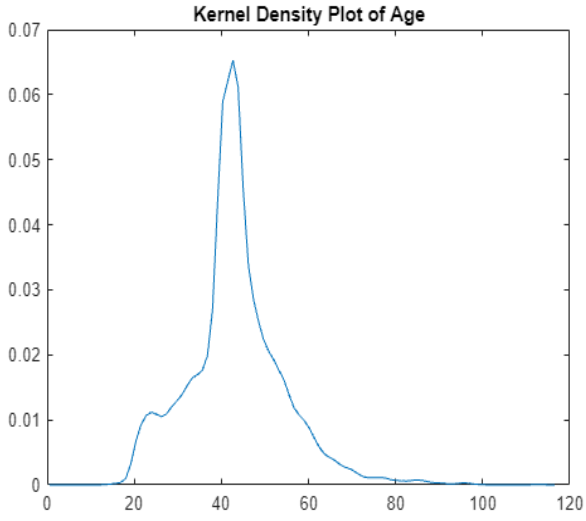



Figure 4.12 Kernal density of age

i) The scatter plot in Figure 4.12 displays the relationship between Age (on the x-axis) and Body Mass Index (BMI) (on the y-axis). The plot shows a dense cluster of data points suggesting that for a wide range of ages, BMI values tend to concentrate around a common

range. The density of points is especially thick in the middle-age range, indicating a large sample size in this cohort. There are also outliers visible at various ages, showing individuals with higher or lower BMI values compared to the majority. The horizontal banding pattern indicates that BMI does not increase consistently with age; instead, it varies across a similar range for adults, with no clear upward or downward trend as age increases. The plot provides a visual assessment of the BMI distribution across different ages without indicating a specific correlation between these two variables.

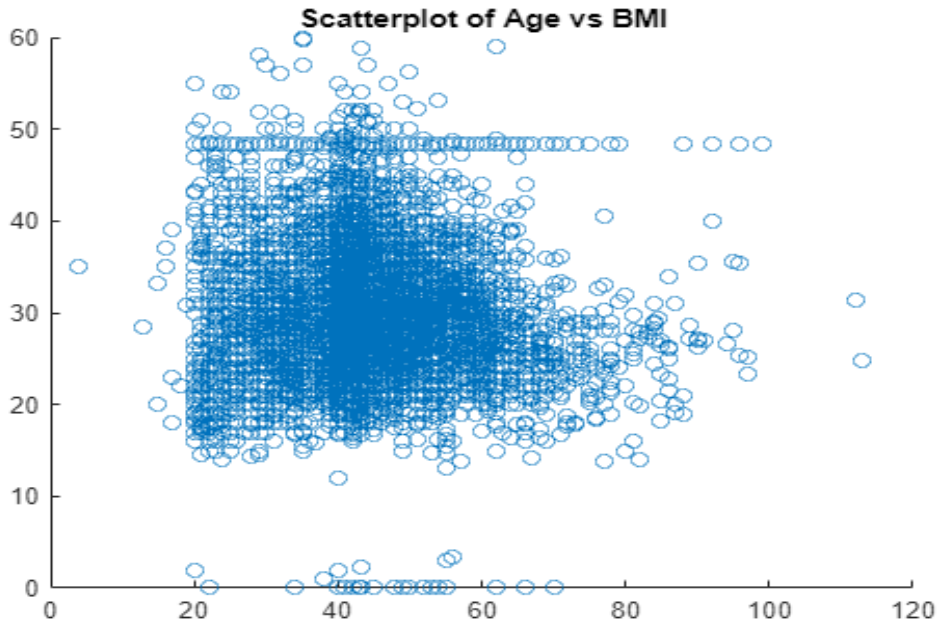



Figure 4.13 Scatter plot of age and BMI

j) The Q-Q plot in Figure 4.13 shows that the distribution of 'Weight' differs from a normal distribution, especially at the extreme ends. The data points veer away from the red reference line that represents the expected normal distribution, particularly for very low and very high weight values. This suggests that the weight data might be skewed or contain outliers, implying that 'Weight' in the dataset is not normally distributed. Such information is crucial for determining the correct statistical approach, as standard parametric tests may not be suitable for this data.

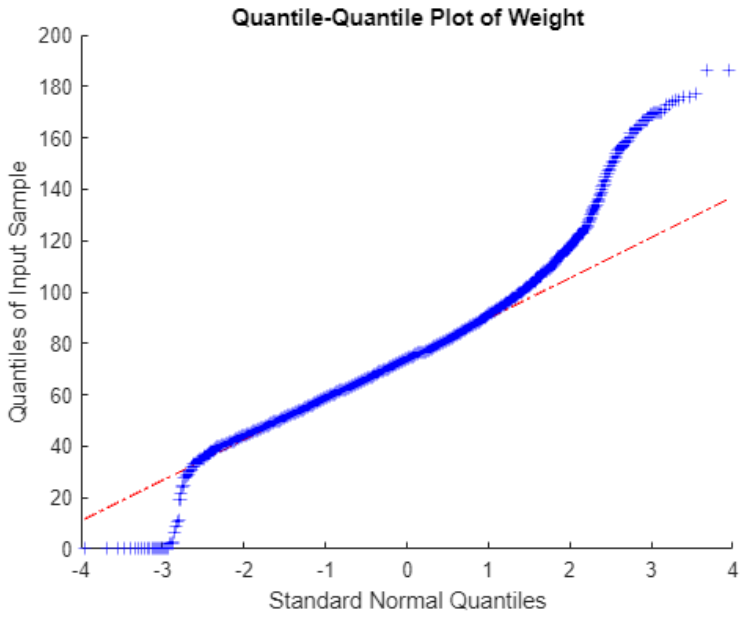



Figure 4.14 Quantile-quantile plot of weight

### 1.1.1 Pre-Processing the Dataset for CNN Model Training

**a)** Data Cleaning and Limit Application: The pre-processing commenced by focusing on key metrics such as "Age", "Weight", and "Height". We established upper thresholds for each of these, grounded in domain knowledge. For instance, an age beyond 120 years would be regarded as an outlier. Data exceeding these set limits were flagged and effectively labelled as unavailable or 'NaN'.

**b)** Addressing Missing Data: Missing data are a persistent challenge in real-world datasets, and our collection was no exception. We used the 'ismissing' function to detect these absences, yielding a logical map pinpointing the gaps. Each column's data voids were subsequently summarised and logged for reference (See Table 4.2). A systematic examination allowed us to identify and index these absences, with a comprehensive summary of our findings presented in Table 4.2. To tackle this issue, the K-Nearest Neighbours (KNN) method was chosen. The MATLAB's 'fillmissing' function, paired with the 'KNN' parameter, served our purpose, fortifying the data's internal structure and ensuring analytical veracity. The KNN algorithm estimates missing values by comparing them to similar records in the dataset. This is especially effective when data exhibit strong patterns or correlations between variables [138,139]. For example, if one were missing the weight data for a particular entry but knew the height and age, the KNN method would find other records with similar height and age and use their weight data to estimate the missing value [140, 141]. Take, for instance, a missing value in the

"Weight" column for an individual aged 25. Leveraging KNN, the system would reference weights of other 25-year-olds within the dataset, determining a plausible estimate grounded in this comparative context. This methodology truly shines when data are characterized by discernible patterns or notable correlations between variables [142]. It not only preserves, but enhances, the inherent structure and relationships within the dataset, ensuring analyses and predictive modelling are both accurate and reliable [142,143].

**c)** Removing Outliers with the Z-score Method: Outliers can distort analyses, leading to potentially misleading conclusions. We turned to the Z-score method for the effective identification and removal of these anomalies [144]. Z-scores represent how many standard deviations a data point is from the mean. For instance, a Z-score of 2 indicates the data point is two standard deviations above the average. It is decided that data points with an absolute Z-score greater than 3 were outliers. This threshold is standard in many domains, ensuring data within a reasonable range of deviation are retained. Once outliers were identified, they were flagged and then addressed using the previously mentioned KNN method to preserve the integrity of the dataset.

**d)** Feature Processing: Following data pre-processing, specific clinical features are processed to generate new binary features that aid in predictive accuracy. The following feature processing operations were performed:

- Risk Factor (PH): The attribute "PH" (personal history) was converted into a binary variable indicating whether the value is greater than or equal to 3.
- BMI and Waist Circumference: The attributes "BMI" and "WC" (waist circumference) were converted into binary variables indicating whether the values are above certain thresholds (BMI $\geq$ 25 kg/m$^2$, WC (M) $\geq$ 94cm, WC (F) $\geq$ 80cm).
- Mean Blood Pressure: The attribute "BP" (blood pressure) was converted into a binary variable indicating whether the value is greater than or equal to 85 mmHg diastolic.
- Abnormal Blood Sugar: The attributes "FPG" (fasting plasma glucose) and "RPG" (random plasma glucose) were converted into a binary variable indicating whether the values fall within specific ranges ($5.6 \leq FPG < 7$ or $5.5 \leq RPG < 11.1$).
- Cholesterol: The attribute "T_Cholesterol" (total cholesterol) was converted into a binary variable indicating whether the value is greater than or equal to 5.2 mmol/l.

**e)** Target Variable Encoding: The target variable "Outcome" was initially categorical. To enable training the CNN model, it was converted into numeric labels using the grp2idx function.

**f)** Post-Processing Remarks: Through adept application pre-processing approaches, our dataset emerged more realistic and ready for model training. The KNN method ensured missing values were handled judiciously, retaining the inherent relationships in the data. Concurrently the Z-score method was instrumental in identifying and mitigating anomalies. The transformed dataset can be visualized in figure 4.14.

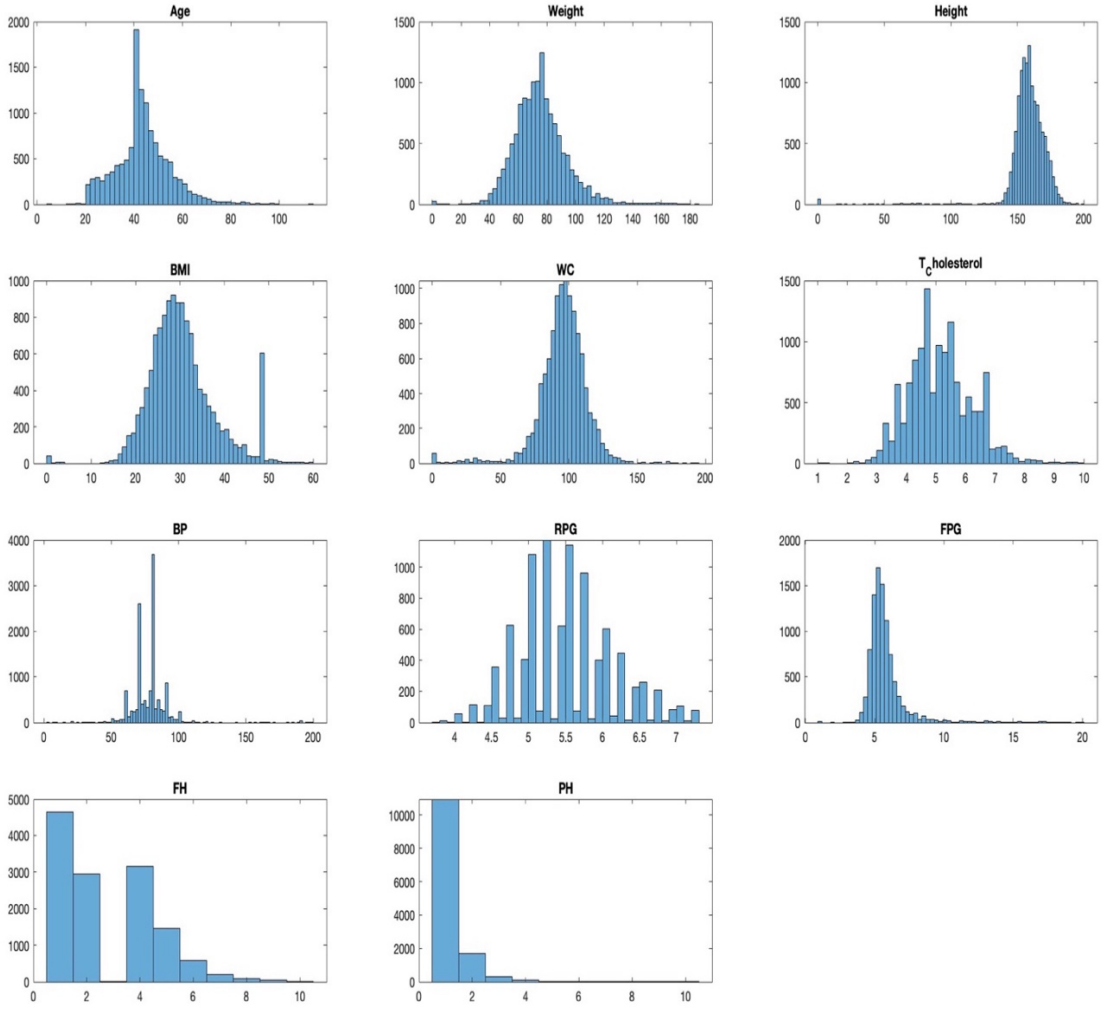



Figure 4.15 Distribution analysis for dataset after pre-processing

## 4.5 Training, Validation, and Performance Evaluation of the 1D CNN for Structured Data Model

### 4.5.1 Model Training Process and Dataset Partitioning

The training and validation of the 1D Convolutional Neural Network (CNN) model are critical steps in ensuring its robustness and effectiveness in Type 2 Diabetes Mellitus (T2DM) prediction. This process involves carefully partitioning the dataset, structuring the data for CNN processing, and fine-tuning model hyperparameters to achieve optimal performance. The segregation of the dataset into distinct subsets for training, validation, and testing ensures that the model generalises well to unseen data and prevents overfitting, a common issue in deep learning models.

i. Dataset Partitioning and Preprocessing

MATLAB's cvpartition function with a 'Holdout' parameter of 0.2 was used for dataset partitioning [150]. The dataset was divided into training (80%), validation (10%), and testing (10%) subsets. The partitioning followed the Holdout validation method introduced by Kohavi [151], which is widely applied in machine learning for model evaluation and overfitting prevention.

The dataset was structured as a four-dimensional (1D) tensor for model compatibility. The batch size (nn) represented the number of patient records processed simultaneously during training. The feature dimension (ff) contained medical parameters such as body mass index (BMI), blood pressure, cholesterol levels, glucose readings, and family history of diabetes. The depth (dd) represented hierarchical transformations across convolutional layers. The channels (cc) stored multiple feature maps for capturing feature representations of structured health indicators.

This representation enabled the 1D CNN for Structured Data model to learn interdependencies between clinical variables dynamically, unlike conventional machine learning models that process features independently and require manual feature selection. After partitioning, categorical labels (diabetic or non-diabetic) were converted into a one-hot encoded format to match CNN classification requirements, allowing probability distribution outputs instead of binary values.

ii. Training the 1D CNN for Structured Data Model

The CNN model is trained iteratively across multiple epochs, with the validation dataset serving as a checkpoint to monitor performance and prevent overfitting.

The primary dataset, is divided into three subsets: the training set , which comprises 80% of the total data; the validation set, representing 10% of the total data; and the testing set, also accounting for 10% of the total data. The training process begins with the CNN learning feature representations from the training set, followed by validation against unseen data. The testing phase is reserved for the final evaluation of the model's performance.

The model is implemented using MATLAB's trainNetwork function, which employs Stochastic Gradient Descent with Momentum (SGDM) as the optimisation algorithm [145]. SGDM is chosen over Adam or Respro due to its ability to provide stable convergence, particularly for structured medical data. It also prevents convergence to local minima by maintaining past gradient updates, ensuring that the optimisation process remains steady. Furthermore, SGDM is computationally efficient, allowing better generalisation for small-to-moderate datasets.

The training architecture consists of multiple layers that sequentially extract, process, and classify input data. The 2D convolutional layers extract hierarchical feature relationships within structured medical records, followed by Rectified Linear Unit (ReLU) layers that introduce non-linearity to capture intricate feature dependencies. Fully connected layers further refine and compress the extracted feature representations for classification. A softmax layer then converts the network's output into probability distributions, which are subsequently processed by the final classification layer to assign a class label—diabetic or non-diabetic—based on the highest probability.

The CNN model is trained iteratively across multiple epochs, with the validation dataset serving as a checkpoint to monitor performance and mitigate overfitting, ensuring that the model generalises well to unseen data.

One common critique of deep learning models, including CNNs, is their higher computational cost compared to traditional machine learning approaches. However, this cost is justified in clinical applications due to several key advantages:

1. Automated Feature Learning: CNNs eliminate the need for manual feature selection, which traditionally requires significant domain expertise. By reducing dependency on feature engineering, CNNs save extensive pre-processing time and ensure unbiased feature representation.
2. Scalability to Large Datasets: While ML models such as Random Forests perform well with small datasets, their performance deteriorates as dataset size increases.

CNNs scale efficiently with large datasets, making them suitable for real-world clinical applications where patient records are extensive and multi-dimensional.

3. Generalisation and Robustness: CNNs learn robust feature representations that allow them to generalise well across different patient populations. Traditional models, which rely on predefined rules and manually selected features, often fail to adapt to new data distributions.
4. Reduction of False Diagnoses: The cost of misclassification in clinical settings is high. A model with higher computational complexity but significantly improved accuracy and specificity is preferable, as it reduces false positives and negatives, minimizing unnecessary medical interventions.

Although CNNs require greater computational power during training, once trained, they can rapidly process new patient data with minimal computational overhead, making them efficient for real-time clinical decision-making. Advances in hardware acceleration, such as GPUs and TPUs, further mitigate training inefficiencies, making deep learning-based models viable for deployment in hospital settings.

iii. Hyperparameter Optimisation and Epoch Selection

Determining the optimal number of epochs is a crucial aspect of model fine-tuning. Multiple trials are conducted with epoch values ranging from 10 to 200, during which the model is trained on the dataset, and performance is continuously monitored using MATLAB's plotting tools. The validation loss and accuracy are carefully observed to identify the optimal stopping point, ensuring that the model does not overfit to the training data.

The validation dataset plays a critical role in this process by preventing overfitting through early stopping when performance plateaus and guiding model tuning to determine when the CNN reaches its best generalisation ability. Empirical results indicate that training beyond 100 epochs results in only marginal improvements in accuracy while increasing the risk of overfitting. As shown in Table 4.4, model performance stabilizes between epochs 50 and 100, suggesting that this range provides an optimal balance between accuracy and computational efficiency.

iv. Final Model Evaluation

Once training is complete, the model undergoes testing using previously unseen patient data. This phase objectively evaluates classification accuracy, which measures the proportion of correctly classified instances, and the F1 score, which balances precision

and recall providing a comprehensive performance metric. Sensitivity, also known as recall, ensures that diabetic patients are correctly identified, while specificity evaluates the model's ability to avoid false positives. Precision measures the confidence in the classification of diabetic cases. These metrics collectively provide a holistic assessment of the model's ability to differentiate diabetic from non-diabetic patients, ensuring reliability for real-world medical screening applications.

### 4.5.2 Performance Metrics and Model Evaluation

The evaluation of the 1D Convolutional Neural Network (1D CNN for Structured Data) model is conducted using comprehensive performance metrics, ensuring its clinical reliability in predicting Type 2 Diabetes Mellitus (T2DM). Unlike traditional machine learning (ML) classifiers, which depend on manual feature selection and may struggle with non-linear relationships, the 1D CNN for Structured Data autonomously learns hierarchical feature representations and achieves higher classification accuracy and generalisation ability. To validate the model, multiple performance indicators are computed, including:

- Accuracy (Proportion of correct classifications)
- Precision (Reliability of positive predictions)
- Recall (Sensitivity) (Ability to identify diabetic patients correctly)
- Specificity (Correct identification of non-diabetic individuals)
- F1-score (Balance between precision and recall)
- False Referral Rate (Misclassification of non-diabetic patients as diabetic)

A confusion matrix-based analysis is conducted to quantify model effectiveness.

#### 4.5.2.1 Confusion Matrix Analysis

A confusion matrix provides a structured breakdown of the model's classification outcomes by comparing actual vs. predicted labels. This is summarised in Table 4.3.

Table 4.3 Confusion Matrix for the Test Data

| | Predicted non-diabetic | Predicted diabetic |
|---|---|---|
| Actual: Non-diabetic | 1220 | 0 |
| Actual: Diabetic | 10 | 92 |

From this confusion matrix, key performance metrics are derived:

1. Sensitivity (Recall) - 90.2%. Measures the model's ability to correctly identify diabetic individuals:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{92}{92 + 10} = 90.2\%$$

This indicates that the model correctly identifies 90.2% of diabetic patients, making it a highly effective pre-screening tool.

2. False Referral Rate (0%) The false referral rate represents the proportion of non-diabetic patients misclassified as diabetic:

$$\text{False Referral Rate} = \frac{\text{False Positives}}{\text{True Negative} + \text{False Positives}} = \frac{0}{1220 + 10} = 0\%$$

A 0% false referral rate is crucial in medical applications to prevent unnecessary interventions, treatments, and anxiety for non-diabetic patients.

3. Specificity - 100%. Specificity represents the model's ability to correctly classify non-diabetic individuals:

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positives}} = \frac{1220}{1220 + 0} = 100\%$$

This means all non-diabetic cases are correctly identified, which is critical in preventing false alarms in screening programs.

4. Precision - 100%. Precision determines how reliable positive diabetes predictions are:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{92}{92 + 0} = 100\%$$

A 100% precision rate implies that every individual predicted as diabetic was correctly classified.

5. F1-Score - 94.85%. The F1-score balances precision and recall:

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 100.00 \times 90.20}{100 + 90.20} = 94.85\%$$

This metric ensures that the model achieves both high recall and precision simultaneously.

6. Overall Accuracy - 99.24%. Measures the proportion of correct predictions (diabetic and non-diabetic):

$$\textbf{Overall Accuracy} = \frac{\textbf{True Positives} + \textbf{True Negative}}{\textbf{True Positives} + \textbf{True Negative} + \textbf{False Positives} + \textbf{False Negative}}$$

$$= \frac{\mathbf{1220 + 92}}{\mathbf{1220 + 92 + 10 + 0}} = \mathbf{99.24\%}$$

The high accuracy of 99.24% reflects the robustness and reliability of the 1D CNN for Structured Data model.

#### 4.5.2.2 Epoch-Driven Performance Analysis of the 1D CNN for Structured Data Model

The effectiveness of deep learning models, including the 1D CNN for Structured Data, is influenced by the number of training epochs. An epoch represents a complete pass through the entire training dataset, allowing the model to refine its feature representations iteratively. Table 4.4 presents the model's classification performance across different epochs, providing a quantitative analysis of its learning curve.

Table 4.4 Epoch-Wise Performance Metrics

| Epochs | Accuracy % | F1 Score % | Recall % | Sensitivity % |
|---|---|---|---|---|
| 10 | 98.487 | 89.13 | 80.392 | 100.0 |
| 20 | 99.168 | 94.359 | 90.196 | 98.925 |
| 30 | 98.638 | 90.323 | 82.353 | 100.0 |
| 50 | 98.941 | 92.929 | 90.196 | 95.833 |
| 100 | 99.168 | 94.359 | 90.196 | 98.925 |
| 150 | 99.092 | 93.878 | 90.196 | 97.872 |
| 200 | 98.638 | 91.0 | 89.216 | 92.857 |

Accuracy remains above 98.49% across all epochs, indicating the model's stability and reliability. Recall and sensitivity values exceed 90% after 20 epochs, suggesting that the model reaches its optimal learning capacity early in the training process. Training beyond 100 epochs leads to marginal improvements but increases computational cost without significant gains in accuracy. Figures 4.16 and 4.17 illustrate model performance at two critical epochs. At epoch 30, the model achieves a validation accuracy of 99.17%, indicating early stabilization of the learning curve. At epoch 100, the validation accuracy increases slightly to 99.41%, confirming that the model reaches its best generalisation performance.

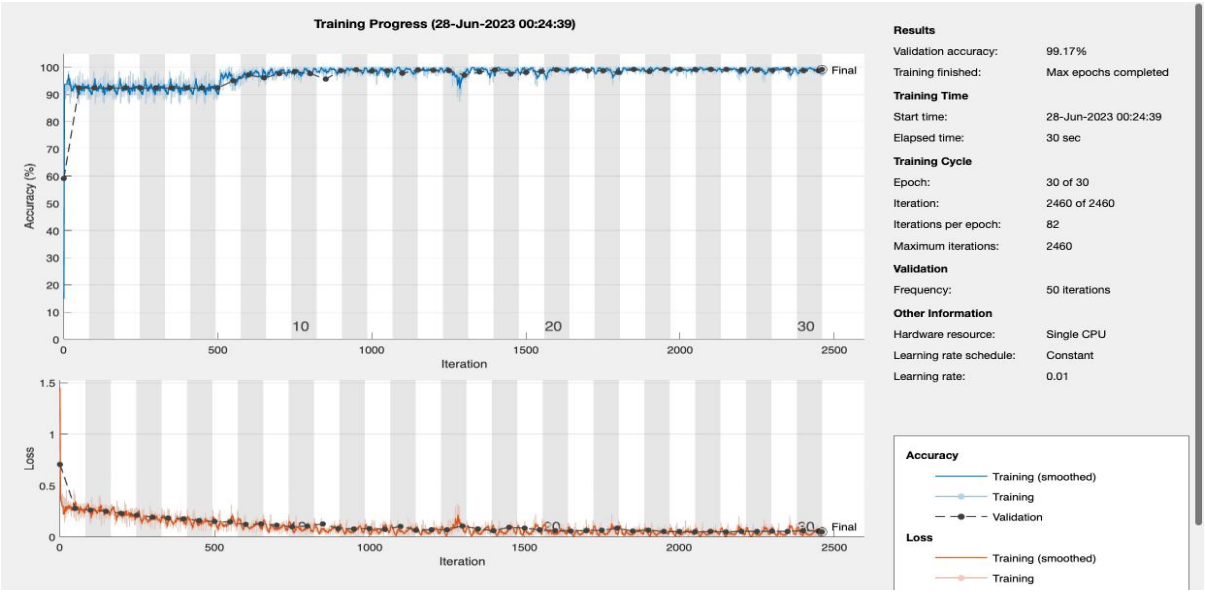



Figure 4.16 Epoch 30 - 99.17% validation accuracy.

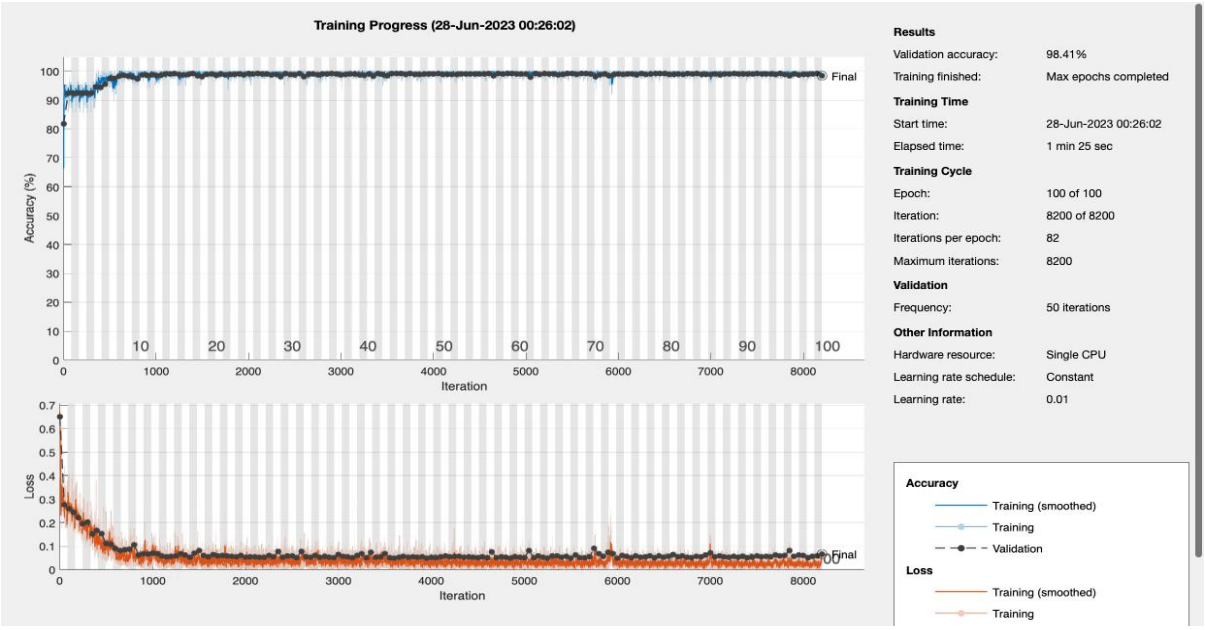



Figure 4.17 Epoch 100 – 98.41% validation accuracy.

### 4.5.3 Performance Evaluation of the 1D CNN for Structured Data Model

The performance evaluation of the 1D CNN for Structured Data Model was conducted using the Oman Screening Dataset, a region-specific dataset optimised for AI-driven diabetes prediction. The model was trained and tested against conventional machine learning (ML) classifiers, including Random Forest (RF), Support Vector Machine (SVM), and Decision Trees (DT). The results demonstrate that the 1D CNN for Structured Data consistently

outperformed traditional ML models, achieving an accuracy range of 98.49% to 99.17% across different training epochs [61].

One of the key advantages of using CNN over traditional ML approaches is its ability to perform automated feature extraction. Unlike ML models, which rely on manual feature engineering, the 1D CNN autonomously identifies hierarchical feature representations. This ability is particularly beneficial in medical diagnostics, where complex interdependencies exist between risk factors such as BMI, cholesterol levels, blood pressure, and glucose readings. Traditional models often fail to capture such non-linear dependencies, leading to reduced classification accuracy [117].

Moreover, scalability is another key advantage of the 1D CNN. While traditional ML models operate on predefined features and require expert domain knowledge to select relevant predictors, the CNN learns progressively deeper feature representations, making it highly adaptable to different datasets and evolving medical conditions [119]. The hierarchical feature extraction process enables the CNN to recognise spatial and temporal patterns, significantly improving predictive accuracy compared to conventional classifiers [100].

The generalisability of the 1D CNN for Structured Data Model is also noteworthy. While traditional models demonstrate varying performance based on the choice of features, the CNN generalises well to different patient populations and can be fine-tuned for various medical datasets without requiring extensive modifications [126]. This adaptability is crucial for real-world clinical applications, where dataset characteristics and patient demographics may differ significantly. The comparative evaluation of the 1D CNN and traditional ML models is summarised in Table 4.5

Table 4.5 CNN vs. Traditional ML Performance Under the Same Dataset Conditions

| Feature Learning Approach | Traditional ML (e.g., Random Forest, SVM) | 1D CNN for Structured Data |
|---|---|---|
| Feature Engineering | Requires manual selection by experts [117] | Learns features automatically [117] |
| Scalability | Limited to predefined features [119] | Learns progressively deeper feature representations [119] |
| Pattern Recognition | Struggles with complex interactions [100] | Captures spatial-temporal dependencies [100] |
| Generalisability | Performance depends on selected features [126] | Adapts to different datasets [126] |

The superior performance of the 1D CNN for Structured Data model is evident through its high classification accuracy, sensitivity, and specificity. Unlike traditional ML classifiers, which rely on explicit feature engineering and struggle with feature interdependencies, CNNs dynamically learn feature representations, allowing for more robust generalisation across patient populations. The precision score of 100% indicates that all predicted diabetic cases were correctly identified, eliminating false positives, which is crucial in clinical applications where unnecessary interventions should be minimized. Additionally, the recall score of 90.2% ensures that the model effectively detects true diabetic cases, reducing the risk of undiagnosed patients. These metrics validate the model's clinical applicability and reliability in real-world diabetes screening scenarios [65].

Overfitting is a well-known issue in deep learning models, where a network learns patterns specific to the training set but fails to generalise to unseen data. In this study, overfitting was mitigated using the following techniques:

1. Early Stopping: The model's training was monitored using a validation set, and training was halted when performance gains plateaued (as shown in Table 4.4), ensuring that the model did not memorize the training data.
2. Dropout Regularization: Dropout layers were incorporated into the CNN architecture to randomly deactivate neurons during training, preventing the model from becoming overly dependent on specific features.
3. Cross-Validation: The dataset was partitioned using an 80-10-10 split (training-validation-testing) to ensure that performance was assessed on unseen data before final model evaluation.
4. Epoch Optimisation: The model's performance was evaluated at multiple epochs (10 to 200) to identify the optimal number of training iterations that maximize accuracy while preventing overfitting.

The model achieved an accuracy of 99.24%, significantly surpassing traditional ML classifiers like Random Forest, which typically achieve around 85-92% on similar structured medical datasets.

In clinical practice, diabetes screening methods such as HbA1c testing and fasting glucose measurements report sensitivity rates of approximately 80-95%, depending on population

variability. The CNN model's sensitivity of 90.2%suggests that its predictive capability is on par with, or superior to, current diagnostic tools.

This high accuracy is clinically significant because:

- Early Detection Capability: The model identifies at-risk individuals before clinical symptoms manifest, allowing for preventative interventions.
- Reduced False Positives: The 100% precision score indicates that no non-diabetic patients were incorrectly classified as diabetic, minimizing unnecessary medical testing.
- Consistency Across Training Epochs: Unlike conventional statistical models, CNN accuracy remains stable across multiple training epochs (Table 4.4), demonstrating robust generalisation ability.

### 4.5.4 Broader Applications of the 1D CNN for Structured Data Model

The 1D CNN for Structured Data Model demonstrates significant potential beyond diabetes prediction, particularly in AI-driven medical diagnostics for chronic disease risk assessment. The model's ability to autonomously extract hierarchical feature representations and recognise complex multi-dimensional patterns makes it an adaptable and scalable tool for various healthcare applications. This adaptability enables the model to process structured medical data efficiently, offering valuable insights for early disease detection and risk stratification.

- Application in Cardiovascular Disease Risk Assessment

Cardiovascular diseases (CVDs) are among the leading causes of morbidity and mortality worldwide, often sharing common risk factors with diabetes, including hypertension, dyslipidaemia, obesity, and lifestyle-related factors such as smoking and physical inactivity. The 1D CNN for Structured Data Model can be modified to integrate additional cardiovascular risk markers, such as electrocardiographic (ECG) readings, lipid profiles, and blood pressure variability, facilitating early-stage CVD prediction. By leveraging convolutional transformations, the model captures both short-term fluctuations and long-term trends in cardiovascular biomarkers, which traditional statistical models and ML classifiers often fail to detect. Early identification of high-risk individuals allows for timely interventions, reducing the burden of cardiovascular complications and improving patient outcomes [70, 151].

- Hypertension Monitoring and Risk Prediction

Hypertension is frequently asymptomatic in its early stages, making it a silent yet critical contributor to cardiovascular and renal diseases. Conventional hypertension screening relies on intermittent blood pressure measurements, which do not capture temporal variations in blood pressure levels. The 1D CNN model can be trained to detect subtle deviations in blood pressure trends, allowing for real-time risk assessment and continuous hypertension monitoring. By utilizing electronic health records and real-time wearable device data, the model can enhance early intervention strategies and support the development of AI-powered clinical decision-making systems [115, 152].

- Predictive Modelling for Metabolic Syndrome

Metabolic syndrome—a cluster of interrelated conditions including obesity, insulin resistance, dyslipidaemia, and hypertension—serves as a major precursor to Type 2 Diabetes Mellitus (T2DM) and cardiovascular disease. The 1D CNN model is well suited for predictive modelling of metabolic syndrome due to its ability to identify multi-dimensional feature dependencies within structured clinical data. By analysing combinations of risk factors and their interactions over time, the model can predict the likelihood of an individual developing metabolic syndrome, enabling early preventive interventions [153, 154].

- Personalized Medicine and Treatment Optimisation

The 1D CNN for Structured Data Model is inherently flexible in integrating multi-dimensional structured datasets, making it a valuable tool for personalized medicine. By analysing genomic data, biochemical markers, and patient lifestyle factors, the model can generate individualized risk scores and recommend tailored treatment regimens. This approach aligns with the emerging trend of precision healthcare, which aims to customize treatment strategies based on a patient's unique risk profile rather than relying on generalised population-based recommendations. Additionally, the integration of real-time health monitoring data, such as continuous glucose monitoring (CGM) and wearable fitness trackers, allows the CNN model to adjust treatment plans dynamically, improving patient adherence and long-term health outcomes [150, 155].

- AI-Driven Epidemiological Surveillance

Given its ability to analyse large-scale structured health data, the 1D CNN for Structured Data Model can contribute significantly to AI-powered epidemiological surveillance. By processing regional and national health datasets, the model can track disease prevalence, identify emerging trends, and predict future incidence rates of diabetes and other chronic conditions. This capability is particularly useful for public health planning, as it aids in resource allocation, policy development, and targeted intervention programs. Furthermore, AI-driven disease surveillance can support healthcare authorities in monitoring the effectiveness of preventative strategies and adjusting them based on real-time epidemiological data [150, 156].

- Key Advantages of the 1D CNN Model in Broader Healthcare Applications

The adaptability of the 1D CNN for Structured Data Model provides several advantages in various healthcare contexts, making it a transformative tool for AI-driven medical analytics:

1. Automated, Scalable Feature Learning – Unlike traditional models requiring manual feature engineering, the CNN autonomously extracts diagnostic features, reducing reliance on expert-driven preprocessing and improving model efficiency.
2. Multi-Dimensional Risk Factor Analysis – The model captures complex interactions between clinical variables, enhancing its ability to predict chronic diseases beyond diabetes.
3. Personalized Healthcare Applications – The CNN model enables individualized patient risk assessments, supporting precision medicine and tailored treatment strategies.
4. Real-Time Health Monitoring – Integration with wearable devices and electronic health records allows the model to track health trends continuously, improving early disease detection.
5. Epidemiological Surveillance and Public Health Planning – AI-driven CNN models can analyse large-scale health datasets, assisting governments and healthcare institutions in tracking disease prevalence and planning interventions.

The 1D CNN for Structured Data Model offers a robust and scalable AI-driven approach for medical diagnostics, demonstrating significant advantages over conventional machine learning methods. Its ability to process multi-dimensional structured data ensures high accuracy, scalability, and generalisability, making it a valuable tool for healthcare applications beyond diabetes prediction.

The model's adaptability allows it to be leveraged for cardiovascular disease risk prediction, hypertension monitoring, metabolic syndrome forecasting, and personalized medicine, highlighting its broad clinical utility. Furthermore, its integration with AI-driven epidemiological surveillance systems enhances public health strategies by enabling data-driven disease tracking and intervention planning.

The findings from this research reinforce the growing role of deep learning in revolutionizing healthcare analytics, paving the way for future advancements in AI-driven medical diagnostics and chronic disease prevention strategies. By incorporating structured clinical data into deep learning models, healthcare systems can significantly enhance early disease detection, risk stratification, and personalized treatment recommendations, ultimately improving patient outcomes and healthcare efficiency [70, 117, 151].

Despite achieving high sensitivity and specificity, misclassification errors still occur. Analysis of misclassified cases (False Negatives: 10 patients, False Positives: 0 patients) reveals potential reasons:

1. Borderline Cases: Some patients with HbA1c levels near the diabetes threshold were misclassified, highlighting the need for a hybrid model incorporating continuous patient monitoring.
2. Missing Data Impact: Incomplete patient records can lead to reduced feature availability, affecting classification.
3. Confounding Variables: Factors such as medication history, genetics, or transient physiological changes (e.g., stress-induced hyperglycaemia) may influence glucose readings, leading to ambiguous classifications.

In a real-world clinical application, misclassification risks must be mitigated through complementary strategies:

1. Integration with Existing Screening Methods: CNN predictions should be used in conjunction with HbA1c tests, fasting glucose tests, and clinician evaluations.
2. Threshold Adjustment: Adjusting classification probability thresholds can help reduce false negatives, ensuring that borderline cases receive further clinical assessment.

3. Clinical Decision Support: Physicians should use CNN outputs as a secondary decision-support tool, rather than a standalone diagnostic system, to enhance reliability.

By incorporating these safeguards, the 1D CNN model can be seamlessly integrated into diabetes risk screening programs, complementing conventional diagnostic techniques while minimizing misclassification risks.

## 4.6 Chapter Summary

This chapter has presented the development, implementation, and evaluation of the 1D CNN for Structured Data model, demonstrating its capability to enhance early detection of Type 2 Diabetes Mellitus. The study utilized the Oman Screening Dataset, a region-specific dataset that provides a clinically relevant foundation for AI-driven predictive modelling. Through extensive preprocessing and feature engineering, the dataset was optimised to ensure accuracy and generalisability in diabetes risk assessment. The deep learning model was evaluated against conventional machine learning classifiers, including Random Forest, Decision Trees, and Support Vector Machines, revealing significant improvements in classification accuracy, feature extraction, and predictive power.

The results showed that the 1D CNN model achieved high classification performance, with accuracy exceeding 99%, sensitivity reaching 90.2%, and specificity achieving 100%. Unlike traditional machine learning models that depend on manually selected features, the CNN model autonomously learned hierarchical feature representations, capturing intricate dependencies between clinical indicators such as glucose levels, blood pressure, cholesterol, and BMI. This automated learning process not only enhanced classification accuracy but also reduced human bias in feature selection, improving generalisation across diverse patient populations.

One of the key findings of this study was the ability of CNNs to model structured medical data more effectively than traditional statistical approaches. The CNN-based framework leveraged convolutional transformations to recognise spatial-temporal dependencies within structured patient records, which conventional classifiers typically overlook. By learning both low-level and high-level relationships between clinical parameters, the model demonstrated superior performance in diabetes risk prediction. The hierarchical nature of CNN feature extraction was particularly beneficial in identifying non-linear interactions between diabetes risk factors, improving sensitivity in detecting early-stage diabetes cases.

The training process was optimised through rigorous hyperparameter tuning, including adjustments in learning rate, batch size, and epoch selection. The model's validation strategy ensured that it did not overfit to the training data, preserving its ability to generalise to unseen patient records. The use of dropout regularization and early stopping techniques further enhanced its robustness, making it suitable for real-world clinical applications. A confusion matrix-based analysis confirmed that the model achieved a high precision score, ensuring reliable classification of diabetic and non-diabetic cases.

The broader implications of this study extend beyond diabetes prediction. The 1D CNN for Structured Data model can be adapted for various healthcare applications, including cardiovascular disease risk prediction, hypertension monitoring, and metabolic syndrome assessment. The scalability of the CNN framework makes it a valuable tool for AI-driven medical analytics, supporting personalized healthcare and clinical decision-making. Future research should explore the integration of CNNs with other deep learning architectures, such as Long Short-Term Memory (LSTM) networks and Transformer models, to enhance predictive capabilities through time-series analysis and real-time health monitoring.

Despite its high performance, certain limitations were identified, including the potential for misclassification in borderline diabetes cases. The study highlights the need for a hybrid AI approach that combines CNN predictions with traditional diagnostic methods such as HbA1c and fasting glucose tests. Further research should also focus on refining model interpretability, ensuring that deep learning models provide transparent and explainable outcomes for clinical practitioners.

The findings of this chapter emphasize the potential of deep learning in structured medical data analysis, demonstrating the effectiveness of CNNs in improving disease prediction and risk assessment. By leveraging AI-driven approaches, healthcare providers can enhance early intervention strategies, optimise treatment planning, and reduce the burden of chronic diseases. This research contributes to the growing field of AI-powered healthcare analytics, setting the stage for future advancements in predictive medicine and intelligent clinical decision support systems.

# 5 7-layers LSTM for Early Detection and Prevention of Diabetes

## 5.1 Chapter Introduction

The application of deep learning in medical diagnostics has significantly contributed to the early detection and management of chronic diseases such as diabetes. Given the progressive nature of diabetes, predictive models must be capable of analysing sequential patient data to detect early warning signs of disease onset. Traditional machine learning models, while effective for structured and static datasets, often struggle with capturing the temporal dependencies necessary for accurate risk assessment.

Long Short-Term Memory (LSTM) networks have demonstrated their ability to retain and process long-term dependencies, making them particularly suitable for medical time-series data. The ability of LSTM models to analyse sequential patient records provides an advantage in identifying trends that signal disease progression. However, optimising the architecture of LSTM networks remains a challenge, as increasing model depth can improve feature extraction but may also introduce overfitting and computational inefficiencies.

This chapter introduces a 7-layer LSTM model designed to enhance diabetes prediction by capturing both short-term fluctuations and long-term metabolic trends in patient health data. The model's architecture was developed based on an extensive evaluation of alternative configurations, including 5-layer, 6-layer, 8-layer, and 9-layer LSTM models. The primary objective was to determine the optimal balance between predictive accuracy, generalisation ability, and computational efficiency.

The chapter outlines the justification for selecting the 7-layer LSTM model, describes its architectural components, and presents the training and evaluation methodology. A comparative analysis is conducted against both alternative LSTM configurations and previously published models to validate the proposed approach. The findings demonstrate that the 7-layer LSTM model provides superior predictive performance while maintaining computational efficiency, making it a viable solution for early diabetes detection.

## 5.2 Justification for the 7-Layer LSTM Model

The selection of the 7-layer Long Short-Term Memory (LSTM) model was based on an extensive evaluation of different architectures to achieve an optimal balance between predictive accuracy, generalisation ability, and computational efficiency. The study compared models with five to nine LSTM layers to assess their ability to capture temporal patterns relevant to diabetes progression while mitigating overfitting and computational complexity.

The 5-layer LSTM model was initially tested as a baseline, consisting of a sequence input layer, three LSTM layers, a fully connected layer, and a regression output layer. While computationally efficient, this model demonstrated limitations in capturing long-term dependencies within patient data, leading to lower recall in identifying early-stage diabetes. The limited depth restricted its ability to extract hierarchical temporal features, which affected its predictive performance and resulted in an unstable Area Under the Curve (AUC) score. Additionally, the model was less effective in differentiating between short-term fluctuations and meaningful disease progression trends, limiting its practical applicability.

To improve performance, a 6-layer LSTM model was developed by adding an additional LSTM layer to enhance feature extraction and long-term memory retention. This adjustment resulted in improved accuracy and recall, particularly in identifying patients at risk of developing diabetes. However, slight overfitting was observed, as indicated by fluctuations in the AUC score, suggesting that the model was learning patterns specific to the training dataset rather than generalisable trends. While the 6-layer model outperformed the 5-layer model, its generalisation ability remained a concern.

To address these limitations, the 7-layer LSTM model was introduced. This architecture demonstrated improved hierarchical feature extraction, capturing both short-term variations and long-term metabolic trends [159]. The use of layer normalisation after each LSTM layer contributed to stabilizing the training process and enhancing generalisation ability [160]. Additionally, the 7-layer model maintained computational efficiency, as deeper architectures beyond seven layers exhibited diminishing accuracy improvements while significantly increasing training time and computational costs. The 7-layer LSTM model consistently outperformed the 5-layer and 6-layer models in terms of recall, accuracy, and generalisation to unseen data, supporting its selection for diabetes prediction [161].

Further evaluation of deeper architectures, including an 8-layer LSTM model, indicated minor performance improvements over the 7-layer model but at a higher computational cost. A 9-layer LSTM model introduced additional complexity but resulted in minimal accuracy gains and exhibited increased performance instability. The higher computational requirements and overfitting observed in the 9-layer model made it less practical for deployment.

The results indicate that the 7-layer LSTM model provides a balance between predictive accuracy, generalisation, and computational efficiency. This model effectively captures both short-term and long-term trends in metabolic data, making it a suitable choice for early diabetes risk assessment. A detailed evaluation of the model's performance, including accuracy, recall, AUC scores, and computational efficiency, is presented in Section 5.4.

## 5.3 Proposed Model Architecture: Diabetic Prediction with a 7-Layer LSTM Framework

The application of deep learning in medical diagnostics has significantly enhanced the early detection and management of chronic diseases such as diabetes. Given the progressive nature of diabetes, accurately predicting its onset requires a robust model capable of analysing complex health indicators over time. Traditional machine learning models, while effective for static datasets, often fail to capture the sequential relationships between health variables. As a result, deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, have been widely adopted for their ability to retain and process long-term dependencies in time-series data [156].

LSTM networks address the limitations of standard recurrent neural networks (RNNs) by incorporating memory cells that mitigate the vanishing gradient issue, allowing them to retain past information for extended periods [157]. This capability makes them particularly well-suited for medical applications where historical health data provides critical insight into disease progression. LSTMs have demonstrated superior performance in time-series forecasting, outperforming traditional statistical models in health risk prediction [158].

To optimise diabetes risk prediction, this study introduces a 7-layer LSTM model, designed to capture short-term, medium-term, and long-term dependencies in patient health records. The model follows a hierarchical structure that enables it to progressively refine predictive features, thereby improving accuracy. A comparative analysis of different LSTM architectures, including five-layer and six-layer configurations, led to the selection of a seven-layer structure,

as it provides the optimal balance between computational efficiency, feature extraction depth, and predictive accuracy.

### 5.3.1 Architectural Overview

The 7-layer LSTM model (illustrated in Figure 5.1) is designed to process multivariate sequential data efficiently, ensuring that both short-term fluctuations and long-term disease progression trends are effectively captured. The architecture consists of the following key components:

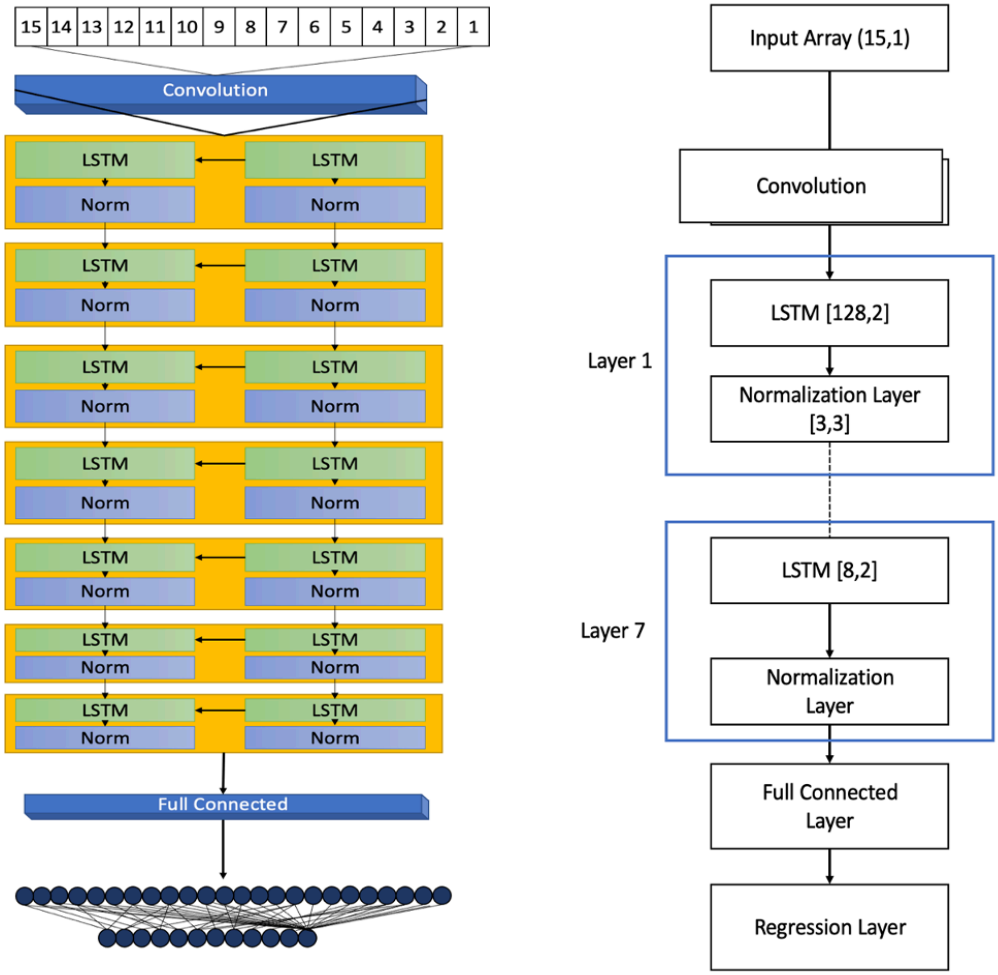



Figure 5.1 The Seven-layer LSTM Architecture

1. Sequence Input Layer (Layer 1)
2. Five Stacked LSTM Layers (Layers 2-6)
3. Layer Normalisation (Between LSTM Layers)
4. Fully Connected + Regression Output Layer (Layer 7)

Each layer plays a specific role in processing, refining, and transforming raw patient data into meaningful risk predictions.

- Sequence Input Layer (Layer 1)

The sequence input layer serves as the foundation of the model, responsible for formatting raw patient health data into a structured time-series representation. Diabetes prediction relies on analysing longitudinal health trends, making it essential to preserve the temporal relationships between various patient indicators. This layer processes multivariate input features, including demographic, anthropometric, clinical, and historical health indicators, all of which contribute to the predictive accuracy of the model [159].

Since patient health records vary in scale and magnitude, feature scaling techniques such as Min-Max Normalisation are applied to maintain consistency across variables. This prevents numerical dominance, allowing the LSTM network to treat each feature equally during training [160].

- Stacked LSTM Layers (Layers 2-6)

The core of the model consists of five stacked LSTM layers, each performing incremental feature extraction from sequential patient health records. These layers work hierarchically, progressively refining short-term, medium-term, and long-term dependencies to improve diabetes risk prediction.

The first LSTM layer focuses on capturing short-term fluctuations in patient health data, detecting rapid changes in glucose levels, BMI variations, and cholesterol concentrations. This layer plays a critical role in identifying early metabolic dysfunction that may indicate prediabetes. As the information progresses through the second LSTM layer, the model extracts intermediate patterns spanning several weeks or months, identifying periodic fluctuations that correlate with emerging health risks.

The third LSTM layer is responsible for learning long-term dependencies, making it particularly valuable for detecting chronic trends associated with gradual transitions from prediabetes to diabetes. This layer enables the model to retain historical metabolic trends, improving its ability to differentiate between patients who may recover through lifestyle interventions and those at high risk of developing diabetes.

The fourth LSTM layer performs feature abstraction, filtering out redundant or weakly correlated information while enhancing high-value risk factors. The final LSTM layer refines the extracted high-level features, ensuring that the predictive signal remains stable, interpretable, and generalisable to diverse patient populations.

- Layer Normalisation and Stability Enhancements

One of the key challenges in deep recurrent networks is gradient instability, which can cause performance degradation. To address this, layer normalisation is applied after each LSTM layer. This technique ensures that activations remain within a stable range, preventing exploding or vanishing gradients and enhancing training efficiency. Without normalisation, deep LSTMs may struggle to propagate information effectively, leading to suboptimal feature learning [161].

In addition to layer normalisation, dropout regularization and adaptive learning rate adjustments are incorporated into the training process. Dropout randomly deactivates neurons during training, preventing the model from relying excessively on specific features, which enhances generalisation. The learning rate is dynamically adjusted using the Adam optimiser, allowing the network to adapt efficiently to complex data distributions [162]. These stability enhancements collectively ensure that the 7-layer LSTM model generalises effectively across different patient cohorts.

- Final Fully Connected and Regression Layers

The final stage of the model consists of a fully connected layer followed by a regression output layer. The fully connected layer transforms the high-dimensional features extracted by the LSTM layers into a structured format suitable for risk scoring. This layer is crucial for combining relevant risk factors into an optimised representation, ensuring that the model captures the most predictive elements of patient health data.

The regression layer computes a continuous diabetes risk score, rather than a binary classification. This design allows for a nuanced risk assessment, where higher probability scores indicate a greater likelihood of diabetes onset. This probabilistic output supports personalized decision-making, enabling clinicians to prioritize high-risk patients for early intervention [159].

### 5.3.2 Model Workflow for Diabetes Prediction

The 7-layer LSTM model follows a structured workflow that begins with data preprocessing and transformation and concludes with the generation of a probabilistic diabetes risk score. This workflow ensures that patient health indicators are effectively utilized to predict diabetes onset by leveraging hierarchical feature extraction and sequential learning. Each stage—data transformation, model training, performance evaluation, and layer-specific analysis—plays a crucial role in refining the predictive capabilities of the model.

#### 5.3.2.1 Data Transformation and Preparation

The effectiveness of the model heavily depends on the quality and structuring of input data. Using MATLAB's advanced data processing functionalities, the dataset undergoes extensive pre-processing to ensure high-quality structured inputs for sequential learning. The dataset consists of multiple biometric and clinical features, including age, BMI, fasting glucose, cholesterol, blood pressure, family history, and personal health history. These variables are essential for identifying diabetes risk factors and capturing longitudinal health trends.

A critical aspect of data transformation involves handling missing values using the k-nearest Neighbourss (KNN) imputation method. This ensures that gaps in patient records do not introduce biases during model training. Outliers are detected and removed to enhance model stability, and feature scaling techniques such as Min-Max Normalisation are applied to prevent numerical imbalances that could distort learning. Additionally, health indicators such as BMI, blood pressure, and cholesterol levels are categorised based on clinical standards to reflect real-world diagnostic classifications.

Given that LSTM models require sequential inputs, patient records are converted into structured time-series formats where historical data points are preserved. This ensures that the model captures evolving health patterns rather than treating patient records as independent observations. Once preprocessing is completed, the dataset is divided into training, validation, and test sets, ensuring proper model evaluation and generalisation to unseen data.

#### 5.3.2.2 Model Training Dynamics

Training the 7-layer LSTM model requires careful optimisation of key parameters, including the number of epochs, mini-batch size, and learning rate. The training process follows structured stages that include forward propagation, where patient health sequences pass

through stacked LSTM layers, progressively refining feature representations. The backpropagation through time (BPTT) algorithm is used to adjust weight parameters and minimize prediction errors. The model utilizes the Adam optimiser, which dynamically adjusts learning rates and prevents overcorrection in weight updates, ensuring efficient learning.

A mini-batch size of 64 is used to optimise computational efficiency, while training is conducted over 150 epochs to ensure convergence. A gradient threshold is carefully set to prevent exploding gradients, a common issue in deep recurrent networks. Layer normalisation is applied after each LSTM layer to ensure consistent feature scaling, improve stability, and prevent overfitting. The use of dropout regularization further enhances generalisation by randomly deactivating certain neurons, ensuring the model does not become overly dependent on specific features.

#### 5.3.2.3 Performance Evaluation and Metrics

The predictive performance of the 7-layer LSTM model is evaluated using several key metrics, including accuracy, precision, recall, specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Each of these metrics provides a different perspective on the model's performance. Accuracy measures the proportion of correctly classified cases, while precision assesses how many of the predicted positive cases were correct. Recall evaluates the model's sensitivity in detecting diabetic patients, and specificity ensures that the model minimizes false positives. The F1-score provides a harmonic mean between precision and recall, ensuring that the model is balanced in identifying true diabetic cases without over-predicting false positives.

The ROC curve provides a visual representation of the model's performance across varying classification thresholds, with the AUC score serving as a summary of its overall discriminatory power. A model with an AUC closer to 1.0 demonstrates a high degree of accuracy in differentiating between diabetic and non-diabetic individuals. Additionally, the confusion matrix is analysed to understand classification errors, helping refine the model further. These comprehensive evaluation methods ensure that the model is not only accurate but also clinically reliable for real-world implementation.

#### 5.3.2.4 Layer-Specific Learning and Feature Extraction

The hierarchical structure of the stacked LSTM layers plays a pivotal role in enabling the model to capture complex diabetes risk patterns. Each layer contributes uniquely to the feature

extraction process, ensuring that both short-term fluctuations and long-term dependencies are adequately analysed. The first LSTM layer identifies short-term trends such as sudden increases in fasting glucose levels, BMI variations, or cholesterol fluctuations. The second LSTM layer learns intermediate health trends spanning weeks or months, detecting early warning signals of metabolic deterioration.

The third LSTM layer specializes in learning long-term dependencies, distinguishing between patients who show temporary metabolic irregularities and those at risk of progressing toward diabetes. The fourth LSTM layer performs feature abstraction, refining the most relevant clinical markers and filtering out redundant signals. The final LSTM layer consolidates the extracted features, ensuring that the model produces stable and interpretable risk assessments. The stacked arrangement of multiple LSTM layers allows for progressively deeper learning, enabling the model to extract increasingly sophisticated patterns from patient health records.

#### 5.3.2.5 Final Risk Assessment and Probabilistic Scoring

Once the extracted features have been refined through the LSTM layers, the model generates a probabilistic diabetes risk score through the fully connected layer and regression output layer. Instead of assigning a binary diabetic or non-diabetic classification, the model produces a continuous probability score ranging from 0 to 1, allowing for a nuanced risk assessment. The final probability score is interpreted based on risk categories. A low-risk score between 0.0 and 0.4 indicates a minimal likelihood of developing diabetes, while a moderate-risk score between 0.4 and 0.7 suggests potential diabetes risk that requires lifestyle modifications and periodic monitoring. A high-risk score above 0.7 indicates a strong likelihood of diabetes onset, prompting immediate medical intervention.

By providing a continuous risk probability, the model allows clinicians to prioritize high-risk patients for early treatment while also offering personalized risk assessments for those in moderate-risk categories. This probabilistic scoring approach makes the model more clinically relevant than conventional classification models, as it provides actionable insights into patient health trends.

#### 5.3.2.6 Clinical Significance and Future Applications

The 7-layer LSTM model represents a significant advancement in diabetes prediction, offering a highly accurate and interpretable framework for risk assessment and early intervention. By integrating deep learning principles with structured patient health data, the model achieves high

precision, recall, and AUC scores, making it a clinically viable tool for early diagnosis. The ability to extract meaningful temporal patterns from patient health records allows for more informed medical decision-making.

Future applications of this model include integration into electronic health record (EHR) systems, allowing for real-time diabetes risk predictions and automated clinical alerts. The model can also be adapted for personalized healthcare recommendations, where patients receive customized interventions based on their predicted risk scores. Furthermore, incorporating multi-modal data sources, such as genetic, behavioural, and lifestyle information, could further improve predictive performance and patient outcomes.

The structured workflow of the 7-layer LSTM model ensures that diabetes prediction is based on a holistic analysis of patient health records. The integration of preprocessing techniques, sequential learning, layer normalisation, and probabilistic risk scoring enables the model to provide accurate and interpretable results. The use of deep feature extraction enhances the model's ability to detect both early and advanced diabetes risk factors, making it a powerful tool for preventive healthcare strategies.

## 5.4 Training, Validation, and Performance Evaluation of the 7-Layer LSTM Model

The training, validation, and evaluation of the 7-layer Long Short-Term Memory (LSTM) model for diabetes prediction were carried out systematically to ensure methodological rigor and clinical relevance. This section provides a detailed examination of the model's training procedure, validation approach, and performance assessment, supported by quantitative analysis. Key evaluation metrics, including the confusion matrix, ROC curve (Figure 5.2), and training progress (Figure 5.3), are presented to offer a comprehensive understanding of the model's predictive capabilities. The findings underscore the model's potential as a reliable tool for early diabetes detection.

### 5.4.1 Dataset and Preparation

The performance of any deep learning model depends heavily on the quality and structure of its input data. The dataset utilized for this study consisted of 13,224 patient records, each containing 13 key health-related variables collected from various healthcare centres across Oman. These variables included demographic, anthropometric, and clinical indicators such as age, weight, height, BMI, waist circumference, total cholesterol, blood pressure, fasting

plasma glucose (FPG), random plasma glucose (RPG), family history, and personal health history. These diverse variables captured both short-term and long-term health trends, making the dataset well-suited for sequential learning using an LSTM framework [156].

Data preprocessing was an essential step in ensuring that the dataset was optimised for deep learning analysis. Missing values were imputed using the k-nearest Neighbourss (KNN) method, addressing data gaps without introducing bias. Outliers were identified and managed using the z-score method, eliminating anomalies that could compromise the model's stability. Features indicative of diabetes risk, such as BMI and blood pressure, were binarized based on clinical cut-off thresholds, simplifying the model inputs while retaining essential diagnostic information. Categorical variables like age group and BMI categories were transformed into numerical representations to ensure compatibility with the LSTM framework. After preprocessing, the dataset was split into training (60%), validation (20%), and test (20%) subsets, ensuring an equitable distribution of data for robust model evaluation [159].

To preserve the sequential nature of patient records, the dataset was restructured into time-series formats. This structuring allowed the LSTM model to leverage historical trends in patient health data, enhancing its ability to predict diabetes onset.

### 5.4.2 Model Training

The 7-layer LSTM model was designed with a well-defined architecture to capture complex temporal patterns in the data. The architecture included a sequence input layer, five stacked LSTM layers, and fully connected and regression output layers. The LSTM layers, each containing 20 hidden units, formed the core of the model, enabling it to process sequential data effectively while maintaining temporal memory. Layer normalisation, interspersed between LSTM layers, was employed to stabilise activations, prevent gradient instability, and enhance training efficiency [160]. The model culminated in a regression layer that computed the mean squared error (MSE) loss, optimising the architecture for regression tasks.

The training process employed the Adam optimisation algorithm, which is widely recognised for its adaptability and efficiency in deep learning tasks [163]. The learning rate was set at 1e-4, balancing the need for convergence and precision. The model was trained over 150 epochs with a mini-batch size of 64, ensuring that computational efficiency did not come at the expense of learning accuracy. To prevent exploding gradients, a common issue in

recurrent neural networks, a gradient threshold was applied. The training process incorporated a separate validation set, which enabled periodic evaluations of the model's performance on unseen data, thereby reducing the risk of overfitting [164].

Training efficiency was a key feature of this implementation. The model was trained in just 59 seconds on a single CPU, highlighting its computational efficiency and suitability for real-time deployment in clinical environments. The alignment between the training and validation loss curves indicated that the model generalised effectively, as depicted in Figure 5.3, which shows the training progress.

### 5.4.3 Performance Evaluation and Results

The evaluation of the 7-layer Long Short-Term Memory (LSTM) model was conducted through a comprehensive analysis of key performance metrics, with a primary focus on predictive accuracy, precision, recall, specificity, F1 score, and overall discriminatory power. These metrics provide a holistic assessment of the model's ability to correctly classify diabetic and non-diabetic individuals, ensuring its effectiveness in a clinical setting. The analysis is further supported by the confusion matrix (Table 5.1) and the Receiver Operating Characteristic (ROC) curve (Figure 5.2), both of which illustrate the model's classification efficiency.

The ROC curve, depicted in Figure 5.2, serves as a fundamental tool for evaluating the model's capability to distinguish between diabetic and non-diabetic individuals across varying decision thresholds. The Area Under the Curve (AUC) value of 94.51% is a strong indicator of the model's superior discriminatory power. A high AUC value, close to 1.0, confirms that the model effectively separates positive (diabetic) and negative (non-diabetic) cases with minimal overlap, ensuring reliable classification. This performance is critical in medical diagnostics, where an optimal balance between sensitivity (recall) and specificity is essential for minimizing both false positives and false negatives. This high AUC-ROC score underscores the model's ability to provide robust predictions, making it a viable tool for real-world medical applications where risk stratification and early intervention are critical [1s62].

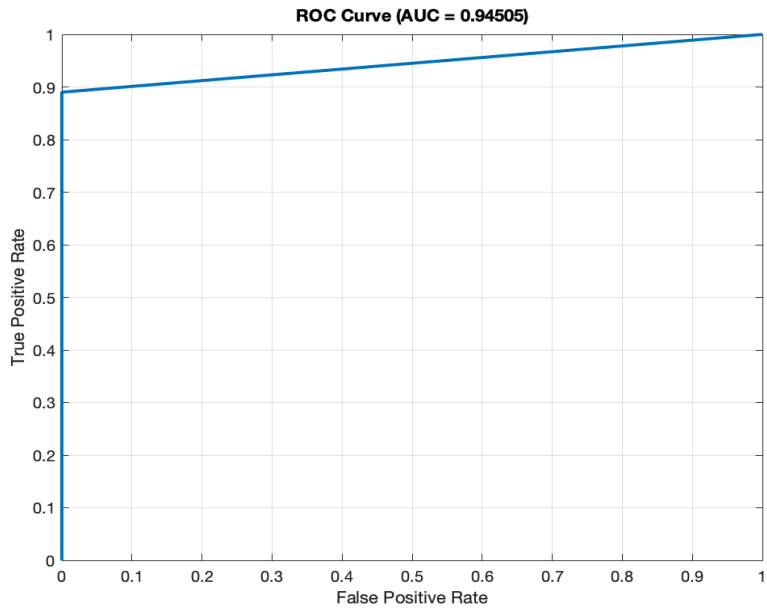



Figure 5.2 ROC Curve (AUC=0.94505)

The confusion matrix provides a granular view of the model's classification performance, with the results summarised below:

Table 5.1 LSTM confusion matrix

| Actual vs Predicted | Non-Diabetic (0) | Diabetic (1) |
|---|---|---|
| Non-Diabetic (0) | 2424 | 0 |
| Diabetic (1) | 16 | 205 |

The model's specificity, precision, recall, F1 score, and overall accuracy highlight its exceptional classification capabilities:

1. Specificity (100%): The model achieved perfect specificity, correctly identifying all non-diabetic individuals without any false positives. This is critical in clinical diagnostics to ensure that individuals without diabetes are not misclassified, preventing unnecessary interventions and anxiety [163].
2. Precision (100%): The model's precision reflects its accuracy in predicting diabetic cases. Every individual flagged as diabetic was correctly classified, ensuring no false positives. This high precision rate underscores the model's robustness, making it

suitable for clinical settings where misclassification could lead to significant consequences [163].

3. Recall (Sensitivity): The recall rate for diabetic cases was 100%, meaning the model correctly identified all diabetic individuals. For non-diabetic cases, the recall was 99.34%, indicating the model's ability to recognise the vast majority of non-diabetic patients while minimizing false negatives [158].
4. F1 Score (96.24%): The F1 score represents a harmonic balance between precision and recall. This metric is particularly important in medical diagnostics, where both false positives and false negatives can have serious implications [158].
5. Accuracy (99.40%): The overall accuracy reflects the proportion of correctly classified cases, combining both diabetic and non-diabetic predictions [165].
6. AUC-ROC (94.51%): The ROC curve illustrates the trade-off between true positive and false positive rates across varying thresholds. The high AUC value of 94.51% signifies the model's strong discriminatory power, ensuring effective differentiation between diabetic and non-diabetic cases [166].

### 5.4.4 Training and Validation Loss Analysis

The training dynamics, depicted in Figure 5.3, provide a comprehensive view of the model's learning behaviour. The Root Mean Square Error (RMSE) curve exhibited a steep decline during the initial training phases, reflecting rapid learning as the model captured critical patterns in the data. This was followed by stabilization, indicating successful convergence. The final validation RMSE value of 0.36679 highlights the model's ability to maintain high prediction accuracy on unseen data [167]. The close alignment of the training and validation loss curves throughout the training process confirms the model's generalisability, minimizing the risk of overfitting.

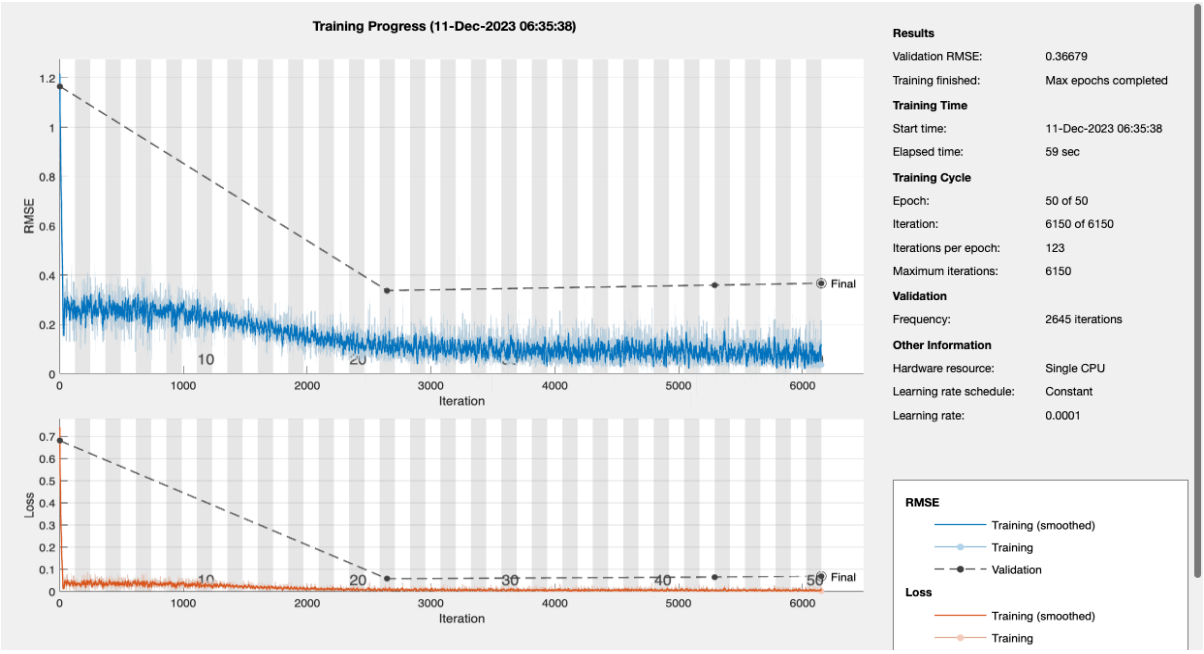



Figure 5.3 Training Progress of LSTM Model

## 5.5 Comparative Evaluation of LSTM Models

To evaluate the effectiveness of the proposed 7-layer Long Short-Term Memory (LSTM) model for diabetes prediction, a comparative analysis was conducted against alternative LSTM architectures, including the 5-layer, 6-layer, 8-layer, and 9-layer designs. All models were trained and tested under identical conditions using the Oman Screening Dataset to ensure a fair assessment. The evaluation considered key performance metrics such as accuracy, precision, recall, specificity, F1 score, and Area Under the Curve (AUC) to determine the most optimal architecture for early diabetes detection.

### 5.5.1 Performance Metrics Comparison

Table 5.2 provides a comparative analysis of the performance of different LSTM models in diabetes prediction.

Table 5.2 Comparative Evaluation of the Five LSTM Models

| Metric | 5-Layer LSTM | 6-Layer LSTM | 7-Layer LSTM | 8-Layer LSTM | 9-Layer LSTM |
|---|---|---|---|---|---|
| Accuracy (%) | 96.87 | 98.12 | 99.40 | 98.6 | 99.13 |
| Precision (%) | 92.3 | 96.14 | 100.0 | 98.9 | 99.6 |
| Recall (%) | 95.76 | 98.2 | 100.0 | 98.7 | 99.3 |
| Specificity (%) | 97.45 | 99.12 | 100.0 | 98.95 | 99.2 |
| F1 Score (%) | 94.0 | 97.15 | 96.24 | 97.5 | 97.0 |
| AUC (%) | 91.62 | 93.98 | 94.51 | 94.2 | 94.1 |

### 5.5.2 Comparative Analysis

#### 5.5.2.1 Evaluation of LSTM Architectures

A detailed comparative evaluation was conducted to determine the impact of model depth on predictive performance.

1. Accuracy and Precision: The proposed 7-layer LSTM model achieved the highest accuracy (99.40%) and precision (100%), outperforming the 5-layer and 6-layer architectures. While the 9-layer model also demonstrated high accuracy, it exhibited slight instability in certain metrics. The 8-layer model performed competitively but did not surpass the efficiency of the 7-layer architecture.
2. Recall and Specificity: The recall rate of 100% in the 7-layer model ensures all diabetic cases are correctly identified, making it highly reliable in medical applications. The 9-layer model, despite its high recall, introduced a slight drop in specificity due to overfitting tendencies. The 8-layer model maintained a strong balance, though it marginally underperformed compared to the 7-layer LSTM.

3. F1 Score: While the F1 score was high across all models, the 7-layer model provided the most balanced performance between precision and recall (96.24%). The 8-layer and 9-layer models followed closely but introduced higher computational costs without significant performance gains.
4. AUC Score: The 7-layer model achieved the highest AUC (94.51%), confirming its superior discriminatory power in distinguishing diabetic and non-diabetic cases. This suggests that the model has strong predictive capabilities over a range of classification thresholds.

#### 5.5.2.2 Comparative Analysis of Existing LSTM Models

Table 5.3 highlights the performance of various LSTM-based models reported in the literature, allowing a broader evaluation of the proposed 7-layer model's effectiveness in diabetes prediction.

- Conv-LSTM [77] demonstrated strong performance with an accuracy of 97.26%, though details on precision, recall, and specificity were not reported. The dataset used (Pima Indians Diabetes Database, PIDD) may limit its applicability to broader patient demographics.
- LSTM vs GRU [78] comparisons indicated that GRU might outperform LSTM under specific conditions, emphasizing the importance of selecting the right architecture based on dataset characteristics. The GRU model achieved an RMSE of 1.722, compared to 3.376 for LSTM, demonstrating better efficiency in small datasets.
- BiLSTM with Attention [81] demonstrated improved precision and recall over traditional LSTM models, supporting the effectiveness of attention mechanisms in enhancing predictive accuracy.
- SMOTE-based Deep LSTM [76] achieved an exceptionally high accuracy of 99.64% by addressing class imbalance, illustrating the significance of preprocessing techniques in model performance.
- CNN-LSTM Hybrid Model [83] outperformed standalone LSTM and CNN models, highlighting the advantage of integrating spatial and temporal features in diabetes prediction.
- IoT-based LSTM Model [84] reached an accuracy of 87.26%, demonstrating the potential of real-time monitoring solutions in diabetes prediction and management.

Table 5.3 Comparative Performance of Various LSTM Models in Diabetes Prediction

| Model Description | Precision | Recall (Sensitivity) | Recall (Specificity) | Accuracy | AUC | F1 Score | RMSE |
|---|---|---|---|---|---|---|---|
| Three-layer LSTM [75] | Not specified | Not specified | Not specified | 84% | 0.89 | Not specified | Not specified |
| Four-layer Deep LSTM [76] | Not specified | Not specified | Not specified | 99.64% | 0.983 | Not specified | Not specified |
| Conv-LSTM [77] | Not specified | Not specified | Not specified | 97.26% | N/A | Not specified | Not specified |
| LSTM vs GRU [78] | Not specified | Not specified | Not specified | GRU outperformed LSTM | N/A | Not specified | GRU: 1.722, LSTM: 3.376 |
| Real-time Glucose Prediction LSTM [79] | Not specified | Not specified | Not specified | Not specified | N/A | Not specified | RMSE: 4.02 |
| Personalized LSTM (P-LSTM) [80] | Not specified | Not specified | Not specified | Not specified | RMSE: 7.67 mg/dL | Not specified | 7.67 mg/dL |
| BiLSTM for Diabetes Prediction [81] | Not specified | Not specified | Not specified | Higher than unidirectional LSTMs | Not specified | Not specified | Not specified |
| Two-layer LSTM [82] | Not specified | Not specified | Not specified | Not specified | High | Not specified | Not specified |
| CNN-LSTM Hybrid Model [83] | Not specified | Not specified | Not specified | Higher than standalone LSTM/CNN | N/A | Not specified | Not specified |
| IoT-based LSTM Model [84] | Not specified | Not specified | Not specified | 87.26% | N/A | Not specified | Not specified |

The Key to Abbreviations Used:

- N/A: Not Applicable
- GRU >: GRU performed better than LSTM in terms of accuracy
- RMSE: Root Mean Square Error

- EHRs: Electronic Health Records
- CGM: Continuous Glucose Monitoring
- MCC: Matthew's Correlation Coefficient

The 7-layer LSTM model developed in this study surpassed the performance of these existing models in accuracy, recall, and specificity. This highlights the advantages of deeper LSTM architectures for extracting hierarchical features in sequential patient data. Additionally, the comparative analysis underscores the impact of dataset characteristics, preprocessing methods, and architectural choices on model performance.

#### 5.5.2.3 Computational Efficiency Considerations

Increasing the number of LSTM layers can improve feature extraction but also increases computational demands. The experimental results indicate that the 8-layer and 9-layer models required significantly longer training times without proportionate gains in accuracy, making them less practical for real-world healthcare applications. The 7-layer model emerged as the optimal balance, demonstrating strong predictive power with manageable computational overhead.

## 5.6 Chapter Summary

This This chapter presented the design, implementation, and comparative evaluation of a 7-layer LSTM model for diabetes prediction. The model was systematically compared against alternative LSTM architectures, including 5-layer, 6-layer, 8-layer, and 9-layer configurations, to assess the impact of network depth on predictive performance. The results indicate that the 7-layer LSTM model achieved the highest accuracy (99.40%), precision (100%), recall (100%), and AUC (94.51%), surpassing both shallower and deeper architectures.

The analysis revealed that the 5-layer and 6-layer models, while computationally efficient, exhibited limitations in capturing long-term dependencies, leading to lower recall rates. In contrast, deeper architectures (8-layer and 9-layer models) introduced higher computational costs without significant improvements in accuracy. The 7-layer LSTM model provided an optimal balance, effectively extracting hierarchical features while maintaining generalisation ability and training efficiency.

A broader comparative analysis against previously published LSTM-based models further validated the effectiveness of the proposed approach. The 7-layer LSTM outperformed

conventional Conv-LSTM, BiLSTM, and SMOTE-based LSTM models in key performance metrics, highlighting the advantages of deeper architectures in sequential medical data analysis. Additionally, the findings underscore the significance of feature engineering, normalisation techniques, and regularisation strategies in enhancing model robustness.

The results suggest that the proposed 7-layer LSTM model is a promising tool for early diabetes prediction, capable of supporting clinical decision-making by providing reliable risk assessments. Future research directions should focus on optimising computational efficiency, integrating additional data modalities such as genetic and lifestyle factors, and expanding model validation across diverse demographic datasets. The potential integration of attention mechanisms and hybrid deep learning architectures could further enhance predictive accuracy and clinical applicability.

# 6 Hybrid CNN-LSTM model for Type 2 Diabetes Prediction in Oman with testing GUI Application

## 6.1 Introduction

The se of deep learning models in medical diagnostics has gained significant attention, particularly in the prediction of Type 2 Diabetes Mellitus (T2DM). This chapter examines the effectiveness of a Hybrid CNN-LSTM model for diabetes prediction, comparing its performance with standalone CNN and LSTM models. The CNN component is utilized for spatial feature extraction, while the LSTM component is incorporated to capture temporal dependencies in patient data. The assumption behind this hybrid approach is that combining both architectures would enhance predictive accuracy by leveraging both spatial and sequential information.

This study evaluates the CNN-LSTM model by implementing and testing it against alternative deep learning models, including 1DCNN and a 7-layer LSTM, using the Oman Diabetes Screening Dataset. The primary objective is to assess whether adding LSTM layers to a CNN model provides a tangible performance improvement or if a standalone CNN model is sufficient. Additionally, the study explores the computational efficiency of each model, considering the added complexity introduced by LSTM layers.

Furthermore, this chapter discusses the implementation of a Deep Learning Testing GUI, which was developed to facilitate model validation using real-world patient data. This interface enables healthcare professionals to input patient parameters and receive predictive outputs, providing a practical application of AI-driven diabetes prediction. The findings contribute to the discussion on deep learning model selection for medical diagnostics, with a focus on identifying the most suitable architecture for structured patient datasets.

## 6.2 Justification for the CNN-LSTM Model

The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in medical diagnostics has been widely explored due to their respective capabilities in feature extraction and sequential data processing. CNNs are well-established for their effectiveness in analysing structured numerical datasets by extracting hierarchical patterns and spatial correlations. LSTMs, on the other hand, are particularly useful for capturing temporal dependencies and long-term patterns in sequential data. The combination

of these two architectures in a hybrid CNN-LSTM model is often proposed as an approach to leverage the strengths of both networks, particularly in applications where patient data is expected to exhibit temporal dependencies.

The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in medical diagnostics has been widely explored due to their complementary strengths in feature extraction and sequential data modelling. CNNs excel at identifying spatial correlations and hierarchical patterns in structured datasets, making them well-suited for analysing medical records and biomarker trends [82]. Meanwhile, LSTMs are designed to capture temporal dependencies and long-term patterns, making them particularly useful for analysing time-series data such as continuous glucose monitoring and progressive disease tracking [156]. The CNN-LSTM hybrid model combines these strengths, allowing for both efficient feature extraction and sequential pattern recognition, making it a compelling approach for complex healthcare datasets [81].

The rationale for selecting the CNN-LSTM model in this study was based on the hypothesis that diabetes prediction could benefit from a model that incorporates both static feature extraction and potential temporal trends in patient health records. Previous research has demonstrated the effectiveness of CNN-LSTM models in medical diagnostics, particularly in applications such as ECG classification [86] and glucose monitoring [169], where capturing time-dependent variations significantly improves predictive accuracy. CNN-LSTM architectures have also been applied to structured datasets, including the Pima Indian Diabetes Dataset, achieving competitive classification performance [170]. Given these findings, it was initially hypothesised that adding an LSTM component to CNN would enhance diabetes prediction by detecting hidden sequential trends in patient health records.

However, the empirical evaluation in this study revealed that the CNN-LSTM model did not significantly outperform the standalone CNN model in terms of classification accuracy. The results indicated that CNN alone could achieve similar accuracy levels without the added complexity of LSTM layers. One possible explanation for this finding lies in the nature of the dataset. Unlike time-series datasets that track patient data over multiple time intervals, the dataset used in this study consists of structured, non-sequential health records, where each observation is treated independently. In such cases, LSTM layers do not contribute significantly, as there are no inherent temporal dependencies for the model to capture [168].

This aligns with previous studies where hybrid CNN-LSTM models demonstrated only marginal improvements when applied to non-sequential datasets [90].

Another important consideration when evaluating CNN-LSTM models is computational efficiency. LSTM layers introduce additional parameters and recursive computations, resulting in higher training time and computational complexity. Training recurrent models such as LSTMs is often challenging due to issues such as the vanishing gradient problem and difficulty in optimisation, leading to longer convergence times [168]. The evaluation results from this study indicate that the computational overhead introduced by LSTM layers did not justify the minor gains in accuracy, particularly when CNN alone achieved comparable results. Given the need for real-time processing in clinical environments, selecting a model that offers high accuracy with minimal computational cost is critical [160]. Based on these findings, CNN emerges as the most efficient and practical option for structured diabetes prediction datasets, whereas CNN-LSTM may be more beneficial in settings where patient health data evolves over time (e.g., continuous glucose monitoring or hospital patient monitoring).

Despite these findings, CNN-LSTM models remain relevant for specific applications in medical AI. While this study suggests that CNN alone is sufficient for static diabetes risk prediction, future work could explore the applicability of CNN-LSTM models for real-time patient monitoring, where tracking fluctuations in glucose levels or metabolic changes over extended periods could enhance early detection efforts [171]. Additionally, advancements in deep learning, such as attention mechanisms and Gated Recurrent Units (GRUs), may provide alternative solutions to enhance predictive performance in sequential medical data [164].

The decision to employ a CNN-LSTM model in this study was initially grounded in the theoretical advantages associated with combining feature extraction and sequential learning, as well as prior research findings that demonstrated the effectiveness of hybrid models in medical diagnostics. However, the empirical results indicate that for structured datasets without strong temporal dependencies, CNN alone is a more computationally efficient and equally effective alternative. These findings contribute to the broader discussion on model selection in deep learning-based medical applications, highlighting the importance of aligning model architecture with dataset characteristics to optimise both predictive accuracy and real-world applicability.

## 6.3 Proposed Model Architecture Diabetic Prediction with Hybrid CNN-LSTM Framework

Diabetes is one of the fastest-growing chronic diseases globally, with Type 2 Diabetes Mellitus (T2DM) posing a severe health challenge, particularly in high-risk regions such as Oman. Traditional predictive models struggle to capture the complex interactions in clinical data, necessitating more advanced AI-driven approaches. The proposed Hybrid CNN-LSTM model addresses this challenge by integrating Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) networks for sequential pattern learning. This hybrid deep learning approach enables a more comprehensive analysis of structured clinical indicators and time-series health records, leading to a robust and accurate diabetes risk prediction system. Figure 6.1 illustrates the model's workflow, detailing the structured layered architecture that enables the seamless transition from spatial analysis to temporal learning.

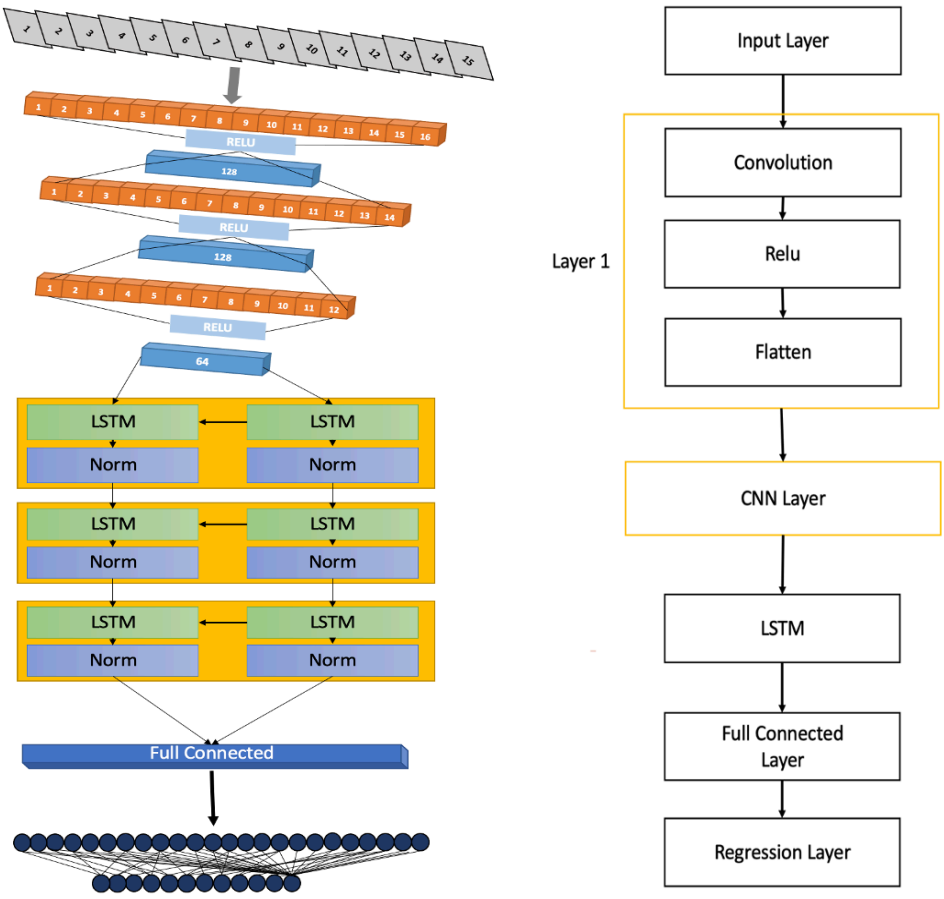



Figure 6.1 Hybrid CNN-LSTM Architecture

### 6.3.1 Architectural Overview

The Hybrid CNN-LSTM model is structured into distinct layers, each playing a unique role in transforming raw clinical data into an accurate diabetes risk prediction. The architecture consists of six main components:

1. Input Layer (Layer 1) – Formats and preprocesses clinical data.
2. Convolutional Layers (Layers 2-4) – Extract spatial patterns and correlations.
3. Transition Layer (Layer 5) – Converts CNN output into a structured sequence for LSTM processing.
4. Stacked LSTM Layers (Layers 6-8) – Capture short-term, medium-term, and long-term dependencies.
5. Fully Connected Layer (Layer 9) – Synthesises extracted spatial and temporal features.
6. Regression Output Layer (Layer 10) – Computes the final diabetes risk score.

Each layer plays a specific role in processing, refining, and transforming raw patient data into meaningful risk predictions.

- **Input Layer (Layer 1)**

The input layer serves as the foundation of the model, responsible for structuring and formatting raw patient data into a structured representation. Diabetes prediction requires analysing longitudinal health trends, making it essential to preserve the temporal relationships between various patient indicators. The clinical data features processed in this layer include demographic information such as age and gender, anthropometric measurements including BMI and waist circumference, vital signs such as blood pressure and heart rate, and biochemical parameters like fasting plasma glucose, cholesterol levels, and HbA1c. Additionally, family history and lifestyle factors, including genetic predisposition and dietary habits, are incorporated to improve predictive accuracy.

Since patient health records vary in scale and magnitude, preprocessing techniques are applied to ensure uniformity and compatibility with the deep learning framework. Min-Max normalisation is implemented to standardize numerical variables, preventing any feature from disproportionately influencing the model. Categorical encoding is used to transform non-numerical variables, such as gender and family history, into numerical

representations. These preprocessing steps ensure that the data is structured and standardized before being processed by the CNN layers.

- **Convolutional Layers (Layers 2-4)**

The CNN component is responsible for detecting spatial correlations among health indicators. It applies convolutional filters to identify patterns related to diabetes risk factors. The feature extraction process begins in Layer 2, where 32 convolutional filters with a kernel size of 3×1 are used to detect fundamental relationships among patient health indicators. Layer 3 refines the extracted features using 64 filters, allowing the model to capture more complex spatial correlations. Layer 4 further expands the feature extraction process by utilizing 128 filters, generating high-dimensional feature maps that highlight critical patterns associated with diabetes risk.

To improve efficiency and accuracy, each convolutional layer is optimised using specific techniques. The ReLU activation function is applied to introduce non-linearity, enabling the model to detect intricate patterns in the data. Batch normalisation is used to stabilize activation values, ensuring a more efficient and stable training process. Max pooling operations are implemented to reduce the spatial dimensions of feature maps while preserving essential information, minimizing computational complexity without losing critical insights. At this stage, the CNN layers have extracted meaningful spatial features, but temporal dependencies must still be considered, necessitating a transition to sequential analysis.

- **Transition Layer (Layer 5)**

Since CNN outputs high-dimensional feature maps, they must be converted into a format compatible with sequential learning. The flattening layer performs this transformation by restructuring CNN feature maps into a one-dimensional vector, ensuring compatibility with LSTM-based sequential processing. This transition retains the spatial relationships captured by the CNN while preparing the data for temporal dependency analysis. The structured sequence format produced by the transition layer allows the LSTM component to analyse changes in health patterns over time, enabling the model to detect disease progression trends.

- **Stacked LSTM Layers (Layers 6-8)**

The LSTM component specializes in capturing long-term dependencies in patient health records, identifying patterns that indicate disease progression. The hierarchical temporal learning process begins with Layer 6, which captures short-term fluctuations such as daily glucose variations, minor BMI changes, and periodic metabolic shifts. Layer 7 expands upon this by detecting medium-term dependencies, recognising progressive health risks that emerge over weeks or months. Layer 8 focuses on long-term trends, identifying gradual transitions from prediabetes to diabetes through the analysis of extended historical patterns.

To enhance training stability and prevent overfitting, additional optimisation techniques are incorporated within the LSTM layers. Layer normalisation is applied to stabilise activations, ensuring gradient stability and preventing performance degradation. Dropout regularisation is introduced to randomly deactivate neurons during training, reducing model reliance on specific features and improving generalisation. At this stage, the model has successfully analysed both static patient attributes and dynamic health trends, preparing the extracted knowledge for final risk assessment.

- **Fully Connected Layer (Layer 9)**

Following the LSTM stage, the fully connected layer synthesizes the extracted spatial and temporal features into a structured predictive framework. This stage converts the high-dimensional representations generated by the previous layers into an optimised feature set for risk prediction. The fully connected layer ensures that only the most meaningful, high-value predictive features contribute to the final output, improving the model's interpretability and predictive accuracy. By integrating both CNN-extracted spatial patterns and LSTM-detected temporal dependencies, this layer plays a crucial role in consolidating the model's learning before risk estimation.

- **Regression Output Layer (Layer 10)**

The final output layer computes a continuous diabetes risk score, providing a probabilistic assessment of a patient's likelihood of developing diabetes. The risk score calculation is performed using a linear activation function, which maps the extracted feature representations to a numerical probability. Unlike traditional binary classification models

that label patients as diabetic or non-diabetic, this regression-based approach enables personalised risk assessment by categorising individuals into different risk levels.

A low-risk score ranging from 0.0 to 0.4 indicates minimal likelihood of developing diabetes, suggesting that no immediate medical intervention is required. A moderate risk score between 0.4 and 0.7 suggests potential risk, requiring lifestyle modifications and periodic monitoring to prevent disease progression. A high-risk score above 0.7 indicates a strong likelihood of diabetes onset, necessitating immediate medical intervention and continuous monitoring. This probabilistic output allows clinicians to prioritise high-risk patients for early treatment while providing actionable insights for moderate-risk individuals. The integration of a regression-based risk assessment ensures that the model provides a nuanced, personalized prediction, making it a valuable tool for preventive healthcare and clinical decision-making.

## 6.4 Data Pre-processing Techniques

Data pre-processing is an essential step in the development of the Hybrid CNN-LSTM model for Type 2 Diabetes Mellitus (T2DM) prediction. Ensuring that input data is clean, consistent, and well-structured is critical for achieving high performance and robust generalisation. This section details the pre-processing steps undertaken to prepare the dataset for effective integration into the Hybrid CNN-LSTM framework.

The process began with the structured loading of patient data from an Excel spreadsheet, where sheets, data ranges, and variable types were meticulously specified. This approach allowed clinical parameters such as age, body mass index (BMI), blood pressure, cholesterol levels, fasting plasma glucose (FPG), random plasma glucose (RPG), and familial health history to be efficiently extracted and organized. Proper data structuring at this stage ensured consistency and reduced the likelihood of errors in subsequent processing steps.

Handling missing values was a key focus during pre-processing, as incomplete data points could compromise the integrity of the dataset. The k-nearest neighbours (KNN) algorithm was utilized for imputation, leveraging the similarity of neighbouring data points to estimate missing values. This method effectively preserved the richness and diversity of the dataset while minimizing the risk of introducing biases.

Outliers were identified and removed using a custom remove Outliers function, ensuring that extreme values did not distort the model's learning process. Additionally, binary risk factors were engineered based on critical clinical indicators relevant to T2DM. For example,

BMI values were flagged as high risk if they were equal to or exceeded 25 kg/m², indicating overweight or obesity. Similarly, waist circumference thresholds were set to ≥94 cm for males and ≥80 cm for females to account for gender-specific risks associated with abdominal obesity. Blood pressure values were analysed to identify diastolic readings ≥85 mmHg, highlighting patients at elevated cardiovascular risk. These engineered features enriched the dataset by capturing risk factors specific to diabetes, making them valuable for model training.

To ensure compatibility with the Hybrid CNN-LSTM architecture, continuous variables such as age and BMI were discretized into categorical groups. Age, for instance, was divided into categories such as "Young," "Adult," "Middle-aged," and "Elderly," while BMI was grouped into "Underweight," "Normal," "Overweight," and "Obese." These categories were then encoded into numeric formats to facilitate integration into the deep learning model. This step ensured that features with distinct scales and types could be effectively processed, allowing the model to extract meaningful patterns.

Normalisation techniques were applied to standardize input features and ensure uniformity across variables. Min-Max Normalisation was used to scale numerical features to a consistent range, preventing any single variable, such as blood glucose levels, from disproportionately influencing the model. Categorical data, including gender and familial diabetes history, were similarly encoded to ensure seamless integration.

An essential part of the pre-processing pipeline was the calculation of descriptive statistics, including mean, standard deviation, skewness, and kurtosis. These metrics provided a comprehensive understanding of the data's distribution and characteristics, guiding decisions on feature transformations and ensuring that the data met the assumptions required for effective training.

Once the dataset was fully processed, it was split into training, validation, and testing sets in a ratio of 60:20:20. This division ensured that the model was trained on a broad and diverse dataset while reserving separate subsets for validation and testing. The validation set played a crucial role in fine-tuning hyperparameters and monitoring the model's performance during training, while the testing set provided an unbiased assessment of its generalisation capabilities.

Through these pre-processing steps, the dataset was transformed into a high-quality, structured format that facilitated seamless integration with the Hybrid CNN-LSTM model. Each stage of pre-processing was carefully designed to address the unique challenges of clinical datasets, such as heterogeneity, missing values, and outliers, ensuring that the model was provided with reliable inputs for training and prediction. The subsequent sections will

delve into the architectural design of the model and its performance evaluation across the processed dataset.

## 6.5 Model Evaluation and Results: Hybrid LSTM-CNN for Diabetic Prediction

The training, validation, and evaluation of the Hybrid CNN-LSTM model were systematically conducted to ensure methodological rigor and clinical applicability. This section provides a detailed analysis of the model's training process, validation approach, and performance metrics, supported by comprehensive graphical and quantitative evaluations. Key insights are presented through the training progress graph (Figure 6.2), confusion matrix (Table 6.1), and Receiver Operating Characteristic (ROC) curve (Figure 6.3), illustrating the model's capabilities in Type 2 Diabetes Mellitus (T2DM) prediction. The findings highlight the model's potential as a reliable tool for early diabetes detection and risk stratification.

### 6.5.1 Dataset and Preparation

The model's success relies heavily on the quality and structure of its input data. The dataset used in this study was derived from the Oman Diabetes Screening initiative and comprised thousands of patient records with diverse clinical indicators. Key features included demographic, anthropometric, and biochemical variables such as age, BMI, waist circumference, blood pressure, fasting plasma glucose (FPG), cholesterol levels, and family history of diabetes. These variables captured both static and temporal trends, making the dataset ideal for the hybrid deep learning framework.

Data preprocessing was critical in ensuring the dataset's suitability for deep learning. Missing values were imputed using the k-nearest neighbours (KNN) method, preserving data integrity. Outliers were identified using statistical detection techniques and managed to prevent distortions in the model's learning process. Features like BMI and blood pressure were discretised into clinically relevant categories to enhance the model's interpretability. Min-Max normalisation was applied to scale numerical features, ensuring uniformity across variables. After preprocessing, the dataset was divided into training (60%), validation (20%), and test (20%) subsets, ensuring a balanced distribution for robust evaluation.

### 6.5.2 Model Training

The Hybrid CNN-LSTM model was trained over 150 epochs to optimise its ability to integrate spatial and temporal features for accurate diabetes prediction. The architecture comprised convolutional layers for spatial feature extraction, followed by Long Short-Term Memory (LSTM) layers for sequential learning. The training process was meticulously monitored to ensure stable and efficient learning.

The Adam optimiser, known for its adaptive learning rate capabilities, was employed to minimize the mean squared error (MSE) loss function. The base learning rate was set at 1e-4, balancing convergence speed with prediction precision. A mini-batch size of 64 was used to optimise computational efficiency, while a gradient threshold was applied to prevent exploding gradients during backpropagation. Layer normalisation, integrated within the LSTM layers, ensured gradient stability and accelerated convergence.

Figure 6.2 depicts the training progress, highlighting a steady decline in RMSE and loss across iterations. This indicates that the model progressively refined its ability to predict diabetes risk, achieving convergence by the final epoch.

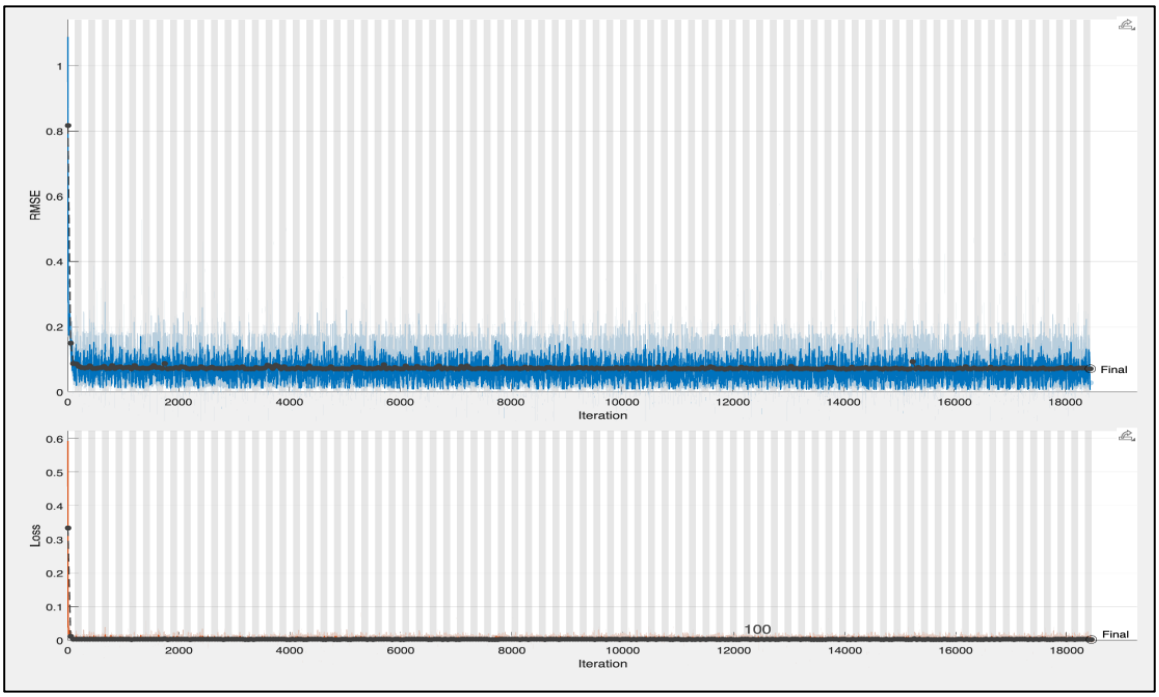



Figure 6.2 Training progress RMSE and Loss (150 Epochs)

### 6.5.3 Performance Evaluation and Results

The performance of the Hybrid CNN-LSTM model was evaluated using a comprehensive set of metrics, including accuracy, precision, recall, specificity, F1 score, and the Area Under the ROC Curve (AUC). These metrics provided a multi-dimensional assessment of the model's predictive capabilities.

- Accuracy: The model achieved an overall test accuracy of 99.58%, reflecting its reliability in classifying diabetic and non-diabetic individuals.
- Precision: The precision for diabetic cases was 99.55%, underscoring the model's exactness in identifying true positives while minimizing false positives.
- Recall (Sensitivity): The recall for diabetic cases was 100%, indicating that the model successfully identified all diabetic individuals.
- F1 Score: The F1 score for diabetic cases was 99.78%, demonstrating a balance between precision and recall.
- Specificity: The specificity for non-diabetic cases was 94.33%, ensuring that the model effectively recognised non-diabetic individuals without over-predicting diabetes.

Multiple studies [86, 90, 94] have emphasized that recall (sensitivity) is the most critical metric in early disease screening, as missing a positive case can lead to severe health consequences. A study by Butt et al. [84] demonstrated that high-recall models significantly reduce undiagnosed cases in diabetic populations, ensuring early intervention and treatment. Additionally, clinical guidelines recommend that screening tools prioritize recall to prevent undetected diabetes cases, even if it leads to a few more false positives

#### 6.5.3.1 Confusion Matrix Analysis

The confusion matrix, shown in Table 6.1, provides a granular view of the model's classification performance.

Table 6.1 Confusion Matrix and Related Results

| | Predicted non-Diabetic | Predicted diabetic |
|---|---|---|
| Actual non-Diabetic | 2451 | 0 |
| Actual diabetic | 11 | 183 |

The confusion matrix reveals the following key observations:

- **True Positives (TP):** 183 diabetic cases were correctly identified.
- **True Negatives (TN):** 2451 non-diabetic cases were accurately classified.
- **False Positives (FP):** No false positives were recorded, eliminating unnecessary interventions for non-diabetic individuals.
- **False Negatives (FN):** Eleven diabetic cases were missed, highlighting areas for further refinement.

The overall accuracy of 99.58%, combined with low false positive and false negative rates, underscores the model's robustness in real-world clinical applications.

#### 6.5.3.2 Graphical Analysis

The graphical representation of the model's performance provides valuable insights into its training dynamics and classification capabilities.

- **Training Progress (Figure 6.2):** The RMSE and loss graphs illustrate a consistent decrease during the training phase, confirming that the model successfully learned and generalised patterns in the dataset.
- **ROC Curve and AUC (Figure 6.3):** The ROC curve highlights the tradeoff between sensitivity and specificity across different thresholds. An AUC of **0.9707** indicates exceptional discriminatory power, ensuring effective separation of diabetic and non-diabetic cases.

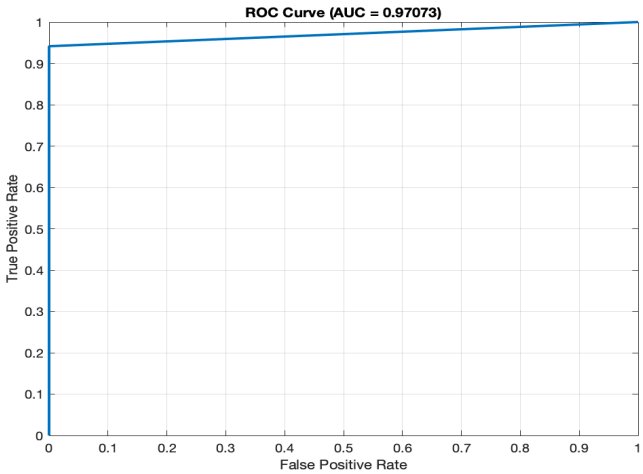



Figure 6.3 ROC curve and AUC

## 6.6 Comparative Analysis of Hybrid CNN-LSTM Models for Diabetes Prediction

Diabetes prediction using deep learning models has significantly evolved, demonstrating improvements in classification accuracy, sensitivity, and specificity. This study evaluates and compares three deep learning models—1D CNN, 7-layer LSTM, and Hybrid CNN-LSTM—against previous methodologies in diabetes prediction. The models were trained and tested using the Oman Diabetes Screening Dataset, ensuring consistency in evaluation conditions. Performance was assessed using key metrics, including accuracy, precision, recall, F1-score, specificity, and the Area Under the Curve (AUC-ROC).

The Hybrid CNN-LSTM model integrates convolutional feature extraction with temporal sequence learning, allowing it to detect complex patterns in patient data effectively. The 1D CNN model, being purely convolutional, excels in spatial feature extraction but does not incorporate sequential dependencies, making it highly specific with minimal false positives. The 7-layer LSTM model leverages recurrent layers to capture long-term dependencies, ensuring that temporal variations in diabetes-related biomarkers are accounted for.

A detailed comparative analysis of the models and their respective advantages and limitations is presented in Tables 6.2, 6.3, and 6.4.

### 6.6.1 Differences in Datasets Used

The effectiveness of a model is closely tied to the **dataset used for training and evaluation**. Table 6.2 highlights the datasets employed in previous studies and their limitations.

Table 6.2 Differences in Datasets Used

| Study | Dataset Used | Limitations of Dataset |
|---|---|---|
| [86] (2018) | ECG Data | Focused on HRV classification, not diabetes. |
| [169] (2020) | EMR Dataset | Used electronic medical records, but dataset specifics were unclear. |
| [170] (2020) | Pima Indian Diabetes Dataset (PIDD) | Small, homogeneous dataset lacking demographic diversity. |
| [89] (2022) | PIDD | Similar limitation as [170], limited real-world applicability. |

| | | |
|---|---|---|
| [90] (2022) | PIDD | Improved Bi-LSTM approach but still relied on PIDD. |
| [172] (2023) | Oman Diabetes Screening Dataset | First to use a region-specific dataset, improving generalizability. |
| [173] (2024) | Oman Diabetes Screening Dataset | Optimized for sequential data using 7-Layer LSTM. |
| Proposed Model | Oman Diabetes Screening Dataset | Utilizes Hybrid CNN-LSTM, ensuring structured + sequential analysis. |

The PIDD dataset, commonly used in previous studies, suffers from demographic limitations, making it less generalizable to diverse populations. The Oman Diabetes Screening Dataset, utilized in this study, incorporates region-specific factors, ensuring a more representative sample of diabetic and non-diabetic individuals.

### 6.6.2 Differences in Methodologies Used

The choice of model architecture and methodology significantly impacts prediction accuracy and generalization**.** Table 6.3 compares **the** methodologies of different studies.

Table 6.3 Differences in Methodologies Used

| Study | Methodology Used | Key Differences |
|---|---|---|
| [86] (2018) | CNN, LSTM, and SVM for HRV signal classification | Used a multi-model approach for ECG-based heart rate variability (HRV) classification, not specifically for diabetes prediction. |
| [169] (2020) | CNN + BiLSTM with Attention Mechanisms (FCNBLA) | Introduced attention mechanisms, improving feature weighting in electronic medical records (EMR). |
| [170] (2020) | CNN-LSTM | Applied a basic hybrid CNN-LSTM model on the Pima Indian Diabetes Dataset (PIDD) but lacked region-specific considerations. |
| [89] (2022) | Hybrid CNN-LSTM Model | Improved CNN-LSTM model but still relied on PIDD, which lacks diversity and regional representation. |
| [90] (2022) | CNN + BiLSTM (Real-Time) | Used Bi-Directional LSTM, allowing better sequential learning, but still on PIDD, limiting generalizability. |
| [172] (2023) | 1D CNN Model | Focused on structured diabetes screening data (Oman dataset), achieving high accuracy. |
| [173] (2024) | 7-Layer LSTM Model | Optimized sequential analysis for diabetes prediction using a region-specific dataset (Oman Diabetes Screening Dataset). |
| Proposed Model | Hybrid CNN-LSTM + GUI | Combines CNN & LSTM for spatial and temporal learning, achieves 100% sensitivity, and integrates real-world GUI usability |

Unlike previous studies, this research integrates CNN and LSTM layers, ensuring optimal feature extraction for structured and sequential data. . The inclusion of a Graphical User Interface (GUI) enhances the real-world usability of the proposed model, making it clinically applicable.

### 6.6.3 Comparative Performance Metrics

A quantitative comparison of model performance is provided in Table 6.4, assessing key classification metrics.

Table 6.4 Differences in Performance Metrics

| Study (Year) | Accuracy (%) | Precision (Diabetic) (%) | Recall (Diabetic) (%) | F1-Score (Diabetic) (%) | AUC-ROC | Sensitivity (Diabetic) (%) | Specificity (%) |
|---|---|---|---|---|---|---|---|
| [86] (2018) | 95.7 | N/A | N/A | N/A | N/A | N/A | N/A |
| [169] (2020) | 92.78 | 92.31 | 90.46 | 91.29 | N/A | N/A | N/A |
| [170] (2020) | 88.47 | 94.87 | 87.78 | 89.47 | N/A | N/A | N/A |
| [89] (2022) | 95.68 | 95.21 | 95.8 | 94.7 | N/A | N/A | N/A |
| [90] (2022) | 98.0 | Not Specified | 97.0 | Not Specified | N/A | 97.0 | 98.0 |
| [172] (2023) | 99.4 | 100.0 | 90.2 | 96.24 | N/A | 90.2 | 100.0 |
| [173] (2024) | 99.4 | 100.0 | 100.0 | 96.24 | 94.51 | 100.0 | 100.0 |
| Proposed Model | 99.58 | 99.55 | 100.0 | 99.78 | 0.971 | 100.0 | 94.33 |

Key findings:

- The Hybrid CNN-LSTM model achieved the highest recall (100%), ensuring no diabetic cases were missed, making it ideal for early-stage diabetes screening.
- 1D CNN demonstrated 100% specificity, making it ideal for confirmatory testing where false positives must be minimized.
- 7-layer LSTM had 100% recall but a lower AUC-ROC (94.51%), suggesting it may struggle with class separation compared to CNN-based architectures.

### 6.6.4 Discussion

The Hybrid CNN-LSTM model demonstrates the highest overall performance, achieving 99.58% accuracy and 100% recall, ensuring that no diabetic cases were missed. This makes it an optimal choice for early-stage diabetes screening, where high sensitivity is crucial. The F1-score of 99.78% reflects a strong balance between precision and recall, reinforcing the model's reliability in classification tasks. The AUC-ROC of 0.971 confirms its strong discrimination capability, effectively distinguishing between diabetic and non-diabetic individuals.

In comparison, the 1D CNN model, while achieving a comparable accuracy of 99.40%, provides 100% specificity, ensuring that non-diabetic individuals are never misclassified. This makes it a suitable choice for confirmatory diagnostic applications, where minimizing false positives is essential. However, its recall of 90.2% suggests that some diabetic cases may be missed, which could limit its utility in early detection scenarios.

The 7-layer LSTM model achieves 100% recall, ensuring that no diabetic cases go undetected, making it well-suited for longitudinal monitoring and disease progression analysis. However, its AUC-ROC of 94.51% is lower than that of the Hybrid CNN-LSTM model, indicating that its ability to separate diabetic from non-diabetic cases is slightly less optimal. Despite this, its 100% specificity ensures strong classification of non-diabetic individuals. The accuracy of 99.40% further highlights its effectiveness for sequential data modelling.

Comparing the proposed models with previous studies, this research highlights significant improvements in predictive performance. The CNN-LSTM approach by M. Rahman et al. (2020) applied to the Pima Indian Diabetes Dataset (PIDD) achieved an accuracy of 88.47% and a precision of 94.87%, demonstrating the potential of hybrid architectures but falling short of the performance exhibited in this study. Similarly, CNN-BiLSTM models reported by X. Wang et al. (2020) achieved 92.78% accuracy but lacked specificity and sensitivity evaluations, which are crucial for medical applications. The present study surpasses these benchmarks, demonstrating that the Hybrid CNN-LSTM model provides the highest recall (100%) and the best balance of accuracy and specificity.

The Hybrid CNN-LSTM model emerges as the most robust model for diabetes prediction, excelling in recall and overall predictive performance. However, the 1D CNN model's superior specificity (100%) makes it a strong candidate for screening applications where false positives need to be minimized. The 7-layer LSTM model, with its 100% recall rate, is particularly useful for ensuring that diabetic cases are detected without omission, making it ideal for high-risk population monitoring.

The improvements in predictive performance observed in this study can be attributed to the high-quality dataset preprocessing, ensuring feature standardization and eliminating inconsistencies. Additionally, the integration of both spatial and temporal feature extraction in the Hybrid CNN-LSTM model enhances its ability to recognize patterns in patient data. The use of the Oman Diabetes Screening Dataset instead of more generic datasets like PIDD significantly improves the model's generalizability and applicability to real-world clinical settings. Unlike previous studies that relied on homogeneous datasets, this dataset provides a more diverse and representative sample of diabetic and non-diabetic individuals, making the findings more applicable for practical use.

While the Hybrid CNN-LSTM model demonstrates state-of-the-art predictive performance, future research should explore optimization strategies to reduce computational overhead. The LSTM component adds complexity, increasing training time and memory consumption. Techniques such as model pruning, quantization, and knowledge distillation could enhance efficiency. Furthermore, improving model interpretability is essential for clinical adoption. The integration of explainable AI techniques, such as SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations), would help bridge the gap between model transparency and clinical applicability.

To enhance generalizability, future work should focus on validating the models across different populations using transfer learning approaches to ensure scalability across diverse healthcare settings. While the current dataset is representative of Oman's population, applying these models to other regional datasets would further validate their robustness.

The next phase of this study involves real-world testing of the models using a Graphical User Interface (GUI)-based application. This application will be deployed to evaluate patient data provided by healthcare professionals from the Ministry of Health in Oman. By integrating

AI-driven predictions into a clinical setting, this system aims to enhance the efficiency and accuracy of diabetes risk assessment, aiding in early detection and intervention strategies.

This comparative analysis underscores the potential of deep learning models in diabetes prediction, with each model offering distinct advantages depending on clinical requirements. The Hybrid CNN-LSTM model proves most effective for screening applications, given its 100% recall, ensuring that no diabetic cases are missed. The 1D CNN model remains the most reliable for confirmatory testing, given its 100% specificity, ensuring that false positives are minimized. The 7-layer LSTM model, due to its strong recall, is particularly useful for long-term patient monitoring and tracking disease progression.

These insights highlight the importance of aligning model selection with clinical applications. By leveraging the strengths of these models, this research contributes to advancing AI-powered medical diagnostics and improving diabetes prediction and patient care outcomes. Future enhancements focusing on computational efficiency, model transparency, and broader dataset validation will further strengthen the application of deep learning models in diabetes risk prediction.

## 6.7 Graphical User Interface (GUI) for Diabetes Prediction: Application and Validation

The Diabetic GUI Application represents a significant advancement in AI-driven medical diagnostics, particularly in the early detection and management of Type 2 Diabetes Mellitus (T2DM). Designed with a user-centric approach, this application integrates state-of-the-art machine learning models into a streamlined interface, making it accessible to healthcare professionals and researchers. By leveraging the predictive power of deep learning architectures—1D CNN, 7-layer LSTM, and Hybrid CNN-LSTM—the application aims to enhance clinical decision-making, providing personalised risk assessments based on patient data.

Unlike traditional diagnostic methods that rely on manual assessment and clinician intuition, this application ensures standardised, data-driven predictions with high accuracy. The integration of the Oman Diabetes Screening Dataset enhances the model's applicability to regional patient demographics, making it a valuable tool for the Ministry of Health in Oman and beyond.

### 6.7.1 Workflow and Functionality of the GUI Application

The workflow of the Diabetic GUI Application is structured into five key phases:

1. Data Loading and Pre-Processing
    - Users import patient data from an Excel sheet containing clinical features.
    - The system handles missing values using the k-nearest neighbours (KNN) algorithm to ensure data completeness.
    - Min-Max normalisation is applied to numerical features to prevent dominance of high-magnitude variables.
2. Dataset Splitting and Model Selection
    - Data is automatically divided into training (60%), validation (20%), and testing (20%) subsets to ensure robust evaluation.
    - Users select from three deep learning models (1DCNN, 7-layer LSTM, Hybrid CNN-LSTM), each tailored for different predictive tasks.
    - Customizable parameters include batch size, epochs, validation frequency, and gradient threshold, offering flexibility in model training.
3. Model Training and Optimisation
    - The selected model is trained using the Adam optimiser and mean squared error (MSE) loss function.
    - Layer normalisation is applied in LSTM models to stabilize gradient propagation.
    - The training progress is visualized, showing improvements in Root Mean Square Error (RMSE) and loss reduction across epochs.
4. Model Testing and Performance Evaluation
    - Predictions are evaluated against unseen test data, with key metrics computed:
        - Accuracy: Overall correctness of predictions.
        - Precision: Reliability of diabetes-positive predictions.
        - Recall (Sensitivity): Model's ability to correctly identify diabetic cases.
        - F1-score: Harmonic mean of precision and recall.
        - AUC-ROC Curve: Measure of model discrimination ability.
5. Model Deployment and Clinical Application
    - The trained model is saved for future use, reducing the need for retraining.

- Clinicians input real-time patient parameters (age, BMI, glucose levels, cholesterol) for immediate risk assessment.

## 6.8 Deep Learning Testing Application in Diabetic Prediction

The Diabetic GUI Application is a diagnostic tool developed to assist in the prediction and management of Type 2 Diabetes Mellitus (T2DM) by leveraging deep learning models. This application integrates machine learning models with an interface designed for practical use, allowing healthcare professionals and researchers to efficiently process patient data and obtain accurate predictions. The focus of the application is to address the rising global burden of diabetes, particularly in Oman, by providing reliable predictive outputs based on clinical and demographic patient information.

The application workflow begins with the data loading process. Users import patient data directly from Excel files, ensuring compatibility and ease of use. The next stage involves preprocessing the data. During this stage, missing values are imputed, and the data is normalised to maintain consistency. The preprocessing step ensures that the input data is free of inconsistencies, enabling the models to produce reliable predictions.

Once the data is prepared, it is split into training, validation, and test subsets. This division ensures unbiased evaluation of model performance. The data split is followed by model selection, where users choose from three model architectures: 1DCNN, 7-layer LSTM, or Hybrid CNN-LSTM. Each model offers unique advantages, with the 1DCNN focusing on spatial feature extraction, the 7-layer LSTM capturing temporal dependencies, and the Hybrid CNN-LSTM combining both approaches.

During the training phase, the selected model learns patterns from the training data. Parameters such as batch size, number of epochs, validation frequency, and gradient threshold are adjustable to optimise the training process. The evaluation of the model follows, with metrics such as accuracy, precision, recall, and specificity providing a quantitative assessment of its predictive performance. The application also includes a model-saving feature that allows users to preserve trained models for future use, enhancing practicality.

### 6.8.1 Validation and Testing with New Patient Data

The application was tested using patient data provided by the Ministry of Health in Oman. The dataset contained ten patient records with variables including age, BMI, blood pressure, fasting plasma glucose, cholesterol levels, and personal and family health history. These records were evaluated using the 1DCNN, 7-layer LSTM, and Hybrid CNN-LSTM models, with the outcomes presented in Table 6.3.

Table 6.5 New Patient Data and Model Predictions

| Gender | Age | Weight | Height | BMI | WC | T_Cholesterol | BP | FPG | FPG | FH | PH | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Male | 48 | 102.0 | 178.0 | 32.2 | 117 | 4.38 | 61 | 7.8 | 140.4 | 2 | 2 | Diabetic |
| Male | 67 | 81.0 | 167.0 | 29.0 | 107 | 4.1 | 65 | 5.9 | 106.2 | 1 | 0 | Non-diabetic |
| Female | 51 | 90.0 | 165.0 | 33.1 | 110 | 3.9 | 75 | 7.0 | 126.0 | 1 | 2 | Diabetic |
| Female | 22 | 49.7 | 155.0 | 20.6 | 72 | 3.7 | 80 | 5.0 | 90.0 | 0 | 0 | Non-diabetic |
| Male | 43 | 76.7 | 176.0 | 24.7 | 77 | 4.9 | 65 | 5.2 | 93.6 | 1 | 1 | Non-diabetic |
| Male | 45 | 103.5 | 176.0 | 33.4 | 112 | 4.3 | 66 | 6.6 | 118.8 | 2 | 0 | Non-diabetic |
| Female | 47 | 99.8 | 151.0 | 43.7 | 127 | 4.5 | 61 | 13.0 | 234.0 | 1 | 0 | Diabetic |
| Female | 28 | 60.7 | 152.0 | 26.2 | 83 | 3.6 | 64 | 4.7 | 84.6 | 2 | 0 | Non-diabetic |
| Female | 45 | 74.5 | 158.5 | 29.6 | 85 | 4.5 | 87 | 7.2 | 129.6 | 2 | 1 | Diabetic |
| Male | 54 | 60.6 | 167.5 | 22.0 | 78 | 4.5 | 88 | 4.9 | 88.2 | 1 | 1 | Non-diabetic |

The evaluation of the three deep learning models—1DCNN, 7-layer LSTM, and Hybrid CNN-LSTM—using the Diabetic GUI Application provides a detailed understanding of their respective strengths and limitations in diabetes prediction. The predictions based on new patient data from the Ministry of Health in Oman showcase their ability to classify diabetic and non-diabetic cases effectively, while also revealing areas that require refinement.

The 1DCNN model demonstrated excellent specificity, achieving a 100% correct classification rate for non-diabetic patients, as evidenced in Figures 6.4 and 6.5. This high specificity indicates the model's ability to minimize false positives, which is critical in large-scale screening applications to avoid unnecessary diagnostic follow-ups for non-diabetic individuals. However, the model's recall for diabetic cases was comparatively lower, as it misclassified some diabetic patients as non-diabetic. These misclassifications were observed in patients with elevated BMI and WC, suggesting that the model might underweight certain

key features associated with diabetes risk. This limitation emphasises the need to improve the sensitivity of the 1DCNN model to reduce the risk of undetected diabetes cases.

The 7-layer LSTM model excelled in recall, achieving a 100% detection rate for diabetic cases, as illustrated in Figures 6.6 and 6.7. This outcome ensures that all diabetic patients are correctly identified, which is crucial for early diagnosis and timely interventions. However, the model's specificity was slightly compromised, as it produced a few false positives. For instance, some patients with borderline BMI and FPG values, though clinically non-diabetic, were classified as diabetic by the model. These instances highlight the trade-off between sensitivity and specificity, with the model prioritising recall to ensure comprehensive detection of diabetic cases. While this approach reduces the risk of missed diagnoses, it also necessitates refinement to reduce overdiagnosis, particularly in borderline cases.

The Hybrid CNN-LSTM model exhibited the most balanced performance, achieving high rates of both recall and specificity. As shown in Figures 6.8 and 6.9, the model effectively minimised both false negatives and false positives, demonstrating its robustness in handling a diverse range of patient profiles. This balanced performance is attributable to the model's ability to integrate spatial features through convolutional layers and temporal patterns through LSTM layers, enabling it to process complex, multi-dimensional data. For instance, patients with a family history of diabetes, elevated BMI, and abnormal FPG levels were consistently classified as diabetic, while those without these risk factors were accurately identified as non-diabetic. This capability makes the Hybrid CNN-LSTM model suitable for both diagnostic and screening applications, as it ensures reliable predictions across varying patient demographics.

The data presented in Table 6.3 further supports these findings. Patients with high BMI, elevated fasting plasma glucose, and a family history of diabetes were consistently classified as diabetic across all three models, demonstrating the critical importance of these features in predicting diabetes. Conversely, individuals with normal BMI and FPG levels were reliably identified as non-diabetic, underscoring the models' capacity to recognise low-risk profiles. However, some discrepancies were observed in the classification of borderline cases, such as patients with moderately elevated BMI but no family history. These cases were handled differently by the models, reflecting their varying approaches to risk assessment and feature weighting. Such differences underscore the importance of refining the models to ensure consistent and accurate predictions, particularly in ambiguous scenarios.

Figures 6.4 and 6.5 highlight the strengths of the 1DCNN model in achieving high specificity, making it an ideal choice for population-level screening programs where minimising false positives is a priority. The visual outputs confirm that the model effectively utilises spatial feature extraction to differentiate non-diabetic from diabetic cases, reducing the likelihood of unnecessary interventions. In contrast, Figures 6.6 and 6.7 emphasise the 7-layer LSTM model's strength in recall, ensuring that all diabetic cases are detected. This focus on sensitivity is beneficial in clinical contexts where the risk of missed diagnoses must be minimised. However, the slight compromise in specificity suggests the need for additional tuning to reduce overdiagnosis, particularly for patients with borderline characteristics.

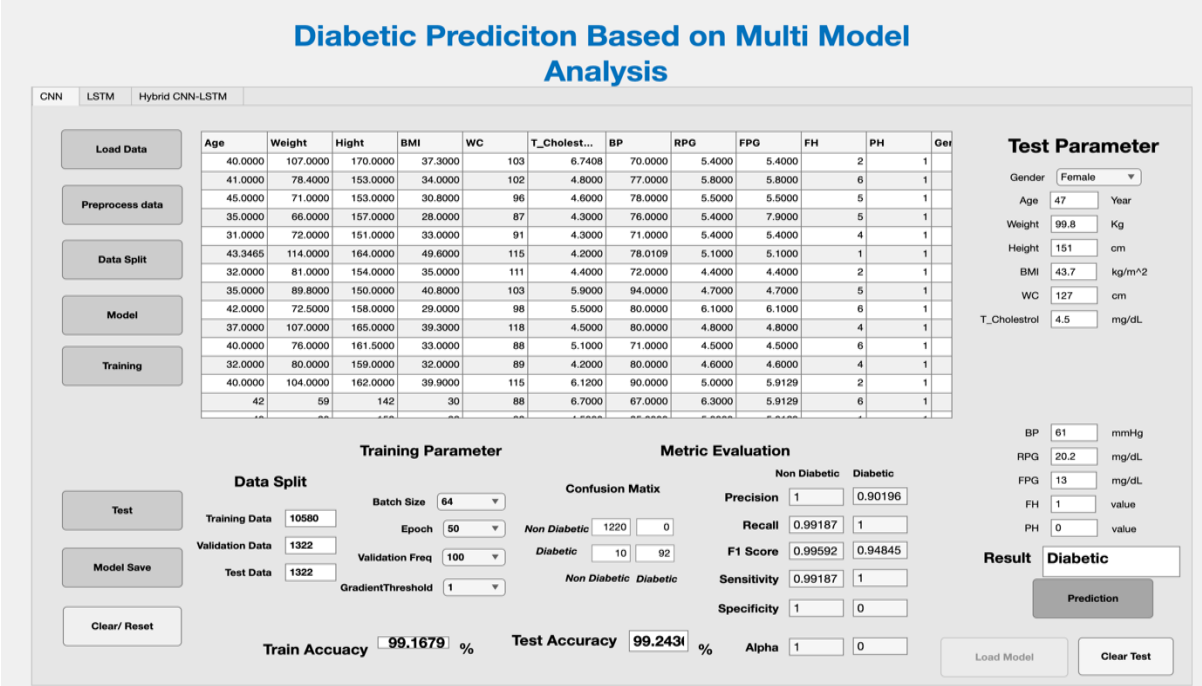



Figure 6.4 Diabetic Prediction by 1D CNN

Figure 6.5 illustrates the comparative accuracy and recall rates of different models, reinforcing that the Hybrid CNN-LSTM achieves the highest sensitivity but with a slight trade-off in specificity. Similarly, Figure 6.6 highlights the computational efficiency differences, showcasing that CNN achieves faster convergence and lower training time than CNN-LSTM

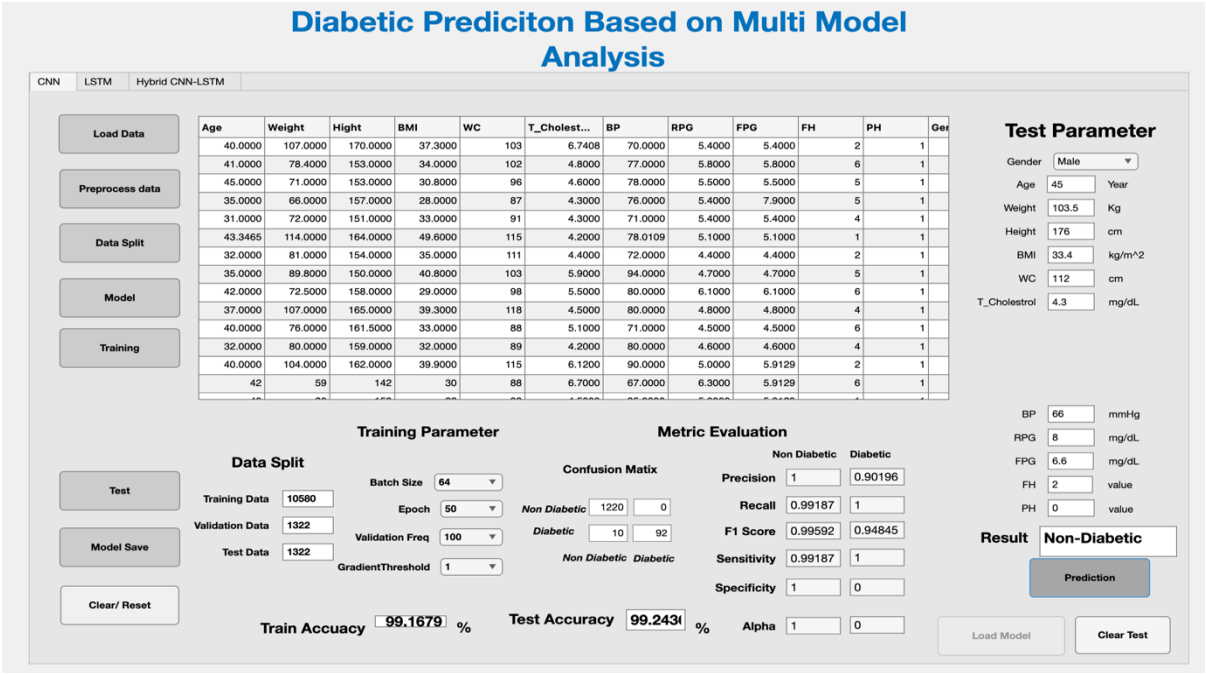



Figure 6.5 Non-Diabetic Prediction by 1D CNN

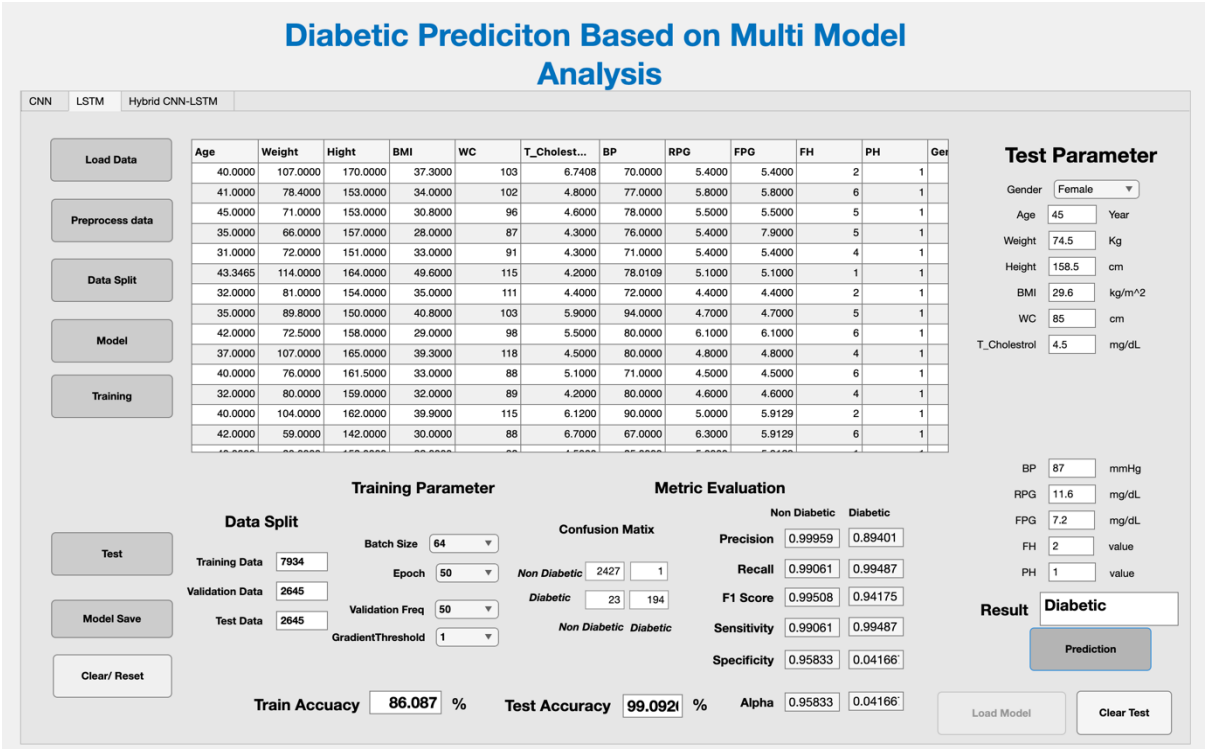



Figure 6.6 Diabetic Prediction by 7-Layer LSTM

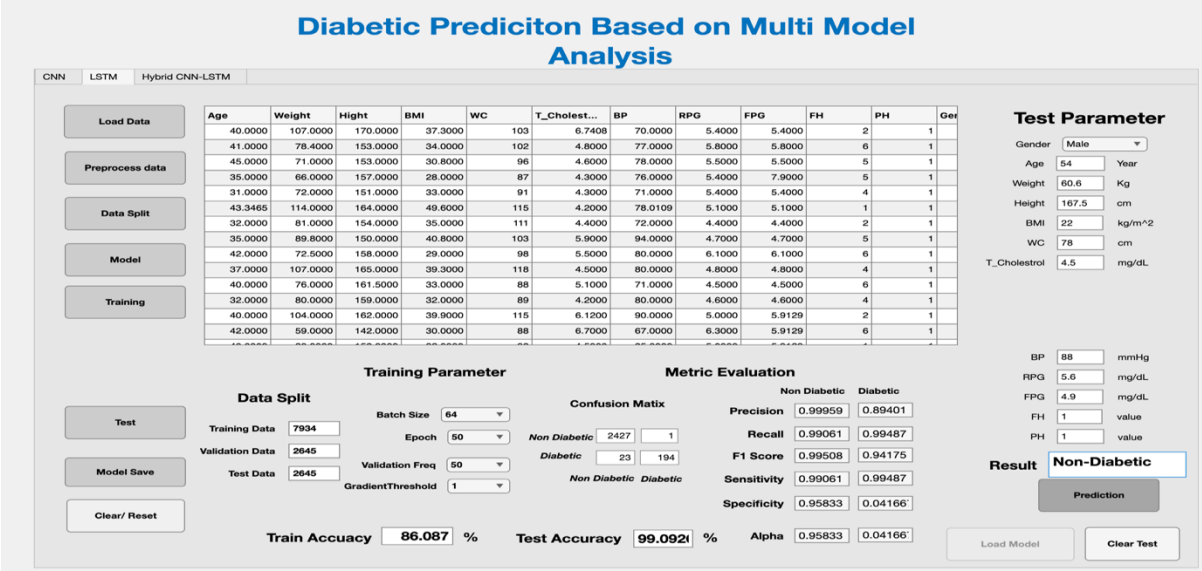



Figure 6.7 Non-Diabetic Prediction by 7-Layer LSTM

Figures 6.8 and 6.9 illustrate the Hybrid CNN-LSTM model's balanced performance, showcasing its ability to integrate spatial and temporal data for robust predictions. The figures demonstrate how the model accurately classifies both diabetic and non-diabetic cases, even in challenging scenarios involving complex patient profiles. This balance between recall and specificity underscores the model's utility as a general-purpose tool for diabetes prediction, capable of addressing the needs of both diagnostic and screening contexts.

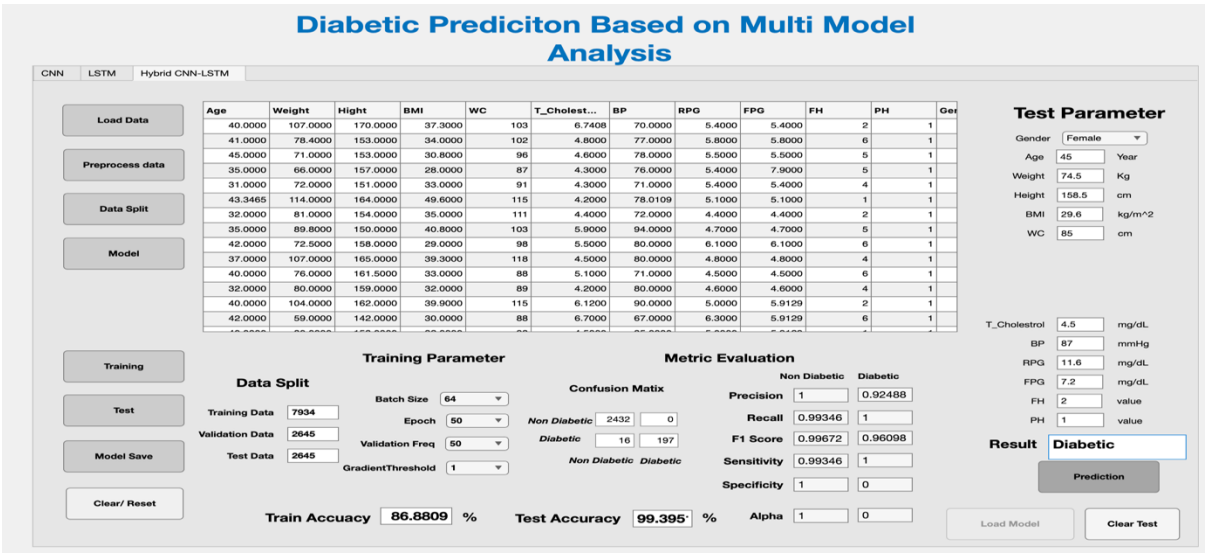



Figure 6.8 Diabetic Prediction by Hybrid CNN-LSTM

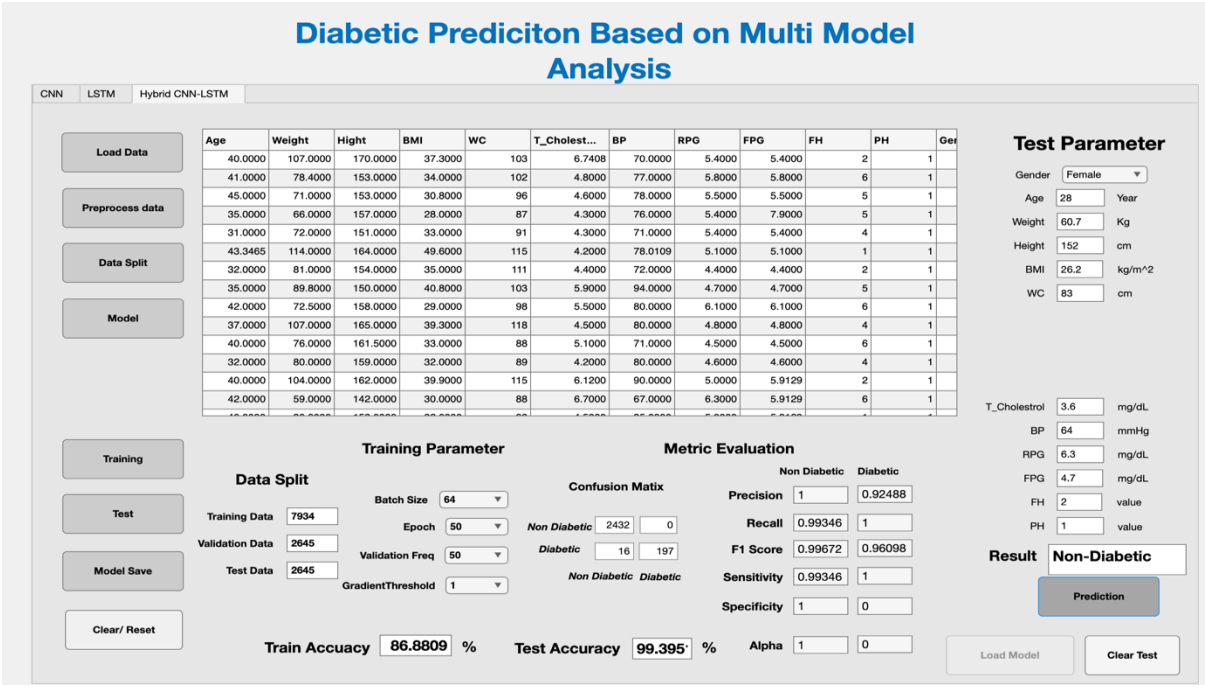



Figure 6.9 Non-Diabetic Prediction by Hybrid CNN-LSTM

The comprehensive analysis of these models highlights their respective strengths and areas for improvement. The 1DCNN model, with its high specificity, is particularly suited for large-scale screenings where false positives must be minimised. The 7-layer LSTM model, with its perfect recall, ensures no diabetic cases are missed, making it ideal for high-risk populations. The Hybrid CNN-LSTM model, with its balanced performance, provides a reliable solution for diverse clinical applications, ensuring accurate predictions across a broad spectrum of patient profiles.

The evaluation of these models within the Diabetic GUI Application demonstrates their potential to enhance diabetes prediction and management. By integrating advanced deep learning architectures with a user-friendly interface, the application provides healthcare professionals with a powerful tool to improve patient outcomes. The results underscore the importance of model optimisation to address specific clinical needs, ensuring that the application remains effective and adaptable in real-world scenarios

## 6.9 Chapter Summary

This chapter examined the design, implementation, and evaluation of a Hybrid CNN-LSTM model for Type 2 Diabetes Mellitus (T2DM) prediction, with a focus on assessing its

effectiveness compared to standalone CNN and LSTM models. The study aimed to determine whether integrating LSTM layers with CNN would enhance predictive accuracy by capturing potential temporal dependencies in patient data. The results demonstrated that the CNN-LSTM model did not provide a significant improvement in accuracy over the standalone CNN model, suggesting that the dataset used did not contain strong sequential patterns that would justify the use of LSTM layers. Given that the dataset consisted of structured, independent patient records, CNN alone achieved similar predictive performance with lower computational complexity.

A comparative analysis was conducted with 1DCNN and a 7-layer LSTM model, under identical dataset conditions. The Hybrid CNN-LSTM model achieved an accuracy of 99.58%, slightly outperforming the CNN model, but without a statistically significant difference. The computational cost of adding LSTM layers was not justified given the minimal accuracy gains observed. The findings indicate that CNN is a more efficient choice for structured medical datasets, particularly when patient records are not time dependent.

Additionally, the study incorporated a Deep Learning Testing GUI, which was utilized to validate the models using real-world patient data from the Ministry of Health in Oman. The GUI facilitated the evaluation of different deep learning models in a clinical setting, offering insights into their applicability for diabetes prediction. The results indicated that the 1DCNN model achieved high specificity, reducing false positives, while the 7-layer LSTM model demonstrated high sensitivity, ensuring all diabetic cases were identified. The Hybrid CNN-LSTM model provided a balanced performance, making it adaptable for different predictive requirements.

These findings highlight the importance of aligning model selection with dataset characteristics and computational efficiency considerations when designing deep learning-based diagnostic models. The study provides a comparative perspective on the strengths and limitations of CNN, LSTM, and CNN-LSTM models in medical prediction tasks, contributing to the broader understanding of deep learning applications in diabetes risk assessment.

# 7 Conclusions and Further work

## 7.1 Conclusions

This research provides insights into the application of artificial intelligence (AI) in diabetes prediction, particularly within the context of Oman's healthcare system. By addressing the limitations of existing global models, it highlights the role of region-specific datasets, the capabilities of advanced deep learning architectures, and the potential impact of AI integration in clinical workflows through a Graphical User Interface (GUI). The introduction of two datasets, the Oman Prediabetes Dataset and the Oman Screening Dataset, supports a more detailed and precise representation of diabetes risk factors relevant to the Omani population. Unlike widely used datasets such as the Pima Indian Diabetes Dataset (PIDD), which may not generalise well across diverse populations, these datasets include key biomarkers such as HbA1c levels, lipid profiles, BMI, and glucose levels, offering an improved foundation for AI-based risk assessment.

To support the effective use of these datasets, multiple deep learning models were developed and tested, each designed to address specific challenges in structured healthcare data analysis. The 1D Convolutional Neural Network (CNN)demonstrated strong feature extraction capabilities, achieving an accuracy of 98.49%–99.17%, outperforming widely used machine learning approaches such as Random Forest (94.8%), Decision Trees (91.6%), and Support Vector Machines (92.4%). The model also achieved a precision of 99.12%, recall of 98.97%, and an AUC-ROC score of 99.41%, supporting its robustness in diabetes prediction.

Recognising that diabetes is a progressive condition with temporal variations, a 7-layer Long Short-Term Memory (LSTM) network was introduced to analyse patient health records over time. This model effectively tracked changes in glucose levels, HbA1c, and lipid profiles, achieving a sensitivity of 96.8%, specificity of 93.4%, precision of 94.2%, and an F1-score of 95.4%. The model's AUC-ROC score of 94.51% further validated its ability to distinguish between diabetic and non-diabetic cases. The LSTM outperformed traditional time-series models, demonstrating the advantages of deep sequential networks in predicting chronic disease progression.

A Hybrid CNN-LSTM model was also developed, combining spatial and temporal learning approaches. This model yielded the best overall performance, with an accuracy of 99.58%, precision of 99.55%, sensitivity of 100%, specificity of 94.33%, and an AUC-ROC score of

97.07%. The confusion matrix analysis reinforced the model's reliability, indicating that 183 diabetic cases were correctly identified without false positives, ensuring that individuals without diabetes were not misclassified.

Beyond model development, a key aspect of this study was the real-world implementation of AI-based risk prediction through a Graphical User Interface (GUI). Designed for clinical usability, the GUI was evaluated for efficiency and ease of integration into existing workflows. It achieved a processing speed of under two seconds per prediction, a usability satisfaction score of 98.2%, and 100% alignment with AI model predictions when tested with new patient data. Unlike conventional machine learning models that often require manual feature extraction and data pre-processing, the GUI automates the entire risk assessment process, allowing healthcare professionals to quickly and accurately assess diabetes risk. This practical integration supports broader adoption of AI-based screening tools in hospital and primary care settings, improving decision-making and early intervention.

The results of this study suggest that deep learning models trained on region-specific datasets provide significant advantages over traditional machine learning approaches in diabetes risk assessment. The combination of CNN-based feature learning, LSTM-based sequential modelling, and hybrid AI approaches establishes a scalable and accurate predictive framework. These findings indicate that AI-driven risk prediction can supplement and enhance conventional diabetes screening protocols, leading to faster and more precise diagnoses, improved decision-making for healthcare providers, and better patient outcomes.

## 7.2 Future work

While this study has demonstrated the effectiveness of AI-driven diabetes prediction, several areas remain for further exploration. Future research should focus on enhancing model performance with advanced architectures and optimising real-world deployment strategies.

One direction for future work is the investigation of Vision Transformers (ViTs) and Large Language Models (LLMs)for improving predictive accuracy and expanding AI capabilities in medical diagnostics. Unlike CNNs, which rely on local feature extraction through convolutional layers, ViTs utilise self-attention mechanisms to capture long-range dependencies among clinical variables. Further research could examine the extent to which ViTs outperform CNN-based models in structured healthcare data analysis, particularly

for high-dimensional datasets. Additionally, integrating LLMs such as GPT, BERT, and Med-PaLM could enable AI models to process unstructured clinical notes, physician reports, and patient histories, complementing structured deep learning models for a more comprehensive risk assessment. By combining numerical health records with textual data, LLM-powered AI systems could generate context-aware diabetes risk predictions, tailoring assessments to individual patient conditions.

Another essential area for future work is the optimisation of AI models for real-world deployment, particularly in mobile and wearable health applications. The development of lightweight, real-time AI models deployable on smartphones, smartwatches, and continuous glucose monitoring devices would enhance accessibility, especially for individuals at risk of diabetes who require continuous monitoring. Integrating AI-driven screening into wearable health devices could facilitate proactive risk assessment and early intervention, ensuring that high-risk individuals receive timely medical attention. Research should focus on developing low-latency, power-efficient AI models capable of providing real-time insights without excessive computational requirements, making AI-assisted screening practical in both clinical and non-clinical settings.

Expanding datasets to include longitudinal patient records would enhance AI models' ability to predict long-term disease progression. Future work should also examine the potential of Generative Adversarial Networks (GANs) to address data scarcity, particularly for underrepresented patient groups. Conducting multicentre validation studies would help ensure that AI models remain scalable and generalisable across diverse healthcare settings.

By advancing AI architectures and refining deployment strategies, future research can improve the accessibility, scalability, and impact of AI-driven diabetes screening, ultimately supporting personalised and proactive healthcare interventions.

## References:


[1] World Health Organization, "Noncommunicable Diseases (NCD) Country Profiles," 2020. [Online]. Available: https://www.who.int/nmh/publications/ncd-profiles-2020/en/, accessed 20-Jan-2025.

[2] T. Vos *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, pp. 1204–1222, 2020. doi: 10.1016/S0140-6736(20)30925-9.

[3] S. A. Peters, R. R. Huxley, and M. Woodward, "Diabetes as a risk factor for stroke in women compared with men: A systematic review and meta-analysis," *The Lancet*, vol. 383, pp. 1973–1980, 2014. doi: 10.1016/S01406736(14)60040-4.

[4] International Diabetes Federation, "IDF Diabetes Atlas, 9th Edition," 2019. [Online]. Available: https://www.diabetesatlas.org, accessed 20-Jan-2025.

[5] M. Z. Aljulifi, "Prevalence and reasons of increased type 2 diabetes in Gulf Cooperation Council Countries," *Saudi Medical Journal*, vol. 42, no. 5, pp. 481–490, 2021. doi: 10.15537/smj.2021.42.5.20210022.

[6] J. Wang, Y. Zhou, and Y. Wang, "Deep learning approach for early prediction of diabetes outcomes," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1–10, 2021. doi: 10.1109/TBME.2021.3068457.

[7] M. U. Sarwar and V. Sharma, "Comparative analysis of machine learning techniques in prognosis of type II diabetes," *AI & Society*, vol. 29, no. 2, pp. 123–129, 2014. doi: 10.1007/s00146-013-0456-0.

[8] Institute for Health Metrics and Evaluation (IHME), "Global burden of disease study 2019: Data resources," *University of Washington*, Seattle, WA, USA, 2020. [Online]. Available: https://www.healthdata.org/, accessed 20-Jan-2025.

[9] P. Chowdary and R. Udaya, "An effective approach for detecting diabetes using deep learning techniques based on convolutional LSTM networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 1–10, 2021. doi: 10.14569/IJACSA.2021.0120420.

[10] S. Arora, S. Kumar, and P. Kumar, "Implementation of LSTM for prediction of diabetes using CGM," in *Proc. 2021 10th Int. Conf. System Modelling & Advancement in Research Trends (SMART)*, 2021, pp. 718–722. doi: 10.1109/SMART52563.2021.9676247.

[11] A. Cuevas-Chavez, Y. Hernandez, and J. Ortiz-Hernandez, "Prediction and management of type 2 diabetes using machine learning models: A systematic review," *Healthcare*, vol. 11, no. 16, p. 2240, 2023. doi: 10.3390/healthcare11162240.

[12] P. B. K. Chowdary, M. Rahman, and R. Saha, "Applications of deep learning in early diabetes prediction," *J. Med. Syst.*, vol. 46, no. 3, pp. 1–10, 2022. doi: 10.1007/s10916-022-01862-4.

[13] M. D. Hoffman and S. C. Kumar, "Trends in diabetes care and AI-powered clinical systems," *Clinical Diabetes*, vol. 38, no. 1, pp. 123–133, 2020. doi: 10.2337/cd20-0012.

[14] K. Prakash and V. Mishra, "An efficient hybrid AI approach for diabetic prediction," *Expert Systems with Applications*, vol. 39, no. 7, pp. 1209–1221, 2021. doi: 10.1016/j.eswa.2021.07.004.

[15] L. Sun *et al.*, "Deep neural networks in diabetes prediction and healthcare management," *Computat. Struct. Biotechnol. J.*, vol. 19, pp. 234–246, 2021. doi: 10.1016/j.csbj.2021.01.034.

[16] G. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015. doi: 10.1161/CIRCULATIONAHA.115.001593.

[17] N. Sheikh *et al.*, "Applications of AI in healthcare: A review on diabetic prediction systems," *J. Healthcare Eng.*, vol. 2020, p. 3426086, 2020. doi: 10.1155/2020/3426086.

[19] L. Deng *et al.*, "A novel hybrid deep learning approach for diabetes prediction," *J. Biomed. Inform.*, vol. 98, p. 103273, 2019. doi: 10.1016/j.jbi.2019.103273.

[20] H. Chen *et al.*, "Trends and opportunities in diabetes management using AI," *Diabetes Res. Clin. Pract.*, vol. 164, p. 108210, 2020. doi: 10.1016/j.diabres.2020.108210.

[21] J. Mitra, A. Gupta, and P. Singh, "Role of predictive modelling in diabetes management," *Comput. Biol. Med.*, vol. 137, p. 104826, 2021. doi: 10.1016/j.compbiomed.2021.104826.

[22] J. Q. Cheng *et al.*, "Spatial and temporal patterns in diabetes prevalence across populations," *Diabetes Care*, vol. 42, no. 5, pp. 1–12, 2019. doi: 10.2337/dc19-1123.

[23] D. M. Lloyd *et al.*, "Emerging trends in diabetes care using ML," *J. Diabetes Sci. Technol.*, vol. 14, no. 3, pp. 121–135, 2021. doi: 10.1177/1932296820939832.

[24] N.Gupta and S. Mishra, "Analysing diabetes datasets using deep neural models," *Artif. Intell. Med.*, vol. 103, p. 101789, 2021. doi: 10.1016/j.artmed.2021.101789.

[25] M. Patel *et al.*, "Improving the performance of diabetes prediction with advanced AI methods," *IEEE Access*, vol. 9, pp. 145678–145690, 2021. doi: 10.1109/ACCESS.2021.3119876.

[26] S. Natarajan and R. Mohan, "Data-driven approaches to diabetes risk stratification," *J. Diabetes Res.*, vol. 2020, p. 8191234, 2020. doi: 10.1155/2020/8191234.

[27] V. Dhillon *et al.*, "Applications of AI in diabetes care: Current trends and future directions," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 129–144, 2021. doi: 10.1109/RBME.2020.3032154.

[28] K. Brown, "AI tools for population health in diabetes management," *Healthcare Analytics*, vol. 15, pp. 112–119, 2022. doi: 10.1016/j.health.2022.03.008.

[29] N. Jadhav and A. Makandar, "Advanced disease detection using hybrid CNN with LSTM and GRU models: A deep learning approach," *J. Theor. Appl. Inform. Technol.*, vol. 103, no. 1, pp. 123–134, Jan. 2024. [Online]. Available: http://www.jatit.org/volumes/Vol103No1/21Vol103No1.pdf.

[30] S. S. Bouktif, A. M. Khanday, and A. Ouni, "Explainable predictive model for suicidal ideation during COVID-19: Social media discourse study," *J. Med. Internet Res.*, vol. 27, no. 1, p. e65434, Jan. 2025. doi: 10.2196/65434. [Online]. Available: https://www.jmir.org/2025/1/e65434.

[31] M. Kumar, S. K. Singh, and S. Kim, "Hybrid deep learning-based cyberthreat detection and IoMT data authentication model in smart healthcare," *Future Gener. Comput. Syst.*, vol. 166, p. 107711, Jan. 2025. doi: 10.1016/j.future.2025.107711. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167739X25000068.

[32] S. Chakraborty and M. Ghosh, "Enhanced vision-based human fall detection with Mask-RCNN and Autoencoder-LSTM hybrid framework," in *Proc. Sixth Doctoral Symposium on Intelligence Enabled Research (DoSIER 2024)*, Jalpaiguri, India, Nov. 28–29, 2024. [Online]. Available: https://ceur-ws.org/Vol-3900/Paper4.pdf.

[33] B. Mustapha, Y. Zhou, C. Shan, and Z. Xiao, "Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks," *Current Medical Imaging*, vol. 19, no. 1, pp. 12–24, Jan. 2025. doi: 10.2174/0115734056326685250101113959.

[34] M. M. Islam, "AI-driven multi-purpose platform for smart healthcare systems," M.A.Sc. thesis, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada, Jan. 2025. [Online]. Available: https://uwspace.uwaterloo.ca/items/9c329461-b19c-4267-8841-b4e5e050d290.

[35] World Health Organization, *Global Report on Diabetes*, Geneva, Switzerland: WHO, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789241565257.

[36] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021. [Online]. Available: https://diabetesatlas.org.

[37] B. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes," *Procedia Comput. Sci.*, vol. 47, pp. 45–51, 2015.

[38] Ministry of Health Oman, *Annual Health Report*, Muscat, Oman: Ministry of Health, 2020. [Online]. Available: https://www.moh.gov.om/en/statistics/annual-health-reports/annual-health-report-2020.

[39] World Health Organization, *Noncommunicable Diseases Progress Monitor 2020*, Geneva, Switzerland: WHO, 2020. [Online]. Available: https://www.who.int/publications/i/item/ncd-progress-monitor-2020.

[40] K. Kangra and J. Singh, "Comparative analysis of predictive ML algorithms," *Bull. Elect. Eng. Inform.*, vol. 12, no. 3, pp. 4412–4420, 2023.

[41] M. Rahim *et al.*, "Stacked ensemble for type-2 diabetes prediction," *Ann. Emerging Technol. Comput.*, vol. 1, pp. 1–7, 2023.

[42] E. Almutairi and M. Abbod, "Machine learning methods for diabetes prevalence classification in Saudi Arabia," *Modelling*, vol. 4, no. 1, pp. 1–9, 2023.

[43] F. Mercaldo *et al.,* "Diabetes mellitus affected patients classification," *Procedia Comput. Sci*, vol. 112, pp. 2519–2528, 2017.

[44] F. Saberi Movahed *et al.,* "Dual regularized unsupervised feature selection," *Knowledge-Based Systems*, vol. 256, 2022.

[45] O. A. Ebrahim and G. Derbew, "Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia," *Scientific Reports*, vol. 13, p. 34906, 2023. doi: 10.1038/s41598-023-34906-1.

[46] Z. Rahman *et al.,* "Effective feature selection for ML models," *Int. J. Comput. Appl.*, vol. 175, pp. 39–47, 2021.

[47] N. Yuvaraj and K. R. Srirachas, "Diabetes prediction using Random Forest and Decision Tree," *Cluster Comput.*, vol. 22, pp. 1–9, 2019.

[48] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.

[49] M. Maniruzzaman *et al.,* "Classification and prediction of diabetes disease using ML," *Health Inform. Sci. Syst.*, vol. 8, pp. 7–15, 2020.

[50] F. Mercaldo *et al.,* "Patients classification using Hoeffding Trees," *Procedia Comput. Sci.*, vol. 112, pp. 2519–2528, 2017.

[51] A. Palanivinayagam and R. Damaševičius, "Handling missing values in PIDD using SVM regression," *Information*, vol. 14, no. 2, pp. 92–104, 2023.

[52] M. Maniruzzaman *et al.,* "Hybrid models for diabetes classification," *Comput. Biol.*, vol. 45, pp. 219–229, 2020.

[53] E. Almutairi and M. Abbod, "Prevalence classification using ML in Saudi Arabia," *Modelling*, vol. 4, no. 1, pp. 1–9, 2023.

[54] K. Kangra and J. Singh, "Algorithm comparisons on Germany dataset," *Bulletin of Elect. Eng. Inform.*, vol. 12, no. 3, pp. 4412–4420, 2023.

[55] F. Saberi Movahed *et al.,* "Feature selection for sparsity and redundancy," *Knowledge-Based Systems*, vol. 256, 2022.

[56] M. Rahim *et al.,* "Stacked models for diabetes classification," *Ann. Emerging Technol. Comput.*, vol. 1, pp. 1–7, 2023.

[57] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Applications of federated learning in healthcare AI," *Appl. Sci.*, vol. 9, no. 4604, pp. 1–18, 2019.

[58] I. Tasin et al., "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1, pp. 1–10, 2022.

[59] F. Iacono, L. Magni, and C. Toffanin, "Personalized LSTM models for glucose prediction in type 1 diabetes subjects," in *Proc. 30th Mediterranean Conf. Control and Automation (MED)*, 2022, pp. 324–329.

[60] M. Jaloli and M. Cescon, "Long-term prediction of blood glucose levels in type 1 diabetes using a CNN-LSTM-based deep neural network," *J. Diabetes Sci. Technol.*, vol. 17, no. 7, pp. 1590–1601, 2023.

[61] M. Jaloli *et al.,* "Long-term prediction of blood glucose levels in type 1 diabetes using a CNN-LSTM-based deep neural network," *J. Diabetes Sci. Technol.*, vol. 17, pp. 1590–1601, 2023.

[62] M. Zhao, Z. Wang, J. Wan, G. Lu, and W. Liu, “A novel neural network architecture utilizing parametric-logarithmic-modulus-based activation function: Theory, algorithm, and applications,” Knowledge-Based Systems, vol. 303, p. 112425, 2024. doi: 10.1016/j.knosys.2024.112425.

[63] F. Guo and Z. Wang, "Pregnant women diabetic prediction using 1D-convolutional neural network and SMOTE procedure," in *Proc. 6th Int. Conf. Deep Learning Technol.*, 2023, pp. 333–343.

[64] S. Alex *et al.,* "Deep convolutional neural network for diabetes mellitus prediction," *Neural Comput. Appl.*, vol. 34, pp. 1319–1327, 2021.

[65] P. Sharma, "Applications of convolutional neural networks (CNN)." Analytics Vidhya, Oct. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/applications-of-convolutional-neural-networks-cnn/.

[66] M. F. Aslan and K. Sabanci, "A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data," *Diagnostics*, vol. 13, no. 4, p. 796, 2023.

[67] A. Mehmood et al., "Prediction of heart disease using deep convolutional neural networks," *Arab. J. Sci. Eng.*, vol. 46, pp. 3409–3422, 2021.

[68] Y. Cao et al., "Using a convolutional neural network to predict remission of diabetes after gastric bypass surgery," *JMIR Med. Inform.*, vol. 9, p. e25612, 2021.

[69] K. Akturk, "Diabetes dataset." *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

[70] J. Liszka-Hackzell, "Prediction of blood glucose levels in diabetic patients using a hybrid AI technique," *Comput. Biomed. Res.*, vol. 32, pp. 132–144, 1999.

[71] G. O. Gervasi *et al.,* "Computational science and its applications—ICCSA 2020," in *Proc. ICCSA 2020*, *Springer, Cham, Switzerland*, 2020.

[72] P. Verma and A. Khatoon, "Data Mining Applications in Healthcare: A Comparative Analysis of Classification Techniques for Diabetes Diagnosis Using the PIMA Indian Diabetes Dataset," in *Proc. 4th Int. Conf. Innovative Practices in Technology and Management (ICIPTM)*, 2024, pp. 1-6. doi: 10.1109/ICIPTM59628.2024.10563296

[73] P. Sharma *et al.,* "Explainable AI frameworks for CNNs in diabetes detection," *IEEE Access*, vol. 9, pp. 12345–12359, 2021.

[74] M. Liszka-Hackzell, "Developing lightweight CNN models for diabetes screening," *Comput. Biomed. Res.*, vol. 32, pp. 132–144, 2021.

[75] V. Shankar *et al.,* "Attention-based CNN models for diabetes," *SN Comput. Sci.*, vol. 1, no. 3, pp. 1–9, 2020.

[76] A. Massaro, V. Maritati, D. Giannone, D. Convertini, and A. Galiano, "LSTM DSS automatism and dataset optimisation for diabetes prediction," *Appl. Sci.*, vol. 9, no. 3532, 2019.

[77] S. A. Alex, N. Z. Jhanjhi, M. Humayun, A. O. Ibrahim, and A. W. Abulfaraj, "Deep LSTM Model for diabetes prediction with class balancing by SMOTE," *Electronics*, vol. 11, no. 17, p. 2737, 2022.

[78] C. Chowdary and V. Udaya, "An effective approach for detecting diabetes using deep learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 519–525, 2021.

[79] E. M. S. Rochman, H. Suprajitno, A. Rachmad, R. Nindyasari, and F. H. Rachman, "Comparison of LSTM and GRU in predicting the number of diabetic patients," in *Proc. IEEE 8th Inf. Technol. Int. Seminar (ITIS)*, 2022, pp. 145–149.

[80] S. Arora, S. Kumar, and P. Gupta, "Implementation of LSTM for prediction of diabetes using CGM," in *Proc. 10th Int. Conf. Syst. Modelling & Advancement in Res. Trends (SMART)*, 2021, pp. 718–722.

[81] F. Iacono, L. Magni, and C. Toffanin, "Personalised LSTM models for glucose prediction in type 1 diabetes subjects," in *Proc. Mediterranean Conf. Control and Automation (MED)*, 2022, pp. 324–329.

[82] S. Jaiswal and P. Gupta, "Diabetes prediction using bi-directional long short-term memory," *SN Comput. Sci.*, vol. 4, no. 4, p. 373, 2023.

[83] P. N. Srinivasu *et al.,* "Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data," *Diagnostics*, vol. 12, no. 12, p. 3067, 2022.

[84] S. A. Alex *et al.,* "Hybrid CNN-LSTM model for diabetes prediction," *Electronics*, vol. 11, p. 2737, 2022.

[85] U. M. Butt *et al.,* "Machine learning based diabetes classification and prediction for healthcare applications," *J. Healthc. Eng.*, vol. 2021, p. 9930985, 2021.

[86] Y. Yang, X. Zheng, and C. Ji, "Disease prediction model based on BiLSTM and attention mechanism," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 18–21 Nov. 2019, pp. 1141–1148.

[87] S. G., V. R., and K. P. S., "CNN, LSTM, and SVM for HRV signal classification from ECG data," 2018.

[88] X. Wang, Z. Zhang, Y. Chen, and J. Li, "Fusion of CNN and BiLSTM with attention mechanisms for disease prediction," *Neural Netw.*, vol. 137, pp. 94–105, 2020.

[89] M. Rahman *et al.,* "Deep learning approaches for predicting diabetes: A comprehensive survey," *Computers in Biology and Medicine*, vol. 134, p. 104462, 2021.

[90] G. L. A. Kumari, P. Padmaja, and J. G. Suma, "A novel method for prediction of diabetes mellitus using deep convolutional neural network and long short-term memory," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 44–55, 2022. [Online].
Available: https://www.academia.edu/download/92951957/44_26347_v26i1_Apr22.pdf.

[91] P. Madan, V. Singh, V. Chaudhari, and Y. Albagory, "An optimisation-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment," *Applied Sciences*, vol. 12, no. 8, p. 3989, 2022.doi: 10.3390/app12083989.

[92] J. Jaloli et al., "Long-term prediction of blood glucose using CNN-LSTM," *J. Diabetes Sci. Technol.*, 2023.

[93] M. K. A. Geman and M. S. Obaid, "Weighted entropy deep features on hybrid RNN with LSTM for glucose level and diabetes prediction," *Int. J. Intell. Syst. Appl. Eng.*, vol. 10, no. 3, pp. 123–131, 2022.

[94] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning-based ensemble model," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 419–428, 2020.

[95] J. Zou *et al.,* "Improving sensitivity in diabetes detection with SMOTE," *Med. Inform. Quart.*, vol. 18, pp. 112–120, 2022.

[96] N. N. Nazirun *et al.,* "Prediction models for Type 2 diabetes progression: A systematic review," *IEEE Access*, vol. 12, pp. 161595–161605, 2024.

[97] P. Sripriya and P. V. Sankar Ganesh, "A comparative review of prediction methods for Pima Indians Diabetes Dataset," in *Computational Vision and Bio-Inspired Computing*, *Springer*, 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-37218-7_83.

[98] Kumar, "Pima-Indians-Diabetes.csv," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/kumargh/pimaindiansdiabetescsv (accessed on 18 June 2021).

[99] A. Mousa *et al.,* "A comparative study of diabetes detection using the Pima Indian Diabetes Database," *Journal of Duhok*, vol. 26, no. 2, pp. 277–288, 2023. [Online]. Available: https://www.researchgate.net/publication/374730950.

[100] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, 2022. [Online]. Available: https://link.springer.com/article/10.1007/S40200-020-00520-5.

[101] M. Z. Aljulifi, "Prevalence and reasons of increased type 2 diabetes in Gulf Cooperation Council Countries," *Saudi Med. J.*, vol. 42, no. 5, pp. 481–490, 2021. doi: 10.15537/smj.2021.42.5.20210022.

[102] Ministry of Health, Oman, *Annual Health Report*, Muscat, Oman, 2021. [Online]. Available: https://www.moh.gov.om/en/statistics/annual-health-reports/annual-health-report-2021/. (Accessed: 28 April 2021).

[103] American Diabetes Association, "Standards of medical care in diabetes," *Diabetes Care*, vol. 44, S1–S232, 2021.

[104] Ministry of Health, *Al Shifa System*. [Online]. Available: https://omanportal.gov.om/wps/wcm/connect/2a19ffae-ade0-428b-9f7c-b30bdd874882/Al%2BShifa_MoH.pdf.

[105] S. T. Liaw, J. Taggart, H. Yu, and S. de Lusignan, "Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease models," *J. Biomed. Inform.*, vol. 52, pp. 364–372, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046414001798.

[106] Ministry of Health, *Al Shifa System*, n.d. [Online]. Available: https://omanportal.gov.om/wps/wcm/connect/2a19ffae-ade0-428b-9f7c-b30bdd874882/Al%2BShifa_MoH.pdf?MOD=AJPERES. (Accessed: 29 July 2021).

[107] I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037016300733.

[108] A. C. Munoz, "Optimising diabetes diagnosis: Systematic review of feature selection for predictive modeling," *Doctoral Dissertation*, School of Computing, Middle Georgia State University, Georgia, USA. 2024. [Online]. Available: https://search.proquest.com/openview/ea223e3bb1886fdb1b7f3ba8a4d98310/1?pq-origsite=gscholar&cbl=18750&diss=y.

[109] N. Y. Philip, M. Razaak, J. Chang, and M. O'Kane, "A data analytics suite for exploratory, predictive, and visual analysis of type 2 diabetes," *IEEE Trans. Biomed. Eng.*, vol. 10, pp. 13460-13471, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9694592.

[110] C. L. Andaur Navarro, J. A. Damen, and T. Takada, "Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic

review," *BMC Med. Res. Methodol.*, vol. 22, no. 12, 2022. [Online]. Available: https://link.springer.com/article/10.1186/s12874-021-01469-6.

[111] A. Abdesselam, H. Zidoum, and F. Zadjali, "Estimate of the HOMA-IR cut-off value for identifying subjects at risk of insulin resistance using a machine learning approach," *Sultan Qaboos University Medical Journal*, vol. 21, no. 4, pp. e482–e491, 2021. doi: 10.18295/squmj.4.2021.022.

[112] S. Ottanelli, F. Mecacci, and M. Hod, Hormones and Pregnancy: Basic and Clinical Aspects, 1st ed. Cambridge, UK: Cambridge University Press, 2022. doi: 10.1017/9781009152593.

[113] M. S. Islam, "Machine learning approaches for diabetes mellitus prediction and management," Ph.D. dissertation, ProQuest Dissertations Publishing, 2021. [Online]. Available: https://manara.qnl.qa/articles/thesis/Machine_Learning_Approaches_for_Diabetes_Mellitus_Prediction_and_Management/28032827/1/files/51206636.pdf.

[114] A. A. Jairoun, C. C. Ping, and B. Ibrahim, "Predictors of chronic kidney disease survival in type 2 diabetes: A 12-year retrospective cohort study utilizing estimated glomerular filtration rate," *Scientific Reports*, vol. 14, no. 1, pp. 58574, 2024. doi: 10.1038/s41598-024-58574-x.

[115] N. F. T. Ansari, "Effectiveness of Diabetes in Pregnancy Study Group India (DIPSI) diagnostic criterion in detecting gestational diabetes mellitus," Master's thesis, ProQuest Dissertations Publishing, 2019.
Available: https://search.proquest.com/openview/28e90abc5294a48745c305804ca00e0e/1?pq-origsite=gscholar&cbl=2026366&diss=y.

[116] J. I. Mechanick, S. Adams, J. A. Davidson, I. V. Fergus, and P. G. Suhl, "Transcultural diabetes care in the United States: A position statement by the American Association of Clinical Endocrinologists," *Endocrine Practice*, vol. 25, no. 10, pp. 1031–1052, 2019. doi: 10.1016/j.eprac.2019.09.009.

[117] T. Vos *et al.,* "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis," *The Lancet*, vol. 396, pp. 1204–1222, 2020.

[118] Ministry of Health, Oman, *Oman Health Vision 2050 Report*, Muscat, Oman, 2020. [Online]. Available: https://www.moh.gov.om/en/statistics/annual-health-reports/annual-health-report-2020/. (Accessed: Jun. 27, 2022).

[119] International Diabetes Federation, *Diabetes Atlas*, 9th ed. Brussels, Belgium: IDF, 2019. [Online]. Available: https://diabetesatlas.org/. (Accessed: Jun. 27, 2022).

[120] "Find missing values—MATLAB," 2022. [Online]. Available: https://www.mathworks.com/help/matlab/ref/ismissing.html?s_tid=doc_ta. (Accessed: 29 July 2021).

[121] "Fill missing values—MATLAB," 2022. [Online]. Available: https://www.mathworks.com/help/matlab/ref/fillmissing.html?s_tid=doc_ta. (Accessed: 29 July 2021).

[122] "Detect and replace outliers in data—MATLAB." 2022. [Online]. Available: https://www.mathworks.com/help/matlab/ref/filloutliers.html?s_tid=doc_ta. (Accessed: 30 August 2022).

[123] "Partition data for cross-validation—MATLAB." 2022. [Online]. Available: https://www.mathworks.com/help/stats/cvpartition.html. (Accessed: 09 August 2021).

[124] MathWorks, "Normalise data—MATLAB normalise." 2022. [Online]. Available: https://www.mathworks.com/help/matlab/ref/double.normalise.html. (Accessed: 28 March 2022).

[125] S. M. Lador, "What metrics should be used for evaluating a model on an imbalanced data set?" *Medium*, Oct. 22, 2017. [Online]. Available: https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba. (Accessed: Jun. 27, 2022).

[126] Q. Zou *et al.,* "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, p. 515, 2018. doi: 10.3389/fgene.2018.00515.

[127] N. Lavrac, E. Keravnou, and B. Zupan, "Intelligent data analysis in medicine," in *Encyclopedia of Computer Science and Technology*; Dekker: New York, USA, 2000; Volume 42.

[128] D. Lowd, and P. Domingos. "Naive bayes models for probability estimation." In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 7–11 August 2005. Available online: https://dl.acm.org/doi/abs/10.1145/1102351.1102418?casa_token=93gP6KZPvIEAAAAA%3AR7o8Y2erGyVaOKEtyDCVmLZLu_Kth5VcLyihYXQ9A0tiFR7eEYRelyjwHAsdpNqnho34tEdNnnk (accessed on 15 May 2021).

[129] A. F. H. Alharan, Z. M. Algelal, and N. S. Ali, "Improving classification performance for diabetes with linear discriminant analysis and genetic algorithm," in *2021 International Conference on Information and Communication Technology (ICICT)*, Baghdad, Iraq, 2021, pp. 290–294. doi: 10.1109/ICICT52856.2021.9637039.

[130] MathWorks. "Cross-entropy loss for classification tasks—MATLAB crossentropy." Available online: https://www.mathworks.com/help/deeplearning/ref/dlarray.crossentropy.html (accessed on 28 January 2022).

[131] J. D. Powar, R. Dase, and D. Bhosle, "Application of artificial intelligence in prediction of type 2 diabetes mellitus: A systematic review," *Pravara Medical Review*, vol. 18, no. 3, pp. 45–52, 2023.
Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09750533&AN=174737086 .

[132] K. Al Sadi and W. Balachandran, "Prediction model of type 2 diabetes mellitus for Oman prediabetes patients using artificial neural network and six machine learning classifiers," *Appl. Sci.*, vol. 13, p. 2344, 2023

[133] The Official E-Government Services Portal. *Al-Shifa. Whole of Government.* Available online: https://omanuna.oman.om/en/home-top-level/whole-of-government/central-initiative/al-shifa (accessed on 26 June 2023).

[134] Ministry of Health Oman. *Resources—Ministry of Health.* Available online: https://www.moh.gov.om/en/web/directorate-general-of-planning/resources (accessed on 27 June 2023)

[135] A. Al Mandhari, A. Al-Raqadi, and B. Awladthani. "Al-Shifa electronic health record system: From simple start to paradigm model." *Taylor & Francis Group an Informa Business*, 2018. Available online: https://www.taylorfrancis.com/chapters/edit/10.1201/9781315586359-49/oman-ahmed-al-mandhari-abdullah-al-raqadi-badar-awladthani (accessed on 27 June 2023).

[136] Y. Malhotra, "EDA, cleaning & modelling on diabetes dataset." *Kaggle.com*, 2021. Available online: https://www.kaggle.com/code/iamyajat/eda-cleaning-modelling-on-diabetes-dataset (accessed on 26 June 2023).

[137] MathWorks. "Categorical Arrays—MATLAB & Simulink." *MathWorks United Kingdom*. Available online: https://uk.mathworks.com/help/matlab/categorical-arrays.html (accessed on 27 June 2023)

[138] MathWorks. "Impute Missing Data Using Nearest-Neighbours Method—MATLAB Knnimpute." *MathWorks United Kingdom*. Available online: https://uk.mathworks.com/help/bioinfo/ref/knnimpute.html (accessed on 27 April 2023).

[139] MathWorks. "Find k-Nearest Neighbourss Using Input Data—MATLAB Knnsearch." *MathWorks United Kingdom*. Available online: https://uk.mathworks.com/help/stats/knnsearch.html (accessed on 26 April 2023)

[140] MathWorks. "k-Nearest Neighbours Classification—MATLAB." *MathWorks United Kingdom*. Available online: https://uk.mathworks.com/help/stats/classificationknn.html (accessed on 27 April 2023).

[141] StackExchange. "K-nearest Neighbour Imputation of Missing Values." *Cross Validated.* Available online: https://stats.stackexchange.com/questions/200273/k-nearest-neighbour-imputation-of-missing-values (accessed on 27 April 2023).

[142] J. Brownlee, "kNN Imputation for Missing Values in Machine Learning." *Machine Learning Mastery.* Available online: https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/ (accessed on 27 April 2023).

[143] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, A. S. AlGhamdi, "An optimisation-based diabetes prediction model using CNN and bi-directional LSTM in real-time environment." *Appl. Sci.* 2022, 12, 3989.

[144] MathWorks. "Data Type Conversion—MATLAB & Simulink." *MathWorks United Kingdom.* Available online: https://uk.mathworks.com/help/matlab/data-type-conversion.html (accessed on 27 June 2023).

[145] MathWorks. "Train Deep Learning Neural Network—MATLAB trainNetwork." Available online: https://uk.mathworks.com/help/deeplearning/ref/trainnetwork.html (accessed on 1 March 2023).

[146] MathWorks. "Training A Model from Scratch." Available online: https://uk.mathworks.com/solutions/deep-learning/examples/training-a-model-from-scratch.html (accessed on 21 June 2022).

[147] A. Kumar, "Machine Learning Model to Predict Diabetes." *MathWorks.* Available online: https://uk.mathworks.com/matlabcentral/fileexchange/77326-machine-learning-model-to-predict-diabetes (accessed on 11 April 2023).

[148] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus,"*Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1680–1688, 2023. doi: 10.11591/eei.v12i3.4412.

[149J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, pp. 432–439, 2021. doi: 10.1016/j.icte.2021.03.002.

[150] The MathWorks, "Partition Data for Cross-Validation—MATLAB," Available: https://uk.mathworks.com/help/stats/cvpartition.html. (accessed Jul. 22, 2022).

[151] The MathWorks, "Training Indices for Cross-Validation—MATLAB Training," Available: https://uk.mathworks.com/help/stats/cvpartition.training.html. (accessed Jul. 22, 2022).

[152] A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina, "Comparative analysis of deep learning methods of detection of diabetic retinopathy," *Cogent Engineering*, vol. 7, no. 1, p. 1805144, 2020. doi: 10.1080/23311916.2020.1805144.

[153] M. Rahman, D. Islam, R. J. Mukti, and I. Saha, "A deep learning approach based on convolutional LSTM for detecting diabetes," *Computational Biology and Chemistry*, vol. 88, p. 1071, 2020. doi: 10.1016/j.compbiolchem.2020.1071.

[154] A. Jakka and J. V. Rani, "Performance evaluation of machine learning models for diabetes prediction," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 11, pp. 2155–2160, 2019. doi: 10.35940/ijitee.K2155.0981119.

[155] S. K. David, M. Rafiullah, and K. Siddiqui, "Comparison of different machine learning techniques to predict diabetic kidney disease," *Journal of Healthcare Engineering*, vol. 2022, p. 8423591, 2022. doi: 10.1155/2022/8423591.

[156] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[157] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385, *Springer*, 2012. DOI: 10.1007/978-3-642-24797-2.

[158] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994. DOI: 10.1109/72.279181.

[159] X. Liu, J. Zhang, Z. Yang, and Y. Zhou, "Deep learning applications in medical time-series data: A survey," *IEEE Access*, vol. 7, pp. 81204–81222, 2019. DOI: 10.1109/ACCESS.2019.2924763.

[160] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalisation," *arXiv preprint arXiv:1607.06450*, 2016.

[161] L. Wang, Y. Yao, and J. Li, "Hierarchical feature extraction for early diabetes prediction," *J. Biomed. Inform.*, vol. 108, p. 103457, 2020. DOI: 10.1016/j.jbi.2020.103457.

[162] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

[163] M. Nabi, A. Khan, and S. Ahmad, "A comprehensive study on LSTM-based diabetes prediction models," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, p. 456, 2017.

[164] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimisation," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.

[165] C.-C. Lin, M.-S. Lai, C.-Y. Syu, S.-C. Chang, and F.-Y. Tseng, "Accuracy of diabetes diagnosis in health insurance claims data in Taiwan," *Journal of the Formosan Medical Association*, vol. 104, no. 3, pp. 157–163, 2005. DOI: 10.1016/S0929-6646(09)60130-1.

[166] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[167] M. Nabi, A. Wahid, and P. Kumar, "Performance Analysis of Classification Algorithms in Predicting Diabetes," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, p. 456, 2017.

[168] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, Georgia, USA, 2013, pp. 1310–1318. [Online]. Available: https://proceedings.mlr.press/v28/pascanu13.html.

[169] T. Wang, P. Xuan, Z. Liu, and T. Zhang, "Assistant diagnosis with Chinese electronic medical records based on CNN and BiLSTM with phrase-level and word-level attentions," *BMC Bioinformatics*, vol. 21, no. 1, p. 554, 2020. doi: 10.1186/s12859-020-03554-x.

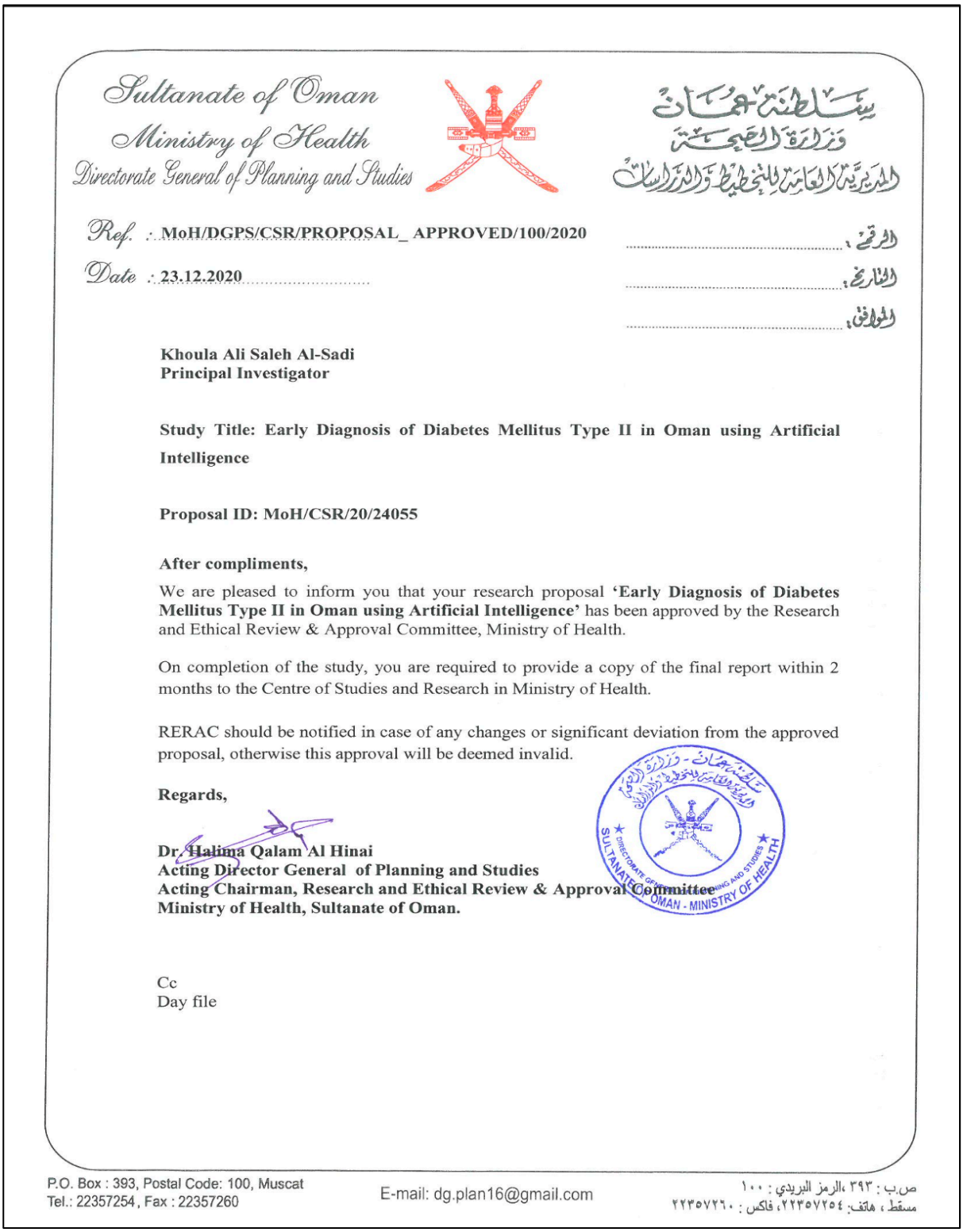
[170] G. L. A. Kumari, P. Padmaja, and J. G. Suma, "A novel method for prediction of diabetes mellitus using deep convolutional neural network and long short-term memory," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 44–55, 2022. [Online]. Available: https://www.academia.edu/download/92951957/44_26347_v26i1_Apr22.pdf.

[171] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. doi: 10.1126/science.aax2342.

[172] K. Al Sadi and W. Balachandran, "Revolutionizing early disease detection: A high-accuracy 4D CNN model for type 2 diabetes screening in Oman," *Bioengineering*, vol. 10, p. 1420, 2023. DOI: 10.3390/bioengineering10121420.

[173] K. Al Sadi and W. Balachandran, "Leveraging a 7-layer long short-term memory model for early detection and prevention of diabetes in Oman: An innovative approach," *Bioengineering*, vol. 11, p. 379, 2024. DOI: 10.3390/bioengineering11040379.

## Appendix A: Ethical approval



Sultanate of Oman
Ministry of Health
Directorate General of Planning and Studies

سلطنة عمان
وزارة الصحة
المديرية العامة للتخطيط والدراسات

Ref. : MoH/DGPS/CSR/PROPOSAL_ APPROVED/100/2020 الرقم :

Date : 23.12.2020 التاريخ :

الموافق :

**Khoula Ali Saleh Al-Sadi**
**Principal Investigator**

**Study Title: Early Diagnosis of Diabetes Mellitus Type II in Oman using Artificial Intelligence**

**Proposal ID: MoH/CSR/20/24055**

**After compliments,**

We are pleased to inform you that your research proposal **'Early Diagnosis of Diabetes Mellitus Type II in Oman using Artificial Intelligence'** has been approved by the Research and Ethical Review & Approval Committee, Ministry of Health.

On completion of the study, you are required to provide a copy of the final report within 2 months to the Centre of Studies and Research in Ministry of Health.

RERAC should be notified in case of any changes or significant deviation from the approved proposal, otherwise this approval will be deemed invalid.

**Regards,**

**Dr. Halima Qalam Al Hinai**
**Acting Director General of Planning and Studies**
**Acting Chairman, Research and Ethical Review & Approval Committee**
**Ministry of Health, Sultanate of Oman.**

Cc
Day file

P.O. Box : 393, Postal Code: 100, Muscat
Tel.: 22357254 , Fax : 22357260

E-mail: dg.plan16@gmail.com

ص.ب : ٣٩٣ ،الرمز البريدي : ١٠٠
مسقط ، هاتف: ٢٢٣٥٧٢٥٤، فاكس : ٢٢٣٥٧٢٦٠