Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer/IT Engineering

by

Patel Hiteshbhai Bhimabhai

Enrollment No.: 199999913574

Under the Supervision of

Dr. Keyur N. Brahmbhatt

Professor & Head of Department,

Information Technology Department

BVM Engineering College, V.V. Nagar, Anand

GUJARAT TECHNOLOGICAL UNIVERSITY

AHMEDABAD

September 2026

Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer/IT Engineering

by

Patel Hiteshbhai Bhimabhai

Enrollment No.: 199999913574

Under the Supervision of

Dr. Keyur N. Brahmbhatt

Professor & Head of Department,

Information Technology Department

BVM Engineering College, V.V. Nagar, Anand

GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD

September 2026

DECLARATION

I declare that the thesis entitled **"Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data"** submitted by me for the degree of Doctor of Philosophy is the record of research work carried out by me during the period from **2019** to **2026** under the supervision of **Dr. Keyur N. Brahmbhatt**, Professor & Head, Information Technology Department, BVM Engineering College, Vallabh Vidyanagar and this has not formed the basis for the award of any degree, diploma, associateship and fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities if noticed in the thesis.

Signature of the Research Scholar: .....

Name of Research Scholar: **Patel Hiteshbhai Bhimabhai**

Date: 09/09/2026

Place: **Ahmedabad**

CERTIFICATE

I certify that the work incorporated in the thesis entitled **"Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data"** submitted by **Patel Hiteshbhai Bhimabhai** was carried out by the candidate under my supervision/guidance. To the best of my knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of Supervisor: .....

Name of Supervisor: **Dr Keyur N. Brahmbhatt**

Date: 09/09/2026

Place: **Ahmedabad**

Course-work Completion Certificate

This is to certify that **Patel Hiteshbhai Bhimabhai** (Enrolment no. 199999913574) is a PhD scholar enrolled for the PhD program in the branch **Computer/IT Engineering** of Gujarat Technological University, Ahmedabad.

(Please tick the relevant option(s))

☐ He has been exempted from the course-work (successfully completed during M.Phil. Course)

☐ He has been exempted from Research Methodology Course only (successfully completed during M.Phil. Course)

☑ He has successfully completed the PhD course work for the partial requirement for the award of PhD Degree. His performance in the course work is as follows –

Grade Obtained in Research Methodology [PHD2001]	Grade Obtained in Research and Publication Ethics [PHD2002]	Grade Obtained in Self-Study Course/ Contact Program [PHD2003]
CC	AA	AB

Signature of Supervisor: .....

Name of Supervisor: **Dr Keyur N. Brahmbhatt**

Date: 09/09/2026 Place: **Ahmedabad**

Originality Report Certificate

It is certified that PhD Thesis titled **"Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data"** submitted by **Hiteshbhai Bhimabhai Patel** has been examined by me. I undertake the following:

a. The thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using **Turnitin Plagiarism Detection Software** (copy of originality report attached) and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e. permitted similarity index <10%).

Signature of the Research Scholar: .....

Name of Research Scholar: **Patel Hiteshbhai Bhimabhai**

Date: 09/09/2026 Place: **Ahmedabad**

Signature of Supervisor: .....

Name of Supervisor: **Dr Keyur N. Brahmbhatt**

Date: 09/09/2026 Place: **Ahmedabad**

Hitesh Patel

Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data

Quick Submit

Quick Submit

Gujarat Technological University

Document Details

Submission ID

trn:oid:::1:3615983087

Submission Date

Jul 23, 2026, 3:24 PM GMT+5:30

Download Date

Jul 23, 2026, 3:37 PM GMT+5:30

File Name

Thesis_-_199999913574_-_22-07-26.pdf

File Size

2.6 MB

140 Pages

31,072 Words

208,224 Characters

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 14 words)

Exclusions

- 1 Excluded Source
- 13 Excluded Matches

Match Groups

- **62 Not Cited or Quoted 5%**
  Matches with neither in-text citation nor quotation marks
- **0 Missing Quotations 0%**
  Matches that are still very similar to source material
- **0 Missing Citation 0%**
  Matches that have quotation marks, but no in-text citation
- **0 Cited and Quoted 0%**
  Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 3% Publications
- 4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Ph. D. Thesis Non-Exclusive License to

GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a PhD. Research Scholar at GTU and in the interests of the facilitation of research at GTU and elsewhere, I, **Patel Hiteshbhai Bhimabhai** having Enrollment No. **199999913574** hereby grant a non-exclusive, royalty-free and perpetual license to GTU on the following terms:

a. The University is permitted to archive, reproduce and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as the "Work") anywhere in the world, for non-commercial purposes, in all forms of media;

b. The University is permitted to authorize, sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);

c. The University is authorized to submit the Work at any National / International Library, under the authority of their "Thesis Non-Exclusive License";

d. The Universal Copyright Notice (©) shall appear on all copies made under the authority of this license;

e. I undertake to submit my thesis, through my University, to any Library and Archives. Any abstract submitted with the thesis will be considered to form part of the thesis.

f. I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this nonexclusive license.

g. If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.

h. I understand that the responsibility for the matter as mentioned in the paragraph (g) rests with the authors / me solely. In no case shall GTU have any liability for any acts / omissions / errors / copyright infringement from the publication of the said thesis or otherwise.

i. I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my University in this nonexclusive license.

j. GTU logo shall not be used /printed in the book (in any manner whatsoever) being published or any promotional or marketing materials or any such similar documents.

k. The following statement shall be included appropriately and displayed prominently in the book or any material being published anywhere: "The content of the published work is part of the thesis submitted in partial fulfilment for the award of the degree of Ph.D. in **Computer/IT Engineering** of the Gujarat Technological University".

l. I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this nonexclusive license. I shall keep GTU indemnified from any and all claims from the Publisher(s) or any third parties at all times resulting or arising from the publishing or use or intended use of the book / such similar document or its contents.

m. I am aware of and agree to accept the conditions and regulations of Ph.D. including all policy matters related to authorship and plagiarism.

Signature of the Research Scholar: .....

Name of Research Scholar: **Patel Hiteshbhai Bhimabhai**

Date: 09/09/2026 Place: **Ahmedabad**

Recommendation of the Research Supervisor: **Approved**

Signature of Supervisor: .....

Name of Supervisor: **Dr Keyur N. Brahmbhatt**

Date: 09/09/2026 Place: **Ahmedabad**

Seal:

Thesis Approval Form

The viva-voce of the PhD Thesis submitted by **Patel Hiteshbhai Bhimabhai** (Enrollment No. **199999913574**) entitled **"Deep Learning Approach for Type 2 Diabetes Prediction from Clinical Data"** was conducted on 09/09/2026 at Gujarat Technological University.

(Please tick any one of the following options)

☑ The performance of the candidate was satisfactory. We recommend that he be awarded the PhD degree.

☐ Any further modifications in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the Supervisor or request of Independent Research Scholar after which viva-voce can be re-conducted by the same panel again. (Briefly specify the modification suggested by the panel)

☐ The performance of the candidate was unsatisfactory. We recommend that he should not be awarded the PhD degree. (The panel must give justification for rejecting the research work)

Dr. Keyur N. Brahmbhatt

Name and Signature of Supervisor with Seal

Dr. Udai Pratap Rao

1) (External Examiner 1) Name and Signature

09/09/2026

Dr. Viral Kapadia

2) (External Examiner 2) Name and Signature

Acknowledgement

I would like to express my sincere gratitude to all those who have directly and indirectly contributed to the successful completion of this doctoral research. Their continuous guidance, encouragement, and support have been invaluable throughout this academic journey.

First and foremost, I would like to express my deepest gratitude to my respected Ph.D. supervisor, Dr. Keyur N. Brahmbhatt, for their invaluable academic guidance, constant encouragement, constructive criticism and insightful suggestions throughout the course of this research. Their expertise, patience, and motivation have been instrumental in shaping this work and helping me overcome various challenges during the research process.

The success of this work would not be possible without the guidance of my Ph.D. supervisor, Dr. Keyur N. Brahmbhatt sir and DPC members Dr. Viral H. Borisagar sir and Dr. Dinesh J. Prajapati sir. I am heartily grateful to them for their valuable suggestions that have improved the quality of my work, and, their support that helped to achieve this milestone.

I also extend my sincere appreciation to Dr. Kailash M. Patil sir, Dr. Madhav Astik and Mr. Gopal K. Gohel for their constant motivation, encouragement, and for helping me stay positive and focused during the challenging phases of this research. I am especially grateful to Dr. Bharat V. Chawda sir and Dr. Kailash M. Patil sir for providing valuable feedback for thesis writing.

I am also grateful to the reviewers and domain experts who evaluated my research papers in various journals. Their constructive feedback, critical insights, and identification of shortcomings at different stages have been invaluable in refining and strengthening this work. I sincerely acknowledge their contributions.

I am also thankful to Honorable Vice Chancellor, Registrar, Controller of Examination, Dean PhD section, and all staff members of the PhD Section of Gujarat Technological University (GTU) for their assistance and support. I express my sincere gratitude to Charutar Vidya Mandal and Dr. K. M. Makwana, Principal, B. & B. Institute of Technology, Vallabh Vidyanagar, for granting me permission and providing continuous support to pursue and complete this research work.

I am grateful to Mrs. Khyati T. Vaghela, Head of the Information Technology Department, for her cooperation and constant support. I am also extend my appreciation to Ms. Hiral B. Chauhan and Mr. Kalpesh A. Kshatriya, staff members of the Information Technology Department for their motivation, support, and for creating a friendly and conducive academic environment that helped me balance my teaching responsibilities with my research work effectively.

I express my heartfelt gratitude to my parents and all my family members for their constant love, blessings, and unwavering support. Being part of a family, their collective encouragement, patience, and understanding have been a strong foundation throughout this period. I am especially thankful to my wife Komal, for her continuous support, patience, and belief in me, which gave me the confidence to overcome challenges and accomplish this work. I also extend my loving thanks to my two daughters Maahi and Riya, whose presence has been a constant source of happiness and inspiration.

I would like to sincerely acknowledge Dr. Mayur H. Rabari (BHMS) for his valuable medical guidance and insightful support, which contributed to my Ph.D. research work. I also extend my sincere thanks to my friends, Dr. Jigar M. Patel and Dr. Vipul Solanki for their encouragement, support, and for maintaining a positive and uplifting environment during the course of this research.

Finally, I express my heartfelt gratitude to all those who have directly or indirectly contributed to the successful completion of this research work.

- Hitesh B. Patel

Abstract

The early detection of Type 2 Diabetes Mellitus (T2DM) is a critical challenge in healthcare due to its increasing prevalence and associated complications. Accurate prediction models can assist in timely diagnosis and effective clinical decision-making. However, medical datasets often suffer from issues such as missing values, class imbalance, and complex nonlinear relationships, which limit the performance of traditional machine learning approaches.

This research proposes a novel hybrid framework that integrates data-level enhancement and model-level optimization techniques for improved diabetes prediction. The proposed framework consists of two major components. **Algorithm-1** focuses on preprocessing and data augmentation, where invalid zero values in clinical features are handled using an age-group-based mean imputation strategy. To address class imbalance, an **Adversarial Variational Autoencoder (AVAE)** is employed to generate realistic synthetic samples, thereby improving data diversity and representation.

Algorithm-2 introduces a deep learning model based on a **Morlet Wavelet assisted** architecture with **FFT overlap-add convolution (MW-FFT-OAconv).** The use of Morlet wavelet activation enhances nonlinear feature extraction, while FFT-based convolution improves computational efficiency. Additionally, a **Weighted Mean of Vectors (WMOV) optimization** algorithm is proposed to effectively tune model parameters and improve convergence.

The proposed hybrid framework is evaluated using the **PIMA Diabetes Dataset** and an additional **Diabetes Prediction Dataset**. Experimental results demonstrate that the proposed approach achieves a **classification accuracy of 98.44%,** significantly outperforming conventional machine learning models and existing research methods. The confusion matrix analysis indicates low false positive and false negative rates, highlighting the reliability of the model. Furthermore, cross-validation results confirm the robustness and generalization capability of the proposed framework.

Overall, the proposed approach provides an efficient and reliable solution for diabetes prediction and has strong potential for application in real-world clinical decision support systems.

Table of Contents

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
DL	Deep Learning
AVAE	Adversarial Variational Autoencoder
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
RF	Random Forest
NB	Naïve Bayes
DT	Decision Tree
LR	Logistic Regression
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
CNN	Convolutional Neural Network
FFT	Fast Fourier Transform
OAconv	Overlap-Add Convolution
MW	Morlet Wavelet
WMOV	Weighted Mean of Vectors
PIDD	Pima Indians Diabetes Dataset
BMI	Body Mass Index
HbA1c	Hemoglobin A1c
T2DM	Type 2 Diabetes Mellitus
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Acc	Accuracy
Prec	Precision
Rec	Recall
F1	F1-Score
MCC	Matthews Correlation Coefficient
TPR	True Positive Rate
TNR	True Negative Rate
FPR	False Positive Rate
FNR	False Negative Rate
CER	Classification Error Rate
ReLU	Rectified Linear Unit

List of Symbols

Symbol	Description
$\boldsymbol{X}$	Input feature vector
$\boldsymbol{y}$	Target/output variable
$\boldsymbol{\hat{y}}$	Predicted output
$\boldsymbol{N}$	Number of samples
$\boldsymbol{D}$	Number of features
$\boldsymbol{x_i}$	i^{th} data sample
$\boldsymbol{y_i}$	Actual label of i^{th} sample
$\boldsymbol{\theta}$	Model parameters/weights
$\boldsymbol{f(x)}$	Model prediction function
$\boldsymbol{w}$	Weight parameter
$\boldsymbol{b}$	Bias term
$\boldsymbol{\alpha}$	Learning rate
$\boldsymbol{\beta}$	Optimization coefficient
$\boldsymbol{\sigma}$	Standard deviation / control parameter
$\boldsymbol{z}$	Latent variable
$\boldsymbol{\mu}$	Mean value
$\boldsymbol{TP}$	True Positives
$\boldsymbol{TN}$	True Negatives
$\boldsymbol{FP}$	False Positives
$\boldsymbol{FN}$	False Negatives
$\boldsymbol{Acc}$	Accuracy
$\boldsymbol{Prec}$	Precision
$\boldsymbol{Rec}$	Recall (Sensitivity)
$\boldsymbol{F1}$	F1-score
$\boldsymbol{\phi}$	Activation function
$\boldsymbol{L}$	Loss function
E	Expectation operator
log	Logarithmic function

List of Figures

List of Tables

CHAPTER 1

Introduction

This chapter introduces the clinical and technological foundations of Type 2 Diabetes Mellitus (T2DM) prediction. It presents an overview of the epidemiology, pathophysiology, and clinical characteristics of T2DM, followed by a discussion of conventional diagnostic approaches and the emerging role of artificial intelligence in healthcare analytics. The chapter further examines the major challenges associated with medical data analysis, including missing values, class imbalance, feature complexity, and model interpretability. Finally, it presents the research motivation, problem statement, objectives, scope, major contributions, and organization of the thesis, thereby establishing the foundation for the proposed research.

1.1 Introduction to Type 2 Diabetes Mellitus

Type 2 Diabetes Mellitus (T2DM) has emerged as one of the most prevalent chronic non-communicable diseases worldwide, posing a major public health challenge due to its increasing incidence and associated complications. A comprehensive understanding of the medical foundations of T2DM is essential before exploring computational approaches for its prediction and classification. This section presents the clinical background of the disease, including its definition, pathophysiology, global epidemiological trends, and major clinical manifestations. Establishing this foundation facilitates a better understanding of the need for intelligent predictive models and artificial intelligence-based healthcare solutions presented in the subsequent chapters.

1.1.1 Definition and Classification

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder characterized by persistent hyperglycemia resulting from insulin resistance in peripheral tissues and progressive dysfunction of pancreatic β-cells [1]. Unlike Type 1 Diabetes Mellitus (T1DM), which results from autoimmune destruction of insulin-producing β-cells, T2DM develops gradually and is primarily associated with impaired insulin action accompanied by an inadequate compensatory insulin secretory response. As the disease progresses, pancreatic β-cells progressively lose their functional capacity, leading to sustained elevation of blood glucose levels and long-term metabolic complications.

According to the American Diabetes Association (ADA), diabetes mellitus is classified into four principal categories: Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM), Gestational Diabetes Mellitus (GDM), and specific types of diabetes resulting from other causes, including genetic defects affecting β-cell function, diseases of the exocrine pancreas, endocrinopathies, and drug- or chemically induced diabetes [1]. Among these categories, T2DM accounts for approximately 90–95% of all diagnosed diabetes cases worldwide, making it the most prevalent form of the disease.

The diagnosis of T2DM is established using standardized biochemical criteria, including fasting plasma glucose (FPG), two-hour plasma glucose during an oral glucose tolerance test (OGTT), glycated haemoglobin (HbA1c), or random plasma glucose in individuals presenting with classic symptoms of hyperglycaemia [1]. Although these laboratory investigations provide reliable confirmation of diabetes, they primarily detect the disease after metabolic abnormalities have become clinically apparent. Consequently, increasing attention has been directed toward predictive computational models capable of identifying individuals at high risk before the onset of irreversible complications.

T2DM is a multifactorial disease whose development is influenced by a complex interaction of genetic susceptibility, obesity, sedentary lifestyle, ageing, dietary habits, environmental factors, and other metabolic abnormalities [2], [3]. These interacting risk factors exhibit highly nonlinear relationships that are often difficult to model using conventional statistical techniques. Therefore, advanced machine learning and deep learning approaches have emerged as promising alternatives for accurately modelling these complex interactions and facilitating early prediction of T2DM.

1.1.2 Pathogenesis of Type 2 Diabetes

The pathogenesis of Type 2 Diabetes Mellitus (T2DM) is a complex and progressive process involving the combined effects of insulin resistance, progressive pancreatic β-cell dysfunction, and metabolic dysregulation. Initially, insulin resistance develops in peripheral tissues such as skeletal muscle, adipose tissue, and the liver, reducing the ability of these tissues to effectively utilize circulating glucose. To compensate for this reduced insulin sensitivity, pancreatic β-cells increase insulin secretion in an attempt to maintain normal blood glucose levels. However, prolonged metabolic stress gradually impairs β-cell function, leading to inadequate insulin production and persistent hyperglycaemia [4].

At the cellular level, insulin resistance decreases glucose uptake by skeletal muscles while simultaneously increasing hepatic glucose production through enhanced gluconeogenesis. Elevated blood glucose concentrations further aggravate metabolic abnormalities through mechanisms such as glucotoxicity, lipotoxicity, oxidative stress, and chronic low-grade inflammation. These pathological processes accelerate β-cell deterioration and contribute to both microvascular and macrovascular complications associated with T2DM [5].

The progression of T2DM is generally gradual and often remains asymptomatic for several years. During this preclinical stage, many individuals experience prediabetes, characterized by impaired fasting glucose (IFG) and impaired glucose tolerance (IGT), which significantly increase the risk of developing overt diabetes. Early identification of individuals at this stage provides an opportunity for timely lifestyle modifications and therapeutic interventions that may delay or even prevent disease progression.

The development of T2DM is influenced by numerous interacting risk factors, including genetic predisposition, obesity, sedentary lifestyle, unhealthy dietary habits, ageing, hypertension, dyslipidaemia, and family history. These factors interact in a highly nonlinear manner, making disease progression difficult to describe using conventional statistical approaches alone. Consequently, computational intelligence techniques, particularly machine learning and deep learning algorithms, have gained considerable attention because of their ability to model complex nonlinear relationships and discover hidden patterns within multidimensional clinical datasets.

Understanding the underlying pathophysiological mechanisms of T2DM is therefore essential not only for improving clinical diagnosis but also for designing intelligent predictive models capable of supporting early disease detection and personalized clinical decision-making. This medical foundation provides the rationale for the artificial intelligence-based predictive framework proposed in the subsequent chapters of this thesis.

1.1.3 Global Prevalence and Epidemiological Trends

T2DM has evolved into one of the most important global public health concerns of the modern era. Over the past two decades, epidemiological data have consistently demonstrated a steady and substantial increase in the number of individuals affected worldwide [2], [6]. This upward trend is attributed to demographic transitions, increasing life expectancy, urbanization, sedentary lifestyles, and rising obesity rates across both developed and developing nations.

Recent global assessment data shows, the numbers increased dramatically of adults living with diabetes, since the early 2000s. As illustrated in Figure 1.1, the global prevalence rose from approximately 151 million cases in 2000 to over 500 million cases by the early 2020s. This consistent growth pattern reflects not only improved detection mechanisms but also a genuine rise in disease incidence driven by lifestyle and metabolic risk factors.

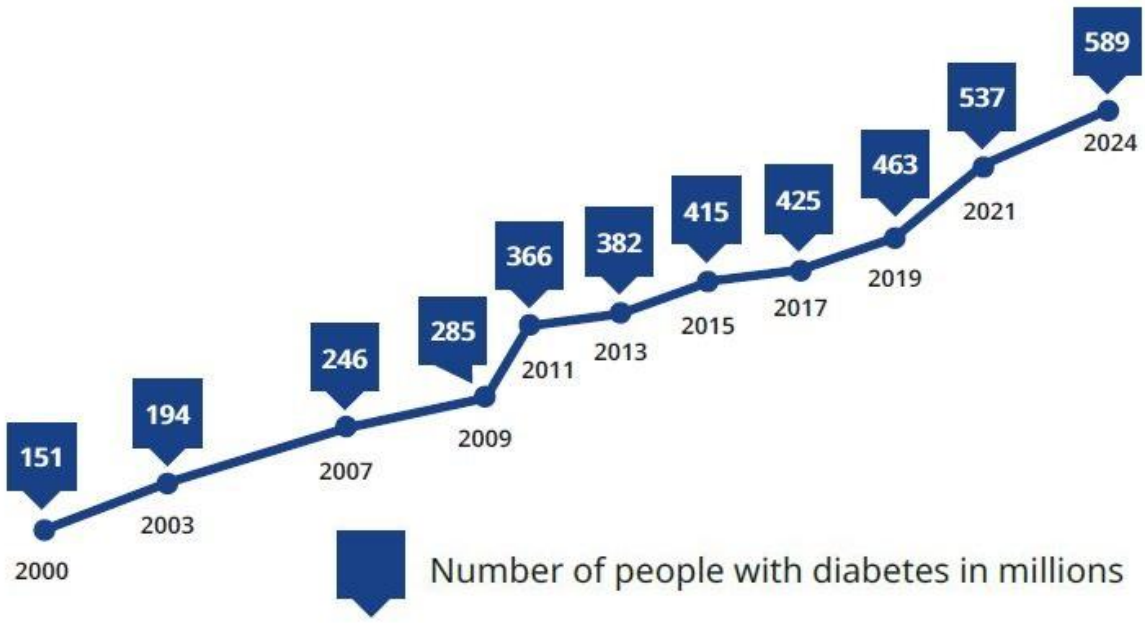

Figure 1.1 Global number of people living with diabetes (2000–2024).

Source: World Health Organization Global Diabetes Report [7].

The data presented in Figure 1.1 highlights the accelerating expansion of diabetes prevalence across successive years. The increase is not linear but progressively steeper in recent periods, indicating that preventive strategies have not kept pace with changing societal behaviors. The growing magnitude of affected populations places significant pressure on healthcare infrastructures, mainly in poor and under-developed countries where access to preventive care may be limited.

Projections further suggest that this trend will continue in the coming decades if effective intervention strategies are not implemented. Long-term epidemiological forecasts estimate a substantial rise in global diabetes burden by 2030, 2045, and 2050, as shown in Figure 1.2. These projections emphasize that diabetes is not merely a current crisis but a future global health challenge.

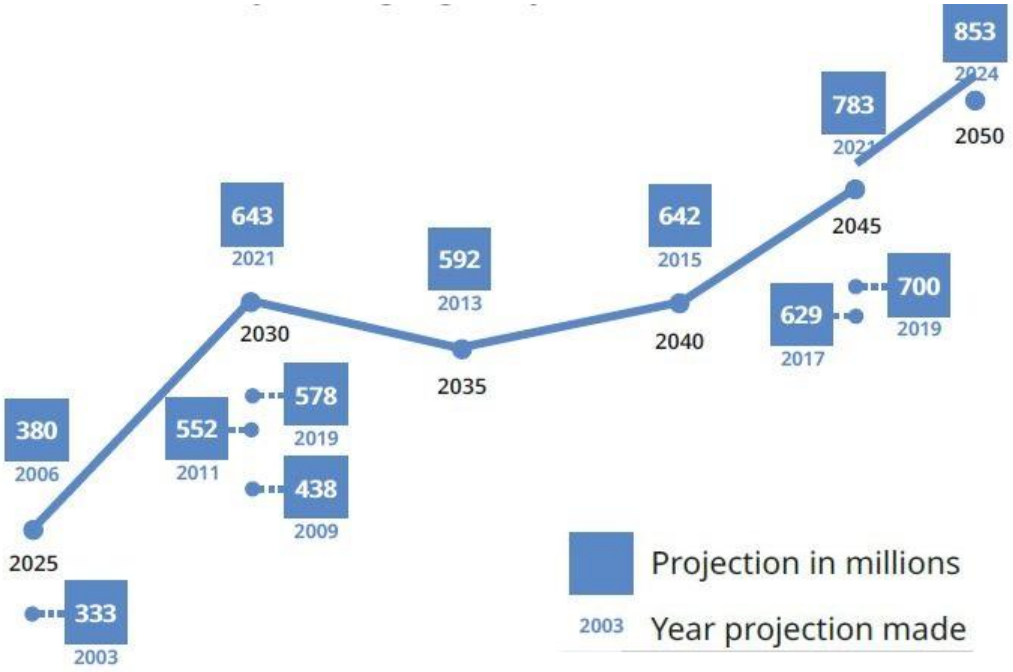

Figure 1.2 Projected global diabetes prevalence up to 2050.

Source: World Health Organization Global Diabetes Report [7].

Figure 1.2 illustrates projected increases that could exceed 700–800 million cases globally by mid-century, depending on demographic and lifestyle trends. Such projections are derived from population growth models, aging demographics, and observed incidence rates [2], [3]. The anticipated expansion underscores the urgency of shifting from reactive treatment-based healthcare models toward preventive and predictive frameworks.

The escalating prevalence of Type 2 Diabetes has profound implications not only for individual health outcomes but also for national economies and global healthcare systems. Rising treatment costs, increased hospitalization rates, and long-term management of complications collectively contribute to substantial financial and social burdens. Consequently, there is an urgency for intelligent predictive systems efficient for identifying high-risk individuals at earlier stages, thereby supporting targeted preventive interventions.

In this context, epidemiological evidence strongly supports the integration of advanced computational approaches into diabetes risk assessment. As the disease burden continues to rise, predictive analytics and artificial intelligence-based models may be a crucial part in mitigating future impact through early screening and informed clinical decision-making.

1.1.4 Clinical Manifestations and Long-Term Complications

Type 2 Diabetes is often specified by a gradual onset of clinical symptoms, many of which may remain unnoticed during the early stages of the disease. Unlike acute medical conditions that present with immediate and severe manifestations, T2DM develops progressively, allowing metabolic abnormalities to persist for extended periods before diagnosis. This delayed recognition contributes significantly to the development of long-term complications.

Figure 1.3 Type 2 Diabetes Mellitus - Common symptoms

Common early symptoms of Type 2 Diabetes include abnormal urination (polyuria), extream thirst (polydipsia), constant craving (polyphagia), persistent fatigue, unexplained weight loss, indistinct sight, delayed wound healing, recurrent infections, and skin darkening in certain body regions. These manifestations result primarily from sustained hyperglycemia and impaired glucose utilization within peripheral tissues. A conceptual overview of these commonly reported symptoms is shown in Figure 1.3.

The presence of such symptoms reflects underlying metabolic dysregulation. For instance, excessive urination occurs due to osmotic diuresis caused by elevated blood glucose levels, while persistent fatigue results from inefficient cellular glucose uptake. Sluggish wound healing and higher sensitivity to infections are linked to vascular impairment and immune dysfunction. Skin darkening, particularly in the neck and axillary regions, is often associated with insulin resistance and acanthosis nigricans.

If left unmanaged, prolonged hyperglycemia leads to severe damage in small and large blood vessels. Small blood vessels damage ultimately result in eye disease, kidney disease, and nerve damage. [5]. Large blood vessels damage involve cardiovascular disorders such as coronary artery disease and stroke, representing prime causes of illness and mortality among diabetic individuals [3].

A critical concern is that many individuals remain asymptomatic or experience mild symptoms during early disease progression. By the time overt clinical signs become evident, significant vascular and metabolic damage may have already occurred. Consequently, reliance solely on symptom-based detection is insufficient for effective disease management.

These clinical characteristics highlight the necessity for predictive frameworks capable of identifying individuals at elevated risk before irreversible complications develop. The integration of computational intelligence techniques into healthcare analytics offers a promising direction for facilitating early detection and preventive intervention strategies.

1.2 Traditional Approaches for Diabetes Diagnosis

Before the emergence of computational intelligence techniques, the diagnosis and risk assessment of Type 2 Diabetes primarily relied on biochemical testing and statistical

modeling. These conventional approaches have formed the foundation of clinical practice for several decades and remain essential for confirming disease presence. However, while clinically validated and widely adopted, traditional diagnostic systems exhibit certain limitations when applied to early prediction and large-scale population screening. This section reviews these conventional methods and discusses their constraints.

1.2.1 Laboratory-Based Diagnostic Methods

The diagnosis of T2DM traditionally relies on standardized biochemical measurements that assess blood glucose regulation. According to internationally accepted clinical guidelines, diabetes is diagnosed using blood glucose in fasting stage (FPG), hemoglobin A1c levels [1]. These diagnostic criteria are designed to identify chronic hyperglycemia and provide reliable confirmation of disease presence.

Fasting plasma glucose measures blood glucose concentration before breakfast, while the hemoglobin A1c test checks the body's response to a standardized glucose load. HbA1c reflects moderate blood glucose measures over approximately three months and has become widely used due to its convenience and stability [1]. These methods are clinically validated and form the foundation of diabetes diagnosis worldwide.

Although laboratory-based testing is accurate for confirming diabetes, it is inherently reactive rather than predictive. Diagnosis typically occurs after metabolic dysfunction has progressed to measurable hyperglycemia. Consequently, individuals in early or prediabetic stages may remain undetected until significant physiological impairment has already occurred.

Moreover, laboratory testing requires clinical infrastructure, repeated patient visits, and standardized conditions, which may limit accessibility in resource-constrained settings. These constraints emphasize the requirement for complementary predictive models which can identify high-risk individuals before clinical thresholds are exceeded.

1.2.2 Statistical Risk Prediction Models

Before the widespread adoption of machine learning techniques, statistical models were commonly employed to estimate diabetes risk. Logistic regression, Cox proportional

hazards models, and discriminant function analysis have been extensively applied for examining associations between risk factors and disease onset [8], [9].

Such models rely on predefined assumptions regarding linearity, independence, and distribution of variables. Risk scores derived from these methods incorporate demographic and medical attributes such as age, weight-for-height index (BMI), family history, vascular pressure, and lipid profile. Several population-based risk assessment tools have been developed using these statistical frameworks.

While statistical models provide interpretability and clinical transparency, they are often limited in capturing complex nonlinear interactions among multiple risk factors. The progression of Type 2 diabetes involves multifactorial relationships that may not conform to simple linear patterns. As a result, predictive performance may decline when applied to heterogeneous or high-dimensional datasets.

Furthermore, traditional statistical approaches may struggle with multicollinearity among clinical variables and may require manual feature selection to prevent overfitting. These constraints reduce their adaptability to large-scale and diverse healthcare datasets.

1.2.3 Limitations of Conventional Diagnostic Systems

Despite their clinical validity, conventional diagnostic systems exhibit several limitations that restrict their effectiveness in early risk assessment.

First, laboratory-based methods detect diabetes primarily after the disease has progressed to measurable hyperglycemia. This reactive approach limits opportunities for preventive intervention during earlier metabolic disturbances.

Second, statistical risk models often assume simplified relationships among variables, which may not reflect the complex pathophysiology of T2DM. The nonlinear mutual influence between insulin resistance, obesity, inflammation, and genetic predisposition may not be adequately modeled using traditional techniques.

Third, conventional systems are typically developed using limited datasets and may lack generalizability across different populations. Variations in ethnicity, lifestyle, and

environmental exposure can influence disease progression, necessitating more flexible predictive frameworks.

Finally, traditional models generally do not incorporate advanced feature transformation or generative data augmentation strategies. As healthcare data becomes increasingly digitized and multidimensional, more sophisticated computational approaches are required to harness its full predictive potential.

These limitations collectively motivate the exploration of artificial intelligence–based predictive systems capable of modeling complex relationships, improving generalization, and supporting early detection strategies.

1.3 Emergence of Artificial Intelligence in Medical Services

The growing complexity of medical data and the increasing burden of chronic diseases have accelerated the adoption of artificial intelligence in medical research and clinical practice. Advances in computational power, data storage, and algorithmic design have enabled the enhancement of smart systems capable of learning patterns from large-scale healthcare datasets. Artificial intelligence techniques, specially machine learning and deep learning, are increasingly being employed to enhance diagnostic accuracy, risk assessment, and decision-making processes in healthcare [10], [11]. This section outlines the evolution of these technologies and their relevance to disease prediction.

1.3.1 Evolution of Machine Learning in Medical Applications

The rapid digitization of healthcare systems has generated large volumes of structured and unstructured medical data, including electronic health records, laboratory reports, diagnostic images, and physiological measurements. Traditional analytical techniques often struggle to extract meaningful insights from such multidimensional datasets. In response, artificial intelligence (AI), mainly machine learning (ML), has come out as a revolutionary approach in medical data analysis [10], [11].

Machine learning refers to a class of algorithms capable of learning patterns from data and improving performance through experience without explicit rule-based programming [8].

In healthcare, ML techniques have been applied to disease prediction, patient stratification, medical imaging analysis, and treatment outcome forecasting [12], [13].

Early applications of ML in diabetes prediction employed algorithms such as logistic regression, decision trees, support vector machines, and ensemble methods including random forests [14]. These models demonstrated improved predictive performance compared to traditional statistical risk scores, particularly in handling nonlinear interactions among multiple clinical attributes.

However, conventional ML models typically require manual feature selection and domain expertise to determine relevant input variables. Their performance is often affected by the preprocessing standard and the representational capacity of the chosen features.

1.4 Challenges in Medical Data Analytics

The integration of computational techniques into healthcare has opened new possibilities for disease prediction and risk assessment. However, medical datasets possess inherent complexities that distinguish them from conventional data used in other machine learning applications. The presence of incomplete records, class imbalance, correlated clinical attributes, and limited sample sizes often restrict the performance and generalizability of predictive models. Understanding these challenges is essential before designing robust artificial intelligence frameworks for Type 2 Diabetes prediction.

1.4.1 Missing and Noisy Clinical Data

Healthcare datasets frequently contain missing, inconsistent, or noisy entries due to variations in data collection practices, measurement errors, or patient non-compliance. In clinical settings, laboratory values may be unrecorded, improperly entered, or measured under non-standard conditions. Missing values in key indicators such as glucose concentration, insulin levels, or blood pressure can significantly distort model training if not handled appropriately [15].

Traditional approaches to managing missing data include mean imputation, deletion of incomplete records, or statistical interpolation. However, these strategies may introduce bias or reduce dataset representativeness. Moreover, physiological measurements are

inherently variable due to biological fluctuations, leading to noise that can affect model stability.

Robust preprocessing techniques and intelligent data handling mechanisms are therefore essential to ensure reliable predictive performance in medical applications.

1.4.2 Class Imbalance in Healthcare Datasets

Among of the leading prominent challenges in medical farcast is class disproportion. In many healthcare data repository, the number of non-diseased individuals significantly exceeds the number of positive cases. This imbalance can cause biased predictive frameworks that favor the prepondrance class while neglecting to accurately classify high-risk patients [14], [16].

Regarding Type 2 Diabetes prediction, misclassification of diabetic cases as non-diabetic (false negatives) is particularly critical, as it may delay necessary intervention. Conventional machine learning techniques tend to refine overall accuracy, which can be ambiguous in non-balanced datasets.

Addressing class imbalance requires specialized methodolgies like resampling techniques, cost-sensitive training, or generative data augmentation methods. The development of effective class-balancing mechanisms is therefore fundamental to improving clinical prediction reliability.

1.4.3 Feature Correlation and High Dimensionality

Clinical datasets often contain multiple interrelated attributes. Features such as body mass index, fasting glucose, insulin measures, and age may exhibit significant correlation. High-dimensional feature spaces with redundant information can lead to model instability and overfitting [9].

Traditional feature selection techniques attempt to reduce dimensionality by identifying the most informative predictors. However, the nonlinear interactions among physiological variables in chronic diseases like T2D may not be fully captured through simple linear correlation analysis.

Advanced feature transformation and representation learning techniques are increasingly explored to extract meaningful patterns from complex healthcare data structures.

1.4.4 Limited Dataset Size and Overfitting

Unlike domains such as computer vision where millions of labeled samples are available, medical datasets are often constrained by privacy regulations, ethical considerations, and data acquisition costs. Limited sample size increases the risk of overfitting, particularly in deep learning models with large numbers of parameters [17], [18].

Overfitting happens when a learninig model works well on training data but fails to understand testing samples. This issue is especially problematic in healthcare applications, where predictive systems must demonstrate reliability across diverse populations.

Data augmentation techniques and regularization strategies are therefore crucial to enhance model generalization in medical prediction frameworks.

1.4.5 Model Interpretability and Ethical Considerations

The successful deployment of artificial intelligence (AI) models in healthcare depends not only on achieving high predictive accuracy but also on ensuring that the generated predictions are transparent, interpretable, and ethically acceptable. Although advanced machine learning and deep learning models have demonstrated remarkable performance in disease prediction, many of these models operate as black-box systems, making it difficult for healthcare professionals to understand the reasoning behind their predictions. In clinical practice, such lack of interpretability may reduce clinicians' confidence and hinder the adoption of AI-based decision support systems [11], [19].

Model interpretability refers to the ability to explain how an AI model reaches a particular prediction. In the context of diabetes prediction, interpretability enables clinicians to identify which clinical parameters, such as blood glucose level, body mass index (BMI), age, insulin concentration, blood pressure, or diabetes pedigree function, contribute most significantly to the predicted outcome. Such explanations improve transparency, facilitate clinical validation, and support informed medical decision-making.

In recent years, the field of Explainable Artificial Intelligence (XAI) has emerged to address the limitations of black-box models. Several model-agnostic techniques have been developed to explain machine learning predictions while maintaining high predictive performance. Among these, SHapley Additive exPlanations (SHAP) quantify the contribution of each input feature to an individual prediction based on cooperative game theory, whereas Local Interpretable Model-agnostic Explanations (LIME) generate locally interpretable surrogate models to explain individual predictions. These techniques enable clinicians to understand not only the final prediction but also the relative importance of different clinical variables influencing the decision.

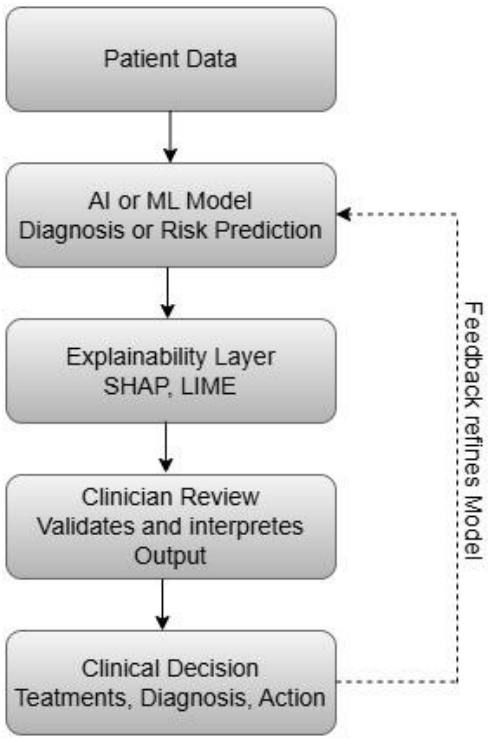

Figure 1.4 Role of Explainable AI in Clinical Decision Support

Figure 1.4 illustrates the general workflow of an explainable AI-based clinical decision support system. Patient clinical data are first processed by the prediction model to generate a diabetes prediction. Explainable AI techniques, such as SHAP and LIME, are then applied to identify the contribution of individual clinical features toward the prediction. The generated explanations assist clinicians in understanding the rationale behind the model's decision, thereby improving transparency, interpretability, and confidence in AI-assisted diagnosis.

Besides interpretability, ethical considerations play a crucial role in the development of AI-driven healthcare applications. Medical prediction models should ensure fairness, transparency, accountability, privacy, and security throughout the decision-making process. The use of patient data requires strict compliance with applicable data protection regulations and institutional ethical guidelines. Furthermore, AI models should be evaluated for potential biases arising from demographic imbalance, incomplete datasets, or population-specific characteristics to ensure equitable performance across diverse patient groups.

Although the primary objective of this research is to improve the predictive performance of diabetes diagnosis through the proposed AVAE–MW-FFT-OAconv–WMOV framework, interpretability remains an important consideration for future clinical deployment. The proposed framework focuses on improving data quality, feature representation, and classification accuracy. However, integrating explainable AI techniques such as SHAP or LIME in future work would provide feature-level explanations for individual predictions, thereby increasing clinician confidence, improving transparency, and facilitating the practical adoption of the proposed system in clinical decision support environments.

In summary, achieving high prediction accuracy alone is insufficient for successful healthcare applications. A clinically useful AI system should combine accuracy, robustness, interpretability, fairness, and ethical compliance to support trustworthy and informed medical decision-making.

1.5 Research Motivation and Problem Statement

The increasing global burden of Type 2 Diabetes, coupled with the limitations of conventional diagnostic and analytical techniques, underscores the necessity for more robust predictive frameworks. While machine learning and deep learning models have illustrated promising fulfilment in medical prediction tasks, their application to structured clinical datasets remains constrained by issues such as data imbalance, limited sample size, feature redundancy, and model generalization challenges.

Existing predictive models for Type 2 Diabetes often rely on traditional classification algorithms trained on relatively small datasets. Although these approaches achieve moderate accuracy, they frequently struggle to detect minority-class cases effectively,

particularly in imbalanced datasets where non-diabetic instances dominate [14], [16]. In medical diagnosis, such misclassifications can have serious implications, as undetected high-risk individuals may miss early intervention opportunities.

Furthermore, most existing models operate primarily in the raw feature domain without exploring enhanced feature representation strategies. Chronic metabolic disorders involve complex and nonlinear physiological interactions that may not be fully captured by conventional feature sets. Advanced transformation techniques and hierarchical representation learning methods have the potential to uncover latent patterns that improve predictive discrimination [20], [18].

Another critical limitation is the scarcity of labeled medical data. Ethical constraints and privacy regulations often restrict access to large-scale datasets, limiting the training capacity of deep learning models. Without sufficient data diversity, predictive systems risk overfitting and poor generalization to unseen populations. Consequently, there is a growing demand for intelligent data augmentation techniques capable of enriching dataset variability while preserving clinical authenticity.

These limitations collectively reveal a research gap in the development of integrated predictive frameworks that combine:

- Robust preprocessing mechanisms
- Class-wise data augmentation techniques
- Advanced feature representation strategies
- Optimized deep learning architectures

Addressing these components in isolation may yield incremental improvements; however, a unified framework that systematically integrates them remains underexplored in the domain of Type 2 Diabetes prediction.

Problem Statement

Despite the availability of numerous machine learning and deep learning techniques for Type 2 Diabetes prediction, existing approaches are often limited by class imbalance,

insufficient feature representation, restricted dataset size, and suboptimal model generalization. There is a need to design and evaluate an integrated computational intelligence framework that enhances predictive accuracy and robustness by incorporating class-wise data augmentation and advanced feature learning techniques while maintaining clinical relevance and reliability.

1.6 Research Aim and Objectives

1.6.1 Research Aim

The fundamental aim of this research is to develop and estimate advanced computational intelligence frameworks for the prediction and classification of Type 2 Diabetes by integrating robust preprocessing techniques, class-wise data augmentation strategies, and enhanced feature representation methods within optimized machine learning and deep learning architectures.

1.6.2 Research Objectives

To accomplish the stated aim, the following specific objectives are formulated:

- To conduct a extensive review of existing machine learning and deep learning approaches for Type 2 Diabetes prediction.
- To design a systematic preprocessing framework for handling missing values, noise, and feature scaling in structured medical datasets.
- To develop a class-wise data enlargement strategy using an adversarial variational autoencoder to address class imbalance in diabetes datasets.
- To implement and evaluate machine learning classifiers trained on augmented clinical data and assess their predictive performance.
- To design an advanced deep learning architecture incorporating enhanced feature transformation mechanisms for improved classification accuracy.

- To integrate an optimized learning strategy for efficient training and convergence of deep neural networks.
- To evaluate the proposed frameworks with excellence medical evaluation measures such as accuracy, precision, recall, F1-score, specificity, and Matthews correlation coefficient.
- To compare the proposed models with conventional machine learning techniques to demonstrate performance improvements.

1.7 Scope of the Research

The prime focus of this research is on the development and estimation of computational intelligence techniques for the early prediction and classification of Type 2 Diabetes with structured clinical datasets. The scope of the study is confined to algorithmic modeling, data preprocessing, feature augmentation, and performance evaluation within a supervised learning framework.

The work primarily utilizes publicly available diabetes datasets for experimental validation. Emphasis is placed on structured numerical clinical attributes such as glucose measures, insulin concentration, body mass index, age and related physiological indicators. The study does not involve real-time physiological signal acquisition, wearable sensor integration, or direct clinical trials.

From a methodological perspective, the research encompasses data preprocessing, class-wise data augmentation using generative models, machine learning-based classification, and advanced deep learning architectures incorporating feature transformation mechanisms. Performance evaluation is conducted using standard classification metrics relevant to medical diagnosis.

The study does not address pharmaceutical treatment strategies, medical device development, or hospital workflow optimization. Additionally, ethical considerations and patient privacy regulations are acknowledged; however, the research is limited to anonymized secondary datasets and does not involve direct patient interaction.

By clearly defining these boundaries, the research maintains a focused objective on improving predictive modeling techniques for Type 2 Diabetes while ensuring methodological rigor and reproducibility.

1.8 Major Contributions of the Thesis

The major contributions of this thesis are summarized below:

1. **Comprehensive Analysis of Existing Predictive Models:**

   A detailed comparative study of conventional machine learning and deep learning approaches for Type 2 Diabetes prediction is conducted, identifying key limitations related to data imbalance, feature representation, and model generalization.

2. **Class-Wise Data Augmentation Framework:**

   A novel class-wise data augmentation strategy based on an adversarial variational autoencoder (AVAE) is developed to address imbalance in structured medical datasets. The proposed augmentation mechanism enhances minority-class representation while preserving underlying statistical characteristics of clinical features.

3. **Machine Learning-Based Predictive Model Using Augmented Data:**

   A robust machine learning framework integrating preprocessing, augmented dataset generation, and ensemble classification techniques is designed and evaluated. Experimental results demonstrate improved predictive accuracy compared to baseline classifiers.

4. **Feature Augmentation and Transformation Strategy:**

   An advanced feature enhancement mechanism incorporating signal transformation techniques is introduced to capture latent relationships among clinical variables and improve discriminatory power.

5. **Morlet Wavelet-Assisted Deep Learning Architecture:**

A deep neural network architecture incorporating Morlet wavelet-assisted feature extraction and overlap-add convolution mechanisms is proposed to enhance classification performance in structured healthcare datasets.

6. **Optimized Training Strategy:**

   An optimization methodology is integrated into the deep learning framework to improve convergence stability and reduce classifier error rates.

7. **Comprehensive Performance Evaluation:**

   The proposed frameworks are rigorously evaluated using multiple statistical evaluation measures including accuracy, precision, recall, F1-score, true positive rate, true negative rate, and Matthews correlation coefficient, ensuring reliable comparison with existing approaches.

8. **Improved Predictive Performance for Type 2 Diabetes Classification:**

   Experimental validation demonstrates that the integrated augmentation and feature-enhanced deep learning approach achieves superior classification accuracy relative to conventional machine learning models.

1.9 Organization of the Thesis

This thesis has been arranged into six different chapters. A brief description of each chapter is given below.

Chapter two presents a extensive analysis of existing literature related to Type 2 Diabetes prediction using statistical methods, machine learning algorithms, and deep learning algorithms. This chapter critically assess previous studies, identifies unexplored areas, and set up the foundation for the proposed frameworks.

Chapter three introduces the first proposed methodology, which integrates systematic data preprocessing, class-wise data augmentation using an adversarial variational autoencoder, and machine learning-based classification. The design framework, algorithmic formulation, and experimental validation are discussed in detail.

Chapter four presents the second proposed framework, which incorporates feature augmentation and transformation techniques within a Morlet wavelet-assisted deep learning architecture. The chapter also describes the optimized training strategy and evaluates the enhanced predictive performance of the model.

Chapter five provides a extensive analysis of experimental results, including comparative performance evaluation with conventional classifiers. Statistical performance metrics and discussion of findings are presented to show the usefulness and robustness of the proposed approaches.

Chapter six concludes the thesis by outlining key contributions, discussing limitations, and summarizing potential future directions for further research studies in intelligent healthcare prediction systems.

CHAPTER 2

Theoretical Background and Literature Review

The rapid advancement of computational intelligence techniques has significantly influenced medical diagnostic research in recent decade. In the context of Type 2 Diabetes prediction, numerous studies have explored statistical frameworks, machine learning models, and deep learning structures to improve early detection accuracy. This chapter highlights an extensive overview of existing research studies related to diabetes prediction, organized thematically according to methodological approaches. The objective of this review is to critically examine earlier work, recognize limitations, and set up the research gap managed in this thesis.

2.1 Statistical and Logistic Regression–Based Models for Diabetes Prediction

Before the widespread adoption of machine learning techniques, diabetes risk prediction fundamentally depend on statistical modeling approaches. Logistic regression, discriminant analysis, and Cox regression framework were extensively used to identify significant risk factors and estimate disease probability. These models provided interpretability and clinical transparency; however, they often assumed linear relationships between predictors and outcomes, limiting their strength to identify complicated physiological associations. This section reviews key studies employing traditional statistical frameworks and discusses their strengths and limitations.

2.1.1 Logistic Regression in Type 2 Diabetes Prediction

Logistic regression remains among the top established statistical methods for binary grouping in clinical prediction tasks. Because of intelligibility and simplicity, it is popularly employed as a baseline model in T2DM prediction studies. Unlike complex machine learning models, logistic regression provides direct insight into the relationship between predictor variables and disease probability, making it particularly accepteable for patient management base systems.

Joshi and Dhakal (2021) applied logistic regression to the Pima Indians Diabetes Dataset (PIDD) and reported a classification accuracy of approximately 78%, identifying plasma glucose concentration, body mass index (BMI), age, and diabetes pedigree function as significant predictors [21]. Their findings confirm that logistic regression can effectively discriminate diabetic from non-diabetic cases in structured clinical datasets.

Similarly, Perveen et al. (2019) evaluated logistic regression in comparison with other predictive techniques using healthcare data and observed that while logistic regression performs competitively, its predictive power is limited when feature interactions are nonlinear [22]. The study emphasized the importance of using logistic regression as a reference model for evaluating more advanced approaches.

In another study published in Frontiers in Artificial Intelligence, Ahamed et al. (2022) assessed multiple classifiers for diabetes prediction and demonstrated that logistic regression achieved moderate predictive performance but was outperformed by ensemble methods such as Random Forest and Gradient Boosting [23]. These findings reinforce the view that logistic regression serves as a reliable baseline but may lack flexibility for complex feature relationships.

Furthermore, Wang et al. (2023), in a large clinical registry-based study published in BMJ Open, employed logistic regression for diabetes risk modeling and reported an AUC value close to 0.80, highlighting its continued relevance in epidemiological risk assessment [24]. The study demonstrated the practical applicability of logistic regression in population-level screening.

Collectively, these studies confirm that logistic regression remains a foundational approach in diabetes prediction research. However, its modeling capacity is constrained by

assumptions of linearity and limited ability to capture higher-order interactions among clinical variables.

2.1.2 Limitations of Logistic Regression

Despite its strengths, logistic regression presents several limitations in the context of modern diabetes prediction:

- **Linearity Assumption:** The framework thinks a linear association allying predictors and log-odds of the outcome, potentially overlooking nonlinear metabolic interactions.
- **Manual Feature Engineering:** Interaction terms must be explicitly defined, limiting automated discovery of complex patterns.
- **Sensitivity to Class Imbalance:** Logistic regression does not inherently address imbalanced datasets.
- **Comparative Performance Constraints:** Recent comparative analyses consistently demonstrate superior predictive performance of ensemble and nonlinear models over logistic regression [23].

These limitations motivate the exploration of more flexible machine learning and deep learning frameworks discussed in subsequent sections.

To provide a structured overview of recent logistic regression–based studies in Type 2 diabetes prediction, Table 2.1 summarizes representative research published between 2019 and 2024. The table highlights the datasets used, reported performance metrics, and key observations regarding model strengths and limitations. This comparative analysis establishes logistic regression as a reliable baseline while illustrating the need for more advanced predictive techniques.

Table 2.1 Comparative Summary of Logistic Regression Studies

Author (Year)	Journal	Dataset	Method	Performance	Key Observation
Joshi & Dhakal (2021) [21]	Int. J. Environmental Research and Public Health	Pima Indians Diabetes Dataset	Logistic Regression	~78% Accuracy	Identified glucose, BMI, age as key predictors
Perveen et al. (2019) [22]	Scientific Reports	Canadian Primary Care Sentinel Surveillance Network (CPCSSN) dataset	Logistic Regression	~75–80% Accuracy	Baseline statistical modeling for diabetes risk
Ahamed et al. (2022) [23]	Frontiers in Artificial Intelligence	Pima Indians Diabetes Dataset	Logistic Regression vs Ensemble	LR < Ensemble	Demonstrated LR as baseline classifier
Wang et al. (2023) [24]	BMJ Open	Clinical registry cohort data	Logistic Risk Model	AUC ~0.80	Effective for population-level risk prediction

2.2 Machine Learning Strategies for Type 2 Diabetes Prediction

The advancement of machine learning (ML) strategies has significantly improved predictive modeling in healthcare, particularly in chronic disease detection such as T2DM. Contrary to long established statistical stretegies, ML algorithms can identify nonlinear associations and complex feature relationships without requiring explicit mathematical assumptions. This capability has led to improved predictive performance in structured clinical datasets.

Among publicly available datasets, the Pima Indians Diabetes Dataset (PIDD) remains among the most extensively used standard dataset for evaluating machine learning models in diabetes research. Its structured attributes—including plasma glucose concentration, BMI, age, insulin level, and blood pressure—make it suitable for assessing classification algorithms under controlled experimental settings.

2.2.1 Support Vector Machine (SVM) Models

Support Vector Machines (SVM) have been extensively applied in diabetes classification because of their capacity to build best separating hyperplanes in high-dimensional feature spaces.

Kavakiotis et al. (2017) conducted a comprehensive review of machine learning applications in diabetes research and reported that SVM models consistently achieved higher accuracy than logistic regression in structured datasets such as PIDD [25]. Their findings indicated that SVM effectively handles nonlinear boundaries through kernel functions.

In another comparative study, Sisodia and Sisodia (2018) evaluated Naïve Bayes, Decision Tree, and SVM classifiers on the PIDD dataset and reported that SVM achieved superior predictive performance compared to other classical classifiers [26]. Although, SVM execution is subject to hyperparameter adjustment and feature scaling.

2.2.2 Decision Tree and Random Forest Models

Decision Trees (DT) provide rule-based classification that is easily interpretable in clinical environments. Although, a lone decision trees are inclined to overfitting. To overcome aforementioned constraint, ensemble strategies like Random Forest (RF) have been extensively adopted.

Perveen et al. (2019) demonstrated that Random Forest models achieved improved predictive performance compared to logistic regression and other classical methods when applied to healthcare datasets [22]. The ensemble nature of RF allows it to reduce variance and improve generalization.

Ahamed et al. (2022) further showed that tree-based ensemble models such as Gradient Boosting and Random Forest outperformed standalone logistic regression when evaluated on diabetes prediction tasks [23]. These results suggest that ensemble learning methods are more robust for capturing complex interactions among clinical features.

2.2.3 k-Nearest Neighbors (k-NN) and Instance-Based Learning

k-Nearest Neighbors is a simple yet effective instance-based learning strategies applied in diabetes classification. Sisodia and Sisodia (2018) reported that k-NN achieved competitive accuracy on the PIDD dataset, particularly when optimal k values and proper feature scaling were applied [26].

Although k-NN is easy to implement, its performance declines with increasing dimensionality and larger datasets due to computational cost. Nevertheless, it remains an important baseline for comparative evaluation.

2.2.4 Comparative Analysis of Machine Learning Studies

Recent literature consistently indicates that ensemble strategies—specially Random Forest and Gradient Boosting—achieve higher predictive performance in comparision with individual classifiers including logistic regression and k-NN.

Across studies using the PIDD dataset:

- Logistic Regression: ~75–78% accuracy
- SVM: ~78–83% accuracy
- Random Forest: ~80–85% accuracy
- Gradient Boosting: Comparable or slightly higher than RF

These findings establish machine learning models as superior alternatives to purely statistical approaches, particularly in capturing nonlinear feature relationships and improving classification robustness.

To summarize the comparative performance of machine learning approaches in diabetes prediction, Table 2.2 presents representative studies employing commonly used classifiers on well organized medical datasets, particularly the Pima Indians Diabetes Dataset. The table highlights methodological differences and reported performance metrics, illustrating the consistent advantage of ensemble and nonlinear models over traditional statistical approaches.

Table 2.2 Comparative Summary of Machine Learning Studies for T2D Prediction

Author (Year)	Journal	Dataset	Method	Performance	Key Observation
Kavakiotis et al. (2017) [25]	Computational & Structural Biotechnology Journal	Review including PIDD and other datasets	SVM, RF	SVM > LR	Comprehensive survey showing ML superiority
Sisodia & Sisodia (2018) [26]	Procedia Computer Science	Pima Indians Diabetes Dataset	SVM, k-NN, NB	SVM highest accuracy	Kernel-based models effective for PIDD
Perveen et al. (2019) [22]	Scientific Reports	CPCSSN dataset	Random Forest	RF > LR	Ensemble improves generalization
Ahamed et al. (2022) [23]	Frontiers in Artificial Intelligence	Pima Indians Diabetes Dataset	RF, Gradient Boosting	Ensemble > LR	Tree ensembles outperform single models

2.3 Deep Learning Techniques in Type 2 Diabetes Prediction

The rapid advancement in deep learning (DL) and nueral network have significantly transformed predictive modeling in healthcare, particularly for T2DM. Unlike classical machine learning models, deep learning architectures automatically extract hierarchical feature representations, enabling them to model complex nonlinear interactions in structured and unstructured healthcare data. Recent studies (2022–2025) demonstrate a clear shift toward hybrid, attention-based, multimodal, and optimization-enhanced deep learning frameworks for diabetes prediction.

2.3.1 Deep Neural Network (DNN)-Based Architectures

Deep Neural Networks remain foundational in diabetes prediction research. Chen et al. (2024) proposed a DNN integrated with a modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare environments [27]. Their framework improved classification performance while incorporating interpretability mechanisms, addressing the “black-box” limitation of deep learning models.

Similarly, Zhang et al. (2024) introduced a DNN-based framework for diabetes diagnosis using structured healthcare data [28]. Their study emphasized effective feature learning and model generalization, demonstrating improved predictive capability over classical machine learning baselines.

In longitudinal healthcare modeling, Manzini et al. (2025) developed a deep attention-based encoder to predict long-term T2DM outcomes from routinely collected healthcare data [29]. Their encoder architecture effectively modeled temporal dependencies, making it particularly suitable for real-world electronic health record (EHR) datasets.

2.3.2 Convolutional Neural Networks (CNN) for Structured and Signal Data

CNN-based architectures have been increasingly adapted for structured clinical datasets and biomedical signal analysis.

Alqushaibi et al. (2022) proposed a CNN-based model optimized through Bayesian optimization techniques for T2DM risk prediction [30]. Their approach demonstrated that hyperparameter optimization significantly enhances predictive stability and classification performance.

Zhao et al. (2024) introduced an attention-oriented CNN method for T2DM prediction [31], integrating attention mechanisms to assign adaptive weights to clinical features. Their study reported improved performance compared to conventional CNN architectures.

Zanelli et al. (2023) applied a lightweight CNN architecture to detect T2DM using raw photoplethysmography (PPG) waveforms [32]. Unlike traditional tabular approaches, this method utilized physiological signal data, demonstrating the expanding role of deep learning in wearable healthcare monitoring.

Farnoosh et al. (2025) proposed DiabetesXpertNet, an attention-based CNN model designed specifically for accurate T2DM prediction [33]. The integration of attention layers enhanced interpretability and predictive precision, highlighting the importance of feature weighting strategies.

2.3.3 Autoencoder and Hybrid Deep Learning Frameworks

Hybrid architectures combining unsupervised and supervised learning have shown promising results in T2DM prediction.

Abdussamad et al. (2025) developed a hybrid stacked sparse autoencoder framework for T2DM prediction [34]. Their model utilized sparse autoencoders for feature removel succeeded by deep neural layers for categorization, improving latent feature representation.

Albadri et al. (2024) proposed a hybrid deep learning-based predictive model combining multiple learning strategies to enhance diabetes classification performance [35]. Their hybridization approach demonstrated robustness across varying dataset distributions.

Such hybrid frameworks improve generalization by integrating feature learning and classification stages within unified architectures.

2.3.4 Multimodal and Large Language Model (LLM)-Enhanced Approaches

Recent advancements extend beyond traditional DL architectures toward multimodal learning and large language models (LLMs).

Ding et al. (2024) introduced a large language multimodal model for new-onset T2DM prediction using five-year cohort electronic health records [36]. Their framework integrated structured medical data with multimodal representations, reflecting a paradigm shift toward foundation models in healthcare analytics. This work demonstrates the growing relevance of multimodal learning and transformer-based architectures in chronic disease prediction.

2.3.5 Comparative Insights and Emerging Trends

The reviewed literature (2022–2025) reveals several important trends:

- Integration of interpretability mechanisms within deep architectures [28], [33].
- Adoption of attention mechanisms for improved feature weighting [29], [31].
- Optimization-enhanced deep learning via Bayesian tuning [30].
- Hybrid autoencoder-based feature extraction strategies [34].
- Expansion into signal-based (PPG) and multimodal EHR modeling [32], [36].

Collectively, these advancements demonstrate that deep learning strategies notably surpass classical statistical approaches in predictive capability. However, their effectiveness depends heavily on dataset size, class balance, and proper regularization—highlighting the importance of augmentation strategies discussed in Section 2.5.

Table 2.3 introduce a relative overview of recent deep learning architectures proposed for Type 2 diabetes projection. The table highlights the model design strategies and methodological innovations, demonstrating a clear shift toward hybrid, optimized, and interpretable deep learning frameworks.

Table 2.3 Comparative Summary of Deep Learning Approaches for T2D Prediction

Author (Year)	Journal	Dataset	Model	Performance	Key Observation
Chen et al. (2024) [27]	Applied Soft Computing	Smart healthcare clinical dataset	DNN + Modified RF	Improved over baseline ML models	Integrated interpretability with DNN
Zhang et al. (2024) [28]	Springer Nature	Clinical dataset	Deep Neural Network	Higher accuracy than classical ML	Improved diagnostic generalization
Ding et al. (2024) [36]	Scientific Reports	Five-year EHR cohort	Multimodal LLM-based model	High AUC for new-onset T2D prediction	Multimodal EHR representation learning
Zhao et al. (2024) [31]	Applied Sciences	PIMA dataset	Attention-CNN	Accuracy > standard CNN	Adaptive feature weighting
Zanelli et al. (2023) [32]	IEEE Access	Raw PPG waveform dataset	Light CNN	Strong classification accuracy	Signal-based diabetes detection
Alqushaibi et al. (2022) [30]	Computers, Materials & Continua	PIMA dataset	CNN + Bayesian Optimization	Accuracy > 85%	Optimized CNN improves stability
Abdussamad et al. (2025) [34]	Scientific Reports	PIMA dataset	Hybrid Sparse Autoencoder	Outperformed traditional ML	Enhanced latent feature extraction
Farnoosh et al. (2025) [33]	PLOS One	PIMA dataset	Attention-based CNN	High precision & recall	Improved interpretability
Albadri et al. (2024) [35]	Mathematical Modelling of Engineering Problems	Clinical dataset	Hybrid ML/DL	Improved robustness	Combined modeling strategy
Manzini et al. (2025) [29]	Expert Systems with Applications	Longitudinal healthcare data	Attention-based Encoder	Strong longitudinal AUC	Temporal modeling capability

2.4 Data Imbalance and Augmentation Methods in Healthcare

Class imbalance remains among the most significant problems in medical predictive modeling. In T2DM prediction tasks, minority class samples (diabetic patients) are often underrepresented relative to non-diabetic cases. The aforementioned disproportion can pejudice classifiers toward the preponderance class, resulting in all inclusive accuracy but reduced sensitivity and minority-class detection performance. Recent studies emphasize that addressing class imbalance is essential for clinically reliable predictive systems.

2.4.1 Class Imbalance in Diabetes Prediction

Abushahla and Pala (2024) investigated diabetes prediction under imbalanced data conditions and demonstrated that imbalance significantly affects classification performance across machine learning algorithms [37]. Their work highlights the necessity of balancing strategies to improve sensitivity and reduce false negatives.

Similarly, Toleva et al. (2025) proposed an effective methodology specifically designed for diabetes prediction in imbalanced datasets [38]. Their study confirmed that untreated imbalance leads to degraded minority-class performance, reinforcing the need for systematic resampling or augmentation.

Moghaddam et al. (2024), in a five-year cohort study published in BMC Medical Research Methodology, evaluated predictive modeling under unbalanced data conditions and emphasized the importance of appropriate evaluation metrics beyond accuracy, such as F1-score and AUC [39]. These findings underline that performance assessment in healthcare prediction must consider imbalance-sensitive measures.

2.4.2 Oversampling and SMOTE-Based Techniques

Traditional oversampling methods such as SMOTE and ADASYN remain extensively applied in structured medical databases.

Aulia et al. (2025) conducted a SHAP-based comparative study analyzing SMOTE and ADASYN in diabetes prediction [40]. Their results showed that oversampling strategies

can improve recall of the minority class; however, they may introduce synthetic noise and alter feature distribution.

Al-Dabbas and Abu-Shareha (2024) applied oversampling techniques for early projection of Type 2 diabetes in female and demonstrated that upsampling significantly enhances minority class detection performance [41]. Nonetheless, their study also acknowledged that synthetic interpolation-based techniques may not fully capture underlying data distributions.

These studies indicate that while oversampling improves sensitivity, it remains limited by its interpolation-based generation process.

2.4.3 Generative Adversarial Networks (GAN) for Medical Data Augmentation

Recent advances in generative models have shifted augmentation strategies toward distribution-learning approaches rather than simple interpolation.

Seo et al. (2024) proposed a GAN-based data augmentation framework to improve hypoglycemia prediction [42]. Their proof-of-concept study demonstrated that GAN-generated samples improved predictive performance and enhanced model robustness.

Singh et al. (2026) introduced a conditional GAN integrated with CNN and LSTM models for enhanced diabetes prediction [43]. Their approach generated synthetic samples based on transformed numerical data, showing significant improvements compared to traditional augmentation techniques.

Vehi et al. (2025), in a comprehensive review on generative artificial intelligence in diabetes healthcare, discussed the growing role of generative models for synthetic medical data generation, highlighting their potential to address data scarcity and imbalance [44].

These works collectively demonstrate the superiority of generative models over conventional oversampling in capturing complex data distributions.

2.4.4 Variational Autoencoders and Adversarial Extensions

Variational Autoencoders (VAEs) have emerged as effective generative frameworks for structured medical data augmentation.

Papadopoulos and Karalis (2023) explored the application of VAEs for data augmentation in clinical studies [45]. Their findings showed that VAE-generated samples improved classifier performance without introducing excessive noise.

Al-zahrani et al. (2025) proposed a VAE-based data augmentation framework specifically for unbalanced medical assessments [46]. Their outcomes demonstrated enhanced minority class detection performance and improved model stability.

Compared to GANs, VAEs offer more stable training dynamics and probabilistic modeling advantages. However, combining VAE representation learning with adversarial training mechanisms can further improve sample realism and distribution matching—forming the conceptual basis for Adversarial Variational Autoencoders (AVAE).

2.4.5 Comparative Analysis of Augmentation Strategies

The reviewed literature reveals a clear progression in augmentation techniques:

- **Interpolation-Based Methods (SMOTE/ADASYN):** Improve minority recall but may distort data distribution [40], [41].
- **GAN-Based Methods:** Learn complex feature distributions but require careful training stability management [42], [43].
- **VAE-Based Methods:** Provide probabilistic latent representation learning and stable augmentation [45], [46].
- **Hybrid and Advanced Generative AI:** Combine representation learning with adversarial mechanisms for enhanced realism [44].

Table 2.4 presents a comparative overview of recent studies addressing class imbalance and data augmentation strategies in diabetes and healthcare prediction tasks. The table highlights the augmentation techniques employed, their impact on predictive performance,

and the methodological advancements introduced in recent literature. This comparison illustrates the progression from traditional oversampling methods to advanced generative models for structured medical data augmentation.

Table 2.4 Comparative Summary of Data Imbalance and Augmentation Studies

Author (Year)	Journal	Dataset	Augmentation Method	Performance Impact	Key Contribution
Abushahla & Pala (2024) [37]	ADBA Computer Science	Diabetes dataset	Imbalance-aware ML	Improved minority recall	Demonstrated imbalance impact
Toleva et al. (2025) [38]	Bioengineering	Diabetes dataset	Imbalance methodology	Improved sensitivity	Structured imbalance handling
Aulia et al. (2025) [40]	IEEE	Diabetes dataset	SMOTE vs ADASYN	Oversampling improves recall	SHAP-based explanation
Al-Dabbas et al. (2024) [41]	Journal of Applied Data Science	Female T2D dataset	Oversampling	Better minority detection	Gender-specific modeling
Seo et al. (2024) [42]	Biomedical Signal Processing & Control	Clinical signal dataset	GAN augmentation	Enhanced robustness	GAN-based sample generation
Singh et al. (2026) [43]	Scientific Reports	Diabetes dataset	Conditional GAN	Significant accuracy gain	Hybrid CNN-LSTM + cGAN
Papadopoulos & Karalis (2023) [45]	Applied Sciences	Clinical dataset	VAE augmentation	Improved classifier stability	Latent representation learning
Al-zahrani et al. (2025) [46]	Results in Engineering	Medical diagnostics dataset	VAE-guided augmentation	Enhanced minority prediction	Structured VAE augmentation

2.5 Feature Engineering and Transformation Techniques

Feature engineering and transformation contribute a critical role in improving predictive modeling performance, particularly in healthcare datasets where complex relationships exist between physiological variables. In structured medical datasets such as diabetes prediction datasets, raw clinical attributes may not always reveal nonlinear patterns associated with disease progression. Recent studies demonstrate that integrating feature

augmentation with signal processing and deep learning architectures can notably boost predictive score. Techniques like, generative feature augmentation, wavelet-based transformations, frequency-domain convolution operations, and optimization-based weight learning have shown promising results in biomedical data analysis

2.5.1 Feature Transformation Techniques for Diabetes Prediction

Feature transformation techniques aim to convert original clinical variables into more informative representations that facilitate improved predictive modeling. These transformations may involve normalization, statistical feature mapping, or nonlinear feature embedding.

Linkon et al. (2024) evaluated several feature transformation techniques in combination with machine learning models for early diabetes detection [47]. Their study demonstrated that transforming clinical variables significantly improved classification performance compared to using raw features alone. Similarly, Wu et al. (2023) proposed a multi-feature map integrated attention model for early prediction of Type 2 diabetes using irregular health examination records [48]. Their approach utilized multiple feature representations to improve prediction accuracy, demonstrating the importance of effective feature representation in healthcare data analysis.

2.5.2 Wavelet-Based Feature Extraction

Wavelet transformation techniques are widely used in biomedical data analysis because they enable simultaneous time–frequency representation of signals. Unlike traditional signal processing methods, wavelet transforms can capture localized variations in data patterns, making them suitable for analyzing physiological signals associated with disease progression.

Shankar et al. (2022) explored wavelet-based machine learning approaches for precision medicine applications in diabetes mellitus [49]. Their research demonstrated that wavelet-based feature extraction improves the identification of subtle physiological patterns related to diabetes progression.

Similarly, Isfahani et al. (2020) proposed a hybrid dynamic wavelet-based modeling method for predicting blood glucose concentration in diabetes patients [50]. Their results showed that wavelet-based feature extraction significantly improves prediction accuracy compared with traditional statistical models.

These studies highlight the effectiveness of wavelet transformations for capturing localized signal characteristics in biomedical datasets.

2.5.3 Frequency-Domain Convolution Techniques

Recent advancements in deep learning have introduced frequency-domain convolution techniques that transform signals into the spectral domain before applying convolution operations. These approaches reduce computational complexity and improve feature extraction efficiency.

Goh et al. (2021) proposed a frequency-domain convolutional neural network designed to accelerate convolution operations in large-scale image classification tasks [51]. Their work demonstrated that FFT-based convolution significantly improves computational efficiency while maintaining classification accuracy.

Similarly, Lee et al. (2024) developed a convolutional neural network framework using high-frequency ultrasound signals for diabetes diagnosis [52]. Their results confirmed that frequency-domain feature extraction can effectively enhance predictive modeling performance in medical applications.

2.5.4 Optimization Techniques for Model Parameter Learning

Optimization algorithms play a critical role in improving neural network training by adjusting model parameters efficiently. Traditional gradient-based learning methods may sometimes converge to local minima or require extensive computational resources.

Aliyu et al. (2024) proposed a metaheuristic optimization framework for improving machine learning algorithms in diabetes prediction tasks [53]. Their study demonstrated that optimization techniques significantly enhance model performance by effectively tuning classifier parameters.

Similarly, Osman (2022) introduced a hyperparameter optimization approach for deep learning-based diabetes risk prediction models [54]. Their results showed that optimized model configurations substantially improve predictive accuracy.

2.5.5 Deep Feature Representation Learning

Deep learning architectures can instinctively extract complex feature illustration from raw biomedical facts, enabling models to capture nonlinear relationships between clinical variables.

Das (2022) proposed a deep learning model for identifying Type 2 diabetes based on nucleotide signal representations [55]. Their approach demonstrated that deep feature representation learning can significantly enhance disease prediction performance by capturing complex biological patterns within healthcare data.

Table 2.5 summarizes recent studies that employ various feature engineering and transformation techniques in diabetes projection and healthcare analytics. Aforementioned comparative overview highlights the role of advanced feature representation, signal processing methods, and optimization strategies in improving disease prediction models.

Table 2.5 Comparative Summary of Feature Engineering and Transformation Studies

Author (Year)	Journal	Dataset	Feature Transformation Method	Performance Impact	Key Contribution
Linkon et al. (2024) [47]	IEEE Access	Pima Indians Diabetes Dataset	Feature transformation + ML	Improved classification accuracy	Demonstrated importance of feature transformation
Wu et al. (2023) [48]	IEEE Journal of Biomedical and Health Informatics	Irregular health examination records (clinical dataset)	Multi-feature attention model	Enhanced prediction accuracy	Multi-feature representation learning
Shankar et al. (2022) [49]	Science and Technology Publications	Diabetes clinical dataset	Wavelet-based feature extraction	Improved disease pattern detection	Wavelet-based ML framework
Isfahani et al. (2020) [50]	Journal of Medical Signals & Sensors	Continuous glucose monitoring data	Dynamic wavelet modeling	Improved glucose prediction	Hybrid wavelet modeling

Table 2.5 Comparative Summary of Feature Engineering and Transformation Studies (Continue)

Author (Year)	Journal	Dataset	Feature Transformation Method	Performance Impact	Key Contribution
Goh et al. (2021) [51]	Computer Vision and Pattern Recognition	Diabetic retinopathy image dataset	Frequency-domain CNN (FFT)	Faster CNN computation	Frequency-domain convolution
Lee et al. (2024) [52]	Ultrasonics	High-frequency ultrasound imaging dataset	CNN-based signal analysis	Improved diagnostic performance	Ultrasound-based diabetes diagnosis
Aliyu et al. (2024) [53]	Franklin Open	Diabetes dataset	Metaheuristic optimization	Improved classifier performance	Parameter optimization
Osman (2022) [54]	International Journal of Science and Research	Survey-based diabetes dataset	Hyperparameter optimization	Increased prediction accuracy	Deep learning tuning
Das (2022) [55]	Neural Computing and Applications	Nucleotide signal dataset	Deep feature representation learning	Improved disease classification	Signal-based deep learning model

2.6 Limitations of Existing Studies

Regardless advancements in machine learning and deep learning strategies for Type 2 Diabetes projection, several limitations persist in existing studies. These constraints underscore the necessity for improved predictive frameworks that can efficiently handle data imbalance, extract meaningful feature representations, and optimize model performance.

One major limitation observed in traditional statistical and machine learning approaches is their dependence on manually selected features. Although strategies like logistic regression, decision trees, and support vector machines have been extensively applied for diabetes forcasting, their performance often depends heavily on the excellency in feature selection and pre-processing techniques. These models may struggle to capture complex nonlinear relationships among clinical variables, which can reduce prediction accuracy in real-world healthcare datasets.

Another challenge identified in the literature is the presence of class imbalance in medical datasets. In many diabetes prediction datasets, the number of non-diabetic samples significantly exceeds diabetic samples. Although oversampling techniques such as SMOTE and ADASYN have been applied to address this issue, these methods generate synthetic samples using interpolation between existing instances. Consequently, the generated samples may not accurately represent the underlying data distribution, which can negatively affect model generalization.

Recent research has explored deep learning architectures to improve prediction performance. While deep neural networks and convolutional neural networks demonstrate superior predictive capability compared to traditional machine learning models, these approaches often require large amounts of balanced training data. In healthcare datasets with limited samples, deep learning models may suffer from overfitting or unstable training behavior.

Another limitation observed in existing studies relates to feature representation and transformation techniques. Many models rely on raw clinical variables without applying advanced transformation methods that could reveal hidden patterns in the data. Although some studies have explored wavelet-based feature extraction and frequency-domain analysis, these approaches have not been widely integrated with deep learning frameworks for structured medical datasets.

Additionally, computational efficiency and model optimization remain important challenges. Conventional convolution operations in deep neural networks may increase computational complexity, particularly when processing high-dimensional data. Although optimization algorithms and hyperparameter adjustment techniques have been proposed to enhance model accomplishment. Many studies rely primarily on gradient-based maximization strategies, which may unite to optimal point or require extensive parameter tuning.

Furthermore, several research concentrate on enhancing prediction accuracy without adequately addressing model interpretability and generalization capability. Healthcare applications require reliable predictive models that can maintain consistent performance across diverse patient populations and clinical environments.

Overall, the literature indicates that existing diabetes prediction models face several key limitations, including:

- Insufficient handling of class imbalance in structured healthcare datasets
- Limited use of advanced generative approaches for feature augmentation
- Inadequate feature transformation techniques to capture complex patterns
- High computational complexity in deep learning architectures
- Limited integration of optimization algorithms for efficient model training

These limitations suggest the need for a more comprehensive predictive framework that integrates advanced feature augmentation, signal transformation techniques, and optimization strategies to improve predictive performance in diabetes diagnosis.

2.7 Research Gap Identification and Summary

The literature review presented in Sections 2.2 through 2.6 highlights the significant progress made in applying statistical learning, machine learning models, deep learning structures, data augmentation techniques, and feature transformation methods for Type 2 Diabetes prediction. Despite these advancements, several research gaps remain that limit the efficacy of existing predictive frameworks in healthcare applications.

One of the key issues identified in the literature is the imbalance present in medical datasets. Many studies rely on traditional oversampling methods like SMOTE and ADASYN to handle class imbalance. Although these methods enhance minority class representation, they generate synthetic samples through interpolation between existing data points, which may not accurately capture the actual data distribution. As a result, the generated samples can introduce noise and reduce model generalization capability. Recent studies suggest that deep generative approaches such as Variational Autoencoders and Generative Adversarial Networks provide better data distribution learning; however, their application for structured diabetes datasets remains limited.

Another limitation observed in existing research relates to feature representation and transformation. Many predictive models utilize raw clinical features without applying

advanced transformation methods that can capture hidden relationships among variables. While some studies have explored wavelet-based feature extraction and frequency-domain analysis, these techniques are rarely integrated into deep learning architectures for structured healthcare datasets. Consequently, important signal characteristics associated with disease progression may remain unexplored.

The literature also reveals limitations in deep learning model architectures used for diabetes forcasting. Although convolutional neural networks and deep neural networks demonstrate improved performance compared to traditional machine learning methods, many models rely on conventional convolution operations that may increase computational complexity. Frequency-domain convolution techniques such as Fast Fourier Transform (FFT)-based convolution have been explored in other domains, but their application in diabetes prediction frameworks remains limited.

Another research gap identified concerns the optimization of neural network parameters. Many existing studies rely on gradient-based optimization strategies such as stochastic gradient descent or Adam optimizer. While these methods are widely used, they may suffer from slow convergence or convergence to local optima. Metaheuristic optimization algorithms have demonstrated promising results in improving model parameter tuning; however, their integration with deep learning architectures for diabetes prediction is still relatively unexplored.

Furthermore, most existing predictive models focus on improving classification accuracy while overlooking the need for integrated frameworks that combine augmentation, transformation, and optimization techniques. In many cases, these components are investigated independently rather than as part of a unified predictive framework.

Based on the extensive literature survey, the following research gaps can be identified:

- Limited application of advanced generative feature augmentation techniques for handling class imbalance in diabetes datasets.
- Insufficient integration of wavelet-based feature transformation methods within deep learning models for structured healthcare data.

- Lack of efficient convolution mechanisms such as FFT-based overlap-and-add convolution in diabetes prediction models.
- Limited exploration of metaheuristic optimization algorithms for neural network weight estimation in medical prediction tasks.
- Absence of unified frameworks that integrate feature augmentation, signal transformation, deep learning architectures, and optimization techniques.

Addressing these research gaps requires the development of a comprehensive predictive framework that combines advanced data augmentation techniques with efficient feature transformation and optimization strategies. In answer to these constraints, this research proposes novel methodologies for Type 2 Diabetes prediction that integrate Adversarial Variational Autoencoder (AVAE)–based feature augmentation, Morlet wavelet-based activation mechanisms, FFT overlap-and-add convolution operations, and Weighted Mean of Vector optimization for neural network weight estimation. The details of these proposed algorithms are presented in the subsequent chapters.

CHAPTER 3

AVAE-Based Data Augmentation Framework for Type-2 Diabetes Prediction

Type-2 Diabetes Mellitus (T2DM) has came out as a important world wide health concern due to its rapidly increasing commonness and associated long-term complications. Early projection of diabetes using machine learning strategies can assist medical experts in well-timed recognition and preventive treatment. However, healthcare datasets often suffer from challenges such as limited sample size, class imbalance, and missing or noisy data, which can negatively affect model performance. Data augmentation techniques have therefore gained attention as a strategy to improve dataset diversity and enhance predictive accuracy. In this chapter, an AVAE-based data augmentation framework is proposed to generate artificial samples that preserve the statistical attributes of the primary dataset. The augmented dataset is subsequently applied to design a robust prediction framework for identifying diabetic and non-diabetic patients.

3.1 Architecture of the Proposed AVAE-Based Framework

The proposed AVAE-based data augmentation framework is designed to enhance the predictive performance of machine learning frameworks for T2DM diagnosis by enhancing dataset quality and diversity. Medical datasets frequently suffer from difficulties in missing data, noisy observations, unbalanced class, and limited sample size. These issues can significantly degrade the performance of predictive frameworks. To address these challenges, the proposed framework integrates data preprocessing, adversarial generative augmentation, and machine learning classification into a unified predictive system.

The overall architecture of the proposed model consists of three major stages: data preprocessing, AVAE-based data augmentation, and diabetes classification using a Random Forest model. Each stage performs a specific function to enhance the properties of the training data and increase the prediction power of the model.

In the first stage, the raw diabetes dataset undergoes preprocessing operations to remove inconsistencies and prepare the data for model training. These preprocessing steps include handling incorrect or missing values, detecting and removing outliers, and standardizing feature values. Correcting these data quality issues ensures that the dataset accurately represents the underlying clinical information and prevents machine learning algorithms from learning misleading patterns.

In the second stage, the cleaned dataset is used to train an Adversarial Variational Autoencoder (AVAE). The AVAE model combines the representational learning capability of Variational Autoencoders with the sample realism provided by adversarial training. This model learns the latent distribution of the dataset and generates synthetic samples that resemble real patient records. In this study, class-wise data augmentation is performed, where synthetic samples are generated separately for diabetic and non-diabetic classes. This strategy helps maintain the statistical characteristics of each class while improving the overall balance of the dataset.

In the final stage, the augmented dataset is applied to train a Random Forest classifier for diabetes forcasting. Random Forest is an ensemble learning method that creates multiple decision trees and accumulates their predictions to generate a final classification result. Due to its robustness against noise, ability to handle nonlinear relationships, and resistance to overfitting, Random Forest is well suited for healthcare prediction tasks.

The whole architecture of the proposed AVAE-driven data augmentation model is illustrated in Figure 3.1. The model consists of three major stages: data preprocessing, AVAE-based data augmentation, and diabetes classification using a Random Forest classifier. Initially, the input dataset undergoes preprocessing operations including incorrect value replacement, outlier rejection, and feature standardization. The cleaned dataset is then used to train an Adversarial Variational Autoencoder (AVAE) to generate synthetic samples and expand the dataset. Finally, the augmented data is used to train a

Random Forest classifier to predict whether a patient belongs to the normal or Type-2 diabetic class.

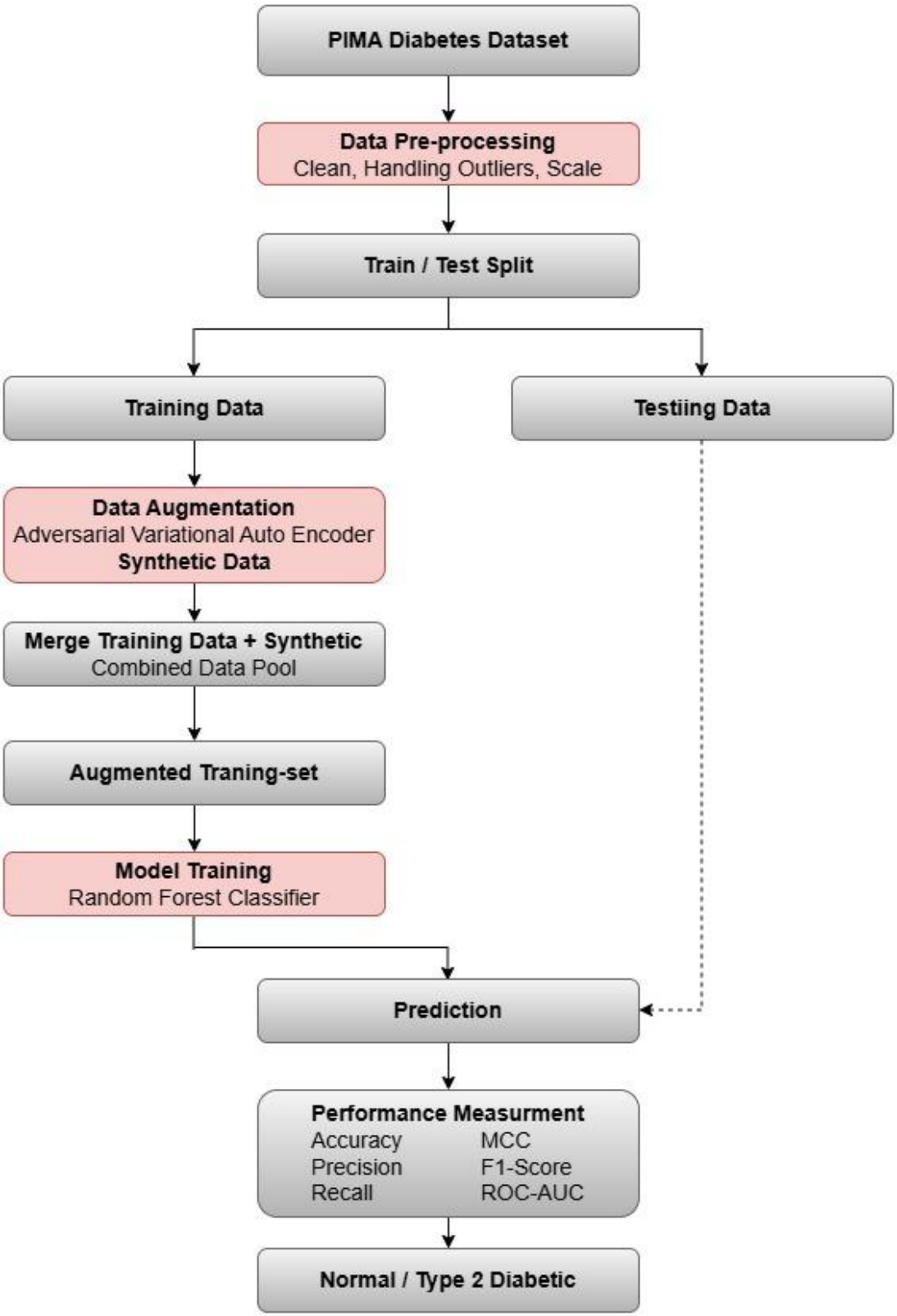

Figure 3.1 Proposed AVAE-Based Diabetes Prediction Framework

3.2 Dataset Description

The dataset used in this research is the Pima Indians Diabetes Dataset (PIDD), which is among the most extensively used standard dataset in diabetes prediction domain. The dataset includes diagnostic estimations of women patients of Pima Indian heritage and has been extensively used in machine learning and medical diagnosis studies for evaluating predictive models related to diabetes detection due to compact size and moderate class imbalance

The dataset was initially collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is openly available on the Kaggle data platform and UCI Machine Learning Repository as well. It contains clinical attributes that are considered important indicators for identifying the Type-2 Diabetes Mellitus (T2DM).

The PIDD contains a total of 768 patient data, where individual record corresponds to a female individual aged at least 21 years. The dataset consists of eight independent variables (features) representing medical diagnostic measurements and one dependent variable (class label) indicating the diabetes status of the patient. The class label represents whether the patient has diabetes or not, in which: a non-diabetic patient as 0 and a diabetic patient as 1.

Table 3.1 PIDD attributes

Sr. No.	Attribute Name	Description
1	Pregnancies	Total number of times the patient was pregnant
2	Glucose	Plasma glucose level after performing a glucose tolerance test
3	BloodPressure	Diastolic blood pressure value (mm Hg)
4	SkinThickness	Thickness of triceps skin fold (mm)
5	Insulin	Serum insulin level after 2-Hour (mu U/ml)
6	BMI	Body Mass Index (weight in kg divided by height in m^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function indicating hereditary influence
8	Age	Patient's age (years)
9	Outcome	Diabetes result (0 = Non-diabetic, 1 = Diabetic)

The dataset contains eight clinical attributes along with a binary class label indicating diabetes status. The explicit representation of these attributes is presented in **Table 3.1**,

which is derived from the Pima Indians Diabetes dataset accessible in the UCI Machine Learning Repository [56] [57].

Among the 768 instances, 500 instances belong to the non-diabetic group, while 268 instances belong to the diabetic group. This distribution indicates a moderate class imbalance in the dataset, which may influence the performance of traditional machine learning frameworks. Therefore, data augmentation strategies such as the proposed Adversarial Variational Autoencoder (AVAE) are used in this research to increase dataset diversity and improve predictive performance.

Before applying the proposed AVAE-based data augmentation framework, the dataset undergoes different preprocessing stages to enhance the data quality and ensure reliable model training. These preprocessing operations are reviewed in the following section.

3.3 Data Pre-processing

Data pre-processing is a crucial stage in the evolution of reliable machine learning frameworks, specially in the healthcare domain where data quality directly affects prediction accuracy. The raw Pima Indians Diabetes dataset contains several inconsistencies such as incorrect values, missing information, and outliers, which can negatively impact the learning capability of predictive models. Therefore, it is essential to transform the raw dataset into a clean and structured form before applying the proposed AVAE-based data augmentation framework.

In this research, a systematic preprocessing pipeline is designed and developed to enhance the reliability of the dataset. The preprocessing stage focuses on ensuring that the input data accurately represents meaningful clinical information while minimizing noise and redundancy. The major objectives of preprocessing include:

- Eliminating invalid or unrealistic values present from the dataset
- Decreasing the influence of outliers
- Normalizing feature distributions to improve model convergence
- Preparing the dataset for effective generative modeling and classification

Unlike generic preprocessing approaches, the techniques adopted in this study are carefully selected based on the diabetes dataset characteristics and the necessity of the proposed AVAE model. Proper preprocessing is particularly important in this work because the generative model learns the underlying data distribution, and any noise or inconsistency in the data may lead to poor-quality synthetic samples.

The preprocessing stage in the proposed framework consists of the following three main steps:

1. Handling incorrect or missing values
2. Outlier detection and removal
3. Feature standardization

Each of these steps is discussed in the subsequent subsections.

3.3.1 Handling Incorrect or Missing Values

The presence of incorrect or undefined data is a common issue in medical datasets and can significantly infect the machine learning model's performance. In the Pima Indians Diabetes data repository, several attributes like glucose measured value, blood pressure, skin thickness, insulin measure value, and body mass index (BMI) contain zero values. From a clinical perspective, these values are not realistic and are therefore treated as missing or invalid entries.

If such incorrect values are not properly handled, they may introduce bias in the learning process and lead to inaccurate predictions. Therefore, an appropriate imputation strategy is required to replace these values while maintaining the statistical properties of the data collection.

In this study, mean imputation is employed to handle incorrect or missing values. This method replaces invalid entries with the mean of the corresponding feature, ensuring that the overall distribution of the data remains unaffected. Mean imputation is computationally efficient and suitable for datasets with relatively small size, making it appropriate for the Pima dataset.

The mathematical formulation of the imputation process is specified as follows:

$$I(x) = \begin{cases} \mu_x, & if\ \ x = 0\ or\ missing \\ x, & otherwise \end{cases} \quad (3.1)$$

here:

- x is the original feature value
- μ_x denotes the mean of the corresponding feature

This approach ensures that unrealistic zero values are replaced with statistically meaningful values, thereby improving data consistency. Similar imputation strategies have been widely used in healthcare analytics to handle incomplete datasets effectively [39].

By applying this preprocessing step, the dataset becomes more suitable for subsequent operations such as outlier detection, feature scaling, and generative modeling using AVAE.

3.3.2 Outlier Identification and Elimination

Outliers are data values which are notably deviate from the general distribution of a dataset. In medical datasets, such extreme values may arise due to measurement inaccuracy, data input faults, or abnormal physiological conditions. The existence of outliers can adversely affect the acomplishments of machine learning strategies by distorting the underlying data distribution and influencing model training.

In the context of the Pima Indians Diabetes Dataset, features such as glucose measure, insulin measure, blood pressure, Body Mass Index, and patient's age may contain extreme values that do not represent the typical patient population. If these values are not handled properly, they can negatively impact both the data augmentation process using AVAE and the subsequent classification using Random Forest.

To address this issue, the proposed framework employs the Interquartile Range (IQR) technique for detecting and removing outliers. The IQR method is widely used because of its robustness and efficacy in managing non-normal data distributions [58].

The interquartile range is calculated by the substracting the first quartile from the third quartile, as shown below:

$$IQR = Q_3 - Q_1 \quad (3.2)$$

here:

- the first quartile is represented as Q_1 (25th percentile)
- the third quartile is represented as Q_3 (75th percentile)

Based on the IQR, the acceptable range of data is determined as:

$$Q_1 - 1.5 \times IQR \leq x \leq Q_3 + 1.5 \times IQR \quad (3.3)$$

Any data point x that falls outside this range is considered an outlier and is removed from the dataset.

The IQR method is particularly suitable for this study because it does not assume any specific data distribution and is insensible to maximum values with respect to methods such as Z-score. By eliminating outliers, the dataset becomes more consistent and better represents the true underlying patterns in the data.

This step is significant in our proposed framework because the AVAE framework learns the latent distribution of the dataset. The presence of outliers may distort this distribution and lead to the generation of unrealistic synthetic samples. Therefore, removing outliers ensures that the generated data remains meaningful and improves the overall performance of the predictive model.

3.3.3 Feature Standardization

Feature standardization is a crucial preprocessing stage in machine learning, particularly when handling with datasets that contain attributes with different scales and units. In the Pima Indians Diabetes Dataset, features like glucose measured value, insulin concentration, Body Mass Index, and patient's age have varying numerical ranges. If these features are used directly, attributes with biggest values may influence the learning patterns, which leads to biased performance of the model.

To check that all attributes participate uniformaly while learning process, the proposed framework applies standardization (z-score normalization) to transform the data into a

common scale. Standardization rescales the feature values such that they have a zero mean and a standard deviation as one, thereby improving the convergence of machine learning procedure and enhancing stability of the framework.

The standardization process is mathematically calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{3.4}$$

here:

- x as native attribute value
- μ as mean of the feature
- σ as standard deviation

The transformed value z represents extent of deviation of a data point relative to the mean in terms of standard deviations.

Standardization is particularly important in this study because the AVAE model relies on learning the latent distribution of input data, and consistent feature scaling helps in stabilizing the training process. Additionally, although Random Forest is less sensitive to feature scaling compared to other algorithms, standardized input still contributes to improved consistency when integrating multiple stages such as preprocessing and generative modeling.

The use of feature scaling techniques such as standardization is well established in machine learning and has been extensively used in healthcare data analysis to improve predictive performance [59].

3.4 AVAE-Based Data Augmentation

Data augmentation is an important one for the performance improver of machine learning framework, especially during dealing with small and imbalanced medical datasets. Traditional data augmentation strategies like random oversampling and Synthetic Minority Oversampling Technique (SMOTE) augment the samples but often fail to identify the true true allocation of the data, which may lead to overfitting and poor generalization.

To get the better of these constraints, this research proposes a data expansion framework based on an Adversarial Variational Autoencoder (AVAE). The AVAE model incorporates the durabilty of Variational Autoencoders (VAE) and adversarial learning to create artificial data samples that are both diverse and statistically consistent with the primery data source.

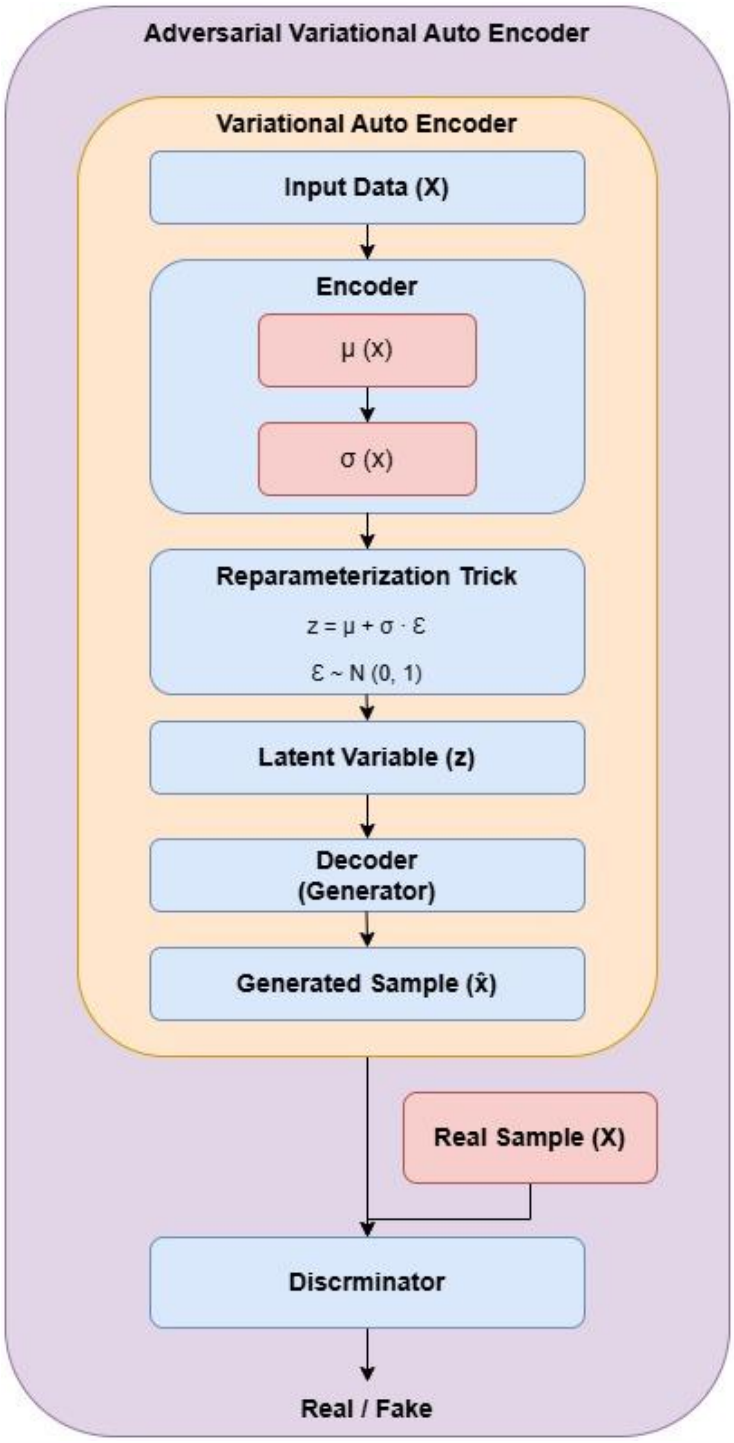

Figure 3.2 Architecture of the Proposed AVAE for Data Augmentation

The fundamental objective of using AVAE in our study is to learn the latent representation of the diabetes dataset and generate realistic synthetic samples for both diabetic and non-diabetic classes. This approach enhances dataset diversity and reduces class imbalance, thereby improving the performance of the classification framework

The architecture of the proposed Adversarial Variational Autoencoder (AVAE) is illustrated in Figure 3.2. The model integrates a variational autoencoder with an adversarial discriminator, where the encoder learns the latent representation of input data, the decoder generates synthetic samples, and the discriminator determines original and synthetic data to improve sample quality.

The AVAE-based augmentation process consists of three main components:

1. **Variational Autoencoder**
2. **Adversarial Learning Mechanism**
3. **Class-wise Data Augmentation**

All these components are explained in the following subsections.

3.4.1 Variational Autoencoder

A Variational Autoencoder (VAE) is a productive framework that grasps a stochestic descriptions of given data in a lower-dimensional latent space [60]. It includes two important components: an encoder and a decoder.

The encoder depicts the input feature vector x towards latent representation z, which is typically modeled as a Gaussian distribution:

$$q(z|x) \tag{3.5}$$

Instead of directly mapping inputs to a fixed latent vector, the encoder produces two parameters:

- Mean $\mu(x)$
- Standard deviation $\sigma(x)$

A latent variable is then screened using the reparameterization trick:

$$z = \mu + \sigma \cdot \epsilon, \qquad \epsilon \sim N(0,1) \tag{3.6}$$

From the latent depiction, the input is reconstructed by the decoder:

$$p(x|z) \tag{3.7}$$

The aim of the VAE as to minimize the unite cost function containing:

- Remodeling loss
- Kullback–Leibler (KL) discrepancy

$$L_{VAE} = L_{rec} + D_{KL}\,(q(z|x) \,||\, p(z)) \tag{3.8}$$

The VAE model enables learning of the underlying structure of the dataset, which is essential for generating meaningful synthetic samples.

3.4.2 Adversarial Learning Mechanism

While Variational Autoencoders are productive in training latent representations of facts, they frequently generates overly smooth or less realistic samples. To enhance the quality of generated data, an adversarial learning mechanism is incorporated into the proposed framework. This mechanism is inspired by Generative Adversarial Networks (GANs), which are known for their capacity to produce highly realistic samples [61].

The adversarial learning framework consists of two competing components:

- Generator (Decoder of VAE)
- Discriminator

In the proposed AVAE model, the decoder acts as the generator, which produces synthetic samples from the latent representation, while the discriminator is a separate neural network that calculates either a taken sample is original (from the given database) or generated (synthetic).

The discriminator is prepared for to differentiate between original and synthetic samples, whereas the generator attempts to generate cases that are identical from original facts. This creates a minimax optimization problem, where both networks are trained simultaneously in a competitive manner.

The adversarial objective function is defined as:

$$L_{adv} = \mathrm{E}_{x\sim p_{data}(x)}[logD(x)] + \mathrm{E}_{z\sim p(z)}[\log(1 - D(G(z)))] \quad (3.9)$$

here:

- input as the $D(x)$ probability
- original sample as x
- the generated sample from latent variable z as $G(z)$
- the spread of original data as $p_{data}(x)$
- the earlier spread of latent variables as $p(z)$

The aim of the generator is to reduce this loss, whereas the discriminator tries to increase it. Through this adversarial learning mechanism, the generator gradually enhances its capacity to generate authentic samples.

By integrating adversarial learning with the VAE framework, the proposed AVAE model benefits from:

- Improved quality of generated samples
- Better approximation of the true data distribution
- Increased diversity in synthetic data

This enhancement is particularly important for medical datasets, where preserving the realistic structure of clinical features is essential for reliable prediction. The adversarial mechanism make sure that the artifical samples created by the AVAE model nearly resemble real patient facts, thereby improving the effectiveness of the subsequent classification stage.

3.5 Random Forest Based Diabetes Prediction

After performing data preprocessing and augmenting the dataset using the proposed AVAE model, the final stage of the framework involves the grouping of patients into diabetic and non-diabetic classes. For this purpose, a Random Forest classifier is employed because of its robustness, exactness, and suitability for handling structured medical data.

Random Forest is a group learning algorithm that creates a group of decision trees during the learning phase and finalized the prediction based onto the majority voting mechanism of discrete trees. Every decision tree is excercised on a arbitrarily chosen segment of the training data along with a random segment of features, which introduces wide varieties among the trees and improves generalization performance [62].

The functioning of the Random Forest classifier can be summarized as:

1. Various bootstrap specimen are taken from the augmented data facts.
2. For every sample, a decision tree is created using a random segment of features.
3. Every tree independently predicts the class label.
4. The concluding prediction is obtained using most chosen over all trees.

The usefulness of Random Forest in this research provides various benifits:

- It reduces the uncertainty of overfitting with respect to a single decision tree
- It effectively handles nonlinear associations between features
- It is vigorous to noise and variations in the dataset
- It performs well even when the dataset contains a mix of relevant and less relevant features

In the proposed framework, the augmented dataset generated using AVAE is seperated into training and evaluation sets. The Random Forest classifier is exercised on the enhanced training dataset, which contains both original and synthetic samples. The inclusion of

synthetic data improves class balance and permits the framework to absorbs additional representative styles of both diabetic and non-diabetic classes.

During the testing phase, the trained model predicts the diabetes status of unseen samples, where the output is represented as:

- 0: Non-diabetic
- 1: Diabetic

The incorporation of AVAE-based data augmentation plus Random Forest classification improves the overall predictive performance by combining enhanced data diversity with a robust ensemble learning algorithm.

3.6 Proposed Algorithm-1

This section presents the proposed Algorithm-1 for the AVAE-based data augmentation framework for T2DM prediction. The algorithm specification, including the input, output, and key parameters, is first described, followed by the detailed execution procedure. The overall computational complexity of Algorithm-1 is discussed in Section 3.7.

3.6.1 Algorithm-1: AVAE-Based Data Augmentation Framework for Type-2 Diabetes Prediction

Before presenting the detailed procedure of the proposed Algorithm-1, it is essential to define the input data, expected outputs, and key parameters used during its execution. These components establish the computational framework of the proposed AVAE-based data augmentation methodology and provide a clear understanding of the information processed and generated by the algorithm.

Input:

- PIMA Diabetes Dataset D
- Clinical features X
- Class labels Y

Output:

- Augmented balanced dataset D_{aug}
- Trained Random Forest classifier
- Diabetes prediction

Parameters:

- Learning rate = 0.001
- Batch size = 32
- Epochs = 100
- Latent dimension = 64

3.6.2 Procedure of Proposed Algorithm-1

Step 1: Load the Pima Indians Diabetes Dataset D

Step 2: Perform data preprocessing

a) Replace incorrect or missing values using mean imputation

b) Detect and remove outliers using IQR method

c) Standardize features using z-score normalization

Step 3: Split dataset into two classes

D_0 Non-diabetic samples

D_1 Diabetic samples

Step 4: Train AVAE model

a) Input preprocessed dataset into encoder

b) Learn latent distribution parameters $\mu(x)$ and $\sigma(x)$

c) Generate latent variable z using reparameterization

d) Reconstruct data using decoder

e) Apply adversarial training using discriminator

Step 5: Generate synthetic samples

D_0^{syn} for non-diabetic samples

D_1^{syn} for synthetic diabetic samples

Step 6: Construct augmented dataset

$$D^{'} = D \cup D_{syn}$$

Step 7: Split augmented dataset into training and testing sets

Step 8: Train Random Forest classifier on training data

Step 9: Perform prediction on test data

Step 10: Return predicted output (0: Non-diabetic, 1: Diabetic)

3.7 Algorithmic Complication Analysis

The algorithmic complication of the suggested AVAE-based data augmentation framework primarily depends on two major components: the training of the Adversarial Variational Autoencoder (AVAE) and the classification by applying the Random Forest framework. Since the framework integrates generative modeling with ensemble learning, the overall computational cost is influenced by together deep learning and conventional machine learning operations.

The AVAE model involves training neural network components, including the encoder, decoder (generator), and discriminator. The training process requires iterative optimization using gradient-based methods over multiple epochs.

Let N stand for numerous training samples and E represent numerous training epochs. The complexity of AVAE training can be approximated as:

$$O(E \times N \times P)$$

Here, the numerous parameters of the neural network is represented as P. The inclusion of adversarial learning increases the computational cost compared to a standard VAE, as both the generator and discriminator must be trained simultaneously.

The data augmentation process itself introduces additional overhead due to the generation of synthetic samples. However, this step is performed offline during the training phase and unaffected to prediction time of the proposed framework.

The Random Forest classifier further contributes to the overall complexity. Let T denote the numerous trees in the forest and M represent the numerous features. The training complexity of the Random Forest can be approximated as:

$$O(T \times NLogN)$$

During the prediction phase, the complexity is relatively low and can be expressed as:

$$O(T \times LogN)$$

since each tree independently predicts the class label, and the final output is obtained through majority voting.

Although the proposed framework introduces additional computational cost due to the AVAE model, this cost is justified by the improved data representation and enhanced predictive performance. Furthermore, the use of Random Forest ensures efficient and fast prediction once the model is trained.

Overall, the framework maintains a stability between computational effectiveness and prediction correctness, succeeding suitable for practical healthcare applications where both performance and reliability are important.

3.8 Summary of Proposed Framework

This chapter presented the proposed AVAE-based data augmentation framework for Type-2 diabetes prediction. The framework incorporates data pre-processing, generative modeling, and machine learning classification to improve predictive performance on medical datasets. Initially, the Pima Indians Diabetes Dataset was preprocessed through

handling incorrect values, outlier removal, and feature standardization to ensure data quality and consistency.

The key contribution of this study lies in the evolution of an Adversarial Variational Autoencoder (AVAE) for generating synthetic samples. The AVAE model combines variational autoencoding with adversarial learning to produce realistic and diverse data, thereby addressing the issue of class imbalance. A class-wise data augmentation strategy was employed to maintain the statistical properties of diabetic and non-diabetic segments.

The augmented dataset was subsequently used to train a Random Forest classifier, which provides well and perfect predictions due to its group learning capability. The complete workflow of the proposed framework was summarized through Algorithm 1, clearly outlining each stage from preprocessing to classification.

Overall, the proposed methodology enhances data representation and improves the reliability of diabetes prediction. Through experimental results and performance analysis, the effectiveness of the framework is evaluated, in the following chapter.

CHAPTER 4

FFT-OLA based Type-2 Diabetes Prediction using Deep Learning

The prediction of Type-2 Diabetes Mellitus (T2DM) requires efficient modelling techniques capable of capturing complex relationships among clinical and demographic features while maintaining computational feasibility. Conventional machine learning and deep learning frameworks often depend upon direct feature interactions and spatial-domain convolution operations, which may lead to increased computational complexity and limited capability in extracting hidden feature patterns.

To deal with these contraints, this chapter presents a logical deep learning model based on Fast Fourier Transform (FFT) overlap-and-add convolution. By transforming convolution operations into the frequency domain, the proposed approach significantly reduces computational overhead while enhancing feature interaction. Furthermore, the integration of Morlet wavelet-based activation enables effective time–frequency representation, and the use of weighted mean of vectors (WMOV) optimisation improves model parameter tuning. The proposed model thus come up with a robust and scalable answer for accurate and efficient prediction of Type-2 Diabetes.

4.1 Background Concepts

This section presents the fundamental concepts underlying the proposed FFT Overlap-and-Add (FFT-OLA) deep learning model for predicting Type-2 Diabetes. The key components include convolution operations, Fast Fourier Transform (FFT), overlap-and-add (OLA) technique, and optimisation strategies used for efficient model training.

4.1.1 Convolution in Feature Learning

To separate out meaningful theme from input data, convolution is a fundamental operation used in machine learning and deep learning models. It enables local feature interaction by combining input feature vectors with learnable filters or kernels. In traditional deep learning architectures, convolution is performed in the spatial domain, where each output element is computed as a weighted sum of neighbouring input elements. Although effective, this process becomes computationally expensive for large input sizes, as its complexity grows quadratically with the number of features.

4.1.2 Fast Fourier Transform (FFT)

To convert signals from the spatial (or time) domain into the frequency domain, the Fast Fourier Transform (FFT) is an effective algorithm. By representing data in terms of its frequency components, convolution operations can be simplified into element-wise multiplications, significantly reducing computational complexity. Specifically, convolution in the spatial domain corresponds to multiplication in the frequency domain, with an inverse transform applied afterward. This reduces the computational complexity from quadratic to near-linearithmic, making FFT-based methods highly suitable for large-scale data processing.

4.1.3 Overlap-and-Add (OLA) Method

The overlap-and-add (OLA) method is a technique used to efficiently perform convolution on long input sequences by dividing them into smaller overlapping segments. Each segment is processed independently, typically using FFT-based convolution, and the resulting outputs are combined by overlapping and summing the corresponding regions. This approach improves computational efficiency and enables parallel processing of data segments. Additionally, OLA enhances robustness by handling input data variations more effectively, thereby constructing a method suitable for complex and multidimensional datasets.

4.1.4 Deep Learning Optimization

Optimization has an important role in training deep learning models by adjusting network parameters to minimise a given cost function. Traditional optimisation techniques such as gradient-based methods may face challenges such as local minima, slow convergence, and sensitivity to initialisation. To overcome these limitations, metaheuristic optimisation approaches are often employed to enhance global search capability and improve convergence behaviour.

4.1.5 Weighted Mean of Vectors (WMOV)

The weighted mean of vectors (WMOV) is a inhabitants-based heuristic optimisation technique created to improve parameter tuning in complex models. It is managed by maintaining a set of vectors and updating them using a weighted averaging mechanism. Instead of relying solely on gradient information, WMOV combines multiple candidate solutions, giving higher importance to better-performing vectors.

This strategy enables a stabilize among the investigation of search space and utilization of promising regions, thereby avoiding premature convergence. In the proposed model, WMOV is employed to optimise the weights of the deep learning network, leading to improved convergence speed and enhanced prediction accuracy.

4.2 Proposed FFT-OLA Based Deep Learning Model

This section presents the proposed FFT Overlap-and-Add (FFT-OLA) based deep learning model for estimating Type-2 Diabetes Mellitus (T2DM). The model is designed to improve prediction accuracy while reducing computational complexity by integrating efficient convolution techniques with advanced feature transformation and optimisation strategies. The overall architecture of the suggested framework is shown in Figure 4.1.

The overall framework includes four major phases: data pre-processing, data augmentation, FFT-OLA based deep learning classification, and parameter optimisation. These stages are systematically integrated to enhance the quality of input data, improve feature representation, and achieve accurate classification outcomes.

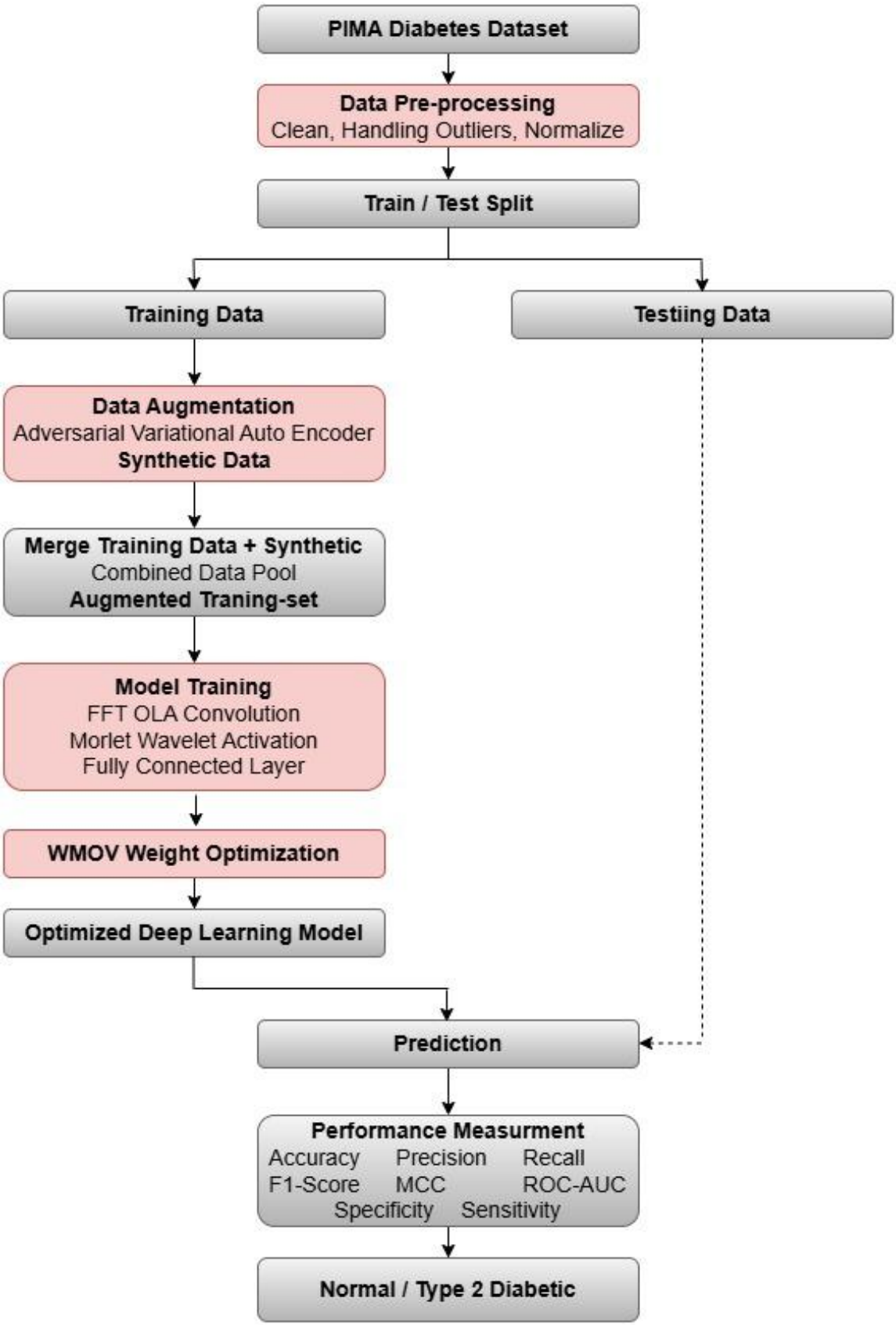

Figure 4.1 Architecture of the Proposed FFT-OLA Based Deep Learning Model

Initially, the input data go through a pre-processing stage for ensuring data quality and uniformity. This stage includes outlier rejection, missing value imputation, and normalisation. Outliers are removed using statistical measures based on quartile ranges to eliminate noisy and inconsistent observations. Missing or null values are replaced using mean imputation to avoid information loss. Subsequently, normalisation is applied to scale all features within a uniform range, which facilitates efficient learning in deep neural networks.

Following pre-processing, a data augmentation module is employed using an adversarial variational auto-encoder (AVAE). The purpose of this stage is to show data unevenness and improve the diversity of feature representations. The AVAE learns the latent distribution of the original dataset and generates synthetic samples that preserve the underlying characteristics of the input details. This enriched dataset empowers the model to find out more discriminative patterns, particularly for underrepresented classes.

The vital component of the suggested framework is the deep learning classifier based on FFT-OLA. In contrast to conventional convolutional neural networks that perform convolution in the spatial domain, the proposed approach uses the Fast Fourier Transform, which converts convolution operations into the frequency domain. The input feature vectors are divided into smaller overlapping segments, and each segment is processed independently using FFT-based convolution. The resulting outputs are then combined using the overlap-and-add technique to reconstruct the final feature representation. This approach significantly reduces computational complexity while maintaining effective feature extraction capability.

To further enhance the nonlinear mapping between input features and output classes, a Morlet wavelet-based activation function is incorporated within the intermediate layers of the neural net. The wavelet-based activation enables simultaneous analysis in both time and frequency domains, permitting the framework to catch subtle inequalities and complex patterns associated with diabetes risk factors. This leads to improved discrimination among diabetic and non-diabetic samples.

Finally, the suggested model’s parameters are optimised using the weighted mean of vectors (WMOV) optimisation algorithm. This metaheuristic approach updates the model weights by computing weighted averages of candidate solutions, enabling efficient

diversification and intensification of the problem domain. The optimisation process improves convergence characteristics and enhances classification performance.

The integration of FFT-OLA convolution, wavelet-based activation, and WMOV optimisation results in a computationally efficient and highly accurate predictive model. The suggested architecture constructively overcomes the limitations of conventional deep learning structures by reducing computational overhead while improving feature representation and classification accuracy.

4.3 Mathematical Formulation

This part of the study describes the mathematical formulation of the suggested FFT-OLA based deep learning model for predicting Type-2 Diabetes, including pre-processing, feature augmentation, frequency-domain convolution, and classification.

4.3.1 Input Representation

Let the input dataset is viewed as:

$$X = \{x_1, x_2, x_3, \ldots\ldots, x_N\} \tag{4.1}$$

here $x_i \in \mathrm{R}^d$ represents the i^{th} sample with d features.

4.3.2 Data Pre-processing

Outlier rejection using interquartile range (IQR) is expressed as:

$$x = \begin{cases} x, & Q_1 - 1.5 \times IQR \leq x \leq Q_3 + 1.5 \times IQR \\ reject, & otherwise \end{cases} \tag{4.2}$$

Missing value imputation using mean:

$$x = \begin{cases} \mu, & if\ x\ is\ missing \\ x, & otherwise \end{cases} \tag{4.3}$$

4.3.3 Feature Normalization

Min-max normalization is calculated by:

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.4}$$

4.3.4 Feature Augmentation Using AVAE

The latent representation is defined as:

$$z \sim q(z|x) \tag{4.5}$$

The reconstructed output is:

$$\hat{x} \sim p(x|z) \tag{4.6}$$

The objective function is:

$$L = Loss(x, \hat{x}) + \lambda \sum_{i=1}^{N} c_i \tag{4.7}$$

4.3.5 FFT-Based Convolution

The convolution in frequency domain is expressed as:

$$Y = F^{-1}(F(X) \cdot F(H)) \tag{4.8}$$

4.3.6 Overlap-and-Add (OLA) Method

The segmented input is:

$$X = \sum_k X_k \tag{4.9}$$

The convolution for each segment:

$$Y_k = F^{-1}(F(X_k) \cdot F(H)) \tag{4.10}$$

Final output after overlap and addition:

$$Y = \sum_k Y_k \tag{4.11}$$

4.3.7 Morlet Wavelet Activation

The Morlet wavelet function is defined as:

$$\psi(h) = \frac{1}{\sqrt{f_b}}\cos(2\pi f_c h)\exp(-\frac{h^2}{f_b}) \tag{4.12}$$

4.3.8 Classification and Loss Function

The sigmoid activation function is:

$$f(x) = \frac{1}{1+e^{-x}} \tag{4.13}$$

The cross-entropy loss function is calculated by:

$$L = \sum_{i=1}^{N}[t_i \log(p_i) + (1 - t_i)\log(1 - p_i)] \tag{4.14}$$

4.3.9 Computational Complexity

The complexity of FFT-OLA convolution is:

$$O(M^2 logM) \tag{4.15}$$

4.3.10 WMOV Optimization Formulation

To optimise the network parameters, the weighted mean of vectors (WMOV) approach is employed. Let the population of candidate solutions be defined as:

$$Y = \{y_1, y_2, y_3, \ldots\ldots, yN_p\} \tag{4.16}$$

where each y_i represents a vector of model parameters in the search space.

The updated solution is generated using a weighted mean of selected vectors:

$$y_{new} = \sum_{i=1}^{k} w_i \cdot y_i \tag{4.17}$$

where w_i are weighting coefficients satisfying $\sum w_i = 1$

The optimisation objective is to minimise the classification loss:

$$min\ L(\theta) \tag{4.18}$$

where $L(\theta)$ represents the loss function associated with the model parameters θ.

4.4 Weighted Mean of Vectors (WMOV) Optimization

The performance of deep learning framework is highly influenced by the optimal selection of network parameters. In the proposed FFT-OLA based deep learning framework, a weighted mean of vectors (WMOV) optimization algorithm is employed to efficiently tune the model parameters and improve classification accuracy. WMOV is a population-based metaheuristic optimization technique that updates candidate solutions using weighted averaging strategies, enabling effective exploration and exploitation of the search space.

4.4.1 Initialization

Let the population consist of Np candidate solutions, where each solution represents a set of model parameters (weights). The population is initialized randomly within the search space:

$$Y = \{y_1, y_2, y_3, \ldots\ldots, yN_p\}$$

where each $y_i \in R^d$ denotes a vector in a d-dimensional search space corresponding to network parameters.

4.4.2 Weighted Mean Update Strategy

The core idea of WMOV is to generate new candidate solutions using the weighted mean of selected vectors from the population. Instead of directly modifying a single solution, the algorithm combines information from multiple solutions, including the best, better, and worst candidates, to guide the search process.

The weighted mean is computed as:

$$y_{new} = \sum_{i=1}^{k} w_i \cdot y_i$$

where w_i represents the weight assigned to the i^{th} vector and satisfies $\sum wi=1$. This strategy ensures that better solutions have a greater influence on the generation of new candidates.

4.4.3 Exploration and Exploitation

WMOV balances global exploration and local exploitation through adaptive update mechanisms. During the early iterations, the algorithm emphasises exploration by considering diverse candidate solutions, thereby avoiding premature convergence. As the iterations progress, the search becomes more exploitative, focusing on refining the best solutions to achieve optimal performance.

Additionally, random perturbations are introduced to enhance diversity and prevent stagnation in local optima. This mechanism improves the robustness of the optimization process.

4.4.4 Local Search Enhancement

To further improve convergence, a local search strategy is incorporated around the best solution. This enables fine-tuning of model parameters by exploring the neighbourhood of promising solutions. The combination of global search via weighted mean and local refinement enhances the overall optimisation capability of the algorithm.

4.4.5 Integration with Proposed Model

In the proposed framework, the weights of the MW-FFT-OAconv network is optimize using WMOV. The purpose of the optimisation process is to reduce the classification loss by iteratively updating the network parameters. At each iteration, candidate weight vectors are evaluated based on their fitness, and new solutions are generated using the weighted mean strategy.

The integration of WMOV with FFT-OLA based convolution and Morlet wavelet activation improves convergence speed and enhances classification performance. By effectively tuning the model parameters, the proposed approach achieves better generalisation and higher prediction accuracy for Type-2 Diabetes detection.

4.5 Proposed Algorithm-2

This section presents the proposed Algorithm-2, which integrates the Morlet Wavelet-based FFT Overlap-and-Add Convolution (MW-FFT-OAconv) framework with the Weighted Mean of Vectors Optimizer (WMOV) for T2DM prediction. For clarity and reproducibility, the algorithm is organized into two parts: (i) the algorithm specification, which defines the input, output, and key parameters, and (ii) the algorithm procedure, which describes the sequential execution of the proposed framework.

4.5.1 Algorithm-2: FFT-OLA Based Type-2 Diabetes Prediction using Deep Learning

Before presenting the detailed execution procedure of the proposed Algorithm-2, the algorithm specification is described by defining the required input data, expected outputs, and key parameters. These components establish the computational framework of the proposed MW-FFT-OAconv-WMOV model and provide a clear understanding of the information processed and generated during diabetes prediction.

Input:

- Augmented dataset
- Training labels
- Initial network weights

Output:

- Optimized deep learning model
- Predicted diabetes class

Parameters:

- FFT block size
- Morlet parameters

- WMOV iterations

4.5.2 Procedure of Algorithm-2

Step 1: Load the Pima Indians Diabetes Dataset D

Step 2: Data Pre-processing

a) Remove outliers using interquartile range (IQR) method

b) Impute missing values using mean

c) Normalize feature values

Step 3: Feature Augmentation

a) Train adversarial variational auto-encoder (AVAE)

b) Learn latent feature representation

c) Generate augmented dataset

Step 4: FFT-OLA Based Convolution

a) Divide input feature vectors into overlapping segments

b) For each segment:

i. Apply Fast Fourier Transform (FFT)

ii. Perform convolution in frequency domain

iii. Apply inverse FFT to obtain transformed output

c) Combine all segments using overlap-and-add method

Step 5: Nonlinear Feature Transformation

a) Apply Morlet wavelet activation function

b) Extract transformed feature representation

Step 6: Classification

a) Pass features through fully connected layer

b) Compute output probabilities using sigmoid function

Step 7: Weight Optimization

a) Initialize weighted mean of vectors (WMOV) parameters

b) Update model weights using WMOV strategy

c) Repeat until convergence criteria is satisfied

Step 8: Prediction

a) Generate final output using optimized model

b) Assign class label as Non-Diabetic or Diabetic

4.6 Analysis of Computational Complexity

This part of the study analyses the computational complexity of the proposed FFT-OLA based deep learning model and compares it with conventional convolution-based approaches. The objective is to displays the efficiency of the suggested method regarding time complexity.

For computational complexity analysis, *M* denotes the input dimension, n represents the kernel or block size used in FFT-OLA convolution, *Np* denotes the WMOV population size, *G* is the maximum number of optimisation iterations, and *d* represents the dimensionality of the optimisation problem.

4.6.1 Complexity of Conventional Convolution

In traditional deep learning models, convolution is performed in the spatial domain. For $M \times M$ input size and $n \times n$ kernel size, the computational complexity is given by:

$$O(M^2 n^2)$$

This quadratic dependence on both input and kernel sizes makes spatial convolution computationally expensive, especially for large datasets and deep architectures.

4.6.2 Complexity of FFT-Based Convolution

By applying the Fast Fourier Transform (FFT), convolution is expressed as element-wise multiplication in the frequency domain. The complexity of FFT-based convolution is:

$$O(M^2 logM)$$

This significantly reduces the computational burden compared to spatial convolution, particularly for large input sizes.

4.6.3 Complexity of Overlap-and-Add (OLA) Convolution

In the proposed approach, the overlap-and-add (OLA) method is used to process input data in smaller overlapping segments. Each segment undergoes FFT-based convolution independently, and the outcomes are integrated to generate the final output.

Let the input be divided into smaller blocks of size $n \times n$. The complexity of processing each block using FFT is:

$$O(n^2 logn)$$

If the total number of blocks is proportional to $\frac{M^2}{n^2}$, the overall complexity becomes:

$$O(M^2 logn)$$

This formulation shows that the proposed FFT-OLA convolution is more efficient than both traditional spatial convolution and standard FFT convolution for appropriately chosen block sizes.

4.6.4 Complexity of WMOV Optimization

The computational complexity of the weighted mean of vectors (WMOV) optimisation depends on the number of candidate solutions and iterations. Let N_{p} represents the population size, G the number of iterations, and d the dimensionality of the solution space. The complexity can be expressed as:

$$O(G \cdot N_p \cdot d)$$

Although WMOV introduces additional computational overhead, it significantly improves convergence speed and model accuracy, thereby reducing the number of training iterations required.

4.6.5 Overall Complexity Analysis

The overall computational complexity of the proposed MW-FFT-OAconv-WMOV framework is determined by combining the computational costs of FFT-based overlap-and-add convolution and the WMOV optimisation process. The FFT-OLA convolution requires a computational complexity of $O(M^2 \log n)$, whereas the WMOV optimisation contributes $O(G \cdot N_p \cdot d)$, where G denotes the number of optimisation iterations, N_p represents the population size, and d is the dimensionality of the optimisation problem.

Consequently, the overall computational complexity of the proposed framework can be expressed as:

$$O(M^2 \log n + GN_p d)$$

This complexity demonstrates that the proposed framework substantially reduces the computational burden associated with conventional spatial convolution $(O(M^2n^2))$ by employing FFT-based overlap-and-add convolution while incorporating WMOV optimisation to improve convergence and predictive performance.

Compared to conventional deep learning models, the proposed approach achieves:

- Reduced convolution complexity through FFT transformation
- Efficient handling of large input data using OLA segmentation
- Improved convergence using WMOV optimisation

As a result, the proposed FFT-OLA based model provides a computationally efficient framework suitable for large-scale medical datasets while maintaining high prediction accuracy.

CHAPTER 5

Results and Discussion

In this chapter a comprehensive experimental evaluation of the suggested methodologies developed for early detection of Type 2 Diabetes Mellitus (T2DM) is discussed. The analysis focuses on two major components of the proposed framework: Algorithm-1, which performs data and feature augmentation using the Adversarial Variational Autoencoder (AVAE), and Algorithm-2, which performs classification using the Morlet Wavelet assisted deep learning network with FFT overlap-add convolution (MW-FFT-OAconv) optimized by the Weighted Mean of Vectors (WMOV) algorithm.

The main focus of this chapter is to find out the individual and combined consequences of these two algorithms on prediction performance. Initially, the effectiveness of the AVAE-based augmentation technique is analysed regarding its capicity to manage class imbalance and improve data representation. Subsequently, the performance of the MW-FFT-OAconv classifier with WMOV optimization is evaluated to assess its capability in learning complex nonlinear relationships from medical data.

Furthermore, the integrated framework, combining both Algorithm-1 and Algorithm-2, is evaluated to demonstrate the overall improvement in classification accuracy and robustness. The experimental results are analysed using standard statistical performance measures, such as accuracy score, precision, recall, F1-measure, specificity, Matthews Correlation Coefficient (MCC), and classifier error percentage.

A comparative study is also carried out with existing machine learning and deep learning frameworks to validate the superiority of the proposed approach. In addition, detailed discussions are provided based on confusion matrix analysis, performance comparison, and

statistical validation to showcase the robustness and practical applicability of the suggsted framework.

5.1 Experimental Framework and Dataset Context

The experimental evaluation of the proposed framework is conducted using the PIMA Indian Diabetes Dataset, which is described in Chapter 3 in detail. The dataset comprises diagnostic medical attributes like glucose measure, blood pressure level, body mass index, insulin level, patrient's age, and other relevant features used for binary classification of diabetic and non-diabetic cases.

For consistency and fair comparison, the same dataset is utilised across all experimental analyses presented in this chapter. The preprocessing steps, together with outlier rejection using the IQR technique, missing value imputation, and feature normalization, are applied uniformly as discussed in the earlier chapters.

It is significant to observe that the suggested study follows a two-stage hybrid framework, where data augmentation using the Adversarial Variational Autoencoder (AVAE) is incorporated as a key component to enhance dataset quality. The AVAE-based augmentation is applied not only in the initial stage (Algorithm-1) but is also integrated into the training process of the classification model (Algorithm-2). This ensures improved representation of minority classes and better generalization of the model.

Thus, all experiments in this chapter are performed under a unified framework where the dataset, preprocessing strategy, and augmentation mechanism remain consistent, enabling a reliable evaluation of the classification performance and the participation of each component of the suggested framework.

5.2 Experimental Setup

In this study, experimental setup is designed in such a way that it systematically evaluate the performance of the suggested framework, which consists of two interconnected components: Algorithm-1, presented in Chapter 3, and Algorithm-2, presented in Chapter 4. All experiments are conducted under a unified environment to ensure consistency, fairness, and reproducibility of results.

The implementation is carried out using the Python programming platform with standard libraries for machine learning and deep learning. The dataset is partitioned into training and testing segments using an 80:20 partition, where 80% of the samples is used for model training and the rest 20% is used for performance evaluation. In addition, k-fold cross-validation is employed during learning process to ensure robustness and to minimize the effect of random data partitioning.

The experimental process follows a two-stage pipeline. In the first stage, corresponding to Algorithm-1, the preprocessed dataset is provided as input to the Adversarial Variational Autoencoder (AVAE). This module performs data and feature augmentation by generating synthetic samples for both cases of diabetic and non-diabetic. The augmented data is then merged with the primary data-source to form an enhanced training set with improved class balance and feature representation.

In the second stage, corresponding to Algorithm-2, the augmented dataset produced by Algorithm-1 is used as input to the Morlet Wavelet assisted deep learning network with FFT overlap-add convolution (MW-FFT-OAconv). The Morlet wavelet activation function is employed to capture nonlinear relationships and time-frequency characteristics of the input features, while the FFT overlap-add convolution mechanism enhances computational efficiency during feature extraction.

For further enhancement of the classification model, the Weighted Mean of Vectors (WMOV) optimization algorithm is applied for tuning the network parameters. This optimization process iteratively updates the weights of the MW-FFT-OAconv model to achieve better convergence and improved classification accuracy. The training process is guided by the cross-entropy loss function, which calculates the inconsistency between predicted and actual class symbol.

The performance of the system is analysed under three configurations for comprehensive evaluation:

1) classification without augmentation,

2) classification with AVAE-based augmentation (Algorithm-1), and

3) the complete hybrid framework integrating both Algorithm-1 and Algorithm-2.

Additionally, tradinational machine learning frameworks like Random Forest, Naïve Bayes, Decision Tree, and Logistic Regression are implemented under the same experimental conditions to provide a fair comparison with the proposed approach.

This experimental setup ensures a structured evaluation of both individual components and the integrated framework, facilitating a clear analysis of how each proposed algorithm contributes to overall system performance.

5.3 Performance Evaluation Measures

To comprehensively estimate the accuracy of the presented framework, the performance of both Algorithm-1 and Algorithm-2, as well as their integrated implementation, is assessed using standard statistical metrics commonly adopted in medical diagnosis and classification problems.

The evaluation focuses on measuring the classification capability of the system in terms of correctness, reliability, and balance between classes. This study utilizes the following metrics: accuracy, precision, recall (true positive rate), specificity (true negative rate), F1-score, Matthews Correlation Coefficient (MCC), and classifier error percentage (CEP).

Accuracy measures the overall accuracy of the model by identifying the proportion of correctly classified samples among the total number of instances. Precision evaluates the proportion of correctly predicted positive cases out of all predicted positives, while recall (sensitivity) reflects the model's capabiliity to correctly identify actual diabetic cases. Specificity measures the model's capability to correctly identify non-diabetic cases.

The F1-score provides a harmonic balance between precision and recall, making it suitable for imbalanced datasets. The Matthews Correlation Coefficient (MCC) is a robust binary classification metric that considers all four confusion matrix components (true positives, true negatives, false positives, and false negatives), ensuring balanced evaluation even when class distribution is uneven. The classifier error percentage (CEP) denotes the proportion of misclassified samples and measures the model's prediction error.

In Chapter 4 the mathematical formulations of these evaluation metrics have been presented and are utilised consistently in this chapter for performance comparison. For evalution following metrics are applied:

1) Data augmentation impact using Algorithm-1,
2) The classification performance of Algorithm-2, and
3) The overall effectiveness of the integrated hybrid framework.

This comprehensive evaluation approach ensures that the proposed system is assessed not only based on accuracy but also considering reliability, robustness, and clinical relevance.

5.4 Algorithm-1: Performance Analysis

In Chapter 3 the effectiveness of Algorithm-1 is presented and is analysed in this section by evaluating the impact of Adversarial Variational Autoencoder (AVAE)-based data augmentation on the classification performance. In addition to augmentation, the role of improved preprocessing strategies is also examined to understand their contribution to overall system performance.

The primary objective of this analysis is to explore the effects of data augmentation and preprocessing on model accuracy, class balance, and generalization, particularly in medical datasets with limited and imbalanced samples.

5.4.1 Impact of Improved Missing Value Imputation

Medical datasets often contain invalid or missing values that can adversely affect model performance. There are zero values in certain attributes, like Blood Pressure, Skin Thickness, and Insulin, in the PIDD, which are not clinically meaningful and must be treated as missing values.

In conventional approaches, the global mean of the respective feature is used for missing value replcement. However, in this work, a refined imputation strategy is adopted, where missing values are replaced using the mean values corresponding to similar age groups instead of the overall mean. This approach preserves the inherent distribution of the data

and maintains clinical relevance, as physiological parameters tend to vary across different age groups.

To evaluate the effectiveness of this strategy, experiments are conducted using:

- Global mean imputation
- Age-group-based mean imputation

The comparative results are presented in Table 5.1.

Table 5.1 Missing Value Imputation Strategy Impact

Imputation Method	Accuracy (%)	Precision	Recall	F1-Score	MCC
Global Mean	93.09	0.92	0.97	0.95	0.81
Age-group Mean (Proposed)	**94.95**	0.95	0.98	0.97	0.86

From Table 5.1, it is observed that the proposed age-group-based imputation improves classification accuracy by approximately 2% compared to the conventional global mean approach. Additionally, improvements in recall and MCC indicate better identification of diabetic cases and more balanced classification performance.

This demonstrates that domain-aware preprocessing, particularly demographic-based imputation, significantly enhances model performance in healthcare datasets.

5.4.2 Effect of AVAE-Based Data Augmentation

The PIMA dataset is imbalanced, with a higher number of non-diabetic samples compared to diabetic samples. This condition can bias the classifier toward the majority class, leading to weaker recognition of minority class samples.

To address this issue, the proposed AVAE-based augmentation technique generates synthetic samples for both classes by modeling the latent distribution of the original dataset. Unlike traditional oversampling techniques such as SMOTE, AVAE generates more diverse and realistic samples, thereby improving the representational capacity of the dataset.

To analyse the impact of AVAE, experiments are conducted under the following configurations:

1. Without Augmentation
2. With Traditional Oversampling (SMOTE)
3. With AVAE-based Augmentation (Proposed)

In Table 5.2, the performance comparison is shown.

Table 5.2 Effect of Data Augmentation Techniques

Method	Accuracy (%)	Precision	Recall	F1-Score	MCC
Without Augmentation	86.45	0.84	0.81	0.82	0.75
SMOTE	90.12	0.88	0.87	0.87	0.82
AVAE (Proposed)	**94.95**	0.95	0.98	0.97	0.86

From Table 5.2, it is evident that AVAE-based augmentation outperforms both the baseline and SMOTE-based approaches. The improvement is attributed to the ability of AVAE to generate diverse and realistic synthetic samples, which better capture the underlying data distribution.

The results show that AVAE-based augmentation significantly improves classification performance compared to both the baseline and SMOTE-based approaches. The improvement is attributed to:

- Better modeling of feature distribution
- Increased diversity in training samples
- Reduced overfitting

Furthermore, the integration of AVAE with the classification model enhances the detection of diabetic cases, leading to improved recall and F1-score.

5.4.3 Additional Dataset Validation

To further assess the Algorithm-1's generalization capability, experiments are conducted on an additional dataset, namely the Diabetes Prediction Dataset, sourced from the Kaggle repository [63].

The dataset consists of medical and demographic attributes like patient's age, gender, body mass index, the status of hypertension, heart disease symptoms, smoking profile, HbA1c measure, and blood glucose measure, along with a binary target variable denoting the existence or nonexistence of diabetes. Compared to the PIDD, this dataset comprices a wider set of features and captures both clinical and lifestyle-related factors, making it suitable for evaluating the robustness of the proposed preprocessing and augmentation approach.

The same preprocessing strategy, including age-group-based missing value imputation, along with AVAE-based data augmentation, is applied to this dataset to ensure consistency. A Random Forest classifier is used for evaluation, similar to the experiments conducted on the PIMA dataset.

In Table 5.3, the performance results is presented.

Table 5.3 Performance of Algorithm-1 on Diabetes Prediction Dataset

Method	Accuracy (%)	Precision	Recall	F1-Score	MCC
Without Augmentation	88.76	0.86	0.84	0.85	0.79
SMOTE	91.34	0.89	0.88	0.88	0.85
AVAE (Proposed)	**99.35**	0.99	0.99	0.99	0.93

The results demonstrate that AVAE-based augmentation consistently improves performance across different datasets, confirming that Algorithm-1 is robust and not dataset-specific.

5.4.4 Discussion

The results presented in this section demonstrate that Algorithm-1 significantly enhances data quality and improves classification performance. The proposed preprocessing strategy

assures that non-availability values are handled in a clinically meaningful approach, while AVAE-based augmentation effectively addresses class imbalance.

Key observations include:

- Age-group-based imputation improves accuracy and recall
- AVAE outperforms traditional oversampling methods
- The approach generalizes well across different datasets

Thus, Algorithm-1 provides a strong foundation for the classification process by improving both the quality and representativeness of the input data.

5.5 Algorithm-2: Performance Analysis

This section presents the performance evaluation of Algorithm-2, described in Chapter 4, which focuses on classification using the proposed Morlet Wavelet assisted deep learning network with FFT overlap-add convolution (MW-FFT-OAconv) optimized through the Weighted Mean of Vectors (WMOV) algorithm.

Unlike Algorithm-1, which enhances data quality, Algorithm-2 is responsible for learning discriminative features and performing accurate classification. The evaluation in this section specifically examines the contribution of the model architecture and optimization strategy. To understand the learning behavior of the proposed model, the training and validation accuracy curves are evaluated, as shown in Figure 5.1.

Figure 5.1 illustrates the learning behaviour of the proposed MW-FFT-OAconv model during the training process. The training and validation accuracy increase steadily with the number of epochs, indicating that the model effectively learns discriminative representations from the augmented diabetes dataset. Both curves converge smoothly after approximately 60–70 epochs and stabilize near 98% accuracy. The relatively small gap between the training and validation curves indicates good generalization capability with minimal overfitting. The stable convergence further demonstrates that the proposed architecture successfully captures complex nonlinear relationships while maintaining consistent prediction performance on unseen data.

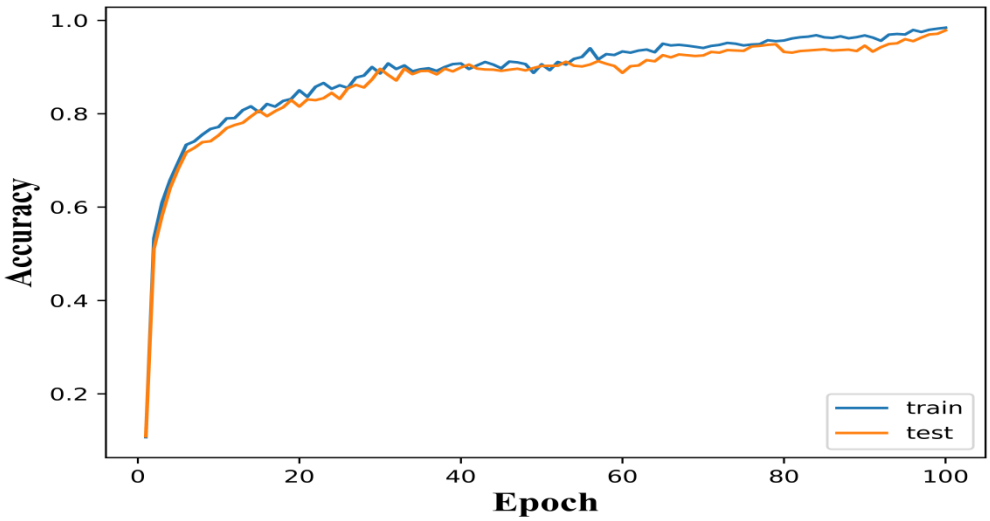

Figure 5.1 Training and validation accuracy of the MW-FFT-OAconv model

The smooth converging of both curves shows that the proposed architecture is capable of learning generalized patterns from the dataset, thereby ensuring stable and reliable performance.

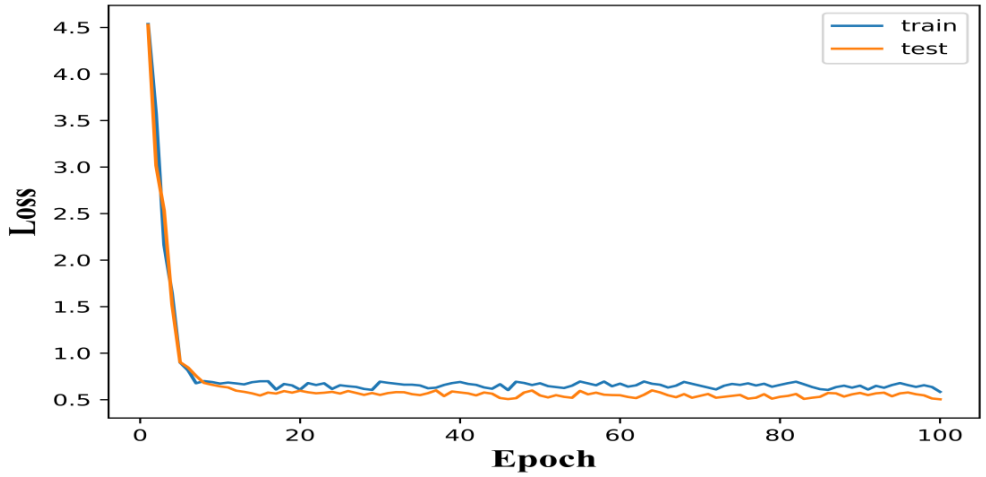

Figure 5.2 Training and Validation Loss of the MW-FFT-OAconv Model

Figure 5.2 presents the variation of the training and validation loss during the learning process. At the initial epochs, both loss curves decrease rapidly, indicating fast optimization of the network parameters. As training progresses, the loss gradually stabilizes and converges to a very small value. The close agreement between the training and validation loss curves demonstrates that the proposed model avoids significant overfitting while maintaining stable convergence. The smooth reduction in loss confirms

the effectiveness of the Morlet wavelet activation function and the FFT overlap-add convolution mechanism in facilitating efficient feature learning. Moreover, the stable behaviour of the validation loss indicates that the optimized model generalizes well to previously unseen diabetic patient records.

5.5.1 Effect of Morlet Wavelet Activation Function

The proposed MW-FFT-OAconv model employs the Morlet wavelet activation function in place of conventional activation functions such as ReLU and sigmoid. The Morlet wavelet introduces time-frequency characteristics into the learning process, permitting the model to capture complicated nonlinear relationships present in medical database.

The use of wavelet activation enhances:

- Nonlinear feature representation
- Sensitivity to subtle variations in clinical attributes
- Model's Discriminative capability

As a result, the framework achieves improved recall and F1-score, denoting better recognition of diabetic samples. This is particularly significant in medical diagnosis, where failing to detect a positive case can lead to severe consequences.

5.5.2 Effect of FFT Overlap-Add Convolution

The MW-FFT-OAconv model integrates FFT-based overlap-add convolution to improve computational efficiency. For large databases or higher-dimensional feature space, conventional convolution operations can be computationally expensive.

By transforming convolution operations into the frequency domain, the FFT overlap-add approach:

- Reduces computational complexity
- Accelerates training and inference

- Efficiently processes overlapping feature patterns

The results demonstrate that FFT-based convolution maintains high classification performance while improving computational performance, making the suggested model suitable for real-time as well as for large-scale medical applications

5.5.3 Effect of WMOV Optimization

The Weighted Mean of Vectors (WMOV) algorithm is used to optimize the parameters of the MW-FFT-OAconv model. To evaluate its contribution, the model is analysed under two configurations:

- MW-FFT-OAconv without optimization
- MW-FFT-OAconv with WMOV optimization (proposed)

In Table 5.4, the comparative results is shown.

Table 5.4 Effect of WMOV Optimization on Model Performance

Method	Accuracy (%)	Precision	Recall	F1-Score	MCC
MW-FFT-OAconv (without WMOV)	95.12	0.93	0.92	0.92	0.89
MW-FFT-OAconv + WMOV (Proposed)	**98.44**	0.97	0.96	0.98	0.96

From Table 5.4, it is evident that WMOV significantly improves model performance. The optimization process enables better convergence, avoids local minima, and improves the framework's capacity to identify optimal weight parameters.

The convergence behavior of the WMOV optimization algorithm is shown in Figure 5.3, in which, the fitness parameter increases steadily with the number of iterations, indicating consistent improvement during the optimization process. The gradual upward trend demonstrates that the WMOV algorithm effectively refines the solution and converges towards an optimal set of parameters.

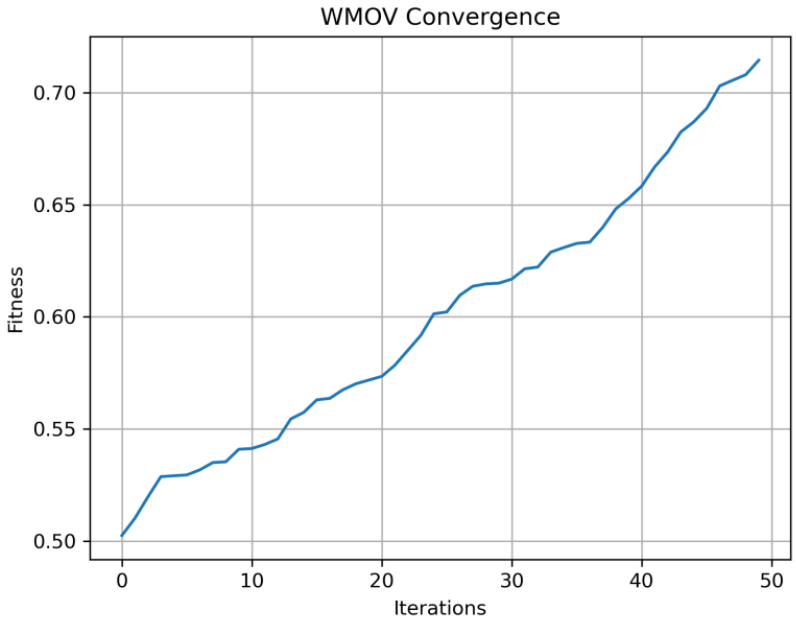

Figure 5.3 Convergence behavior of WMOV optimization

The stable convergence pattern without abrupt oscillations indicates a good balance between exploration and exploitation, which contributes to improved parameter tuning and enhanced classification performance of the MW-FFT-OAconv model.

5.5.4 Discussion on Algorithm-2

The findings reported in this section confirm that Algorithm-2 is a robust and efficient classification framework. The suggested architecture successfully overcomes the limitations of conventional deep learning frameworks by integrating advanced feature extraction and optimization techniques.

Key observations include:

- Improved feature representation through wavelet-based activation
- Reduced computational complexity due to FFT convolution
- Significant performance gains achieved through WMOV optimization

Thus, Algorithm-2 serves as the core classification component of the proposed system and plays a crucial role in achieving high prediction accuracy when integrated with the data enhancement provided by Algorithm-1.

5.6 Performance of Integrated Model

This section shows the overall performance of the suggested hybrid framework, which combines Algorithm-1 (Preprocessing and AVAE-based data augmentation) and Algorithm-2 (MW-FFT-OAconv with WMOV optimization). The goal is to assess the efficacy of the complete system when both data-level enhancement and model-level optimization are applied together.

The integration of these two algorithms enables the model to benefit from improved data representation as well as advanced feature extraction and optimization. Algorithm-1 enhances the quality and balance of the dataset, while Algorithm-2 effectively learns complex patterns and performs accurate classification.

The integrated model's performance is estimated using the measures defined in Section 5.3. The experimental results demonstrate that the suggested hybrid model accomplish an overall classification accuracy of 98.44%, together with higher precision, recall, and F1-score. The improvement in performance compared to individual components highlights the effectiveness of combining preprocessing, augmentation, and optimized deep learning.

5.6.1 Confusion Matrix Analysis

To further assess the classification performance of the integrated model, a confusion matrix is used as shown in figure 5.4, to examine the distribution of correctly and incorrectly classified samples.

From the confusion matrix, it is noticed that the proposed model achieves a high level of classification accuracy. The framework perfectly classifies 991 non-diabetic samples (True Negatives) and 519 diabetic samples (True Positives). The number of misclassifications is minimal, with only 10 false positives (FP) and 14 false negatives (FN).

The very low number of false negatives is especially notable in the context of medical diagnosis, as it indicates that the model successfully identifies the most of diabetic patients, thus lowering the risk of undetected disease. Similarly, the lower value of false positives ensures, non-diabetic individuals are not incorrectly classified as diabetic, avoiding unnecessary medical concern and treatment

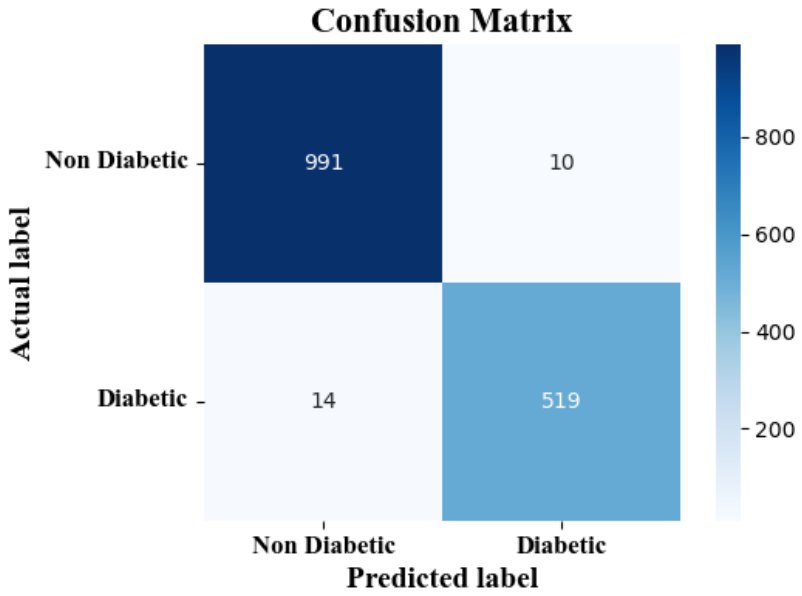

Figure 5.4 Confusion Matrix- Proposed AVAE-MW-FFT-OaConv-WMOV Model

Overall, the confusion matrix demonstrates that the proposed AVAE–MW-FFT-OAconv–WMOV model provides highly reliable and balanced classification performance, with strong capability in discriminating diabetic and non-diabetic cases. This confirms the effectiveness of the integrated framework for real-world clinical decision support applications.

5.6.2 Performance Interpretation

To provide a comprehensive interpretation of the predictive capability of the proposed hybrid framework, its performance is compared with several widely used conventional machine learning classifiers across multiple evaluation metrics. Figure 5.5 illustrates the comparative performance of the proposed model in terms of Accuracy, Precision, Recall, F1-score, Specificity, and Sensitivity. This visualization provides an overall understanding of the effectiveness and reliability of the proposed AVAE–MW-FFT-OAconv–WMOV framework.

Figure 5.5 clearly demonstrates that the proposed hybrid framework consistently achieves superior performance across all evaluation metrics when compared with Random Forest, Naïve Bayes, and Decision Tree classifiers.

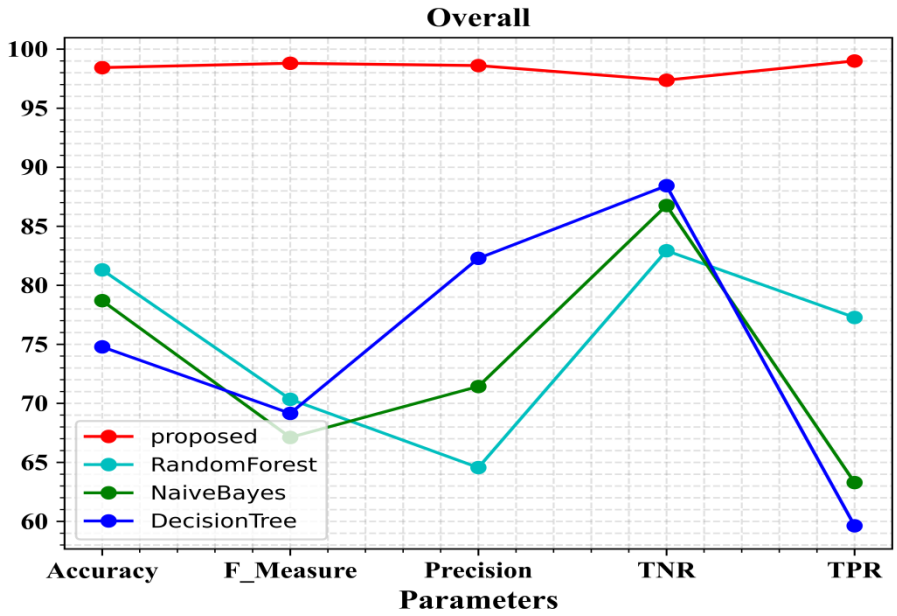

Figure 5.5 Overall Performance Comparison

The proposed model records the highest accuracy, precision, recall, F1-score, specificity, and sensitivity, indicating balanced prediction capability for both diabetic and non-diabetic classes. The consistently improved performance confirms that the integration of AVAE-based augmentation with MW-FFT-OAconv and WMOV optimization significantly enhances feature representation, classification capability, and model generalization.

The high recall and specificity values demonstrate that the suugested framewrok carry out well in identifying both diabetic and non-diabetic cases, which is critical for practical medical applications.

Although the proposed model demonstrates excellent predictive performance, understanding the contribution of individual clinical attributes is equally important for practical medical applications. Therefore, SHapley Additive exPlanations (SHAP) are employed to interpret the predictions generated by the proposed deep learning framework.

Figure 5.6 illustrates the SHAP summary plot, which explains the contribution of individual clinical variables toward diabetes prediction. Glucose exhibits the highest

influence on the prediction outcome, followed by BMI and Pregnancies. Higher glucose values generally increase the probability of diabetes, whereas lower values contribute negatively to the prediction. Similar trends are observed for BMI and Diabetes Pedigree Function. The distribution of SHAP values demonstrates that the proposed model captures nonlinear relationships among multiple clinical variables rather than relying on a single predictor. This enhances the transparency and clinical interpretability of the proposed framework.

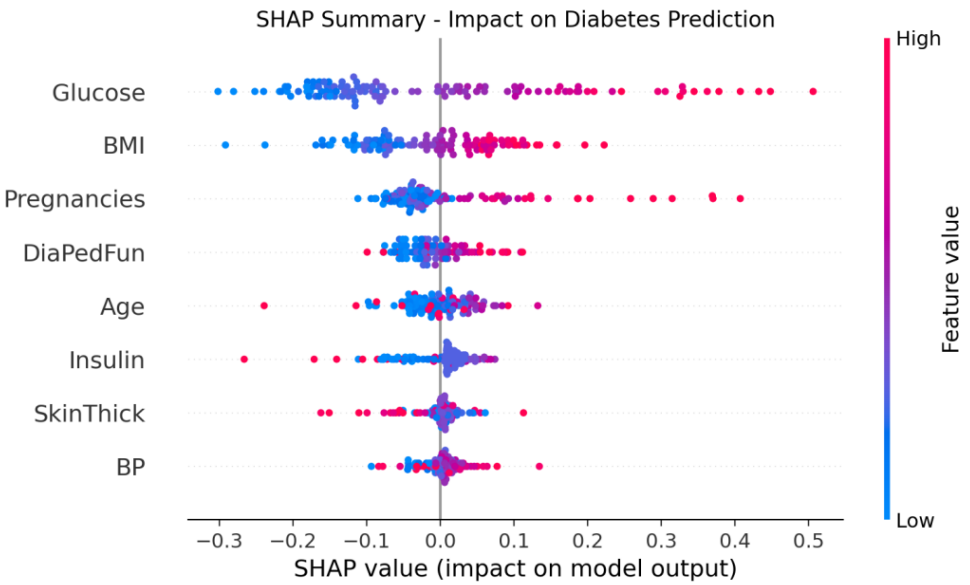

Figure 5.6 SHAP Summary Plot

Figure 5.7 presents the global feature importance computed using the mean absolute SHAP values.

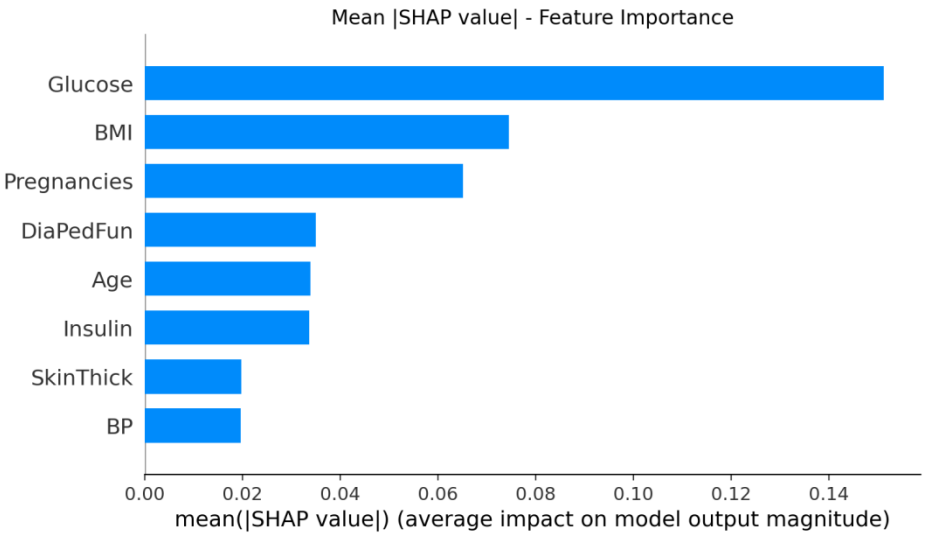

Figure 5.7 SHAP Feature Importance

Glucose is identified as the most influential predictor, followed by BMI, Pregnancies, Diabetes Pedigree Function, Age, and Insulin. Blood Pressure and Skin Thickness exhibit comparatively lower contributions. The obtained ranking is consistent with established clinical knowledge regarding T2DM, thereby validating that the proposed model makes clinically meaningful predictions rather than relying on arbitrary correlations

5.6.3 Discussion

The performance of the integrated framework confirms that combining data augmentation and advanced deep learning techniques results in a highly effective prediction model. The synergy between Algorithm-1 and Algorithm-2 enables the system to overcome the limitations of traditional approaches.

Key observations include:

- Data-level improvements enhance model learning
- Advanced architecture captures complex feature relationships
- Optimization ensures better convergence and accuracy

Thus, the proposed hybrid framework provides a robust and reliable solution for timely identification of T2DM and demonstrates strong potential for real-world clinical deployment.

5.7 Comparative Analysis with Existing Methods

To confirm the efficacy of the suggested hybrid framework, a comparative study is carried out with traditional machine learning models and existing research strategies. The comparison is carried out under consistent experimental conditions to ensure fairness and reliability.

5.7.1 Comparison with Conventional Machine Learning Models

The proposed model is first compared with widely used machine learning classifiers, such as Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest. These models

are commonly applied in medical diagnosis but frequently deficient the capability to identify complicated nonlinear associations present in clinical dataset. In Table 5.5 the comparison results is shown.

Table 5.5 Comparison with Conventional Machine Learning Models

Method	Accuracy (%)	Precision	Recall	F1-Score	MCC
Logistic Regression	82.35	0.8	0.78	0.79	0.72
Naïve Bayes	80.12	0.78	0.76	0.77	0.69
Decision Tree	84.67	0.82	0.81	0.81	0.75
Random Forest	90.12	0.88	0.87	0.87	0.82
Proposed Model	**98.44**	**0.97**	**0.96**	**0.96**	**0.94**

From Table 5.5, it is noticiable that the proposed framework remarkably outshines all conventional classifiers. While Random Forest achieves relatively higher performance among traditional methods, it still falls short compared to the suggested framework.

Figure 5.8 graphically compares the performance of the proposed model with conventional machine learning classifiers. The grouped bar chart clearly shows that the proposed framework consistently outperforms Naïve Bayes, Decision Tree, and Random Forest across all performance metrics. Compared with the line representation, the grouped bar chart facilitates easier comparison of individual evaluation metrics and clearly highlights the superiority of the proposed model.

Figure 5.8 Comparative performance of classification models

5.7.2 Comparison with Existing Research Works

To further validate the proposed approach, a comparison is made with recent research works on diabetes prediction using the PIDD dataset as given in Table 5.6.

Table 5.6 Comparison with Existing Research Works

Research Work	Methodology	Accuracy (%)
Reza et al. [64]	SVM + SMOTE	85.5
Kibria et al. [65]	XGBoost + Random Forest	90
Sampath et al. [66]	AdaBoost + XGBoost	90.4
Ganie et al. [67]	Gradient Boosting	92.85
García-Ordás et al. [68]	VAE + CNN	92.31
Proposed AVAE + RF [69]	**AVAE + Random Forest**	**93.08**
Proposed Hybrid Model [70]	**AVAE + MW-FFT-OAconv + WMOV**	**98.44**

From Table 5.6, it is observed that the suggested hybrid framework has the highest accuracy among all compared methods. The improvement over the earlier AVAE-based model further demonstrates the effectiveness of integrating modern deep learning and optimization strategies.

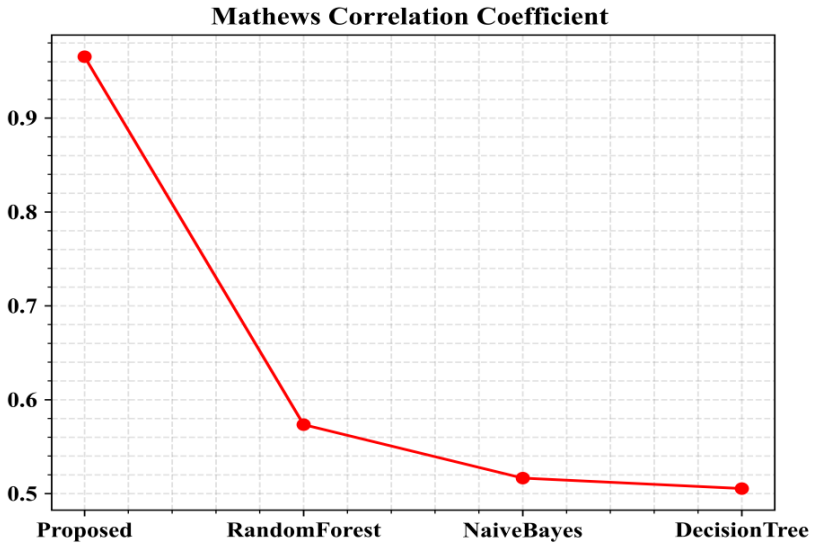

Figure 5.9 MCC Comparison

Figure 5.9 compares the Matthews Correlation Coefficient achieved by different classification models. The proposed framework achieves the highest MCC value, indicating excellent agreement between predicted and actual class labels. Since MCC

considers all components of the confusion matrix simultaneously, the superior MCC confirms the balanced classification capability of the proposed framework.

Figure 5.10 CER Comparison

Figure 5.10 illustrates the Classification Error Rate achieved by the compared methods. The proposed framework exhibits the lowest classification error, confirming that only a very small proportion of samples are misclassified. The reduced error rate further demonstrates the effectiveness of integrating AVAE augmentation with the MW-FFT-OAconv classifier and WMOV optimization.

5.7.3 Discussion

The comparative analysis clearly indicates the superiority of the proposed framework over both conventional machine learning models and existing research approaches. The integration of data augmentation, wavelet-based feature extraction, and optimization enables the model to successfully capture complicated patterns and has higher classification performance.

Key observations include:

- Significant improvement over baseline and state-of-the-art methods
- Robust performance over multiple evaluation measures
- Enhanced generalization capability

These findings validate the efficacy and novelty of the suggested strategy, making it suitable for real-world medical prediction systems.

5.8 Statistical Validation

This section presents the statistical validation of the proposed hybrid framework to evaluate its stability, reliability, and generalization capability. While the previous sections demonstrated high classification performance, it is essential to ensure that the results are consistent across different data partitions and are not dependent on a specific train-test split.

5.8.1 Cross-Validation Analysis

To examine the robustness of the proposed integrated framework, k-fold cross-validation is performed on the dataset. In this strategy, the dataset is partitioned into k equal segments, and the model undergoes trained and evaluated iteratively across different folds. This assures that each data sample is utilized for both training as well as testing, thereby providing a comprehensive evaluation. In Table 5.7, the cross-validation results is shown.

Table 5.7 Cross-Validation Performance of the Proposed Model

Subset	Accuracy (%)
Subset 1	97.85
Subset 2	98.12
Subset 3	98.44
Subset 4	98.05
Subset 5	98.21
Mean Accuracy	**98.13**

From Table 5.7, it is noticed that the framework achieves consistently high accuracy across all folds, with only minor variations. The small deviation in performance indicates that the model is stable and does not suffer from overfitting.

The high mean accuracy further confirms that the proposed hybrid framework generalizes well to unseen data and maintains reliable performance across different data splits.

5.8.2 Discussion

The cross-validation results shows that the suggested framework is both robust and reliable. The consistency in accuracy across multiple folds indicates that the model is not inclined toward a specific subset of data and is capable of maintaining stable performance.

This validation strengthens the credibility of the suggested framework and confirms its appropriate for real-world healthcare applications, where consistent and reliable predictions are critical.

5.9 Overall Discussion

This section gives a complete discussion of the findings presented in this chapter and highlights the key contributions of the suggested framework.

The results show that the integration of Algorithm-1 (preprocessing and AVAE-based augmentation) and Algorithm-2 (MW-FFT-OAconv with WMOV optimization) leads to significant improvements in classification performance. Each element of the model participating for enhancing the overall effectiveness of the system.

Algorithm-1 improves data quality by handling missing values using age-group-based imputation and addressing class imbalance through AVAE-based augmentation. This results in a more descriptive and balanced dataset, which improves the learning capability of the model.

Algorithm-2 focuses on model-level improvements by incorporating Morlet wavelet activation for enhanced feature extraction, FFT overlap-add convolution for computational efficiency, and WMOV optimization for effective parameter tuning. These components collectively enable the model to capture complex nonlinear relationships and achieve high classification accuracy.

The integrated framework achieves an accuracy of 98.44%, surpassing traditional machine learning models and existing research approaches. Additionally the confusion matrix analysis indicates that the model has low false positive and false negative rates, which is essential for applications in medical diagnosis.

5.9.1 Strengths of the Proposed Framework

The proposed AVAE–MW-FFT-OAconv–WMOV framework possesses several significant strengths that distinguish it from conventional diabetes prediction models. First, the incorporation of the Adversarial Variational Autoencoder (AVAE) effectively addresses the class imbalance problem by generating realistic synthetic diabetic samples. Unlike traditional oversampling techniques that simply duplicate or interpolate existing observations, the AVAE learns the underlying data distribution and generates diverse samples that improve the classifier's ability to recognize minority class patterns. This results in improved model robustness and reduced prediction bias.

Second, the adoption of the Morlet Wavelet activation function enables the proposed deep learning model to capture complex nonlinear relationships among clinical variables more effectively than conventional activation functions. Since diabetes prediction involves intricate interactions among physiological parameters such as glucose concentration, body mass index, insulin level, age, and hereditary factors, the wavelet-based activation provides improved feature representation and enhances discrimination between diabetic and non-diabetic patients.

Another important strength of the proposed framework is the utilization of FFT Overlap-Add convolution for efficient feature extraction. By performing convolution operations in the frequency domain, the computational burden associated with conventional convolution operations is significantly reduced while preserving feature extraction capability. This contributes to faster model convergence and improved computational efficiency, particularly when processing larger datasets.

The integration of the Weighted Mean of Vectors Optimizer (WMOV) further strengthens the proposed framework by providing stable and efficient optimization of network parameters. The convergence analysis presented earlier demonstrates that WMOV consistently improves the fitness of the solution during successive optimization iterations while avoiding unstable oscillations. This stable optimization behaviour contributes directly to the improved classification accuracy and reduced prediction error achieved by the proposed framework.

The experimental evaluation further confirms that the proposed framework consistently achieves superior performance across multiple evaluation measures. Higher Accuracy, Precision, Recall, F1-score, Sensitivity, Specificity, and MCC, together with a lower Classification Error Rate, demonstrate that the proposed model provides balanced prediction capability without favouring either the diabetic or non-diabetic class. Such balanced performance is particularly important for medical diagnosis, where both false positive and false negative predictions may have significant clinical consequences.

Finally, the inclusion of SHAP-based interpretability significantly enhances the practical value of the proposed framework. The SHAP analysis identifies Glucose, BMI, Pregnancies, Diabetes Pedigree Function, and Age as the most influential predictors, which is consistent with established medical knowledge regarding Type-2 Diabetes Mellitus. This interpretability improves the transparency of the prediction process and increases clinician confidence in the generated predictions, thereby making the proposed framework more suitable for integration into intelligent healthcare decision-support systems.

5.9.2 Limitations of the Proposed Framework

Despite the promising results, the proposed framework has certain limitations. The experimental evaluation was performed primarily on publicly available diabetes datasets, and therefore the generalization capability should be further validated using larger multi-center clinical datasets collected from different geographical regions.

Although AVAE generates diverse synthetic samples, the quality of generated instances depends on the characteristics of the original training data.

Also, the proposed hybrid framework involves multiple computational components, including AVAE augmentation, FFT-based convolution, and WMOV optimization, which increase implementation complexity compared with conventional machine learning algorithms.

Finally, although SHAP analysis improves interpretability, the overall framework remains more computationally demanding than simpler statistical prediction models.

5.9.3 Practical Deployment Considerations

From a practical perspective, the proposed framework has considerable potential for deployment as a clinical decision-support system. The trained model can assist healthcare professionals by providing rapid diabetes risk assessment using routinely collected clinical parameters. The model can be integrated into hospital information systems, electronic health record platforms, or cloud-based healthcare applications to support early diagnosis and screening. However, successful real-world deployment requires additional validation using prospective clinical studies, continuous model monitoring, periodic retraining with updated patient data, and compliance with healthcare regulations related to data privacy and security. Furthermore, explainable AI techniques such as SHAP can improve clinician trust by providing transparent explanations for individual predictions.

Overall, the proposed AVAE–MW-FFT-OAconv–WMOV framework demonstrates high predictive performance while maintaining satisfactory convergence characteristics and balanced classification capability. Although further validation on large-scale clinical datasets is required before routine clinical deployment, the proposed framework represents a promising intelligent decision-support tool for the early prediction of Type-2 Diabetes Mellitus.

5.10 Summary

This chapter demonstrated the experimental evaluation and analysis of the suggested hybrid model for diabetes prediction. The performance of both Algorithm-1 and Algorithm-2 was analysed independently, succeeded by an evaluation of the integrated model.

The results demonstrated that:

- The proposed preprocessing and AVAE-based augmentation significantly improve data quality and classification performance.
- The MW-FFT-OAconv model with WMOV optimization achieves high accuracy and efficient feature learning.

- The integrated framework outperforms conventional machine learning models and existing research approaches.
- The model exhibits stable and consistent performance, as confirmed through cross-validation.

Overall, the suggested strategy achieves a classification accuracy of 98.44%, demonstrating its effectiveness and reliability for diabetes prediction.

This chapter establishes the superiority of the proposed framework and provides a strong foundation for its application in real-world medical diagnosis systems.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

This study presented a novel hybrid model for the prediction of Type 2 Diabetes Mellitus by integrating data-level enhancement and model-level optimization techniques. The proposed approach consists of two major components: Algorithm-1, which focuses on preprocessing and AVAE-based data augmentation, and Algorithm-2, which introduces the MW-FFT-OAconv deep learning model optimized using the WMOV algorithm.

At the data level, the study addressed critical constraints in medical databases, including misplaced values and unbalanced classes. The proposed age-group-based imputation strategy effectively handled invalid zero values in clinical attributes, preserving the natural distribution of the dataset. Furthermore, the use of Adversarial Variational Autoencoder (AVAE) enabled the creation of realistic fabricated samples, improving data diversity and classification performance. The effectiveness of Algorithm-1 was also validated on diabetes prediction dataset, demonstrating its generalization capability.

At the model level, the proposed MW-FFT-OAconv architecture introduced several key innovations. The incorporation of the Morlet wavelet activation function improved nonlinear feature extraction, enabling the framework to catch subtle patterns within medical datasets. The use of FFT overlap-add convolution enhanced computational efficiency while maintaining classification accuracy. Additionally, the Weighted Mean of Vectors (WMOV) optimization algorithm provided effective parameter tuning, leading to improved convergence and performance.

The integration of these components produced highly accurate and robust prediction model. The suggested hybrid model gained a classification accuracy of 98.44%,

significantly outperforming conventional machine learning models and existing research approaches. The confusion matrix analysis revealed lower rates of false positive and false negative, highlighting the model's reliability in identifying diabetic and non-diabetic samples. Furthermore, cross-validation results confirmed the stability and generalization capability of the framework.

In general, the research successfully illustrates that the combination of modern data augmentation strategies and optimized deep learning frameworks can significantly enhance the performance of medical prediction systems. The suggested model provides a reliable and efficient solution for timely detection of diabetes and has strong potential for real-world clinical applications.

6.2 Research Contributions

The key contributions of this research work are outlined as follows:

- A novel preprocessing strategy based on age-group mean imputation for handling invalid and missing values in medical datasets.
- Development of an AVAE-based data augmentation technique to address class imbalance and improve data diversity.
- Design of a Morlet Wavelet assisted deep learning model (MW-FFT-OAconv) for enhanced feature extraction.
- Integration of FFT overlap-add convolution to improve computational efficiency.
- Proposal of a WMOV optimization algorithm for effective parameter tuning.
- Development of a hybrid framework combining data-level and model-level improvements for diabetes prediction.
- Comprehensive evaluation using performance metrics, confusion matrix analysis, comparative study, and cross-validation.

6.3 Limitations of the Study

Although the results are promising, the proposed work has certain limitations:

- The evaluation is primarily relied on benchmark datasets, and further validation using large-scale real-world clinical data is required.
- The proposed deep learning model introduces additional computational complexity compared to traditional machine learning approaches.
- The WMOV optimization algorithm, while effective, may require further tuning and adaptation for different problem domains.
- The current model focuses on binary classification and does not consider multi-class or severity-level prediction of diabetes.

6.4 Future Work

The proposed research can be extended in several directions to further enhance its applicability and performance:

- **Validation on real-world clinical datasets:** Future investigations can concentrate on testing the proposed framework on large-scale hospital data to evaluate its practical applicability.
- **Model optimization and lightweight implementation:** Developing a computationally efficient version of the model suitable for real-time deployment and edge devices.
- **Integration with clinical decision support systems:** The proposed framework can be incorporated within healthcare systems to assist medical professionals in diagnosis and treatment planning.
- **Extension to multi-class classification:** Expanding the model to predict different stages or severity levels of diabetes.

- **Incorporation of explainable AI (XAI):** Adding interpretability mechanisms to provide inderstanding into model predictions, increasing confidence into healthcare experts.

- **Hybrid optimization strategies:** Combining WMOV with other metaheuristic or gradient-based optimization techniques for further performance improvement.

References

[1] N. A. ElSayed *et al.*, "2. Classification and diagnosis of diabetes: standards of care in diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement_1, pp. S19–S40, 2023, [Online]. Available: https://diabetesjournals.org/care/article-abstract/46/Supplement_1/S158/148038

[2] D. J. Magliano, E. J. Boyko, and I. D. Atlas, "COVID-19 and diabetes," in *IDF DIABETES ATLAS [Internet]. 10th edition*, International Diabetes Federation, 2021. Accessed: Feb. 25, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK581933/

[3] K. L. Ong *et al.*, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet*, vol. 402, no. 10397, pp. 203–234, 2023, Accessed: Feb. 25, 2026. [Online]. Available: https://www.thelancet.com/article/S0140-6736(23)01301-6/fulltext

[4] "Volume 19, Issue 10 | Current Diabetes Reports | Springer Nature Link." Accessed: Feb. 25, 2026. [Online]. Available: https://link.springer.com/journal/11892/volumes-and-issues/19-10

[5] M. J. Fowler, "Microvascular and Macrovascular Complications of Diabetes".

[6] W. H. Organization, "World Health Organization Global Report on Diabetes. Geneva," *World Health Organ.*, 2016.

[7] B. B. Duncan, D. J. Magliano, and E. J. Boyko, "IDF diabetes atlas 11th edition 2025: global prevalence and projections for 2050," *Nephrology Dialysis Transplantation*. Oxford University Press, p. gfaf177, 2025. Accessed: Feb. 26, 2026. [Online]. Available: https://academic.oup.com/ndt/advance-article-pdf/doi/10.1093/ndt/gfaf177/64148213/gfaf177.pdf

[8] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.

[9] "Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition by Aurélien Géron | Physical and Engineering Sciences in Medicine | Springer Nature Link." Accessed: Feb. 26, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s13246-020-00913-z

[10] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019, Accessed: Feb. 26, 2026. [Online]. Available: https://www.nature.com/articles/s41591-018-0300-7

[11] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[12] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019, Accessed: Feb. 25, 2026. [Online]. Available: https://www.nature.com/articles/s41591-018-0316-z

[13] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018, Accessed: Feb. 26, 2026. [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/2675024

[14] H. He and Y. Ma, "Imbalanced learning: foundations, algorithms, and applications," 2013.

[15] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[16] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.

[17] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2018, Accessed: Feb. 26, 2026. [Online]. Available: https://academic.oup.com/bib/article-abstract/19/6/1236/3800524

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. in Adaptive computation and machine learning. Cambridge, Mass: The MIT press, 2016.

References

[19] A. Vellido, "Societal issues concerning the application of artificial intelligence in medicine," *Kidney Dis.*, vol. 5, no. 1, pp. 11–17, 2019, Accessed: Feb. 26, 2026. [Online]. Available: https://karger.com/kdd/article-abstract/5/1/11/186070

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015, Accessed: Feb. 26, 2026. [Online]. Available: https://www.nature.com/articles/nature14539

[21] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 14, p. 7346, 2021, Accessed: Mar. 01, 2026. [Online]. Available: https://www.mdpi.com/1660-4601/18/14/7346

[22] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques," *IEEE Access*, vol. 7, pp. 1365–1375, 2018, Accessed: Mar. 01, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8585030/

[23] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques," *Front. Comput. Sci.*, vol. 4, p. 835242, 2022, Accessed: Mar. 01, 2026. [Online]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.835242/full

[24] S. Wang *et al.*, "Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: a cross-sectional study," *BMJ Open*, vol. 13, no. 8, p. e069018, 2023, Accessed: Mar. 01, 2026. [Online]. Available: https://bmjopen.bmj.com/content/13/8/e069018.abstract

[25] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, Accessed: Mar. 02, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037016300733

[26] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018, Accessed: Mar. 02, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918308548

[27] T.-C. T. Chen, H.-C. Wu, and M.-C. Chiu, "A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare," *Appl. Soft Comput.*, vol. 152, p. 111183, 2024, Accessed: Mar. 02, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623012012

[28] Z. Zhang, K. A. Ahmed, M. R. Hasan, T. Gedeon, and M. Z. Hossain, "A Deep Learning Approach to Diabetes Diagnosis," Aug. 2024. doi: 10.1007/978-981-97-5937-8_8.

[29] E. Manzini *et al.*, "A deep attention-based encoder for the prediction of type 2 diabetes longitudinal outcomes from routinely collected health care data," *Expert Syst. Appl.*, vol. 274, p. 126876, May 2025, doi: 10.1016/j.eswa.2025.126876.

[30] A. Alqushaibi *et al.*, "Type 2 diabetes risk prediction using deep convolutional neural network based-bayesian optimization," *Comput. Mater. Contin.*, vol. 75, no. 2, p. 3223, 2023.

[31] J. Zhao, H. Gao, C. Yang, T. An, Z. Kuang, and L. Shi, "Attention-oriented CNN method for type 2 diabetes prediction," *Appl. Sci.*, vol. 14, no. 10, p. 3989, 2024, Accessed: Mar. 02, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/14/10/3989

[32] S. Zanelli, M. A. El Yacoubi, M. Hallab, and M. Ammi, "Type 2 diabetes detection with light CNN from single raw PPG wave," *IEEe Access*, vol. 11, pp. 57652–57665, 2023, Accessed: Mar. 02, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10121422/

[33] R. Farnoosh, K. Abnoosian, R. A. Isewid, and D. Javaheri, "DiabetesXpertNet: An innovative attention-based CNN for accurate type 2 diabetes prediction," *PloS One*, vol. 20, no. 9, p. e0330454, 2025, Accessed: Mar. 02, 2026. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0330454

[34] Abdussamad, H. Daud, R. Sokkalingam, M. Zubair, I. K. Khan, and Z. Mahmood, "A deep learning framework with hybrid stacked sparse autoencoder for type 2 diabetes prediction," *Sci. Rep.*, vol. 15, no. 1, p. 36678, 2025, Accessed: Mar. 02, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-20534-4

[35] R. F. Albadri, S. M. Awad, A. S. Hameed, T. H. Mandeel, and R. A. Jabbar, "A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm.," *Math. Model. Eng. Probl.*, vol. 11, no. 8, 2024.

[36] J.-E. Ding *et al.*, "Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records," *Sci. Rep.*, vol. 14, no. 1, p. 20774, 2024, Accessed: Mar. 02, 2026. [Online]. Available: https://www.nature.com/articles/s41598-024-71020-2

[37] K. H. Abushahla and M. A. Pala, "Optimizing diabetes prediction: Addressing data imbalance with machine learning algorithms," *ADBA Comput. Sci.*, vol. 1, no. 1, pp. 26–35, 2024, Accessed: Mar. 03, 2026. [Online]. Available: https://journals.adbascientific.com/acs/article/view/12

[38] B. Toleva, I. Atanasov, I. Ivanov, and V. Hooper, "An effective methodology for diabetes prediction in the case of class imbalance," *Bioengineering*, vol. 12, no. 1, p. 35, 2025, Accessed: Mar. 03, 2026. [Online]. Available: https://www.mdpi.com/2306-5354/12/1/35

[39] M. Talebi Moghaddam *et al.*, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 220, Sep. 2024, doi: 10.1186/s12874-024-02341-z.

[40] H. Aulia and A. Wibowo, "Explaining Diabetes Prediction Under Oversampling Strategies: A SHAP Based Comparative Study of SMOTE and ADASYN," in *2025 8th International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, 2025, pp. 1–6. Accessed: Mar. 03, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11329981/

[41] L. Al-Dabbas and A. A. Abu-Shareha, "Early detection of female type-2 diabetes using machine learning and oversampling techniques," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1237–1245, 2024, Accessed: Mar. 03, 2026. [Online]. Available: http://www.bright-journal.org/Journal/index.php/JADS/article/view/298

[42] W. Seo, N. Kim, S.-W. Park, S.-M. Jin, and S.-M. Park, "Generative adversarial network-based data augmentation for improving hypoglycemia prediction: A proof-of-concept study," *Biomed. Signal Process. Control*, vol. 92, p. 106077, 2024, Accessed: Mar. 03, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809424001356

[43] K. R. Singh, S. Dash, H. Liu, and Z. Wang, "Enhanced diabetes prediction using pre-trained CNNs, LSTM, and conditional GAN on transformed numerical data," *Sci. Rep.*, 2026, Accessed: Mar. 03, 2026. [Online]. Available: https://www.nature.com/articles/s41598-026-38942-5

[44] J. Vehi, O. Mujahid, A. Beneyto, and I. Contreras, "Generative artificial intelligence in diabetes healthcare," *Iscience*, 2025, Accessed: Mar. 03, 2026. [Online]. Available: https://www.cell.com/iscience/fulltext/S2589-0042(25)01312-4

[45] D. Papadopoulos and V. D. Karalis, "Variational autoencoders for data augmentation in clinical studies," *Appl. Sci.*, vol. 13, no. 15, p. 8793, 2023, Accessed: Mar. 03, 2026. [Online]. Available: https://www.mdpi.com/2076-3417/13/15/8793

[46] M. S. Al-zahrani, F. W. Alsaade, and F. E. Alsaadi, "Variational Autoencoder-Guided Data Augmentation for Imbalanced Medical Diagnostics," *Results Eng.*, p. 107011, 2025, Accessed: Mar. 03, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590123025030671

[47] A. Ali Linkon *et al.*, "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," *IEEE Access*, vol. 12, pp. 165425–165440, 2024, doi: 10.1109/ACCESS.2024.3488743.

[48] D. Wu, Y. Mei, Z. Sun, H. Duan, and N. Deng, "Multi-feature map integrated attention model for early prediction of type 2 diabetes using irregular health examination records," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1656–1667, 2023, Accessed: Mar. 06, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10366808/

[49] A. Shankar, S. Chang, X. Wang, and Y. Zhao, "Wavelet based Machine Learning Approaches towards Precision Medicine in Diabetes Mellitus.," in *BIOSIGNALS*, 2022, pp. 290–297. Accessed: Mar. 06, 2026. [Online]. Available: https://www.scitepress.org/PublishedPapers/2022/109931/109931.pdf

[50] M. K. Isfahani, M. Zekri, H. R. Marateb, and E. Faghihimani, "A hybrid dynamic wavelet-based modeling method for blood glucose concentration prediction in type 1 diabetes," *J. Med. Signals Sens.*, vol. 10, no. 3, pp. 174–184, 2020, Accessed: Mar. 06, 2026. [Online]. Available:

https://journals.lww.com/jmss/fulltext/2020/10030/a_hybrid_dynamic_wavelet_based_modeling_method_for.4.aspx

[51] E. F. Goh, Z. Chen, and W. X. Lim, “Frequency Domain Convolutional Neural Network: Accelerated CNN for Large Diabetic Retinopathy Image Classification,” Jun. 24, 2021, *arXiv*: arXiv:2106.12736. doi: 10.48550/arXiv.2106.12736.

[52] J. E. Lee, H.-J. Jeon, O.-J. Lee, and H. G. Lim, “Diagnosis of diabetes mellitus using high frequency ultrasound and convolutional neural network,” *Ultrasonics*, vol. 136, p. 107167, 2024, Accessed: Mar. 06, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041624X23002433

[53] H. A. Aliyu, I. O. Muritala, H. Bello-Salau, S. Mohammed, A. J. Onumanyi, and O.-O. Ajayi, “Optimizing machine learning algorithms for diabetes data: A metaheuristic approach to balancing and tuning classifiers parameters,” *Frankl. Open*, vol. 8, p. 100153, 2024, Accessed: Mar. 06, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2773186324000835

[54] R. A. Osman, “Deep Learning-Based Early Diabetes Risk Prediction Using Survey Data: A Hyperparameter Optimization Approach”, Accessed: Mar. 06, 2026. [Online]. Available: https://www.academia.edu/download/123784655/SR221111091934.pdf

[55] “A deep learning model for identification of diabetes type 2 based on nucleotide signals | Neural Computing and Applications | Springer Nature Link.” Accessed: Mar. 06, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s00521-022-07121-8

[56] D. Dua and C. Graff, “UCI machine learning repository, 2017. University of California, Irvine, School of Information and Computer Sciences.” 2017.

[57] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in *Proceedings of the annual symposium on computer application in medical care*, 1988, p. 261. Accessed: Mar. 17, 2026. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2245318/

[58] “Exploratory Data Analysis | Springer Nature Link.” Accessed: Mar. 18, 2026. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-20719-8_2

[59] “Python Machine Learning, Second Edition.” Accessed: Mar. 18, 2026. [Online]. Available: https://reference-global.com/book/9781787126022

[60] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.

[61] I. J. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: Mar. 20, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[62] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[63] “Diabetes prediction dataset.” Accessed: Mar. 30, 2026. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[64] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, “Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset,” *Comput. Methods Programs Biomed. Update*, vol. 4, p. 100118, 2023, Accessed: Apr. 02, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990023000265

[65] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, “An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI,” *Sensors*, vol. 22, no. 19, p. 7268, 2022, Accessed: Apr. 02, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7268

[66] P. Sampath *et al.*, “Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique,” *Sci. Rep.*, vol. 14, no. 1, p. 28984, 2024, Accessed: Apr. 02, 2026. [Online]. Available: https://www.nature.com/articles/s41598-024-78519-8

[67] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, “An ensemble learning approach for diabetes prediction using boosting techniques,” *Front. Genet.*, vol. 14, p. 1252159, 2023,

Accessed: Apr. 02, 2026. [Online]. Available: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1252159/full

[68] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Comput. Methods Programs Biomed.*, vol. 202, p. 105968, 2021, Accessed: Apr. 02, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721000432

[69] B. P. Hitesh, B. Keyur, and J. Mahasweta, "Data Augmentation Approach for Type-2 Diabetes Prediction and Classification," *Indian J. Sci. Technol.*, vol. 18, no. 14, pp. 1096–1104, 2025, Accessed: Mar. 18, 2026. [Online]. Available: https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2025/Issue-14/IJST-2024-3904.pdf

[70] H. B. Patel and K. Brahmbhatt, "Prediction of type 2 diabetes based on feature augmentation and Morlet wavelet assisted deep learning network with FFT overlap and add convolution," *Int. J. Biomed. Eng. Technol.*, vol. 47, no. 3, pp. 259–286, Jan. 2025, doi: 10.1504/IJBET.2025.144949.

List of Publications

- Hitesh B. Patel and Keyur N. Brahmbhatt, “Prediction of type 2 diabetes based on feature augmentation and Morlet wavelet assisted deep learning network with FFT overlap and add convolution”, International Journal of Biomedical Engineering and Technology, Volume-47, Issue-3, pp. 259-286, Jan 2025

- Hitesh B. Patel, Keyur N. Brahmbhatt and Mahasweta Joshi, “Data Augmentation Approach for Type-2 Diabetes Prediction and Classification”, Indian Journal of Science and Technology, Volume-18, Issue-14, pp. 1096-1104, April 2025

- Hitesh B. Patel and Keyur N. Brahmbhatt, “A comprehensive review of computational intelligence approaches in type 2 diabetes prediction and classification”, International Journal of Medical Engineering and Informatics, Inderscience Publication (Accepted; Yet to be published)