People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
A. Mira University of Béjaïa
Faculty of Exact Sciences
Department of Computer Science

Final Study Thesis

In view of obtaining the Master of Research in Computer Science

Specialization : Advanced Information Systems

Topic

Diabetes Mellitus Prediction

Prepared by

M^{iss} Melissa CHAKIR,
M^{iss} Asma ABDELGHAFOUR.

In front of the jury composed of

President :	Mr. ALLEM Khaled	M.C.B	University of Béjaïa
Examiner :	Mrs. KHOULALLEN Nadjet	M.C.B	University of Béjaïa
Supervisor :	Mrs. EL BOUHISSI BRAHAMI Houda	M.C.A	University of Béjaïa

Promotion 2023 - 2024

Acknowledgements

As we conclude this work, we would like to express our deep gratitude and sincere appreciation.

Firstly, we thank Almighty God for providing us with the strength, determination, and perseverance needed to complete this endeavor..

Our heartfelt thanks go to **our parents**, whose unwavering support has been a constant source of inspiration and motivation throughout this journey.

We are deeply grateful to our advisor, **Mrs. Houda El Bouhissi**, for her continuous support, insightful guidance, and trust in us.

Her invaluable collaboration and encouragement throughout the writing of this thesis have been instrumental.

We extend our appreciation to the members of the jury for taking the time to evaluate our work.

Finally, we express our heartfelt gratitude to all the faculty and staff of the Computer Science Department at the University of Béjaïa for their invaluable support and contributions to our academic journey and professional development.

Dedication

In grateful recognition of the Almighty Allah's guidance and blessings
and with heartfelt appreciation for the unwavering support of my beloved family,

I dedicate this thesis.

To the memory of my eldest sister **H**amama, and the memory of my father, whose presence is deeply missed.

To my dear mother , whose endless love and sacrifices have been the foundation of my journey,
I am forever indebted to you.
Your strength and resilience inspire me every day.

To my sisters, brothers, and cherished ones,
your encouragement fuels my ambition and drives me to strive for excellence.

Last but not least I wanna thank **ME**,
for believing in me,
i wanna thank me for doing all this hard work,
and for never quitting.

Melissa.

Dedication

In grateful recognition of the Almighty Allah's guidance and blessings
and with heartfelt appreciation for the unwavering support of my mother,

I dedicate this thesis.

To my beloved mother,
Your unwavering support, endless encouragement, and boundless love have been my guiding light through every step of this journey. This thesis is a testament to your strength, sacrifices, and belief in me. Thank you for always being my rock and my inspiration.
With all my love and gratitude,

To our dear mentor Nour El Houda Elbouhissi
Thank you for your mentorship and dedication, which have greatly enriched my academic journey.
Your guidance, support, and encouragement, has been invaluable.

Last but not least I wanna thank **ME**,
for believing in me,
i wanna thank me for doing all this hard work,
I want to thank me for having no days off
I want to thank me for never quitting.
I want to thank me for just being me at all times.

Asma.

Contents

Liste des figures vii

List of tables viii

List of algorithms ix

List of abbreviations x

Abstract xi

Résumé xii

1 General Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Motivation . . . 2
- 1.3 Work methodology . . . 2
- 1.4 Thesis organisation . . . 3

2 Background and Related Concepts 4
- 2.1 Introduction . . . 4
- 2.2 Diabetes . . . 4
  - 2.2.1 Diabetes definition . . . 4
  - 2.2.2 Diabetes complications . . . 5
- 2.3 Diagnosis of diabetes . . . 6
  - 2.3.1 Diabetes Screening . . . 6
  - 2.3.2 Signs of Diabetes . . . 7
  - 2.3.3 Blood Glucose Analysis . . . 7
- 2.4 Classification of Diabetes Types . . . 7
  - 2.4.1 Type 1 Diabetes . . . 7
  - 2.4.2 Type 2 Diabetes . . . 7
- 2.5 Deep Learning . . . 8
  - 2.5.1 Long Short-Term Memory (LSTM) . . . 9
  - 2.5.2 Gated Recurrent Unit (GRU) . . . 9
- 2.6 Swarm Intelligence . . . 9

2.6.1 Particle Swarm Optimization (PSO) . . . 10
2.6.2 Ant Colony Optimization (ACO) . . . 10
2.6.3 The Grey Wolf Optimizer (GWO) . . . 11
2.6.4 Enhanced Harris Hawk Optimization (EHS) . . . 11
2.7 Conclusion . . . 11

3 The State of the Art 12
3.1 Introduction . . . 12
3.2 Related Works . . . 12
3.3 Analysis and Comparison . . . 19
3.3.1 Comparison criteria . . . 19
3.3.2 Comparative table . . . 19
3.3.3 Discussion . . . 22
3.4 Conclusion . . . 23

4 Proposed Approach for Diabetes Prediction 24
4.1 Introduction . . . 24
4.2 Contribution . . . 24
4.3 Proposed approach . . . 25
4.3.1 Data Collection . . . 26
4.3.2 Data Preprocessing . . . 26
4.3.3 Feature Selection . . . 27
4.3.4 Prediction Process . . . 29
4.4 Conclusion . . . 33

5 Experimental Setup and Evaluation 34
5.1 Introduction . . . 34
5.2 Dataset Description . . . 34
5.2.1 Definition . . . 34
5.2.2 Dataset Overview: . . . 35
5.2.3 Dataset Content: . . . 35
5.2.4 Data Distribution Plots . . . 36
5.2.5 Data Cleaning . . . 38
5.3 Development Environment . . . 42
5.3.1 Hardware Environment . . . 42
5.3.2 Software Environment . . . 42
5.4 Description of the Tool . . . 43
5.4.1 Homepage . . . 43
5.4.2 Prediction Interface . . . 44
5.4.3 Diagnosis Results . . . 46
5.4.4 Algorithm Results . . . 47
5.5 Evaluation . . . 47
5.5.1 Evaluation Metrics . . . 47
5.5.2 Evaluation of the Proposed Model . . . 49
5.5.3 Comparison of different metrics across algorithms . . . 57
5.6 Conclusion . . . 62

6 General conclusion and perspectives **63**
6.1 Introduction . . . 63
6.2 Methodology . . . 63
6.2.1 General introduction and motivation . . . 63
6.2.2 Background and Related Concepts . . . 64
6.2.3 The State of the Art . . . 64
6.2.4 Proposed Approach for Diabetes Prediction . . . 64
6.2.5 Experimental Setup and Evaluation . . . 64
6.2.6 General conclusion and perspectives . . . 64
6.3 Limits . . . 64
6.4 Perspectives . . . 65
6.5 Final Conclusion . . . 65

Bibliographie **66**

List of Figures

2.1 Effect of Insulin on Glucose Uptake[1]. 5
2.2 Diabetes screening [2]. 6
2.3 Scope of Deep Learning[3]. 8
2.4 Ant Colony Optimisation[4]. 10

4.1 Proposed approach 25

5.1 Overview of the dataset 35
5.2 Smoking history Distributions: 36
5.3 Diabetes distribution for each gender 37
5.4 Data Balance in the Dataset 38
5.5 Overview of the dataset. 39
5.6 The data before encoding 40
5.7 The data after encoding/ 40
5.8 removing duplicates 41
5.9 Correlation Matrix of the Dataset 41
5.10 Homepage of the Tool 44
5.11 Diagnosis Interface 1 44
5.12 Diagnosis Interface2 45
5.13 Diagnosis Interface 3 45
5.14 Diagnosis Interface 4 46
5.15 Diagnosis result Interface 47
5.16 LSTM roc curve 50
5.17 ACO-LSTM roc curve 52
5.18 PSO-LSTM roc curve. 53
5.19 GWO-LSTM roc curve. 54
5.20 EHHO-LSTM roc curve. 55
5.21 ACO-GRU roc curve. 56
5.22 Accuracy comparison. 57
5.23 RMSE comparison. 58
5.24 MAE comparaison. 59
5.25 ROC comparison. 60
5.26 R2 Score comparison 61

List of Tables

3.1 Comparative Analysis of Approaches 21

5.1 Algorithm Results 47
5.2 Performance Metrics for LSTM Model 49
5.3 Performance Metrics for ACO-LSTM Model 51
5.4 Performance Metrics for PSO-LSTM Model 52
5.5 Performance Metrics for GWO-LSTM Model 54
5.6 Performance Metrics for EHHO-LSTM Model 55
5.7 Performance Metrics for ACO-GRU Model 56

List of Algorithms

List of abbreviations

ACC	Accuracy.
ACO	Ant Colony Optimization.
ANN	Artificial Neural Network.
BMI	Body Mass Index.
CNN	Convolutional Neural Networks.
DL	Deep learning.
DT	Decision Tree.
EHS	Enhanced Harris Hawk Optimization.
GRU	Gated Recurrent Unit
GWO	The Grey Wolf Optimizer.
IA	Artificial Intelligence .
KNN	K-nearest neighbors.
LSTM	Long Short-Term Memory.
ML	Machine Learning.
PIMA	Pima Indians Diabetes Database.
PSO	Particle Swarm Intelligence.
RF	Random Forest.
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks.
SVM	Support Vector Machines.
T1D	Type 1 Diabetes.
T2D	Type 2 Diabetes
WHO	World Health Organization

Abstract

Diabetes is a long-term condition resulting from malfunction of the pancreas, causing high blood sugar levels that can eventually damage multiple body systems like the cardiovascular, nervous, renal, ocular, integumentary, gastrointestinal, and endocrine systems... etc. Timely and precise detection is crucial to prevent complications.

This study aims to enhance the identification of diabetes through the utilization of deep learning and swarm intelligence strategies, relying on medical information and past records.

Several deep learning and swarm methods can aid in achieving our objective of avoiding tardy diagnosis of the illness. To achieve this objective, we employed various feature selection algorithms such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Enhanced Harris Hawk Optimization (EHHO), and Grey Wolf Optimizer (GWO) to determine the optimal parameters for training deep learning models, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, for predicting diabetes.

The experimental results demonstrate that the ACO-LSTM model provides the highest accuracy rate of 97.1%. This highlights the potential of integrating deep learning and swarm intelligence to create a robust and reliable system for efficient diabetes diagnosis and treatment.

Keywords: Diabetes, Prediction, Deep learning, Swarm intelligence, LSTM, GRU, PSO, ACO, GWO, EHHO.

Résumé

Le diabète est une condition chronique causée par un dysfonctionnement du pancréas, entraînant des niveaux élevés de sucre dans le sang qui peuvent finalement endommager plusieurs systèmes corporels, tels que les systèmes cardiovasculaire, nerveux, rénal, oculaire, tégumentaire, gastro-intestinal et endocrinien, etc. Une détection rapide et précise est cruciale pour prévenir les complications. Cette étude vise à améliorer l'identification du diabète grâce à l'utilisation de stratégies d'apprentissage profond et d'intelligence collective, en s'appuyant sur des informations médicales et des dossiers historiques.

Plusieurs méthodes d'apprentissage profond et d'intelligence collective peuvent aider à atteindre notre objectif d'éviter un diagnostic tardif de la maladie. Pour atteindre cet objectif, nous avons employé divers algorithmes de sélection de caractéristiques tels que l'Optimisation par Colonie de Fourmis (ACO), l'Optimisation par Essaim Particulaire (PSO), l'Optimisation Avancée du Faucon Harris (EHHO) et l'Optimiseur de Loups Gris (GWO) pour déterminer les paramètres optimaux pour l'entraînement des modèles d'apprentissage profond, notamment les réseaux de Mémoire à Long Court Terme (LSTM) et les Unités Récurrentes à Portes (GRU), pour prédire le diabète.

Les résultats expérimentaux démontrent que le modèle ACO-LSTM offre le taux de précision le plus élevé, soit 97.1%. Cela met en évidence le potentiel de l'intégration de l'apprentissage profond et de l'intelligence collective pour créer un système robuste et fiable pour un diagnostic et un traitement efficaces du diabète.

Mots-clés: Diabète, Prédiction, Apprentissage profond, Intelligence collective, LSTM, GRU, PSO, ACO, GWO, EHHO.

Chapter 1

General Introduction

1.1 Introduction

Diabetes mellitus belongs to the top 10 chronic diseases of the developed world at present and gravely affects overall health on a global scale. It is a metabolic disease in which there is low insulin production by the pancreas or malfunction of the body cells in the use of the produced insulin. Insulin regulates blood sugar levels, but its deficiency or ineffective use leads to hyperglycemia; with time, it causes much organ damage.

World Health Organization recorded that the number of people with diabetes around the world had reached over 422 million, against 108 million. It causes an estimated 4.2 million deaths annually, most of which are recorded in low and middle-income countries where resources for health care are generally scanty.

The present human lifestyle, including inadequate dietary habits, unhygienic living conditions, stress, and a lack of physical activity, has been considerably blamed for the increasing incidence of diabetes. This fact is particularly applicable to countries like Algeria, which, within the last decade, have seen an explosive rise in the rate of diabetes due to such factors as obesity, sedentary lifestyles, and genetic predisposition.

Diabetic screening is so crucial for early detection, which in turn helps in preventing or reducing complications. Early detection of diabetes makes it possible for timely interventions, which subsequently leads to significantly improved patient outcomes.

Artificial innovation of late includes sophisticated techniques that pertain to diabetes prediction. AI techniques that involve deep learning and swarm intelligence can be applied to analyze large volumes of medical data and help in identifying patterns that may improve the accuracy of predictions.

This study aims to focus on using AI techniques to improve diabetes predictions. It, therefore, uses deep learning algorithms and swarm intelligence to develop more accurate predictive models. Such predictive models can support interventions and probably the early detection of diseases for improved health outcomes for those at risk of developing diabetes.

1.2 Motivation

Diabetes mellitus is a chronic condition that poses significant health risks and economic burdens globally. The prevalence of diabetes has been increasing steadily, making it one of the most pressing public health concerns of our time. Moreover, diabetes significantly reduces the quality of life and increases mortality rates.

One of the critical challenges in managing diabetes is early detection. Many individuals with diabetes remain undiagnosed until they develop severe complications. Early diagnosis and timely intervention are crucial in preventing the progression of the disease and mitigating its associated risks. Traditional diagnostic methods often require invasive procedures and frequent medical visits, which can be a barrier for many individuals, particularly those in remote or underserved areas.

The motivation behind our study is to address the need for more accessible and non-invasive methods for predicting diabetes. By developing a predictive model that utilizes easily obtainable health metrics, such as age, body mass index (BMI), glucose levels, blood pressure, and family history, we aim to facilitate early detection of diabetes. This predictive model can be integrated into a user-friendly application, allowing individuals to assess their risk of developing diabetes conveniently and efficiently.

Such an application would empower users to take proactive measures in managing their health. By providing personalized risk assessments, the application can encourage users to adopt healthier lifestyles, seek medical advice, and undergo further testing if necessary. This approach not only helps in early detection but also in prevention, as individuals can make informed decisions about their health before the disease progresses.

Through this study, we aim to demonstrate the potential of predictive modeling in improving public health outcomes and reducing the burden of diabetes on individuals and healthcare systems.

1.3 Work methodology

Our goal is to contribute to the global effort in combating diabetes by leveraging technology to enhance early diagnosis and promote preventive healthcare.The methodology employed in this endeavor is structured into the following phases:

1. **Literature Review**: Extensive literature review was conducted to understand the multi-faceted nature of diabetes, encompassing its etiology, risk factors, symptoms, and existing predictive models. Additionally, efforts were made to identify pertinent datasets for subsequent analysis.

2. **Analysis of Literature:** The findings from the literature review were systematically analyzed to distill key insights and determine the primary areas of focus for the research. This phase involved synthesizing information from diverse sources to inform the subsequent stages of the study.

3. **Problem Identification:** A critical examination of the challenges surrounding diabetes diagnosis and management was undertaken. The emphasis was on identifying gaps and limitations in existing approaches, with a view to developing innovative solutions to enhance early detection and prevention efforts.

4. **Solution Development:** Various methodologies for predicting diabetes were explored, with a particular emphasis on deep learning techniques and optimization algorithms. Through rigorous experimentation and comparative analysis, the most promising approach was selected for further development.

5. **Implementation:** The selected methodology was translated into a functional predictive

model through systematic implementation and coding. Careful attention was paid to ensuring the scalability, efficiency, and usability of the developed solution.

6.**Evaluation:** The final phase of the methodology involved comprehensive testing and evaluation of the developed predictive model. Performance metrics such as accuracy, sensitivity, specificity, and predictive value were assessed to gauge the effectiveness of the model in accurately predicting diabetes onset and guiding preventive interventions.

1.4 Thesis organisation

The subsequent chapters of this thesis are organized as follows:

Chapter 2: Background and Related Concepts This chapter provides an overview of key concepts and topics relevant to our study, including an in-depth discussion of diabetes and its various types, and foundational knowledge on deep learning and swarm intelligence.

Chapter 3: The state of art This chapter presents a comprehensive review of existing research on diabetes prediction. We discuss various approaches and methodologies employed in previous studies, highlighting their outcomes and contributions to the field. This chapter serves as a state-of-the-art overview that contextualizes our research within the broader scientific discourse.

Chapter 4: Proposed Approach for Diabetes Prediction In this chapter, we introduce our proposed methodology for predicting diabetes. We detail the techniques and algorithms used, followed by an extensive evaluation of the method's effectiveness. This section encapsulates our original contributions to the domain of diabetes prediction.

Chapter 5: Experimental Setup and Evaluation This chapter describes the experimental phase and the evaluation process of our proposed prediction model. It includes a thorough description of the dataset, details about the hardware and software environment used, a step-by-step explanation of the implementation phases, and shows the application that we developed for predicting Diabetes. This chapter provides a clear and detailed account of how our approach was tested and validated.

Chapter 6: Conclusion and perspectives The concluding chapter presents our final reflections on the thesis. We summarize the key findings, discuss the implications of our research, and offer perspectives on potential directions for future work. This chapter aims to encapsulate the significance of our study and propose avenues for further investigation.

Chapter 2

Background and Related Concepts

2.1 Introduction

The prediction of diabetes presents a captivating example of the utility of technology and data science in the service of better health in today's world that is characterized by personalized medicine and a data driven revolution. In this regard, ongoing efforts towards diabetes prediction play a critical role in the war against this dangerous disease.
This chapter discusses the basic concepts necessary to be understood for predicting diabetes using newly developed computational methods. The chapter defines diabetes and its complications, how to diagnose it, and the classification of its types. Additionally, other vital things that were introduced are deep learning and swarm intelligence, two cutting-edge methodologies being employed in our predictive model. The overall aim of this overview is to equip the readers with a strong base in the core underpinning topics from which we will study more advanced techniques.

2.2 Diabetes

2.2.1 Diabetes definition

Diabetes is a chronic design of disease that occurs as a result of high sugar levels in the blood. This happens because the body does not produce enough insulin hormone or the body cells do not respond to the insulin appropriately. According to WHO: "Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to severe damage to the heart, blood vessels, eyes, kidneys, and nerves."[5].
The IDF projections estimate that in 2045, 1 out of every eight adults an equivalent of about 783 million people will be living with diabetes, up by 46%. Over 90% of individuals with diabetes have type 2 diabetes, and this finding is associated with socioeconomic, demographic, environmental, and possibly genetic factors. Major driving forces behind the dramatic rise of type 2 diabetes include urbanization, an aging population, decreasing levels of physical activity, and increasing overweight and obesity prevalence. However, it is possible to avert Diabetes to a greater extent by allowing risk factors for type 2 diabetes to be addressed and by ensuring that all diabetes is early diagnosed and adequately treated. The chronic complications of people with diabetes can be avoided or delayed [6].
This chronic hyperglycemia is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. In this respect, glucose needs

to be uptake in the body cells for energy and metabolic purposes.[7] The uptake of glucose by different cell types in the body is essential for maintaining energy levels and metabolic functions. For example, liver cells store glucose as glycogen and manage its metabolism. Muscle cells also convert glucose into glycogen, which acts as a critical energy source during physical activity[8]. Insulin mediates a centered role in glucose regulation depicted from its role in homeostasis towards peripheral tissue. For a graphic view of how glucose uptake should be,figure2.1.

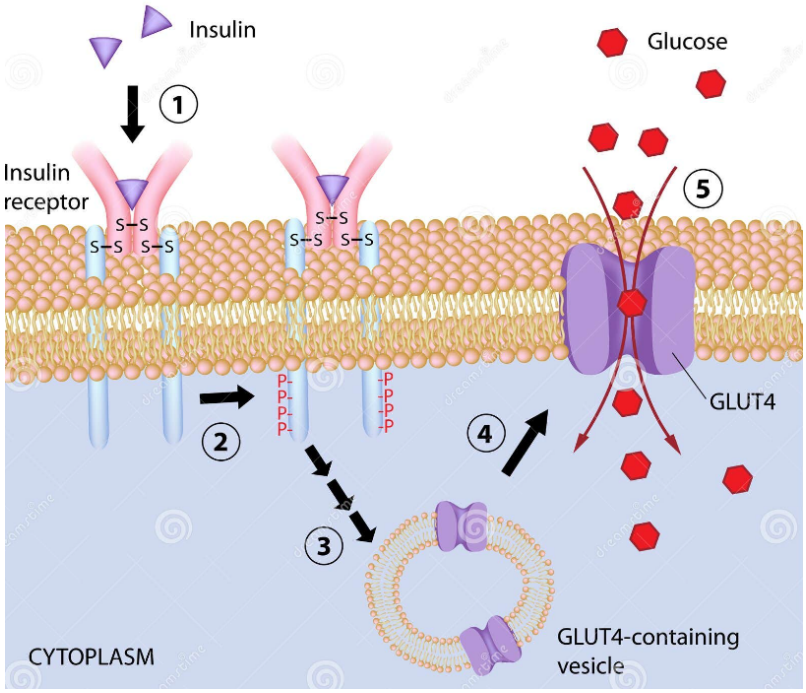


Figure 2.1: Effect of Insulin on Glucose Uptake[1].

2.2.2 Diabetes complications

Like all other diseases, diabetes can seriously cause many health problems if not well managed. These health complications are usually divided into two: microvascular and macrovascular complications. The microvascular complications include retinopathy, nephropathy, and neuropathy, which affect the eyes, kidneys, and nerves, respectively. The result may be blindness, kidney failure, and excruciating pain that leads to numbness. The long-term effects of high blood glucose levels, as in macrovascular complications, can lead to cardiovascular disease, stroke, and peripheral arterial diseases because of the damage they do to blood vessels and nerves throughout the body[9]. Early intervention and effective management play critical roles in preventing or delaying these complications.

2.3 Diagnosis of diabetes

Diabetes is usually discovered when long-term complications become evident. Type 1 diabetes is diagnosed when symptoms appear quickly. On the other hand, type 2 diabetes is often found by accident.

2.3.1 Diabetes Screening

Screening aims to test people for an increased level of glucose in their blood and detect individuals who are at risk. Its ultimate goal is to identify asymptomatic individuals who may have diabetes but are not aware of it. Some standard screening tests include the fasting plasma glucose test, which measures blood sugar levels on an empty stomach in the morning; the oral glucose tolerance test, which measures blood sugar levels before and after drinking a sugary solution; and HbA1c, which is a measure of the average amount of glucose over the previous two to three months [10]. These diagnostic tests help to catch diabetes at an early stage, which could then be prevented promptly without causing more complexities. Illustrated on figure 2.2.

Figure 2.2: Diabetes screening [2].

2.3.2 Signs of Diabetes

The symptoms of diabetes can be diverse. Still, many times are characterized by increased thirst (polydipsia), frequent urination (polyuria), fatigue, blurry vision, and an unexplained gain or loss in weight. Less common symptoms of diabetes could be a slow healing of cuts and other sores, frequent infection, and tingling or numbness of feet and hands. Early detection of these signs allows for a timely diagnosis and proper management. Recognizing these signs early can significantly reduce the risk of developing serious complications associated with diabetes [11].

2.3.3 Blood Glucose Analysis

Blood glucose analysis is the key to diabetes diagnosis and follow-up. The significant tests will be the fasting blood glucose test, where measurement of sugar in the blood is taken after a person has taken an overnight fast; the random blood glucose test, which can be done at any random time, to measure sugar in the blood; and the HbA1c test, which gives a broad spectrum of blood glucose levels over the past two to three months. These tests help assess current blood sugar levels and their average over some time, giving one an all-around view of how glucose is managed in the body. Close monitoring is the key to effective diabetes management and avert complications that may come from this condition[12].

2.4 Classification of Diabetes Types

2.4.1 Type 1 Diabetes

Type 1 diabetes is an autoimmune condition that you develop when your immune system decides to start attacking and killing your own body's beta cells. kills off the beta cells of the pancreas that make insulin. This condition can occur inup in childhood, more commonly in adolescence but also at any age. Will People With Type 1 Diabetes need to be treated with insulin for the rest of their lives and will hopefully be able to carry on keeping their blood sugars in check. Although the reason for Type 1 diabetes has not been determined, researchers believe that some sort of Immune system attacks itself called begin attack the middle of it is genetics likely a viral yield sites kills off the wings of the pancreas that produce insulin [13]. Type 1 diabetes is managed with dietary management and regular physical exercise, in combination with insulin therapy.

There are several different elements to the treatment of Type 1 Diabetes. Insulin therapy is vital and aimed at administering insulin via the subcutaneous route (injections) or infusion pumps to maintain blood sugar.examples of subcutaneous route: (Insulin syringe, insulin pump) Dietary monitoring, following a low carbohydrate diet, and consuming nutritional foods are crucial. This is not an encouragement to be sedentary, as regular physical activity is recommended for improved insulin sensitivity and overall health. Regular monitoring of blood glucose is required to determine if blood glucose levels are under control.Equally important is support from your healthcare team, peer groups, and diabetes education for effective diabetes management [14].

2.4.2 Type 2 Diabetes

Type 2 diabetes is the most common form of the disease, representing consequences of insulin resistance and relative insulin deficiency; it is commonly associated with obesity, physical inactivity,

and poor dietary habits. In the case of type 1 diabetes, there is simply an incapability of the body to produce insulin, but for type 2 diabetes, it is a case of a body that cannot make good use of the insulin it produces. Treatments are primarily through lifestyle modification, proper diet, regular exercise, oral medication, and, depending on the metabolic profile of the patient, insulin therapy. Most commonly, with lifestyle changes and frequent monitoring, Type 2 diabetes can be delayed or prevented from occurring.[13]. [15]

Treatment for Type 2 Diabetes focuses on lifestyle changes such as following a healthy diet full of whole grains, fruit, vegetables, and lean proteins while minimizing processed foods and sugars. Regular physical activities with both aerobic and resistance training will enhance insulin sensitivity. Management of blood sugar levels includes oral medications, such as metformin and sulfonylureas. In some cases, insulin may be needed for treatment. It is essential to follow up regularly with monitoring blood glucose levels as a part of informed decision-making about diet, exercise, and medication. Following up with routine blood glucose monitoring is crucial for making educated decisions about medicine, exercise, and diet [16].

2.5 Deep Learning

A subfield of machine learning, deep learning (DL) uses artificial neural networks to model high-level abstractions in data through multiple processing layers. It can be defined as a class of learning where data representation enables abstraction at two or more levels. These layered structures model designs that are developed to approximate the functioning of the physical structure of the human brain. This simulates enabling machines to learn from data hierarchically and abstractly. Notably, these techniques have been used for image and speech recognition, natural language processing, and predictive analytics. In the healthcare domain, deep learning is also being used to predict certain diseases—for example, diabetes—by considering a large amount of patient data to find some patterns within it based on positive or negative cases with high accuracy [17].

Figure2.3 illustrates the scope of deep learning.

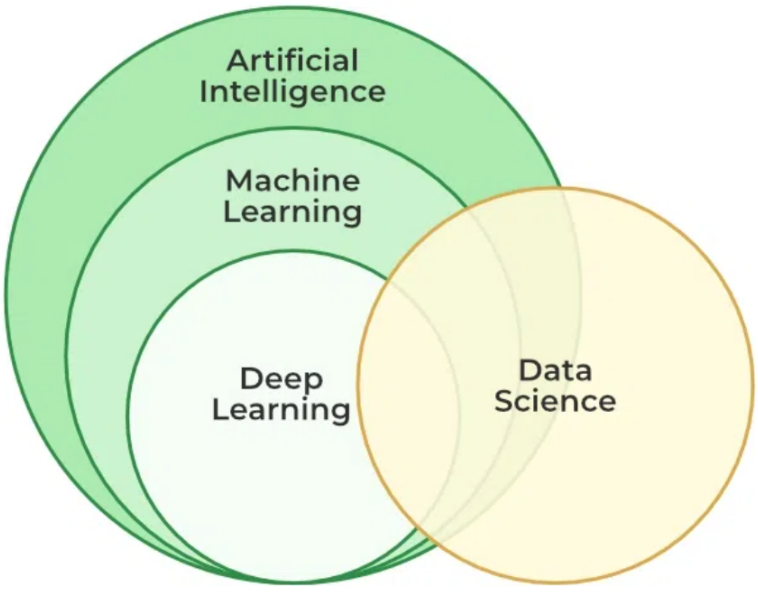


Figure 2.3: Scope of Deep Learning[3].

2.5.1 Long Short-Term Memory (LSTM)

A specialized type of recurrent neural network that excels in capturing sequential long-term dependencies is the Long Short-Term Memory (LSTM) network. LSTMs mitigate the vanishing and exploding gradient problems through a unique architecture that includes input, forget, and output gates. This architecture allows LSTMs to maintain information over extended sequences, making them suitable for tasks such as speech recognition and language modeling. LSTMs are also widely used in time series forecasting, video analysis, and other applications that require understanding and predicting temporal patterns [18].

2.5.2 Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture that is designed to handle sequential data and time series prediction tasks efficiently. Introduced by Cho et al., the GRU is an enhancement of the traditional RNN, addressing the vanishing gradient problem that often plagues deep learning models with long-term dependencies. The GRU architecture simplifies the structure of the long short-term memory (LSTM) units by combining the forget and input gates into a single update gate, and merging the cell state and hidden state. This streamlined design reduces the computational complexity while maintaining the ability to capture important temporal dependencies. GRUs have shown remarkable performance in various applications, including natural language processing, speech recognition, and time series forecasting, making them a popular choice for tasks requiring sequence modeling and prediction.

2.6 Swarm Intelligence

Swarm intelligence is a subcategory of artificial intelligence resting on the base of collective behavior of decentralized, self-organized systems, such as ant colonies, bird flocking, and fish schooling. These systems achieve problem-solving capabilities by using large numbers of relatively simple agents/approximators following simple behavioral rules[19].

Swarm intelligence algorithms such as Particle Swarm Optimization and Ant Colony Optimization mimic these natural processes to optimize solutions in robotics, telecommunications, and predictive modeling. In predicting diabetes, swarm intelligence can be put to use in parameter optimization of predictive models to increase their efficiency and accuracy[20].

They may be applied to solve optimization problems, either continuous, discrete, or multi-objective. Hence, it should have a wide range of applications in many areas. For example, they can be applied to water resources engineering, wireless networks, cloud-based Internet of Things, optical systems, recommendation systems, anomaly detection systems, and supply chain management. In addition, these algorithms are being widely applied in the process of clustering, feature selection, and solution for traveling salespeople problems. Further, these algorithms have numerous other applications in optimal designs, networking, electrical engineering, mechanical engineering, machine learning, resource allocation, and digital image processing, among many different fields[21].

2.6.1 Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a stochastic optimization technique that relies on populations. In PSO, the search space has a number of candidate solutions or particles that are moving according to straightforward mathematical equations guided by their own best-known position and the global best known position in the swarm[22] [23].

2.6.2 Ant Colony Optimization (ACO)

Ant Colony Optimization imitates an approach used by ants for solving computational problems whose essence is to locate good routes in graphs Inspired by bird flocks or fish schools' social behavior [24].

One of the most widely regarded instances of swarm intelligence is observable in real ants. With a collective goal of locating food, the ants will leave their colony and head off in any random direction. When one ant discovers food, it will return to the colony, leaving behind a trace of a specific chemical, pheromone, as it moves back. Other ants will then be able to sense this pheromone and head in the same direction. Interestingly, the frequency at which ants trail that path is determined by the pheromone concentration of that particular route. Since pheromones will naturally evaporate over time, the length of the path is, therefore, a factor. Thus, under all these considerations, a shorter route will be favored because there is an ant following that path, so they keep adding pheromone, thereby making the concentration strong enough against evaporation. In this manner, the shortest path from the colony to the food source is created [25]. Illustrated on figure 2.4

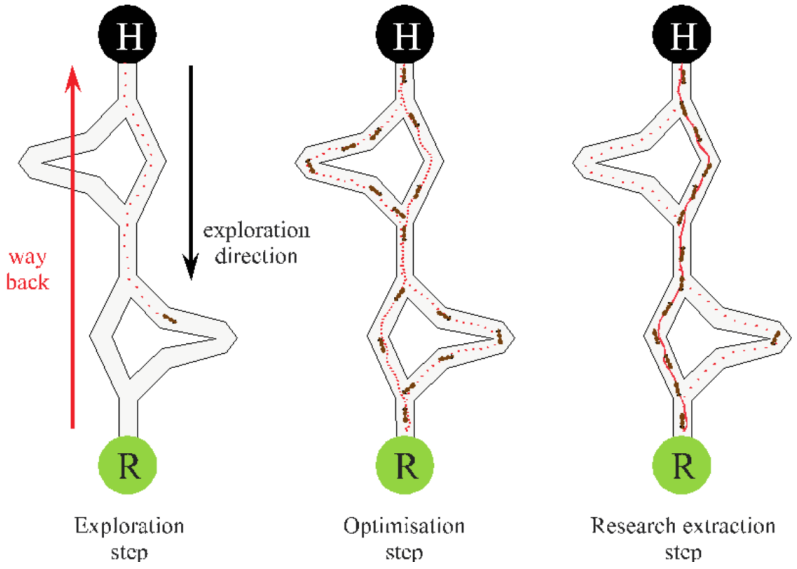


Figure 2.4: Ant Colony Optimisation[4].

2.6.3 The Grey Wolf Optimizer (GWO)

Nature-inspired metaheuristic algorithms, such as the Grey Wolf Optimizer (GWO), draw inspiration from the behaviors of animals in nature to solve optimization problems efficiently. Developed by Mirjalili et al.,[26] the GWO algorithm is specifically designed to emulate the collaborative hunting behavior of gray wolves in a pack. By mimicking how wolves coordinate and adapt their strategies during hunting, the GWO algorithm effectively explores and exploits search spaces to find optimal solutions.

Initial studies have demonstrated the effectiveness of the GWO algorithm across various optimization tasks. Its ability to balance exploration and exploitation, coupled with its simplicity and computational efficiency, has made it a promising tool in solving complex optimization problems. As such, the GWO algorithm has garnered attention in both academic research and practical applications, offering a valuable alternative to traditional optimization techniques.

2.6.4 Enhanced Harris Hawk Optimization (EHS)

Enhanced Harris Hawk Optimization (EHS) is an advanced variant of the Harris Hawk Optimization (HHO) algorithm, inspired by the cooperative hunting strategies of Harris hawks in nature. Developed to improve the performance of the original HHO, EHS incorporates various enhancements that augment its convergence speed, robustness, and accuracy. By simulating the intelligent predation behaviors of Harris hawks, EHS excels in exploring and exploiting complex search spaces to identify optimal solutions. Its adaptive parameters and integration with additional optimization techniques, such as chaos theory or differential evolution, prevent premature convergence and improve diversity. This has positioned EHS as a highly effective and efficient tool for tackling complex optimization problems, making significant contributions in both academic research and practical applications.

2.7 Conclusion

This chapter has provided a comprehensive overview of the fundamental concepts related to diabetes and its diagnosis, as well as the advanced computational techniques of deep learning and swarm intelligence. . These foundational elements are critical in understanding the methodology and analyses of the following chapters presented in this thesis.

Chapter 3

The State of the Art

3.1 Introduction

The current era witnesses a profound transformation driven by the widespread integration of artificial intelligence (AI) across diverse sectors, catalyzing a notable revolution in healthcare. Particularly, this technological surge is reshaping chronic disease management and treatment paradigms, with significant implications for diabetes prediction and care.

AI's continuous evolution equips healthcare professionals with potent diagnostic and treatment tools, while empowering researchers to develop intelligent systems aimed at augmenting patient care and forecasting diseases, with a specific emphasis on diabetes. Given the paramount importance of early intervention in mitigating diabetes-related complications, the focus on diabetes prediction research is instrumental in identifying individuals at heightened risk and implementing timely preventive measures. This proactive approach holds the potential to significantly enhance disease management and improve patient outcomes.

3.2 Related Works

Safial I and **Milon I** [27] present a comprehensive exploration of utilizing deep learning techniques, specifically deep neural networks, for predicting diabetes. By leveraging the Pima Indian Diabetes dataset, the study showcases the efficacy of deep neural networks in accurately predicting diabetes, achieving a remarkable accuracy rate of 98.35% for five-fold cross-validation.

The research delves into various aspects of the study, including a thorough discussion on related works in the field, a detailed background study on deep neural networks, the methods and materials employed, and the evaluation criteria to assess the predictive performance.

Another approach proposed by **Amani Y et al.** [28] which presents a novel Decision Support System designed for predicting diabetes through the utilization of advanced Machine Learning and Deep Learning techniques. The research conducted a comparative analysis of various algorithms, including Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN), using the Pima Indians Diabetes dataset as the basis for evaluation.

The results revealed that Random Forest (RF) exhibited superior performance in diabetes prediction when contrasted with both deep learning and SVM methodologies. Specifically, Random

Forest achieved an impressive overall accuracy rate of 83.67%, surpassing the accuracy rates of SVM at 65.38% and deep learning at 76.81%.The paper also outlines the future direction of the research, which will concentrate on enhancing feature extraction techniques and refining model fitting processes to further enhance the accuracy of diabetes prediction using Machine Learning and Deep Learning techniques.

This future work aims to optimize the predictive capabilities of the Decision Support System, ultimately contributing to more effective and reliable diabetes prediction models.

Haneen Q and **Mohammed A** [29] focus on the development of a model using machine learning techniques to classify and provide predictive analysis on the diagnosis of diabetes, specifically differentiating between types 1 and 2 diabetes.

The study utilizes a Palestinian dataset called DataPal, collected from the Palestinian Institute of Diabetes, consisting of 314 diabetic females aged between 5 and 89 years. Various machine learning models such as SVM, K-NN, DA, NB, DT, RF, and PSO-MLPNNs were applied to the DataPal dataset, with the PSO-MLPNNs model achieving an accuracy of 97.77%, sensitivity of 99.48%, and specificity of 93.12%.

The study also employed two-fold and four-fold cross-validation methods to evaluate the model's performance, using metrics like accuracy, sensitivity, and specificity to assess the effectiveness of the classification models. Future plans include developing a medical application for early detection and preventive treatments for diabetes, building on the success of the PSO-MLPNNs model in accurately predicting and diagnosing diabetes types.

Jyoti R's [30] article delves into the pressing issue of diabetes, a chronic disease posing a significant global health challenge. With statistics from the International Diabetes Federation indicating a doubling of diabetes cases by 2035, the urgency to develop effective predictive measures is paramount. Diabetes manifests through elevated blood glucose levels, leading to symptoms like frequent urination, increased thirst, and heightened hunger, while also being a major contributor to various complications such as blindness, kidney failure, and heart disease.

To address this challenge, the study employs machine learning techniques, a burgeoning field in data science, to devise a system for early diabetes prediction with enhanced accuracy. The dataset utilized comprises 2000 cases sourced from Kaggle, aiming to predict diabetes based on a range of measures. Five distinct machine learning algorithms - K nearest neighbor, Logistic Regression, Random forest, Support vector machine, and Decision tree - are harnessed and assessed for their predictive capabilities.
The study meticulously evaluates each algorithm's performance, analyzing their accuracy metrics to determine the most effective model. Notably, the Decision tree algorithm emerges as the frontrunner, boasting a remarkable accuracy rate of 98% in training and 99% in testing. This outcome underscores the potential of machine learning in revolutionizing medical diagnostics and early intervention strategies.

Looking ahead, the article posits avenues for future research and development in the realm of diabetes prediction and beyond. It advocates for the exploration of additional machine learning algorithms and the expansion of the designed system to encompass the prediction and diagnosis of other diseases. By leveraging the insights gained from this study, future endeavors aim to further refine and automate the analysis of diabetes and related medical conditions, ultimately advancing

healthcare outcomes on a broader scale.

Maniruzzaman et al. [31] their study aims to develop a machine learning (ML)-based system for predicting diabetic patients. Logistic regression (LR) identifies 7 risk factors for diabetes, including age, education, BMI, systolic BP, diastolic BP, direct cholesterol, and total cholesterol. Four classifiers (naïve Bayes, decision tree, Adaboost, and random forest) are employed, with a focus on three partition protocols (K2, K5, K10) and 20 trials. The ML-based system achieves an overall accuracy of 90.62%.

Notably, a combination of LR-based feature selection and the random forest classifier attains an impressive 94.25% accuracy and a 0.95 area under the curve (AUC) for the K10 protocol. The LR and random forest-based classifier combination is identified as highly effective for predicting diabetic patients.

Arianna et al. [32] utilized a data mining pipeline to develop predictive models for chronic microvascular complications in patients with type 2 diabetes mellitus (T2DM) based on electronic health record data of nearly 1,000 patients from the ICSM gùhospital in Italy. The pipeline involved clinical center profiling, predictive model targeting, construction, and validation, with logistic regression and random forest used for prediction.

The study focused on predicting the onset of retinopathy, neuropathy, or nephropathy at different time scenarios, considering variables such as gender, age, BMI, HbA1c, hypertension, and smoking habit. The final models achieved an accuracy of up to 0.838, with distinct models tailored for each complication and time scenario. SVMs and RF showed higher AUC values for predicting microvascular complications in T2DM patients based on electronic health record data.

Sandip M and **Vijay K.** [33] conducted a comprehensive study aiming to develop an advanced diabetes prediction model leveraging a multitude of machine learning techniques. Their model incorporated a wide array of algorithms, including Logistic Regression (LR), K-nearest neighbors (KNN), Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), Adaboost (AB), Decision Tree (DT), and ensemble methods such as XGBoost, LightGBM, CatBoost, Adaboost, and Bagging. By harnessing the power of these diverse algorithms, the researchers aimed to create a robust predictive framework capable of accurately identifying individuals at risk of developing diabetes.

The study utilized real-world datasets sourced from Kaggle, ensuring the relevance and applicability of the model to clinical settings. Python programming language was employed for the development and training of the model, highlighting the versatility and widespread adoption of Python in data science and machine learning communities. Performance evaluation of the developed model was conducted using a range of metrics, including the confusion matrix, sensitivity, and accuracy measurements. These metrics provided insights into the model's effectiveness in accurately predicting diabetes and distinguishing between diabetic and non-diabetic individuals. Of particular note, CatBoost emerged as the most effective ensemble technique, achieving an impressive accuracy rate of 95.4% and a higher AUC-ROC score of 0.99 compared to other ensemble methods like XGBoost.

This underscores the importance of selecting appropriate algorithms and ensemble techniques to enhance the predictive performance of the model. The study underscored the significance of

data analysis, personalized medicine, and AI models in providing accurate predictions and personalized recommendations for diabetes management.

By leveraging machine learning and data mining techniques, the researchers aimed to streamline clinical decision-making processes and alleviate the workload of healthcare professionals. In conclusion, the study presented a comprehensive approach to diabetes prediction and management, showcasing the potential of advanced machine learning techniques in improving patient outcomes and reducing the burden of diabetes on healthcare systems. The findings of the study contribute valuable insights to the field of diabetes research and pave the way for future advancements in predictive analytics and personalized medicine.

Khondokar et al. [34] tackle the global challenge of diabetes, focusing on early detection and intervention, particularly among women. Their study aims to develop accurate prediction models using various machine learning algorithms such as Random Forest, XGBoost, NGBoost, Bagging, LightGBM, and AdaBoost. Preprocessing of the dataset ensures optimal performance and reliability by cleaning data, normalizing, and engineering features.

The results are promising, with the model achieving a competitive accuracy rate of 92.91%. The integration of Shapley Additive Explanation (SHAP) techniques enhances interpretability, providing valuable insights into diabetes prediction factors. The study's benefits extend beyond academia, aiding healthcare providers, stakeholders, students, and researchers in diabetes prediction and management. By offering valuable insights and methodologies, the research advances predictive analytics and personalized medicine, potentially revolutionizing diabetes care globally.

Aditya et al. [35] introduced a pioneering hybrid machine learning model aimed at effectively categorizing diabetes by integrating the non-dominated sorting genetic algorithm (NSGA-II) with ensemble learning techniques. The researchers focused on robust feature selection and model optimization to enhance accuracy and efficiency in diabetes classification. Their methodology began with meticulous data preprocessing, addressing missing data and normalization to ensure dataset integrity. Leveraging NSGA-II's evolutionary principles, salient features relevant to diabetes classification were extracted from the dataset, navigating the vast search space to select the most indicative features.

Subsequently, an ensemble learning-based extreme gradient boosting (XGBoost) model was constructed using the selected features to enhance predictive performance. Performance assessment involved rigorous comparison with existing models, utilizing statistical parameters to evaluate accuracy, specificity, sensitivity, and F-score. Results showcased the superiority of the NSGA-II-XGB approach, achieving an average accuracy of 98.86% and outperforming existing models. With further validation through statistical metrics such as specificity 88.6%, sensitivity 96.36%, and F-score 97.84%, the proposed methodology demonstrates effectiveness and reliability in early diabetes diagnosis.

This research offers valuable insights and methodolog-ies with the potential to revolutionize diabetes diagnosis and management, advancing predictive analytics for enhanced patient care and improved health outcomes.

Kalpana and **Booba.** [36] proposed a comprehensive framework for predicting diabetes, leveraging machine learning techniques like the FireFly algorithm, Fuzzy C Mean, and Support Vector Machine (SVM) to enhance accuracy and enable early diagnosis. Recognizing the potential of (ML) in healthcare, especially in diabetes prediction, the researchers aimed to revolutionize detection and management.

Their methodology begins with meticulous data collection from the Pima Indian diabetes database, ensuring dataset integrity through preprocessing to handle missing values and ensure consistency. This phase lays the foundation for subsequent analysis and model development, enabling accurate predictions.

Key to the framework is the innovative use of the FireFly algorithm for feature selection, aiming to identify relevant features while minimizing complexity. Fuzzy C Means clustering groups similar data points, enhancing interpretability and prediction accuracy. Once features are selected and grouped, SVM, known for handling high-dimensional data and nonlinear relationships, is used for diabetes classification, offering superior performance.

The proposed algorithm's effectiveness is evaluated using metrics like accuracy, sensitivity, and specificity, comparing it with other ML approaches. Results show the FireFly-SVM hybrid classifier outperforms others, achieving an 84.76% accuracy rate. This research offers valuable insights, potentially transforming diabetes diagnosis and management. By combining innovative ML techniques with traditional methods, the framework represents a significant advancement in predictive analytics, leading to improved patient outcomes and healthcare delivery.

Addressing the formidable global health challenge posed by diabetes, **Huma N** and **Sachin A.** [37] propose an innovative solution for early detection, leveraging the capabilities of machine learning algorithms. Recognizing the critical importance of timely intervention and management in combating diabetes, the researchers embarked on a mission to develop a robust predictive model capable of identifying individuals at risk of developing the disease in its early stages.

In their study, the authors addressed the global health threat of diabetes, which currently affects 382 million people and is projected to reach 629 million by 2045. Recognizing the critical need for early detection and lifestyle intervention, the authors proposed a methodology for diabetes prediction using various machine learning algorithms on the PIMA dataset.
They employed Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL) classifiers to identify hidden patterns in healthcare data, achieving accuracies ranging from 90% to 98%. Among these, DL demonstrated the highest accuracy at 98.07%, establishing it as the most effective tool for early diabetes diagnosis. The results of this study pave the way for developing an automatic prognostic tool to assist healthcare professionals in the early detection of diabetes. The authors suggest that incorporating omics data could further enhance the accuracy of the DL approach, indicating a promising direction for future research.

In their study **Surabhi** and **Yogesh K.** [38] explore the application of machine learning algorithms for diabetes prediction, focusing specifically on the widely recognized PIMA diabetic dataset. They conduct a comparative analysis of various algorithms, such as decision tree, genetic algorithm, and evolutionary algorithm, to gauge their effectiveness in accurately predicting the onset of diabetes. The researchers emphasize the crucial role of dataset quality in achieving dependable results. For example, the genetic algorithm achieves an impressive 84% accuracy in predicting diabetes onset using the PIMA dataset.

In contrast, the decision tree algorithm yields a slightly lower accuracy of 79.9% when applied to a real-world male heart disease dataset. This highlights the significant influence of data quality on algorithmic performance and stresses the importance of refining both aspects to improve predictive models in diabetes prognosis.

KAMRUL H et al. [39] their paper proposes a robust framework for diabetes prediction, leveraging machine learning (ML) techniques to address challenges such as limited labeled data and outliers in diabetes datasets. The approach includes outlier rejection, missing value imputation, data standardization, feature selection, and the utilization of various ML classifiers, such as k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost, along with Multilayer Perceptron (MLP). Weighted ensembling of different ML models is employed to improve prediction accuracy, with weights estimated from the corresponding Area Under ROC Curve (AUC).

The experiments were conducted on the Pima Indian Diabetes Dataset, yielding an ensembling classifier with superior performance, outperform-ing state-of-the-art results by 2.00% in AUC. The proposed framework demonstrates better results and holds promise for enhanced diabetes prediction. The proposed ensembling classifier achieved the following individual performance metrics: Sensitivity: 0.789, Specificity: 0.934, False omission rate: 0.092, Diagnostic odds ratio: 66.234%.Additionally, the source code for diabetes prediction is publicly available.

Nagaraj and **Deepalakshmi.** [40] presented an innovative approach for early screening of diabetes using advanced machine learning techniques. Recognizing the critical need for accurate prediction of diabetes to prevent serious health complications, the authors propose a method that combines the strengths of Support Vector Machine (SVM) and Deep Neural Network (DNN) models. Their approach involves enhancing the SVM model, a powerful tool in traditional machine learning, and integrating it with a DNN, known for its ability to learn complex patterns from data. By leveraging the features learned by the SVM, the integrated model achieves a high accuracy of 98.45% in predicting diabetes status. The experiment, conducted using Python programming language, demonstrates the effectiveness of the proposed method in screening for diabetes. This integrated approach offers a promising solution for early detection of diabetes, paving the way for timely intervention and prevention of associated health risks.

Chetan A and **Ajay K.** [41] In this study, an optimization approach called the Improved Invasive Weed Bird Swarm Optimization Algorithm (IWBSOA) is proposed for predicting diabetes using gene expression data. The paper focuses on optimizing the prediction process by refining the input data and selecting relevant features to enhance classification accuracy while minimizing computational costs. The IWBSOA combines the Improved Invasive Weed Optimization (Improved IWO) and Bird Swarm Algorithm (BSA) for feature selection. Additionally, two classifiers, namely Recurrent Neural Network (RNN) and Support Vector Machine (SVM), are employed for the prediction process. The RNN classifier is trained using the Rider Optimization Algorithm (ROA) and Chicken Swarm Optimization (CSO), forming a hybrid deep learning model. The performance of the developed IWBSOA is evaluated using three metrics: specificity, sensitivity, and accuracy. The results demonstrate significant improvements in accuracy, sensitivity, and specificity, with the IWBSOA achieving an accuracy of 0.9619, sensitivity of 0.9711, and specificity of 0.9439. Additionally, the Mean Squared Error (MSE) is reported as 0.1887, indicating the model‘s effectiveness in predicting diabetes using gene expression data.

Dinesh C and **Harikumar R.** [42] In their study, they aimed to enhance diabetes detection using microarray gene data. Four techniques:Detrend Fluctuation Analysis, Chi-square probability density function, Firefly algorithm, and Cuckoo Search, are used for dimensionality reduction. Metaheuristic algorithms like Particle Swarm Optimization and Harmonic Search are employed for feature selection. Seven classifiers, including Support Vector Machine with Radial Basis Function, are utilized for classification. Performance metrics such as accuracy, recall, and precision are analyzed. The SVM (RBF) classifier with Chi2pdf reduction and PSO feature selection achieves 91% accuracy and a Kappa of 0.7961, outperforming others. This approach enhances type II diabetes mellitus detection. The study focuses on timely diabetes identification. The research addresses the global impact of diabetes mellitus. It proposes novel methods for diabetic detection. Microarray gene data is central to the analysis. Feature selection plays a crucial role in model enhancement. Metaheuristic algorithms optimize the feature selection process.

Yassine A et al. [43] 's paper presents a novel methodology for diabetes classification by integrating Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models. Leveraging the unique capabilities of both architectures, the approach aims to capture sequential patterns and extract meaningful features from input data efficiently. Through the utilization of a comprehensive dataset containing pertinent features related to diabetes patients, the classifiers are trained and evaluated rigorously. Evaluation metrics, including kappa score, F1-score, accuracy, precision, and recall, are employed to gauge the performance of each model. Results showcase the superiority of the CNN-LSTM model over traditional approaches such as LR, RF, SVM, and KNN, achieving an impressive accuracy rate of 97%. This methodology underscores the potential of combining CNN and LSTM architectures for accurate diabetes classification, marking a significant advancement in diagnosis and treatment approaches and offering promising prospects for personalized healthcare solutions.

Ganesan M et al. [44] Thier paper introduces a novel approach to diabetes disease classification, leveraging an optimized deep neural network (DNN) model. The methodology involves the integration of a multilayer perceptron (MLP) to enhance classifier efficiency by identifying and correcting misclassified instances within the dataset. Specifically, the DNN is utilized for effective disease diagnosis, while the MLP serves to refine the data by removing inaccuracies. The proposed model is rigorously evaluated using the PIMA Indians Diabetes dataset, containing medical details of 768 patients across 8 attributes per record. Experimental results demonstrate the superior performance of the proposed methodology compared to existing approaches, underscoring its efficacy in accurate disease classification.

Suja A. et al [45] introduces a novel approach for early detection of diabetes using artificial intelligence (AI) techniques. Specifically, a new SMOTE-based deep LSTM system is developed to address class imbalance in diabetes datasets and improve prediction accuracy. The methodology involves an investigation of various deep learning architectures, including CNN, CNN-LSTM, ConvLSTM, and deep 1D-convolutional neural network (DCNN). Through extensive experimentation and analysis, the proposed SMOTE-based deep LSTM method emerges as the most effective in diabetes prediction. The model achieves an impressive prediction accuracy of 99.64%, surpassing other machine learning and deep learning approaches evaluated in the study. These findings underscore the significance of the proposed methodology in enhancing diabetes prediction accuracy and highlight its potential for improving patient outcomes through early detection and intervention.

3.3 Analysis and Comparison

3.3.1 Comparison criteria

Previously, we have outlined the primary prediction methodologies within the healthcare domain. In the following section, we will conduct a comparative analysis of the approaches based on the following 8 criteria:

- **Article**: designates the proposed approach.
- **Approach**: designates the used techniques.
- **Dataset**: indicates the data sources used for the implementation of the approach for the prediction of diabetes.
- **Used techniques**: describe the specific techniques or models applied to predict diabetes.
- **Software tools**: indicates whether the approach is supported by any specific software tools.
- **Performance evaluation**: results of the effectiveness and accuracy of advanced techniques, algorithms, or software tools that help determine how well the tool or model performs (we have used the accuracy).
- **Target**: refers to the specific type of diabetes that is the focus of the evaluation or prediction.
- **Advantages**: advantages of the approach discussed.

3.3.2 Comparative table

Article	**Approach**	**Dataset**	**Used Techniques**	**Tools**	**Performance Evaluation**	**Target**	**Advantages**
Safial I and Milon I	Deep Learning	PIMA	DNN	YES	Accuracy = 98.35%, F1-Score = 0.98, MCC = 0.97	Diagnosing diabetes	High accuracy
Amani et al.	Machine Learning	PIMA	SVM, RF, CNN	YES	RF = 83.67%, SVM = 65.38%, DL = 76.81%	Medical Decision Support	Diverse techniques
Haneen Q and Mohammed A	Swarm, DL	DataPal	PSO, MLPNNs	YES	Accuracy = 97,77%	Early detection	High sensitivity and specificity
Jyoti R	Machine Learning	John Diabetes Database	KNN, LR, RF, SVM, DT	YES	DT = 99%	Early diabetes prediction	High accuracy
Maniruzzaman et al.	Machine Learning	Yes	LR, NB, DT, Adaboost, RF	YES	LR + RF = 94.25%	Predicting diabetic patients	Flexible classifiers
Arianna et al.	Machine Learning	Yes	RF, LR	NO	Accuracy = 0.838	Type 2 diabetes	Predictive analytics
Sandip M and Vijay K	Machine Learning	Kaggle dataset	LR, KNN, SVM, NB, RF,DT, Adaboost, XGBoost, LightGBM, CatBoost,	YES	CatBoost = 95.4%	Diabetes prediction	Ensemble methods
Khondokar et al.	Machine Learning	Yes	RF,XGBoost,NO NGBoost,Bagging, LightGBM, AdaBoost		Accuracy = 92.91%	Early diabetes detection	Stacked ensemble approach
Aditya et al.	Hybrid Model (NSGA-II + XGBoost)	Hybrid dataset	NSGA-II, XGBoost	NO	Accuracy = 98.86%; specificity = 88.6%, sensitivity = 96.36%, F-score = 97.84%	Categorizing diabetic patients	Superior accuracy and performance in early diabetes diagnosis

Kalpana and Booba	Machine Learning, SWARM	PIMA	FireFly, Fuzzy C Mean, SVM	YES	Accuracy = 84.76%	Predicting diabetes	Fuzzy logic-based approach
Huma N and Sachin A	Deep learning techniques	PIMA	ANN, NB, DT, DL	NO	DLC accuracy = 98.07%	Early disease detection	High accuracy rates.
Nagaraj and Deepalakshmi	Machine Learning	Yes	SVM, DNN	YES	SVM-DNN = 98.45%	Early diabetes detection	Hybrid approach
Kamrul H et al.	Machine Learning	PIMA	KNN, DT, RF, AdaBoost, NB, XGBoost, MLP	YES	Sensitivity = 0.789, Specificity = 0.934	Early diabetes screening	Robust framework,superior results.
Surabhi K and Yogesh K	Machine Learning	PIMA	DT,GA,EA, RF, LR, SVM, NB	YES	Predict the likelihood of diabetes	predict the likelihood of diabete	Using a genetic algorithm for disease risk classification.
Chetan A and Ajay K	SWARM, ML, DL	Gene expression data	IIWO, BSA,SVM,RNN,	YES	Accuracy = 0.9619	Advanced algorithms	Improved accuracy
Dinesh C and Harikumar R	Machine Learning, Swarm	Yes	SVM-RBF	YES	SVM-RBF + Chi2pdfR + PSO = 0.91	Improving accuracy	Custom algorithms
Yassina A et al.	Deep learning	PIMA	CNN, LSTM	NO	Accuracy = 0.97	Diabetes classification	Superior performance over traditional approach
Ganesan M et al.	Deep Learning	PIMA	DNN, MLP	YES	Accuracy = 0.996	High accuracy	Very high accuracy
Suja A. et al.	Deep learning	PIMA	CNN-LSTM, ConvLSTM, DCNN,SMOTE-LSTM	YES	Accuracy = 0.996	Early detection of diabetes	Surpassed performances of other models

Table 3.1: Comparative Analysis of Approaches

3.3.3 Discussion

The analysis of the works presented in the previous section reveals a diverse array of artificial intelligence concepts employed for the early prediction of diabetes. The primary approaches involve leveraging machine learning and deep learning techniques, with a notable but lesser emphasis on swarm intelligence.

The studies reviewed highlight a variety of methods and techniques for diabetes prediction, where the effectiveness of these models is heavily influenced by the selection of algorithms and data pre-processing methods. The demonstrated methodologies and evaluation metrics emphasize the capability of AI algorithms in accurately predicting and managing diabetes. This ongoing research underscores a dynamic field focused on utilizing AI techniques to enhance prediction accuracy and facilitate early detection.

Deep Learning and Hybrid Models have demonstrated remarkable potential in improving diagnostic precision and early disease detection. For instance, Safial I and Milon I achieved an impressive accuracy of 98.35% and an F1-Score of 0.98 using a Deep Neural Network (DNN) on the PIMA dataset, highlighting the effectiveness of deep learning in diagnostic tasks. Similarly, Huma N and Sachin A developed a predictive model with deep learning techniques, achieving an accuracy of 98.07% for early disease detection, showcasing the capability of deep learning in identifying high-risk individuals at an early stage. Additionally, Nagaraj et al. employed a hybrid approach by combining Support Vector Machine (SVM) and Deep Neural Network (DNN) models, resulting in an accuracy of 98.45% in predicting diabetes status, demonstrating the enhanced predictive performance obtained by leveraging both methods. These findings underscore the potential of deep learning and hybrid approaches in advancing the field of diabetes prediction and management.

Additionally, swarm intelligence approaches have shown promise in diabetes prediction. Chetan A. and Ajay K. used Improved Intelligent Water Drops (IIWO) and Binary Swarm Algorithm (BSA), achieving an accuracy of 96.19%. Kalpana and Booba employed a Firefly Fuzzy C-Means approach and achieved an accuracy of 84.76%. Haneen Q. and Mohammed A. utilized Particle Swarm Optimization (PSO) for feature selection, resulting in an accuracy of 97.77%. These results highlight the potential of swarm intelligence methods in enhancing predictive accuracy and providing robust solutions for diabetes prediction. The reviewed studies indicate that advanced models, particularly those involving deep learning, hybrid approaches, and ensemble methods, generally offer superior performance in accuracy and early detection capabilities. These models are essential for improving predictive capabilities and patient outcomes in diabetes management.

Furthermore, the integration of AI in diabetes prediction is not only limited to accuracy improvement but also involves enhancing the interpretability and transparency of the models used. This is crucial for gaining the trust of healthcare professionals and patients. The development of explainable AI (XAI) techniques is increasingly becoming a focus, as it allows for the understanding of how AI models make predictions, ensuring that these models can be reliably used in clinical settings.

Moreover, ongoing research is exploring the use of real-time data and wearable technology to continuously monitor and predict diabetes risk. This approach can lead to more personalized and timely interventions, ultimately reducing the incidence and complications associated with diabetes. The synergy between AI and healthcare technologies presents a promising avenue for revolutionizing diabetes management, making it more proactive and patient-centric.

Overall, the use of AI in predicting and managing diabetes is a quickly developing area that has the potential to greatly impact healthcare outcomes. The dynamic and impactful nature of this

research field is emphasized by the use of advanced AI techniques, such as deep learning and hybrid models, as well as efforts to enhance model transparency and incorporate real-time monitoring.

3.4 Conclusion

In conclusion, the review and analysis presented in this chapter underscore the critical role of innovative predictive methodologies in advancing diabetes diagnosis and management. The integrat-ion of deep learning, hybrid models, and ensemble techniques shows significant promise in enhancing predictive accuracy and early detection. Future research should focus on incorporating these advanced methods into real-world clinical settings, addressing challenges related to data diversity, model interpretability, and computational efficiency to fully realize their potential in healthcare applications.

Given the promising results from combining deep learning with swarm intelligence, our future approach will leverage these advanced techniques to develop a robust and accurate system for early diabetes prediction. By integrating deep learning models with swarm intelligence algorithms, we aim to harness the strengths of both methodologies, potentially leading to breakthroughs in predictive performance and clinical applicability, which will be presented in detail in the next chapter.

Chapter 4

Proposed Approach for Diabetes Prediction

4.1 Introduction

Diabetes remains a pervasive health concern globally, affecting millions of individuals annually and presenting significant challenges for both individuals and healthcare systems. In this introductory chapter, we provide an overview of our research objectives, highlighting the rationale behind our chosen methodologies and outlining the structure of the thesis. Subsequent chapters will delve into the conceptual underpinnings of ACO and LSTM algorithms, elucidating their roles in feature selection and predictive modeling for diabetes risk assessment. Through a comprehensive analysis of our approach, we endeavor to contribute valuable insights to the ongoing pursuit of effective diabetes prediction strategies.

4.2 Contribution

The ability to accurately predict diabetes risk is paramount for early intervention and personalized healthcare, thereby mitigating potential health complications associated with undiagnosed or poorly managed diabetes. In response to this imperative, this thesis proposes an innovative approach to diabetes prediction, leveraging Ant Colony Optimization (ACO) for feature selection and Long Short-Term Memory (LSTM) networks for predictive modeling.

The complexity of diabetes prediction stems from the multifaceted nature of the disease, necessitating the analysis of diverse variables such as medical history, lifestyle factors, and genetic predisposition. Addressing this complexity requires advanced computational techniques that can effectively identify relevant features within datasets and construct robust predictive models. Our approach seeks to meet this challenge by harnessing the power of ACO, a metaheuristic optimization algorithm inspired by the foraging behavior of ants, to select the most informative features for diabetes prediction.

Furthermore, we employ LSTM networks, a type of recurrent neural network renowned for its ability to capture long-term dependencies in sequential data, for predictive modeling. By leveraging the temporal dynamics inherent in medical data, LSTM networks offer a promising avenue for accurate and timely diabetes risk assessment.

This thesis represents a significant contribution to the field of diabetes prediction, building upon prior research endeavors. We aim to extend the current understanding by exploring the efficacy of ACO for feature selection and LSTM networks for predictive modeling in the context of diabetes risk assessment. Through this research, we aspire to enhance the accuracy of diabetes

prediction, facilitate early interventions, and ultimately improve health outcomes for individuals at risk of developing diabetes.

4.3 Proposed approach

The overall architecture of the ACO-LSTM proposal is presented in 4.1.

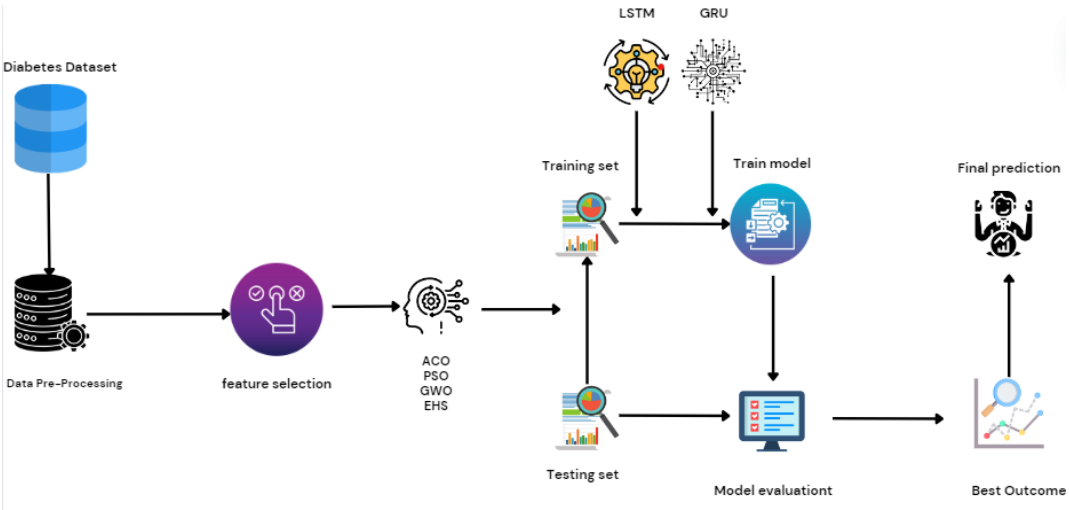


Figure 4.1: Proposed approach

Figure 4.1 above shows the overall structure of the LSTM-ACO model which constitutes four major primary stages:

Stage 1: Data collection.

Stage 2: Data pre-processing. This phase includes various processes for managing and supervising data.

Stage 3: Feature selection. Using the ACO algorithm as the main focus, optimal features are selected.

stage 4: The prediction process using the LSTM is the final step to consider.

The subsequent sections will provide a detailed presentation of the steps involved in the proposed approach.

4.3.1 Data Collection

A dataset, as defined by the Oxford Dictionary, is a collection of data treated as a single unit by a computer. It comprises structured data organized for analysis, research, or other purposes. Typically, datasets consist of individual data points, such as numbers, text, images, or other information types, organized into rows and columns. Each row represents a single observation or record, while each column denotes a specific attribute or variable [46].

Data collection is the primary and essential step in any diabetes prediction process, it provides the base for an ideal predictive model by analyzing the given population sincerely if the implementation is done correctly. This is the key to better evaluation and best-quality results.

4.3.2 Data Preprocessing

From the data collected, the next important part is pre-processing the data to derive a clean and complete dataset devoid of missing values and incoherent data. Incomplete or unclean data usage makes the model have poor performance. , normalizing the data, and splitting it into training and testing sets. Data pre-processing involves cleaning the dataset to handle missing values and outliers.

4.3.2.1 Handling Missing Values and Outliers

Missing data is a prevalent issue in medical datasets that can compromise the integrity of predictive models. To mitigate this, we employ robust strategies such as imputation and data removal:

• **Imputation:** When feasible, missing values are filled using statistical methods such as mean or median imputation. This approach ensures that the dataset remains comprehensive while preserving statistical integrity.

• **Data Removal:** In cases where missing data is extensive or systematic, records with significant missing information are removed. This ensures that the analysis is based on complete and reliable data points, preventing biases in subsequent modeling.

4.3.2.2 Normalization

Normalization is essential to standardize the scale of continuous variables across different features, ensuring fair contribution to the predictive model:

• **Min-Max Normalization:** This method scales the values of each feature to a range typically between 0 and 1. It is suitable when the distribution of data does not follow a Gaussian (normal) distribution and preserves the relationships between different data points.

• **Standardization:** Alternatively, standardization transforms the data to have a mean of 0 and a standard deviation of 1. It is effective when the data distribution is Gaussian, making it suitable for algorithms that assume a normal distribution in the input variables.

4.3.2.3 Data Splitting

To evaluate the performance of our predictive model effectively, the preprocessed dataset is divided into distinct subsets for training and testing purposes:

•**Training Set:** Typically comprising 70-80% of the data, this subset is used to train the predictive model. The model learns patterns and relationships within the data during this stage.

• **Testing Set:** The remaining 20-30% of the dataset is reserved for testing the performance of the trained model. This independent subset helps assess the model's ability to generalize to new, unseen data and provides critical insights into its predictive accuracy and robustness.

4.3.3 Feature Selection

Feature selection is a critical preprocessing step in machine learning that aims to identify and retain the most relevant attributes from the dataset. Effective feature selection not only reduces the dimensionality of data but also enhances model interpretability and predictive accuracy. This section explores various feature selection algorithms—Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), Enhanced Harris Hawk Optimization (EHHO), and Ant Colony Optimization (ACO)—applied in the context of diabetes prediction. Each algorithm's methodology, optimization process, and contributions to improving the robustness of predictive models are discussed, providing insights into their respective strengths and applications.

4.3.3.1 Grey Wolf Optimization (GWO)

The feature selection process employs the Grey Wolf Optimization (GWO) algorithm, selected for its ability to efficiently explore and exploit the solution space. GWO mimics the social hierarchy and hunting behavior of grey wolves to optimize feature subsets that enhance classification accuracy.

In the context of GWO, the algorithm initializes a population of grey wolves (agents), each representing a potential solution (feature subset). During each iteration, wolves collaborate and compete to improve the fitness of their solutions. The position update equation (Equation 1 described) adjusts each wolf's position based on its personal best (pbest) and the global best (gbest) among all wolves, using predefined acceleration coefficients.

$$\text{Position Update in GWO: } \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \mathbf{A} \odot \mathbf{D} \tag{4.1}$$

Where: - $\mathbf{x}^{(t)}$ is the current position (solution) of a wolf at iteration t. - $\mathbf{A}$ and $\mathbf{D}$ represent adjustment matrices derived from the alpha, beta, and delta positions of the wolves. - The matrices control the magnitude and direction of movement towards the optimal solution.

GWO iteratively refines feature subsets by balancing exploration and exploitation, ultimately selecting the most discriminative features for subsequent classification tasks.

4.3.3.2 Enhanced Harris Hawk Optimization (EHHO)

The Enhanced Harris Hawk Optimization (EHHO) algorithm is employed for feature selection due to its enhanced ability to converge towards optimal solutions in complex datasets. EHHO builds

upon the original Harris Hawk Optimization by incorporating additional mechanisms to diversify the search and improve solution quality.

In EHHO, each hawk (solution candidate) dynamically adjusts its position and velocity based on the observed behavior of other hawks in the population. The optimization process iteratively updates the position of each hawk towards better solutions, influenced by both personal experience and collective intelligence.

The position update in EHHO (Equation 2 described) integrates personal best (pbest) and global best (gbest) metrics, leveraging acceleration coefficients to regulate exploration and exploitation efforts effectively.

$$\mathbf{v}_i(t+1) = \omega \cdot \mathbf{v}_i(t) + c_1 \cdot r_1 \cdot (\mathbf{pbest}_i(t) - \mathbf{x}_i(t)) + c_2 \cdot r_2 \cdot (\mathbf{gbest}(t) - \mathbf{x}_i(t)) \tag{4.2}$$

Where: - $\mathbf{v}_i(t)$ is the velocity vector of the i-th hawk at iteration t. - $\mathbf{x}_i(t)$ represents the current position of the i-th hawk at iteration t. - $\mathbf{pbest}_i(t)$ is the previous best position of the i-th hawk. - $\mathbf{gbest}(t)$ is the global best position among all hawks in the population at iteration t. - ω, c_1, and c_2 are positive acceleration coefficients controlling the balance between local and global search efforts.

EHHO effectively optimizes feature subsets by integrating diverse search strategies, making it a suitable choice for enhancing the discriminative power of diabetes prediction models.

4.3.3.3 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a population-based optimization algorithm inspired by the social behavior of birds flocking or fish schooling. PSO leverages the collective intelligence of a swarm of particles to navigate a complex search space and identify optimal solutions.

In PSO, each particle represents a potential solution and has a position and velocity within the search space. The particles move through the search space, adjusting their positions based on their own experience and the experience of neighboring particles. This adjustment is influenced by two factors: cognitive and social components.

The PSO algorithm iteratively updates the velocity and position of each particle as follows:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (p_{best,i} - x_i(t)) + c_2 \cdot r_2 \cdot (g_{best} - x_i(t)) \tag{4.3}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{4.4}$$

Where: - $v_i(t)$ denotes the velocity of particle i at iteration t. - $x_i(t)$ represents the position of particle i at iteration t. - w is the inertia weight, controlling the influence of the previous velocity. - c_1 and c_2 are acceleration coefficients that weigh the influence of the cognitive and social components, respectively. - r_1 and r_2 are random numbers between 0 and 1. - $p_{best,i}$ is the best position encountered by particle i. - g_{best} is the best position encountered by the entire swarm.

PSO balances exploration of the search space and exploitation of the best solutions found by dynamically adjusting particle velocities and positions. This collective strategy enables the identification of high-quality feature subsets, enhancing the accuracy of diabetes prediction.

4.3.3.4 Ant Colony Optimization (ACO)

Feature selection employs the Ant Colony Optimization (ACO) algorithm, renowned for its ability to discover high-quality solutions through the collective intelligence of ant-like agents. ACO models

the foraging behavior of ants to navigate through a complex search space and identify optimal feature subsets.

In ACO, artificial ants construct solutions by iteratively selecting features based on pheromone trails and heuristic information. Pheromone trails represent the attractiveness of features, updated dynamically based on solution quality. The solution construction mechanism in ACO (Equation 5 described) balances exploitation of known good solutions and exploration of new feature combinations.

$$\mathbf{p}_{ij}(t+1) = \frac{\tau_{ij}(t)^{\alpha} \cdot \eta_{ij}}{\sum_{j \in J_i(t)} \tau_{ij}(t)^{\alpha} \cdot \eta_{ij}} \tag{4.5}$$

Where: - $\mathbf{p}_{ij}(t)$ denotes the probability of selecting feature j by ant i at iteration t. - $\tau_{ij}(t)$ represents the pheromone level on the trail connecting ant i and feature j at iteration t. - η_{ij} is the heuristic information associated with feature j, aiding in decision-making. - α controls the relative importance of pheromone trails versus heuristic information.

ACO iteratively refines feature subsets by leveraging collective exploration and exploitation strategies, ultimately selecting the most informative features for accurate diabetes prediction.

4.3.4 Prediction Process

The prediction phase focuses on utilizing Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) neural networks to predict the likelihood of diabetes onset based on selected features. These algorithms are chosen for their ability to capture temporal dependencies and sequence information, critical for analyzing longitudinal health data and making accurate predictions.

4.3.4.1 Long Short-Term Memory (LSTM)

LSTM is a specialized type of recurrent neural network (RNN) designed to address the vanishing gradient problem in traditional RNNs. It achieves this by introducing a memory cell that maintains information over extended time periods, selectively updating and retaining information through gates—input gate, forget gate, and output gate.

•Methodology: LSTM networks are structured to process sequences of data efficiently, allowing them to learn dependencies over time steps. The architecture includes memory cells that enable the network to remember previous states, crucial for modeling long-term dependencies.

•Training Process: LSTM networks are trained using backpropagation through time (BPTT), where gradients are propagated backward through time steps to update weights. This enables the network to learn from historical data and improve its predictive accuracy.

•Application in Diabetes Prediction : LSTM is applied to the feature-selected dataset to capture temporal patterns and relationships. It excels in scenarios where the sequence of events and their timing are essential for accurate predictions of diabetes onset.

4.3.4.2 Gated Recurrent Unit (GRU)

GRU is another variant of RNN that simplifies the architecture compared to LSTM, yet maintains effectiveness in capturing long-term dependencies. It combines the update and reset gates into a single gate mechanism, reducing computational complexity and enhancing training speed.

• Methodology: GRU networks utilize fewer parameters compared to LSTM, making them computationally efficient while still capable of learning complex temporal dynamics. The simplified gating mechanism allows for adaptive learning of sequence data.

• Training Process: Similar to LSTM, GRU networks undergo BPTT during training to optimize weights based on historical data. This iterative process adjusts the network's internal state to improve predictions over successive time steps.
• Application in Diabetes Prediction: GRU is integrated into the prediction process to evaluate its performance relative to LSTM. It excels in scenarios requiring efficient processing of sequential data while maintaining predictive accuracy in diabetes prediction tasks.

4.3.4.3 Comparative Analysis

Both LSTM and GRU offer distinct advantages in modeling temporal data for diabetes prediction. The choice between them hinges on factors such as computational efficiency, training speed, and the ability to capture long-term dependencies effectively. A comparative analysis is conducted to evaluate their performance based on metrics such as accuracy, sensitivity, specificity, and computational resources required.

4.3.4.4 Prediction Process with Feature Selection and LSTM

1. Feature Selection Algorithms (ACO, PSO, GWO, EHHO):

• Objective: These algorithms aim to identify a subset of features from a larger dataset that optimizes the predictive performance of the LSTM model for diabetes prediction.

• Initialization:
Pheromone levels (ACO), particle positions (PSO), wolf positions (GWO), or heuristic-driven strategies (EHHO) are initialized based on the features' relevance and initial assessments.

• Search and Evaluation:
Iteratively, these algorithms explore different combinations of features.
- They select features probabilistically or based on heuristic information.
- Each subset is evaluated using LSTM as the classifier to determine its predictive accuracy.
- Evaluation metrics (e.g., accuracy, sensitivity) are computed to measure how well each subset performs in predicting diabetes.

2. Integration with LSTM:

• Selected Feature Subset:
- The feature selection algorithms output the subset of features that demonstrate the highest predictive performance when evaluated by LSTM.
- These selected features are integrated into the input layer of the LSTM model.

3. LSTM Training and Prediction:

• Training Phase:
- LSTM is trained using sequences of data points where each sequence corresponds to the selected feature subset.
- LSTM learns to capture temporal dependencies and patterns in the data related to diabetes.

• Prediction Phase:
- Once trained, LSTM acts as the predictive model.
- It takes new data points, represented by the selected features, and predicts whether a patient is likely to have diabetes or not.
- LSTM's prediction outputs are based on its learned patterns from the training data, which have been optimized through the selected features identified by the feature selection algorithms.

4. Evaluation and Refinement:
• Performance Evaluation:
- The performance of LSTM, using the selected features, is evaluated using standard metrics such as accuracy, MEDAE, RMSE, MAE, R2 Score and ROC-AUC.
- These metrics assess how well LSTM predicts diabetes compared to other methods or using all available features.
• Iterative Improvement:
- Based on evaluation results, adjustments may be made to the feature selection process or LSTM model to further enhance predictive accuracy.
- This iterative process continues until satisfactory predictive performance is achieved.

4.3.4.5 GRU with ACO for Feature Selection

1. Ant Colony Optimization (ACO) Overview:
• Objective: ACO is an optimization algorithm inspired by the foraging behavior of ants.
• Initialization:
- Pheromone levels are initialized on features, reflecting their attractiveness.
- Heuristic information is calculated based on feature relevance.
• Feature Subset Construction:
- Ants iteratively construct feature subsets:
- They select features probabilistically using pheromone levels and heuristic values.
- Each subset's performance is evaluated using a predictive model (in this case, GRU).
• Pheromone Update:
- Based on the subset's performance, pheromone levels are updated to reinforce paths (features) that contribute to better predictive accuracy.
2. Integration of GRU:
• Selected Feature Subset:
- ACO identifies the feature subset that exhibits the highest predictive performance when evaluated by GRU.
- This subset typically includes features that are most relevant and informative for predicting diabetes.
- The selected features are then fed into the GRU model for training and prediction.

3. GRU Training and Prediction:
• Training Phase:
- GRU is trained using sequences of data points where each sequence corresponds to the selected feature subset identified by ACO.
- GRU learns to capture temporal dependencies and patterns in the data related to diabetes.
• Prediction Phase:

- Once trained, GRU serves as the predictive model.
- It takes new data points represented by the selected features and predicts whether a patient is likely to have diabetes or not.
- GRU's predictions are based on its learned patterns from the training data, optimized through the selected features identified by ACO.

4. Evaluation and Optimization:

• Performance Evaluation:
- The performance of GRU using the selected feature subset is evaluated using metrics such as accuracy, sensitivity, specificity, and ROC-AUC.
- These metrics gauge how effectively GRU predicts diabetes compared to using all available features or other methodologies.

• Iterative Improvement:
- Based on evaluation results, adjustments may be made to the feature selection process or GRU model parameters to further enhance predictive accuracy.
- This iterative process continues until satisfactory predictive performance is achieved.

Integrating ACO with GRU for feature selection in diabetes prediction optimizes the selection of relevant features that GRU uses to make accurate predictions. ACO's ability to systematically explore feature subsets and GRU's capability to model temporal relationships in data synergistically enhance the overall predictive power of the system. This combined approach exemplifies a robust methodology for leveraging both optimization algorithms and deep learning models in medical data analysis and predictive modeling tasks.

4.4 Conclusion

In this section, we have comprehensively defined and detailed the feature selection algorithms (ACO, PSO, GWO, EHHO) and prediction algorithms (LSTM, GRU) that we utilized in our research for predicting diabetes onset. By outlining their methodologies, training processes, and the integration of feature selection with neural network architectures, this study sets a robust foundation for identifying the most effective approach for accurate and timely diabetes prediction. Each algorithm has been explained in detail, highlighting their unique contributions to enhancing the predictive performance of our models.

In the next chapter, we will transition to the implementation stage, where we will apply these models to our dataset. This phase will involve rigorous testing and validation to empirically verify which combination of feature selection and neural network gives the best predictive performance. Through this process, we aim to determine the optimal model for diabetes prediction, providing valuable insights for early diagnosis and intervention. This comprehensive evaluation will not only validate our theoretical approach but also contribute significantly to the field of medical data analysis and predictive modeling.

Chapter 5

Experimental Setup and Evaluation

5.1 Introduction

The primary aim of this project is to develop a sophisticated predictive model for the early detection of diabetes. To accomplish this, we are utilizing the Long Short-Term Memory (LSTM) networks, a type of recurrent neural network known for its exceptional ability to handle sequential data and capture long-term dependencies. LSTM networks are particularly well-suited for this task, given the temporal nature of the health data involved in predicting diabetes.

To enhance the model's efficiency, we incorporate Ant Colony Optimization (ACO) for feature selection. ACO, inspired by the foraging behavior of ants, is effective in solving combinatorial optimization problems. By using ACO, we can identify the most relevant features from the dataset, ensuring our model remains both accurate and efficient.

In this chapter, we will detail the experimental setup and evaluation of our predictive model. We will start with an introduction and provide a comprehensive description of the dataset, including its definition, overview, content, statistical summary, distribution plots, and data cleaning process.

Following this, we will describe the development environment, covering both hardware and software aspects. Next, we will outline the tool's features, such as the homepage, prediction interface, diagnosis results, and algorithm results.

Finally, we will discuss the evaluation process, including the metrics used and the performance of our proposed model, and conclude with a comparison of different metrics across algorithms.

5.2 Dataset Description

5.2.1 Definition

The dataset used in our study is a diabetes dataset downloaded from Kaggle. The dataset is in CSV format, with a total of 100,000 records. This great dataset will give the entire training and testing of our model. It comprises nine attributes : Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose Level, and Diabetes Status. This feature is classed as :

- Class 0 for non-diabetic person.
- Class 1 for diabetic person.

5.2.2 Dataset Overview:

The provided dataset's origines are from Kaggle and containes 99999 patient records, each representing a unique individual. It is structured in a CSV format, chosen for ease of handling in Python. The dataset's size is approximately 3722 Ko, facilitating efficient processing and analysis.

5.2.3 Dataset Content:

Below is a description of the attributes present in the dataset:

• Gender: This attribute represents the gender of the individuals, categorized as male, female, or other.

• Age: The age attribute denotes the age of each individual in years.

•Hypertension: This binary attribute indicates whether an individual has hypertension (1) or not (0).

• Heart Disease: Similarly, this binary attribute signifies the presence (1) or absence (0) of heart disease in the individual.

-• Smoking History: This attribute captures the smoking history of individuals, categorizing them as smokers, non-smokers, or ex-smokers.

• BMI (Body Mass Index): The BMI attribute quantifies the body mass index of each individual, calculated using their weight (in kilograms) divided by the square of their height (in meters).

• HbA1c Level: HbA1c (glycated hemoglobin) level is a measure of average blood glucose over the past two to three months. It indicates the percentage of hemoglobin that is bound to glucose.

• Blood Glucose Level: This attribute represents the blood glucose level of individuals, measured in milligrams per deciliter (mg/dL).

• Diabetes: The diabetes attribute serves as the target variable, indicating whether an individual has diabetes (1) or not (0).

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1	80.0	0	1	4	25.19	6.6	140	0
1	1	54.0	0	0	0	27.32	6.6	80	0
2	0	28.0	0	0	4	27.32	5.7	158	0
3	1	36.0	0	0	1	23.45	5.0	155	0
4	0	76.0	1	1	1	20.14	4.8	155	0

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
99995	1	1.000000	0	0	0	0.202031	0.490909	0.045455	0
99996	1	0.024024	0	0	0	0.085901	0.545455	0.090909	0
99997	0	0.824825	0	0	3	0.207983	0.400000	0.340909	0
99998	1	0.299299	0	0	4	0.296569	0.090909	0.090909	0
99999	1	0.712212	0	0	1	0.144958	0.563636	0.045455	0

Figure 5.1: Overview of the dataset

5.2.4 Data Distribution Plots

5.2.4.1 Data visualization

• Smoking history distribution Figure 5.2 shows the smoking history Distribution

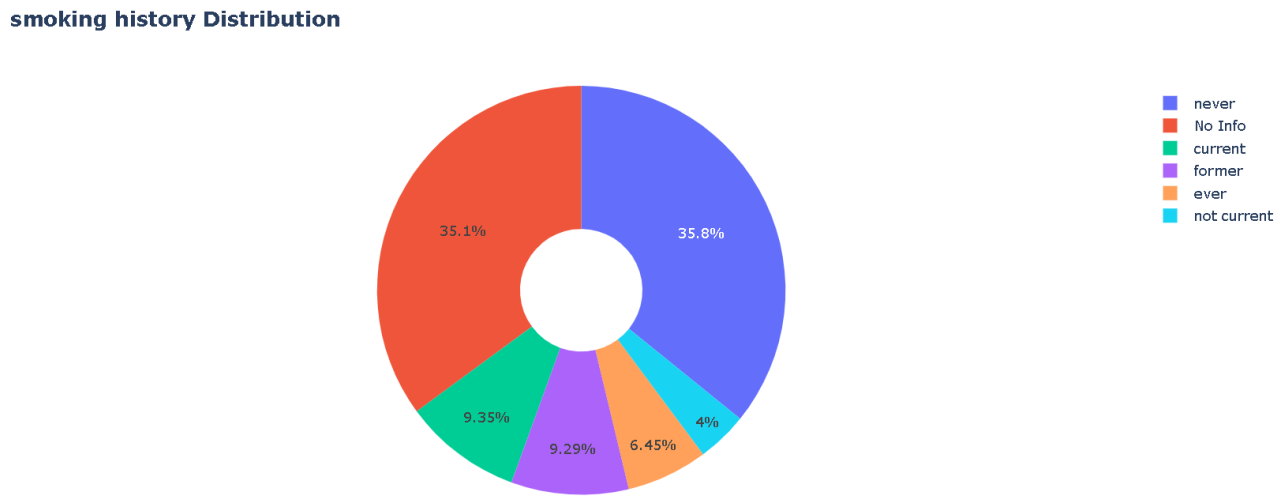


Figure 5.2: Smoking history Distributions:

Figure 5.2 illustrates the distribution of smoking history among individuals in the dataset. The largest groups consist of those who have never smoked (35.0%) and those who are not currently smoking (35.31%). A smaller percentage of individuals are either current smokers (6.45%) or former smokers (15.0%), while a minority fall under the "ever" smoked category (2.39%). Additionally, 9.39% of the data lacks information on smoking history. This distribution provides a clear overview of smoking behavior, which may be a key factor in further analysis.

5.2.4.2 Histogram

Figure 5.3 shows the histograms for each feature, illustrating the data distribution.

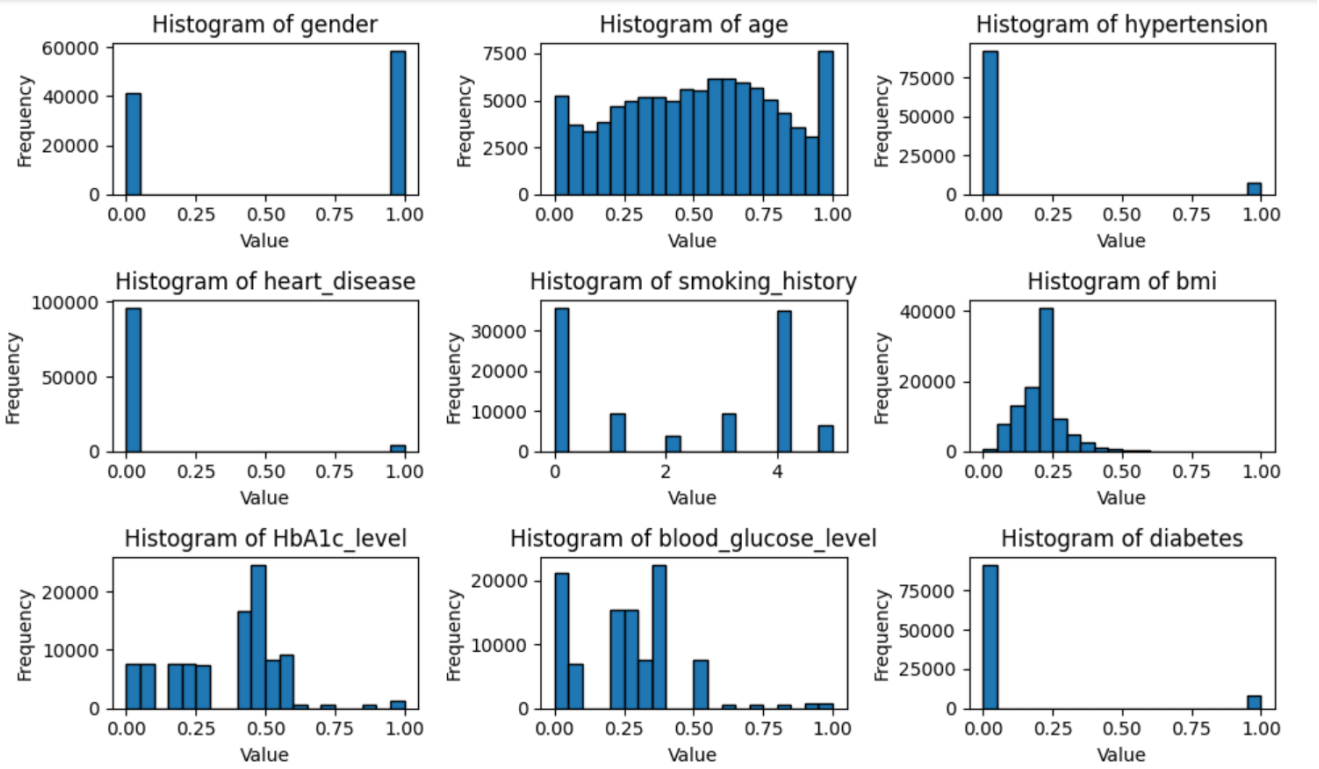


Figure 5.3: Diabetes distribution for each gender

Figure 5.3 presents histograms for various features in the dataset, providing insight into the distribution of key variables. The histograms cover a range of factors including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c levels, blood glucose levels, and diabetes. These visualizations highlight the frequency distribution of each feature, helping to identify patterns, outliers, and the overall structure of the data. This is crucial for understanding how these factors may relate to diabetes risk, which is further explored in the analysis.

5.2.4.3 Data Balance

Figure 5.4 depicts the balance of the dataset, indicating the proportion of diabetic and non-diabetic cases.

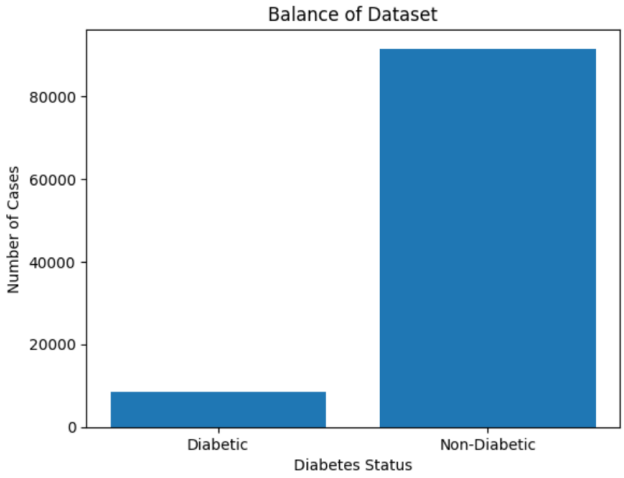


Figure 5.4: Data Balance in the Dataset

The image shows a bar chart titled "Balance of Dataset," which visualizes the distribution of diabetic and non-diabetic cases in a dataset. On the x-axis, two categories are represented: "Diabetic" and "Non-Diabetic." The y-axis represents the "Number of Cases," with values ranging from 0 to over 80,000.

The bar for the "Diabetic" category is significantly smaller, representing a much lower number of cases, around 10,000.

The bar for the "Non-Diabetic" category is much larger, showing a dataset imbalance, with over 80,000 cases.

This chart highlights the imbalance between the two classes, with the non-diabetic cases heavily outnumbering the diabetic ones. This kind of imbalance is typical in medical datasets and suggests the need for balancing techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), to improve model performance during training.

5.2.5 Data Cleaning

5.2.5.1 Handling NaN Values and Outliers

The dataset underwent a thorough cleaning process to handle missing values and outliers. This involved [specific methods]. As explained in Chapter Four, the preprocessing step is crucial to maximize the effectiveness of our model. To achieve this, we tried to remove the NaN values (null values), but as the figure shows below our dataset did not have any NaN values, figure 5.5shows an overview of the dataset.

```
gender                0
age                   0
hypertension          0
heart_disease         0
smoking_history       0
bmi                   0
HbA1c_level           0
blood_glucose_level   0
diabetes              0
dtype: int64
```

Figure 5.5: Overview of the dataset.

5.2.5.2 Converting Nominal Data to Numerical Data

The conversion of nominal data to numerical data is a crucial step in data preprocessing, especially for machine learning and statistical analysis. Nominal data, which represents categories without any intrinsic order, needs to be converted into a numerical format that can be easily interpreted by algorithms. This process ensures that the data can be effectively utilized in various predictive models and analytical tools.

• Handling the "Other" Category in Gender

In our dataset, the "gender" column contains a category labeled "Other," which has only 18 records. Given the size of our dataset, which includes nearly 100,000 records, the presence of these 18 records is statistically insignificant. To maintain the integrity and simplicity of our analysis, we have decided to remove the "Other" category. This step ensures that the dataset remains focused and free of noise that could potentially skew the results.

• Gender Encoding

To facilitate the conversion of the "gender" column, we have mapped the categories to numerical values:

-Female = 1

-Male = 0

This binary encoding allows for efficient processing and analysis of the gender data, enabling the machine learning algorithms to better understand the distinctions between the two categories.

• Smoking History Encoding

Similarly, the "smoking history" column contains various categories that need to be converted into numerical values. The following mapping has been applied:

-No Info = 0

-Current = 1

-Ever = 2

-Former = 3

-Never = 4

By assigning numerical values to these categories, we ensure that the smoking history data is interpret-able by our models, allowing for more accurate predictions and insights.

In summary, converting nominal data to numerical data enhances the compatibility, performance, and analytical capabilities of machine learning models. By addressing the "Other" category in gender and applying appropriate numerical mappings, we ensure that our dataset is ready for advanced analysis and predictive modeling.Figure 5.6 and figure 5.7 illustrates the data before and after encoding

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Figure 5.6: The data before encoding

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	1	80.0	0	1	4	25.19	6.6	140	0
1	1	54.0	0	0	0	27.32	6.6	80	0
2	0	28.0	0	0	4	27.32	5.7	158	0
3	1	36.0	0	0	1	23.45	5.0	155	0
4	0	76.0	1	1	1	20.14	4.8	155	0

Figure 5.7: The data after encoding/

5.2.5.3 Removing duplicates

Removing duplicates is an essential data preprocessing step in data analysis and machine learning. Duplicate entries can lead to misleading results, inflated statistics, and overfitting in models. By ensuring that each row in the dataset is unique, we can maintain the integrity and accuracy of our analyses.

In our dataset, we performed this step to ensure data quality. As shown in the figure below, the number of columns in our dataset remained the same after this process. This indicates that there were no duplicate rows, confirming the uniqueness of our data entries and ensuring that our subsequent analysis is based on clean and reliable data.

```
df.drop_duplicates(inplace=True)
df.tail()
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	d:
99994	1	36.0	0	0	0	24.60	4.8	145	
99996	1	2.0	0	0	0	17.37	6.5	100	
99997	0	66.0	0	0	3	27.83	5.7	155	
99998	1	24.0	0	0	4	35.42	4.0	100	
99999	1	57.0	0	0	1	22.43	6.6	90	

Figure 5.8: removing duplicates

5.2.5.4 Correlation Matrix

A correlation matrix is a chart that shows the correlation values among multiple variables. Every cell in the chart displays the connection between two factors. Correlation coefficients measure how strong and in which direction a linear connection exists between two variables. Figure 5.9 presents the correlation matrix, highlighting the relationships between different features.

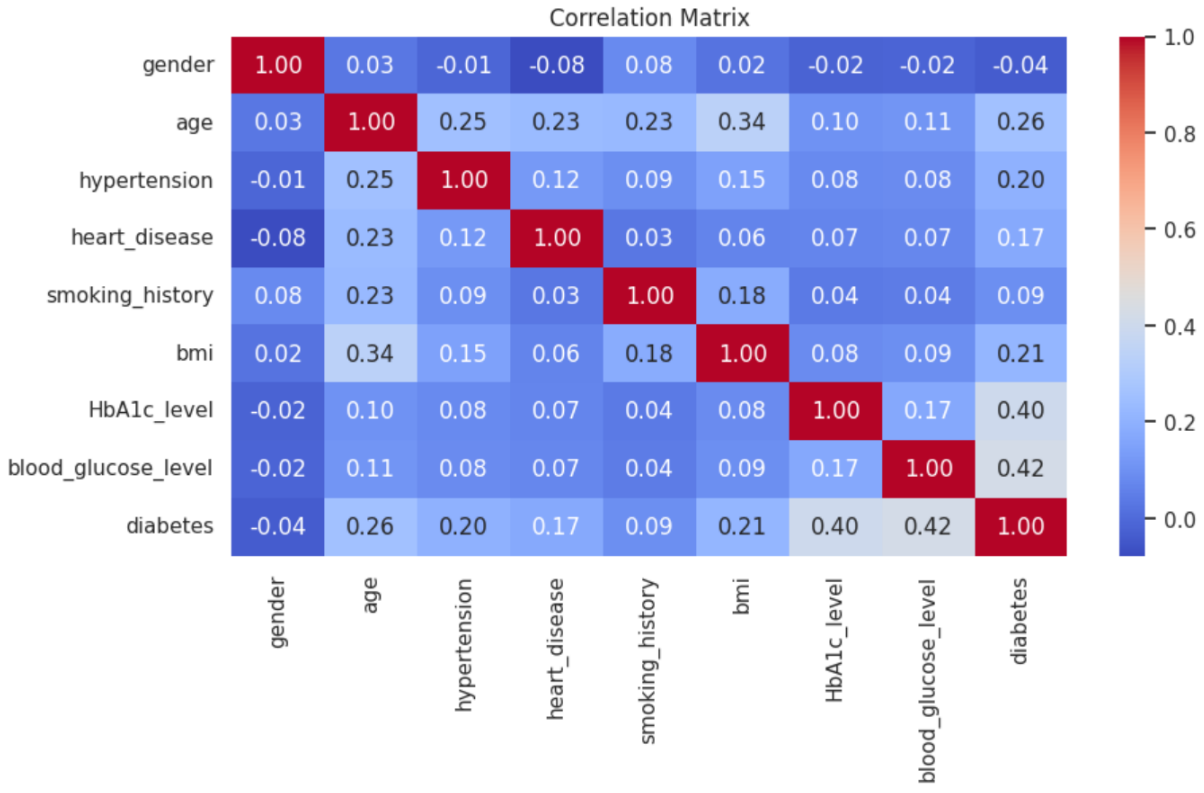


Figure 5.9: Correlation Matrix of the Dataset

5.3 Development Environment

5.3.1 Hardware Environment

The experiments were conducted on a machine with the following listed specifications:
PC1
• Machine type: Dell XPS 13 9365
• Processor: Intel(R) Core(TM) i5-8200Y CPU @ 1.30GHz 1.60 GHz
• RAM: 8,00 Go
• Operating system: Windows 11 Professional
PC 2
• Machine type: HP Pro book 430 G4
• Processor: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz
• RAM: 8Go
• Operating system: Windows 10 Professional

5.3.2 Software Environment

5.3.2.1 Programming Language

Python was the primary programming language used for implementing the models.

It is a popular programming language recognized for its ease of use and clear syntax. It backs various programming paradigms such as procedural, object-oriented, and functional programming styles. Python is extensively utilized across different fields including web development, scientific computing, data analysis, artificial intelligence, and automation[47].

5.3.2.2 Python Libraries

The implementation leveraged several Python libraries including :

Numpy: - NumPy is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.[48].

Pandas : - Pandas is a powerful library for data manipulation and analysis. It offers data structures like DataFrames and Series, which allow you to handle labeled and relational data easily. Pandas provides tools for reading/writing data, reshaping, merging, and performing calculations on data.[49].

Sklearn: - Scikit-learn (sklearn) is a comprehensive library for machine learning in Python. It includes a wide range of tools for data preprocessing, model selection, evaluation, and many machine learning algorithms like classification, regression, clustering, and dimensionality reduction.[50].

Keras: - Keras is an open-source deep learning library that provides a high-level interface for neural networks. It allows for easy and fast prototyping, supports both convolutional networks and recurrent networks (like LSTM,GRU), and can run on top of TensorFlow, CNTK, or Theano.[51].

Tensorflow: - TensorFlow is an end-to-end open-source platform for machine learning. 'tensorflow.keras' provides an interface for defining and training deep learning models using Keras API within TensorFlow. It includes support for various neural network architectures, optimization algorithms, and utilities for data preprocessing.[52].

Matplotlib: - Matplotlib is a plotting library for Python that provides a MATLAB-like interface for creating static, animated, and interactive visualizations. It can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.[53].

Scipy : - SciPy is a library used for scientific and technical computing in Python. It builds on NumPy and provides additional functions that operate on NumPy arrays and are useful for different types of scientific and engineering applications.[54].

Scikeras : -Scikeras is a library that bridges the gap between scikit-learn and Keras, allowing you to use Keras models as scikit-learn estimators. It provides wrappers that enable Keras models to be used seamlessly within scikit-learn pipelines, enabling easy integration of deep learning models with traditional machine learning workflows.[55].

Seaborn -Also known as SNS, is a data visualization library in Python based on Matplotlib, providing a high-level interface to draw great deals of informative and attractive statistical graphics. It is beneficial for visualizing complex data sets and statistical relations concisely and attractively[56].

These libraries together provide a comprehensive ecosystem for data handling, machine learning, deep learning, scientific computing, and visualization in Python, supporting a wide range of tasks from data pre-processing and model training to evaluation and visualization.

5.3.2.3 Strapi

Strapi is an open-source headless content management system (CMS) designed to facilitate the creation and management of APIs. It provides a customizable and user-friendly interface for content management, allowing developers to define content types, manage data, and integrate seamlessly with various front-end frameworks. Strapi supports RESTful and GraphQL APIs, making it a versatile choice for modern web applications and enabling efficient data handling and delivery.

5.4 Description of the Tool

5.4.1 Homepage

The tool's homepage provides an overview and access to different functionalities, as shown in Figure 5.10.

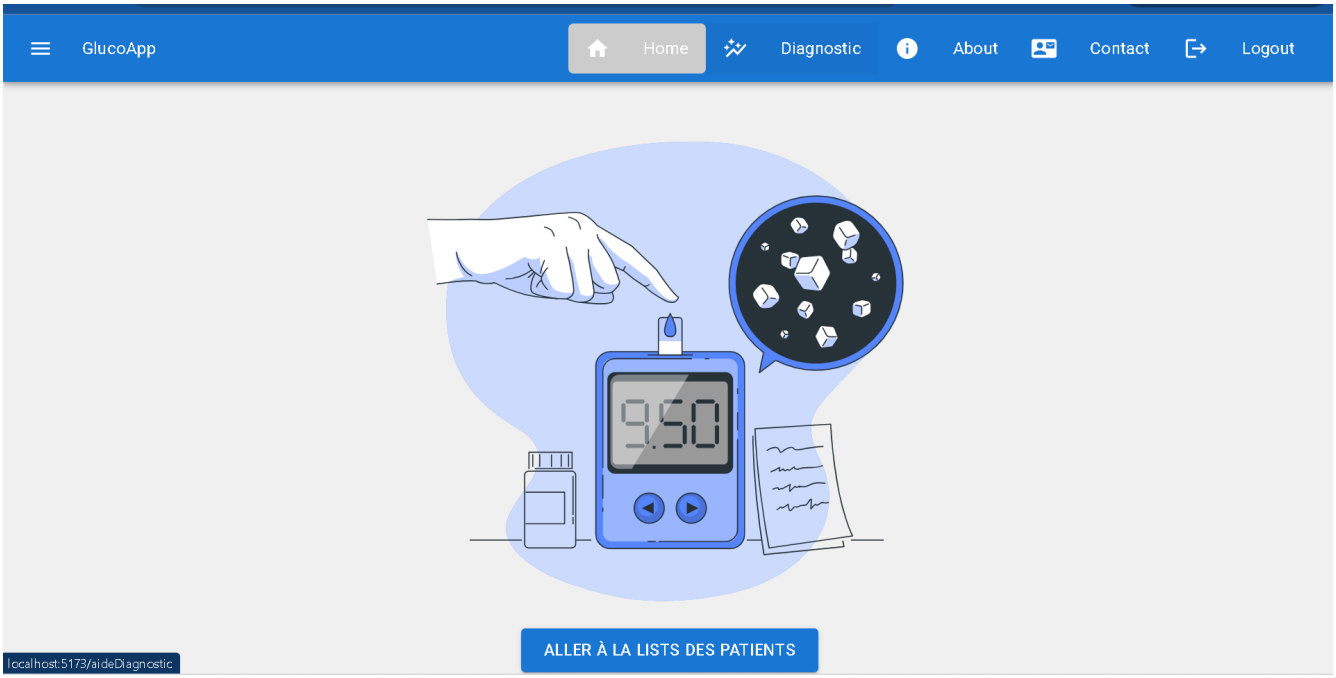


Figure 5.10: Homepage of the Tool

5.4.2 Prediction Interface

The prediction interface allows users to input relevant data and obtain diabetes risk predictions, illustrated in Figures bellow.

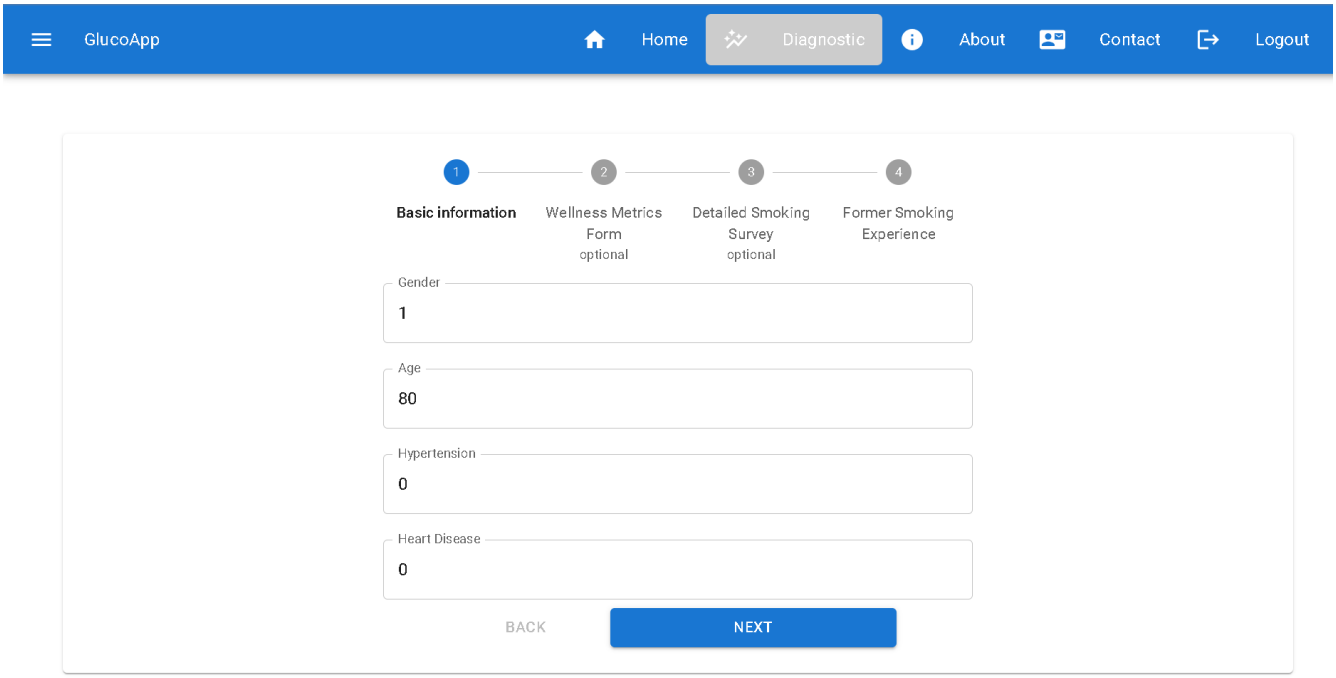


Figure 5.11: Diagnosis Interface 1

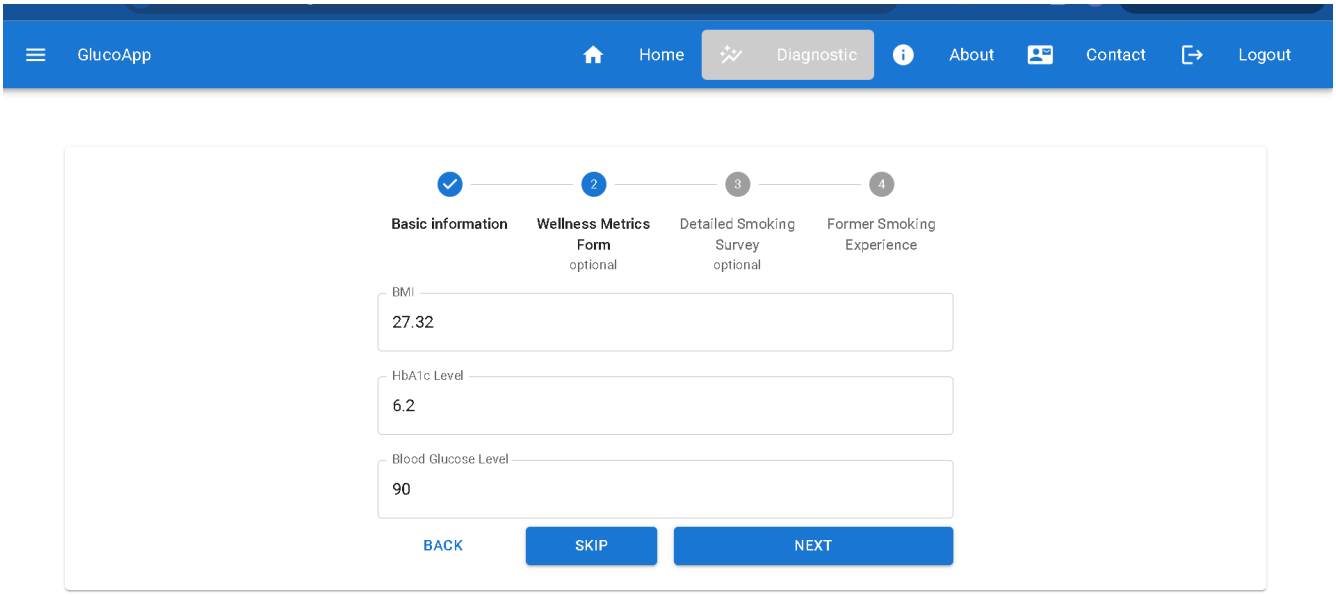


Figure 5.12: Diagnosis Interface2

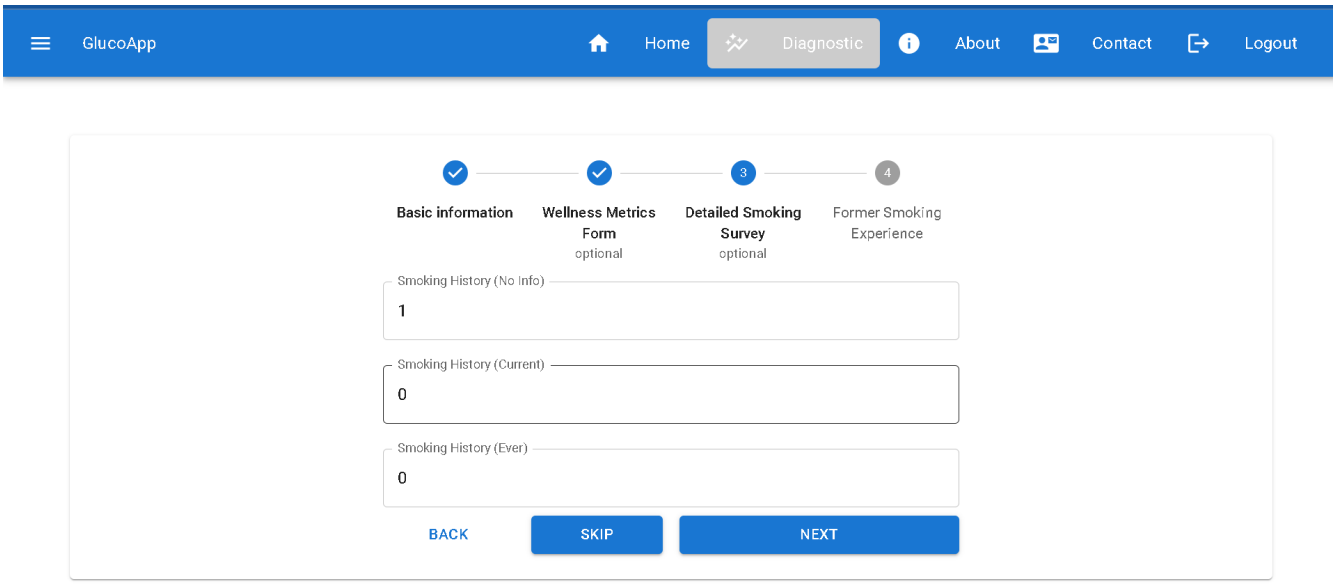


Figure 5.13: Diagnosis Interface 3

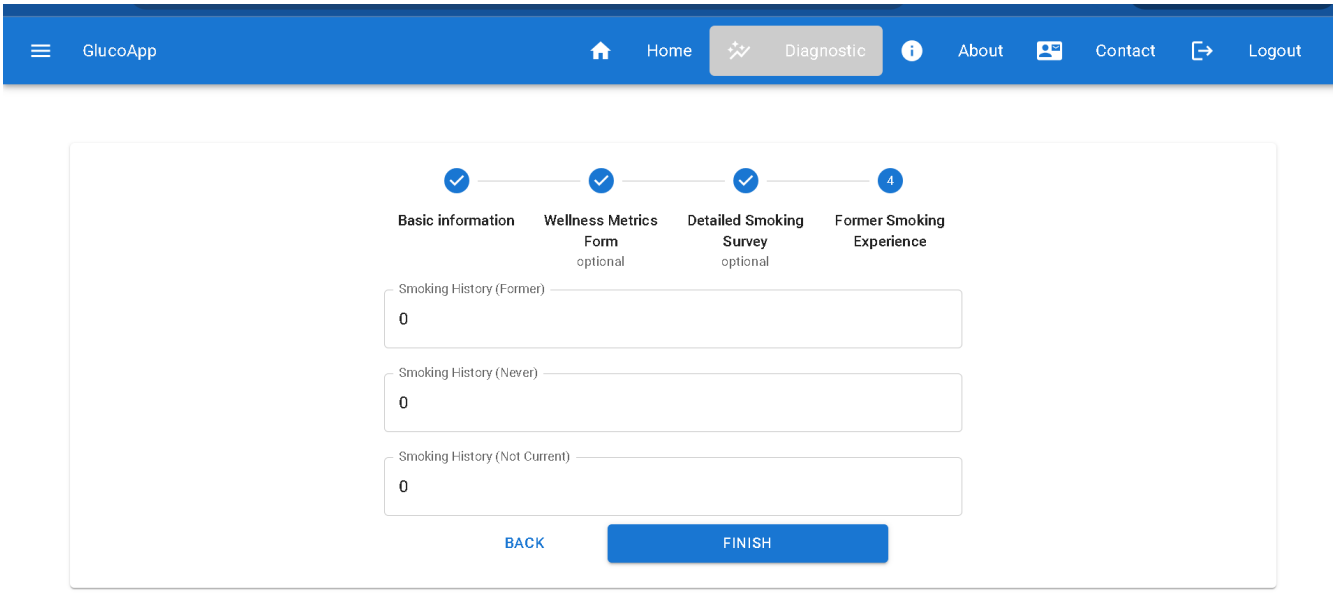


Figure 5.14: Diagnosis Interface 4

Here, we present the implementation of the prediction interface for our application, GlucoApp. On the left, you see the main dashboard, where users can navigate through different sections such as home, diagnostics, and contact. This intuitive layout ensures easy access to all functionalities. On the right, we showcase the diagnostic process, which includes several optional steps like entering basic information, wellness metrics, and smoking history.

5.4.3 Diagnosis Results

The diagnosis result Interface is illustrated in Figure 5.15

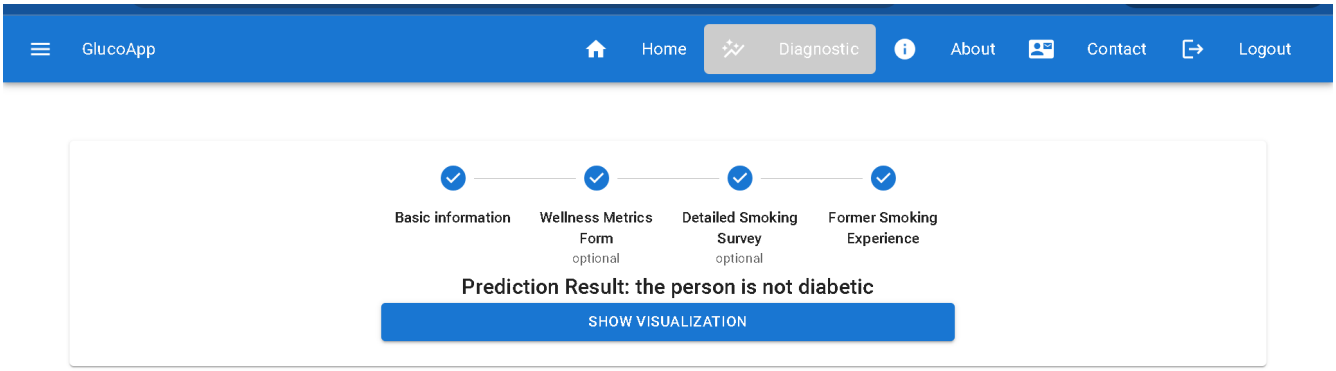


Figure 5.15: Diagnosis result Interface

The prediction result indicates whether a person is diabetic or not, with an option to view detailed visualizations. This user-friendly interface aims to simplify diabetes risk assessment for users.

5.4.4 Algorithm Results

The performance of different algorithms is summarized in Table 5.1, comparing accuracy, precision, recall, and ROC AUC.

Table 5.1: Algorithm Results

Model	Accuracy	MAE	RMSE	MEDAE	R2 Score	ROC AUC
LSTM	0.95	0.041	0.20	0.0	0.47	0.96
ACO-LSTM	0.971	0.029	0.17	0.0	0.64	0.98
PSO-LSTM	0.938	0.061	0.24	0.0	0.32	0.86
GWO-LSTM	0.931	0.068	0.26	0.0	0.39	0.90
EHHO-LSTM	0.911	0.0888	0.2979	0.0	0.08	0.78
ACO-GRU	0.938	0.089	0.22	0.011	0.398	0.924

5.5 Evaluation

5.5.1 Evaluation Metrics

In the following subsection, we will define each classification metric: Accuracy, Precision, and the ROC (Receiver Operating Characteristic) Curve. Additionally, we will cover regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Median Absolute Error (MEDAE), and R-Squared (r2 Score).

5.5.1.1 Accuracy

Accuracy is the general correctness of a model in classification. It shows what part of the total of the correct predictions came out. This is quite common for a classification problem[57]. In binary or two-class issues, it can be calculated as follows:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of made predictions}}$$

The breakdown of these components is as follows:

1. **True Positive (TP)**: These are instances correctly predicted as positive by the model.

2. **True Negative (TN)**: These are instances which the model has correctly predicted as unfavorable.

3. In general, the model's total predictions are divided into true positives, true negatives, false positives, and false negatives.

An overall estimation of how effective a classification model is at making the correct positive and negative predictions.

5.5.1.2 Receiver Operating Characteristic (ROC)

The ROC curve plots the sensitivity of a test against one minus its specificity for all possible threshold values of the marker under study. Sensitivity is the percentage of correctly identified ill cases, while specificity is the percentage of correctly identified healthy people[58].

5.5.1.3 Mean Absolute Error (MAE)

MAE is the average of absolute differences between predicted and actual values[59], which can be calculated as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{5.1}$$

5.5.1.4 Root Mean Squared Error (RMSE)

RMSE is the square root of the average of the squared differences between predicted and actual values[60]. It can be found using the formula:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5.2}$$

5.5.1.5 Median Absolute Error (MEDAE)

It is defined as the median of the absolute differences between predicted and actual values. Equivalently, it can be considered the middle value of the absolute error for a datum[61].

5.5.1.6 R-Squared (r2Score)

R-Squared is a statistical measure of the proportion of the variance for a dependent variable that's explained by an independent variable[61]. It can be calculated using the formula:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{5.3}$$

5.5.2 Evaluation of the Proposed Model

In this subsection, we evaluate the performance of the proposed model using various metrics. The performance of different variations of the LSTM model, including ACO-LSTM, PSO-LSTM, GWO-LSTM, EHHO-LSTM, and ACO-GRU, are compared and analyzed. The following tables and figures present the detailed results for each model.

5.5.2.1 LSTM Model

Table 5.2 presents the performance metrics for the LSTM model.

Table 5.2: Performance Metrics for LSTM Model

Metric	Value
Accuracy	0.95
MAE	0.041
RMSE	0.20
MEDAE	0.0
R2 Score	0.47
ROC AUC	0.96

Here's the roc curve of the LSTM model

```
Overall Accuracy: 0.9589929580688477
2188/2188 [==============================] - 4s 2ms/step
Mean Absolute Error: 0.04100702977653312
Root Mean Squared Error: 0.2025019253650027
Median Absolute Error: 0.0
R2 Score: 0.4767354723107037
ROC AUC Score: 0.9600698932952072
```

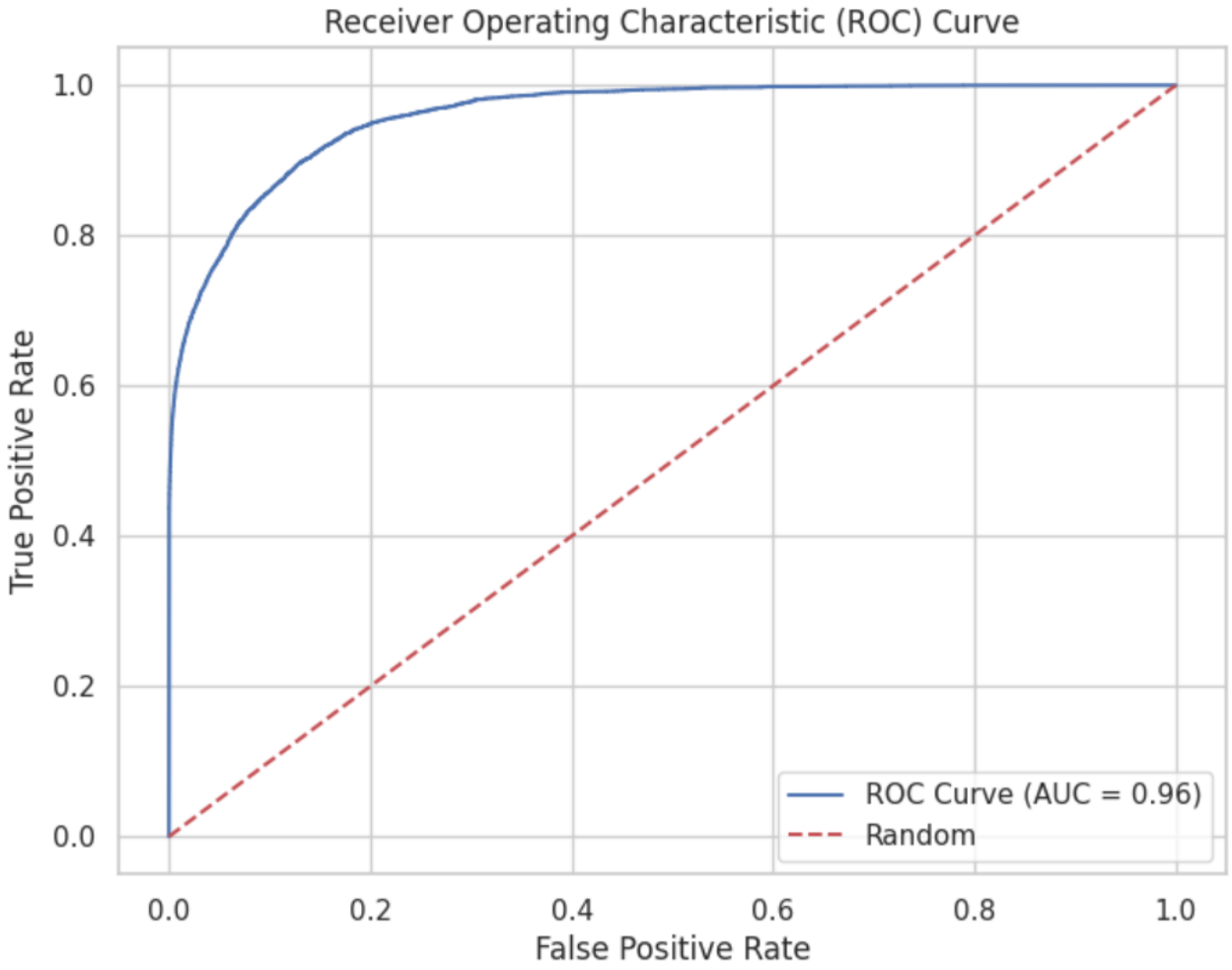


Figure 5.16: LSTM roc curve

This figure presents the Receiver Operating Characteristic (ROC) curve for an LSTM model. Key information displayed at the top includes:

Overall Accuracy: 95.89%

Mean Absolute Error: 0.041

Root Mean Squared Error: 0.202

Median Absolute Error: 0.0

R^2 Score: 0.476

ROC AUC Score: 0.9607

Below this is the ROC curve, which plots the True Positive Rate (y-axis) against the False Positive Rate (x-axis). The curve is shown in blue, with an Area Under the Curve (AUC) of 0.96, indicating a strong performance. The red dashed line represents a random classifier, where the True Positive Rate equals the False Positive Rate (AUC = 0.5).

The curve's proximity to the upper left corner reflects the model's excellent discrimination ability, with an AUC of 0.96, demonstrating that the LSTM model has a high capacity to distinguish between the positive and negative classes.

5.5.2.2 ACO-LSTM Model

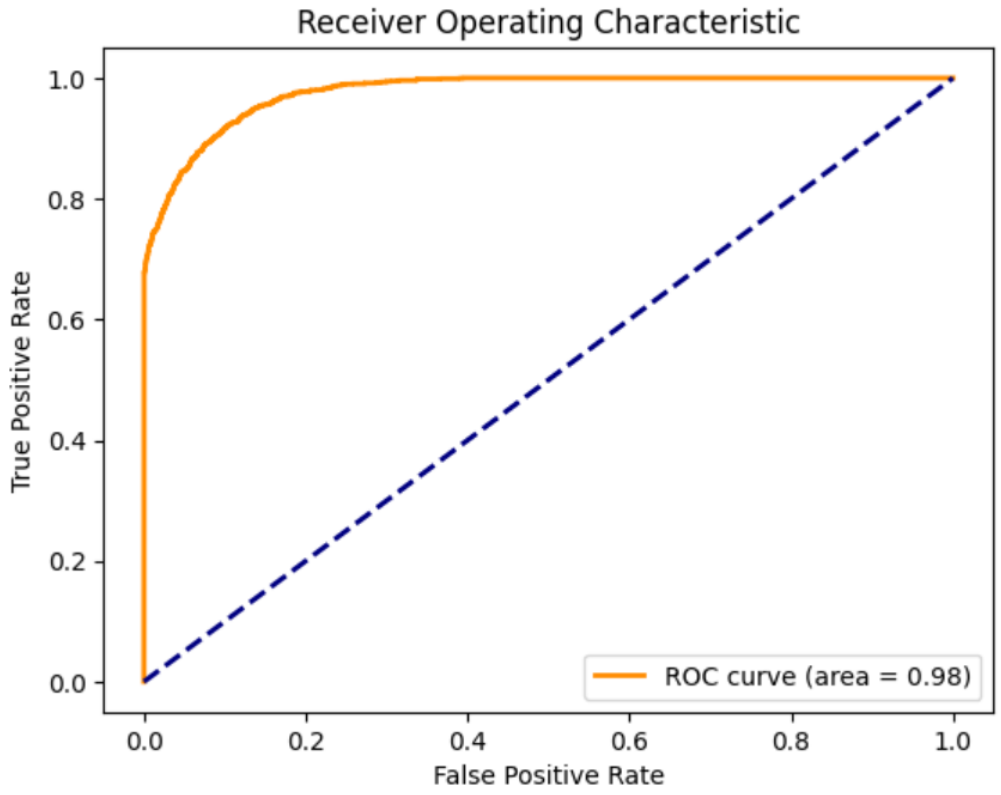
Table 5.3 shows the performance metrics for the ACO-LSTM model.

Table 5.3: Performance Metrics for ACO-LSTM Model

Metric	Value
Accuracy	0.971
MAE	0.029
RMSE	0.17
MEDAE	0.0
R2 Score	0.64
ROC AUC	0.98

Here's the roc curve of the ACO-LSTM model

```
Best Accuracy: 0.9710
```




```
Mean Absolute Error: 0.0290
Median Absolute Error: 0.0000
R2 Score: 0.6420
Root Mean Squared Error: 0.1702
```

Figure 5.17: ACO-LSTM roc curve

5.5.2.3 PSO-LSTM Model

Table 5.4 details the performance metrics for the PSO-LSTM model.

Table 5.4: Performance Metrics for PSO-LSTM Model

Metric	Value
Accuracy	0.938
MAE	0.061
RMSE	0.24
MEDAE	0.0
R2 Score	0.32
ROC AUC	0.86

Here's the roc curve of the pso-lstm model

```
2812/2812 [==============================] - 8s 3ms/step
Accuracy: 0.9389891536273115
Mean Absolute Error (MAE): 0.06101084637268848
Root Mean Squared Error (RMSE): 0.24700373756825722
Median Absolute Error (MEDAE): 0.0
R^2 Score: 0.323935840146232
```

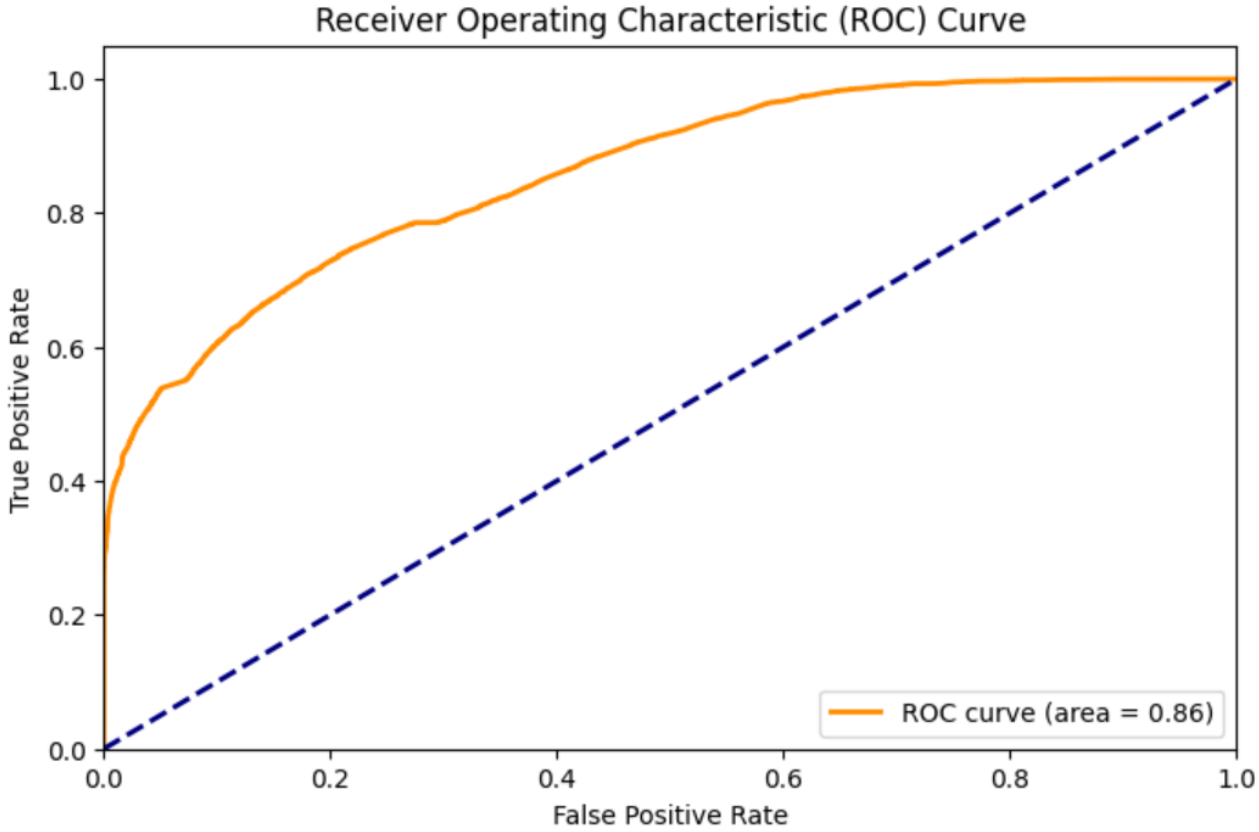


Figure 5.18: PSO-LSTM roc curve.

This image displays the ROC curve for the PSO-LSTM (Particle Swarm Optimization paired with LSTM) model. Key metrics shown at the top include:

Accuracy: 93.89

Mean Absolute Error (MAE): 0.0461

Root Mean Squared Error (RMSE): 0.247

Median Absolute Error (XEDAE): 0.0

R^2 Score: 0.323

The ROC curve, in orange, plots the True Positive Rate against the False Positive Rate, with an AUC (Area Under the Curve) of 0.86. The dotted blue line represents the performance of a random classifier (AUC = 0.5). The ROC curve's AUC value of 0.86 indicates good but not perfect classification performance.

This figure, labeled "Figure 5.18: PSO-LSTM ROC Curve," highlights that while the PSO-LSTM model is effective, its classification ability is not as strong as models with higher AUC scores, like the LSTM in the previous figure.

5.5.2.4 GWO-LSTM Model

Table 5.5 provides the performance metrics for the GWO-LSTM model.

Table 5.5: Performance Metrics for GWO-LSTM Model

Metric	Value
Accuracy	0.931
MAE	0.068
RMSE	0.26
MEDAE	0.0
R2 Score	0.39
ROC AUC	0.90

Here's the roc curve of the GWO-LSTM model

```
Final model accuracy: 0.9319999814033508
Mean Absolute Error (MAE): 0.068
Median Absolute Error (MedAE): 0.0
Root Mean Squared Error (RMSE): 0.260768096208106
R2 Score: 0.3908882791400822
AUC-ROC: 0.9016935759204396
```

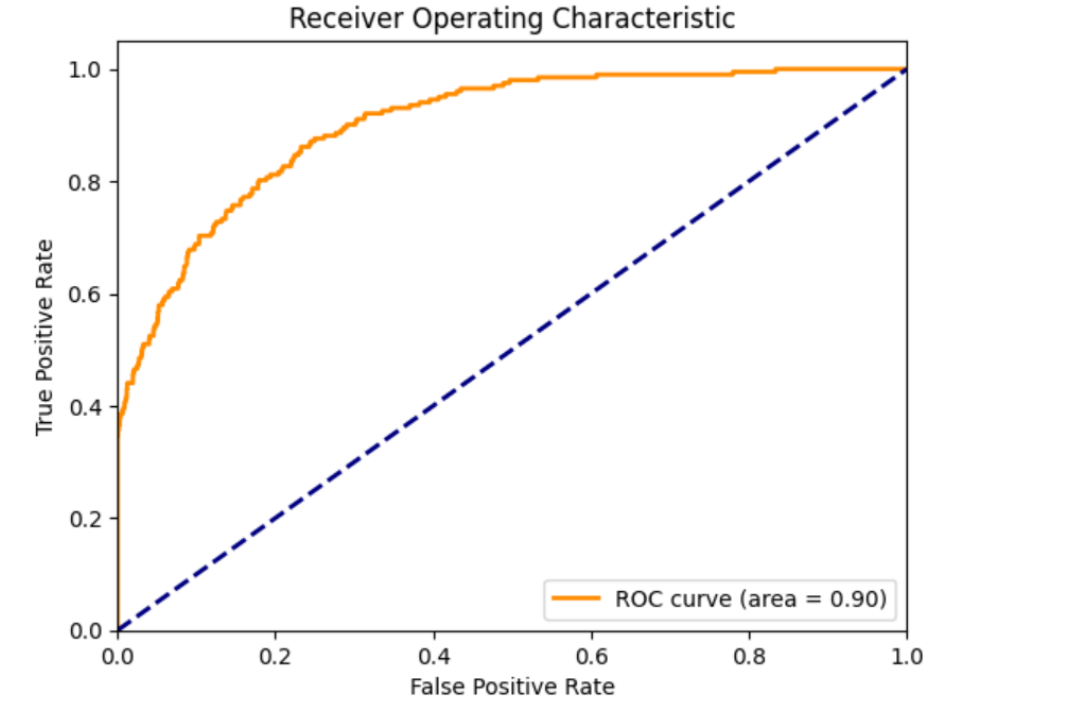


Figure 5.19: GWO-LSTM roc curve.

5.5.2.5 EHHO-LSTM Model

Table 5.6 provides the performance metrics for the EHHO-LSTM model.

Table 5.6: Performance Metrics for EHHO-LSTM Model

Metric	Value
Accuracy	0.911
MAE	0.0888
RMSE	0.297
MEDAE	0.0
R2 Score	0.08
ROC AUC	0.78

Here's the roc curve of the EHHO-LSTM model

```
Accuracy: 0.9112367033958435
625/625 [==============================] - 2s 2ms/step
Mean Absolute Error (MAE): 0.0888
Root Mean Squared Error (RMSE): 0.2979
Median Absolute Error (MEDAE): 0.0000
R2 Score: 0.0800
```

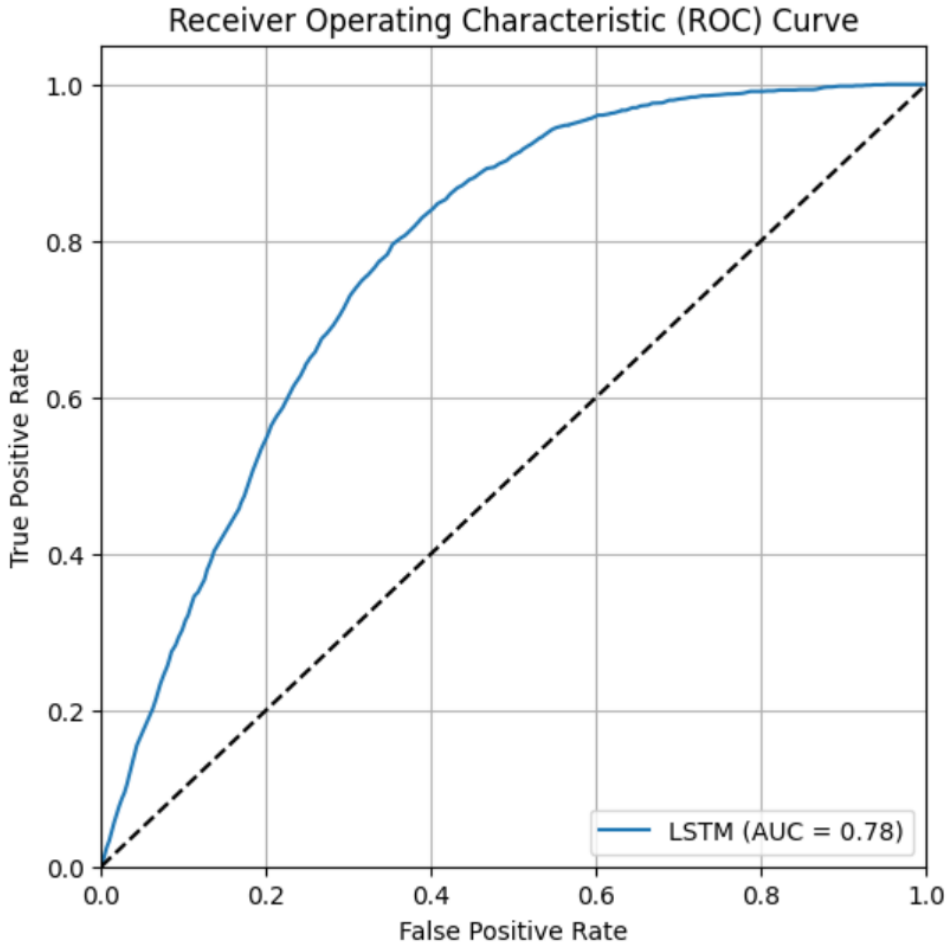


Figure 5.20: EHHO-LSTM roc curve.

5.5.2.6 ACO-GRU Model

Table 5.7 provides the performance metrics for the ACO-GRU model.

Table 5.7: Performance Metrics for ACO-GRU Model

Metric	Value
Accuracy	0.938
MAE	0.089
RMSE	0.22
MEDAE	0.011
R2 Score	0.398
ROC AUC	0.924

Here's the roc curve of the ACO-GRU model

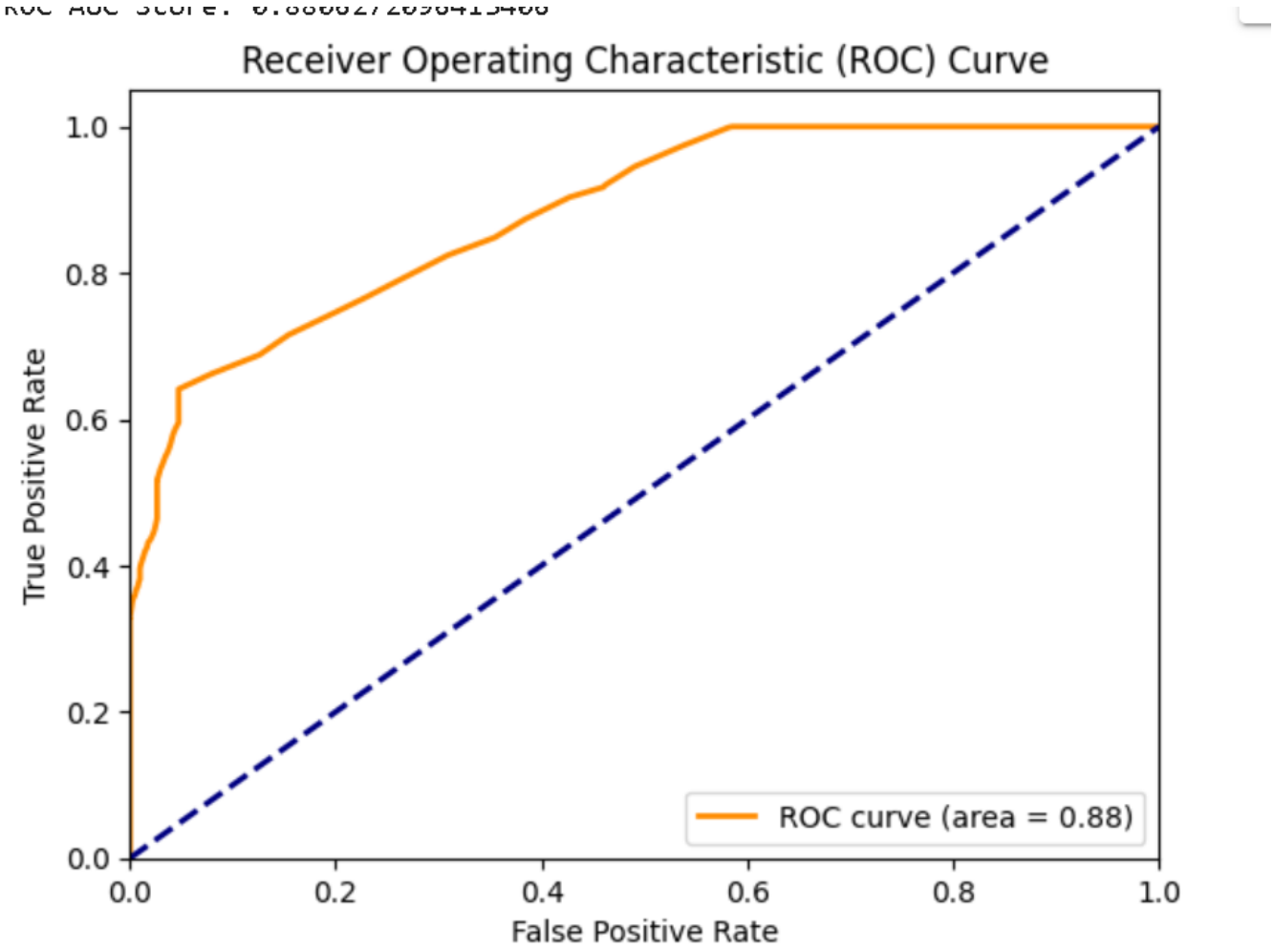


Figure 5.21: ACO-GRU roc curve.

The ROC curve, or Receiver Operating Characteristic curve, is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is

varied. The closer the ROC curve is to the top left corner of the plot, the better the model is performing. This indicates a higher true positive rate (sensitivity) and a lower false positive rate (1-specificity). The area under the curve (AUC) quantifies the overall ability of the model to discriminate between positive and negative classes. An AUC of 1.0 represents a perfect model, while an AUC of 0.5 suggests no discriminative ability, equivalent to random guessing.

In our analysis, we compared six different models using their ROC curves. Among these, the ACO-LSTM model demonstrated the highest AUC, indicating superior performance. This means that the ACO-LSTM model is more effective at distinguishing between classes than the other models. The closer proximity of its ROC curve to the top left corner further validates its higher accuracy and reliability in classification tasks. Thus, based on these ROC curve evaluations, we can confidently conclude that ACO-LSTM is the best-performing model among those tested.

5.5.3 Comparison of different metrics across algorithms

In this subsection, we present a comprehensive comparison of various performance metrics across different algorithms. The figures below illustrate the accuracy, root mean square error (RMSE), mean absolute error (MAE), receiver operating characteristic (ROC), and R2 score for each algorithm evaluated.

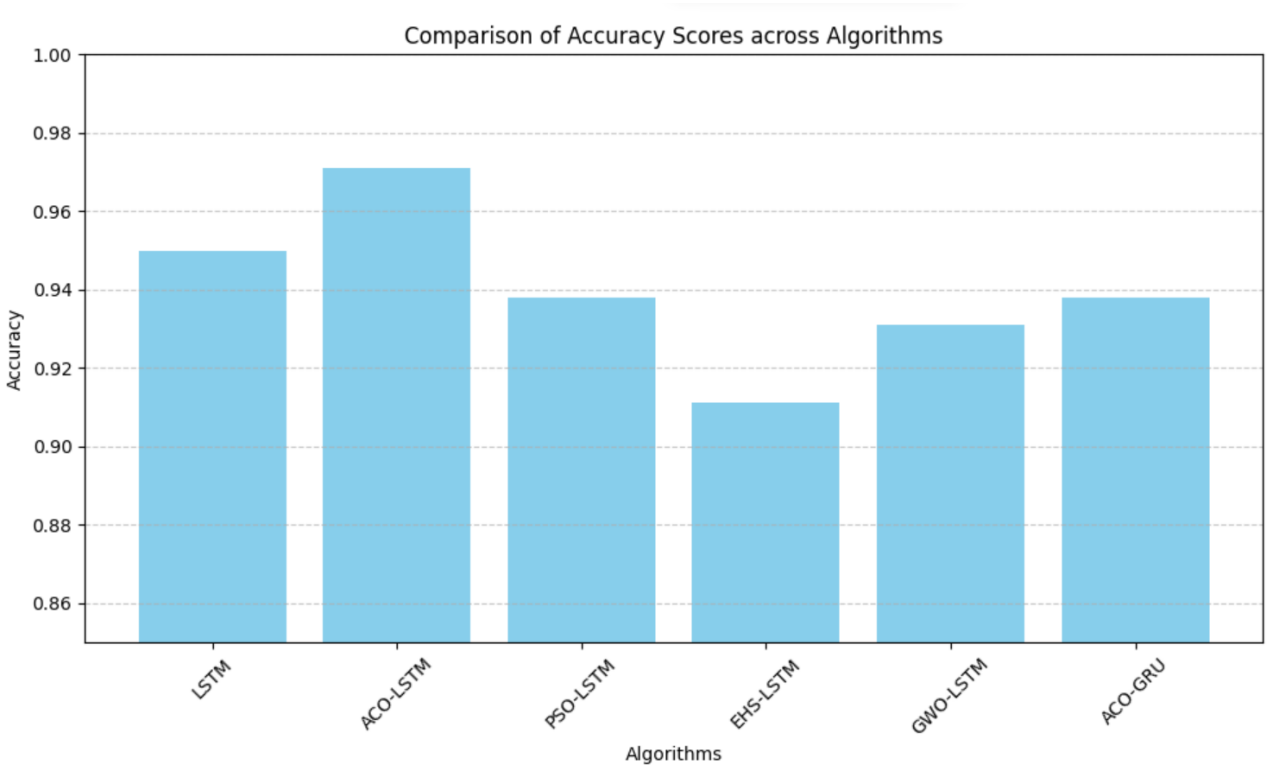


Figure 5.22: Accuracy comparison.

The first chart illustrates the accuracy scores of different algorithms for diabetes prediction. ACO-LSTM demonstrates the highest accuracy, indicating its strong predictive capabilities by combining Ant Colony Optimization with Long Short-Term Memory networks. This approach outperforms the others, including LSTM alone and ACO-GRU, highlighting the effectiveness of hybrid models in improving performance. The results suggest that incorporating swarm intelligence techniques with deep learning models enhances the predictive power significantly.

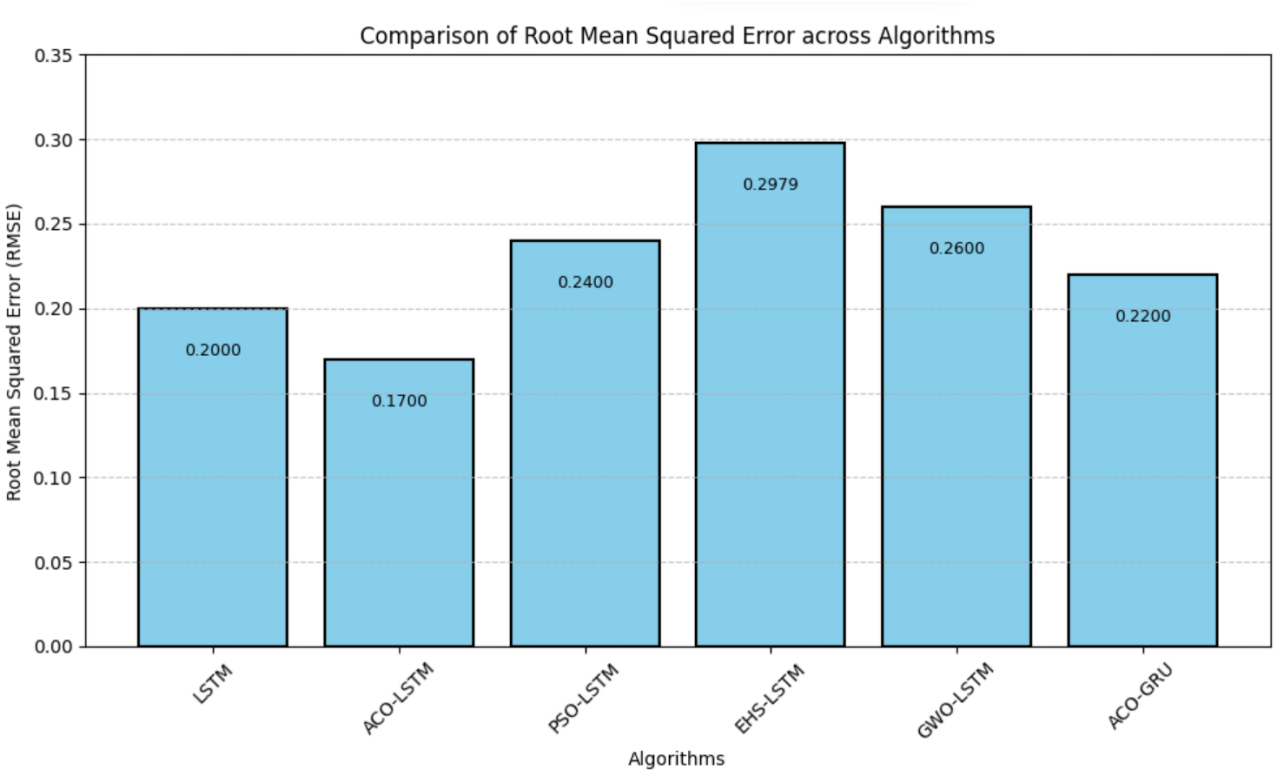


Figure 5.23: RMSE comparison.

The second chart shows the Root Mean Squared Error (RMSE) for the same algorithms. ACO-LSTM again exhibits the lowest RMSE, confirming its high accuracy and reliability in predictions. Lower RMSE values correspond to better predictive performance, indicating that ACO-LSTM not only predicts with high accuracy but also maintains consistency across predictions. Other models like ACO-GRU also perform well, but ACO-LSTM stands out as the most effective in minimizing error, further emphasizing the advantage of hybrid approaches.

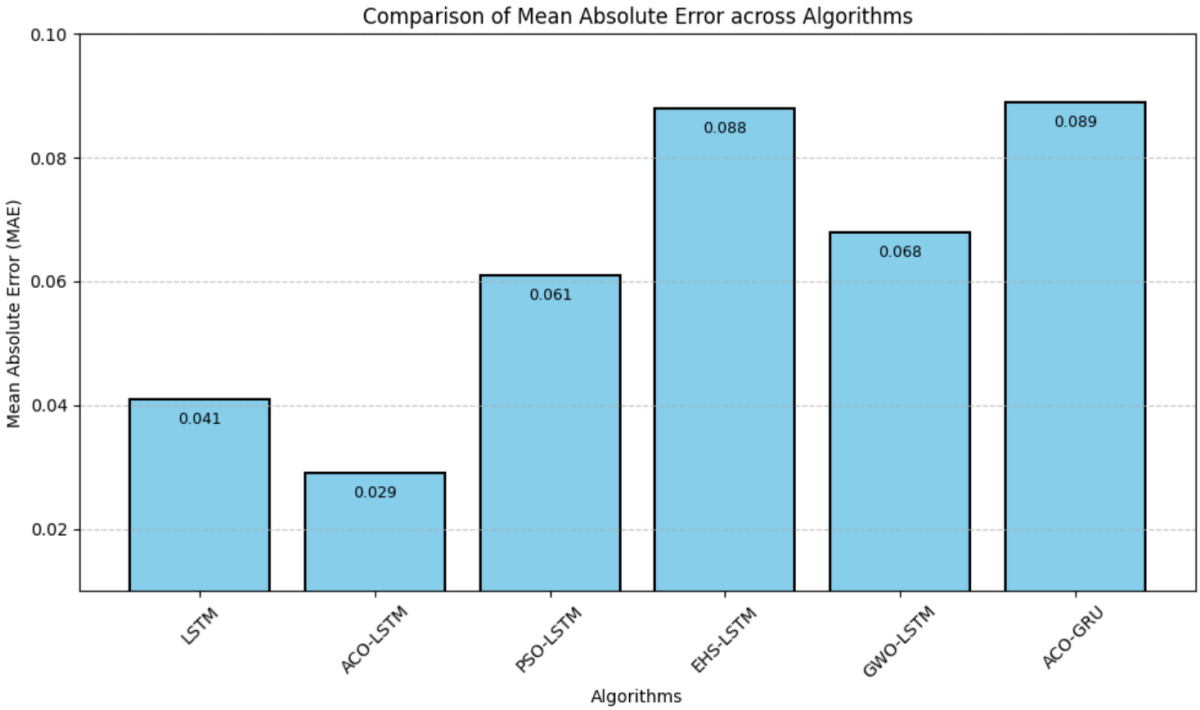


Figure 5.24: MAE comparaison.

In the MAE chart, ACO-LSTM displays the lowest Mean Absolute Error, confirming its superior accuracy in diabetes prediction. This low error rate underscores the model's precision and consistency. Following closely is the LSTM model, demonstrating effective performance but slightly less optimal compared to ACO-LSTM. Conversely, ACO-GRU shows the highest MAE, indicating that while it is still a viable model, it lacks the precision seen in ACO-LSTM and LSTM. This comparison highlights the importance of model selection in achieving minimal error in predictive tasks.

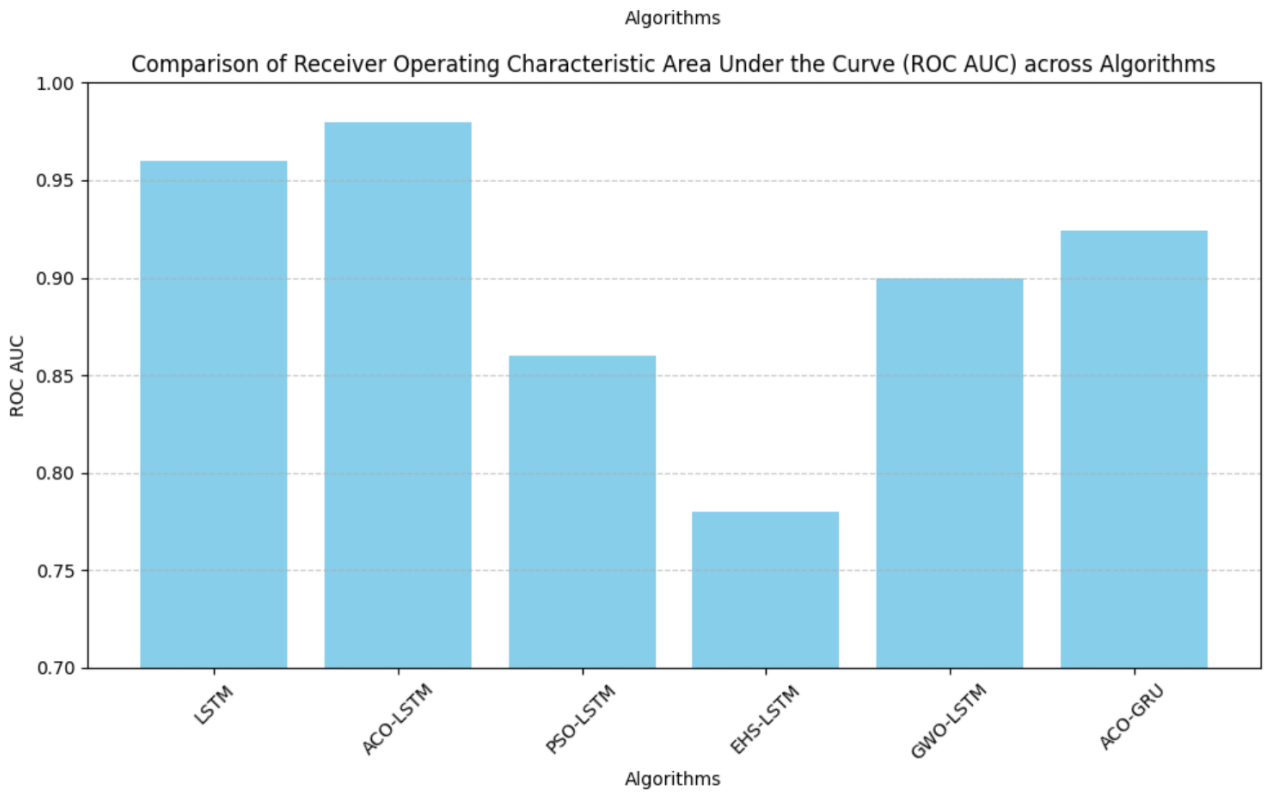


Figure 5.25: ROC comparison.

In the ROC comparison, ACO-LSTM achieves the highest value, approaching 1, which indicates its excellent capability in distinguishing between classes for diabetes prediction. This high ROC value signifies strong predictive performance and accuracy. Other models show varying levels of effectiveness, but ACO-LSTM stands out as the most reliable and precise. This highlights the importance of selecting models that maximize the area under the ROC curve for optimal classification outcomes in healthcare applications.

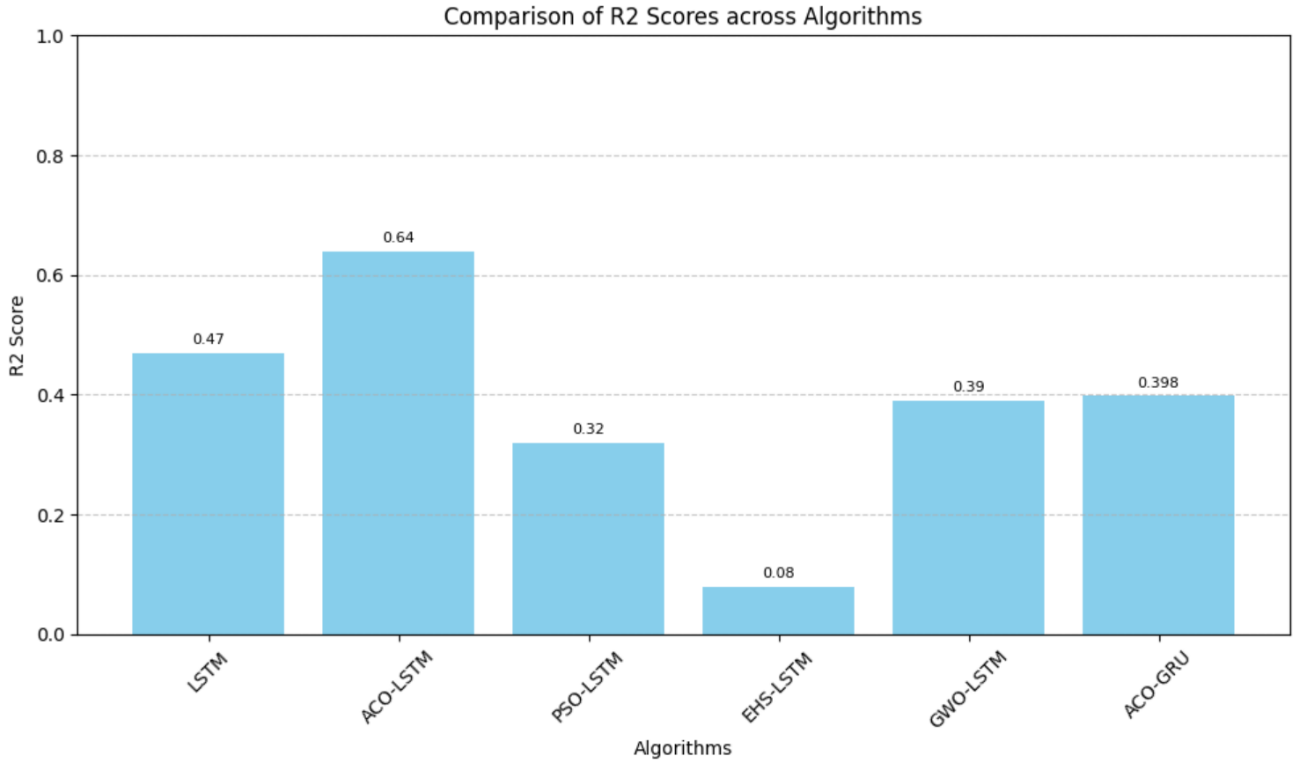


Figure 5.26: R2 Score comparison

In the R2 score comparison, ACO-LSTM achieves the highest value, indicating superior predictive accuracy and model fit. Following closely is LSTM, which also demonstrates strong performance. On the other hand, EHHO-LSTM has the lowest R2 score, suggesting less accuracy in capturing the variability of the data. This highlights the effectiveness of ACO-LSTM in modeling complex relationships within the dataset, making it the preferred choice for diabetes prediction tasks.

5.6 Conclusion

In this chapter, we conducted a comprehensive evaluation of various hybrid models designed for the early prediction of diabetes. Our primary objective was to ascertain the most accurate model by combining different swarm intelligence techniques with advanced machine-learning algorithms. Specifically, we developed and tested six models: LSTM, EHHO-LSTM, GWO-LSTM, PSO-LSTM, ACO-LSTM, and ACO-GRU, with accuracies of 0.95, 0.9055, 0.911, 0.938, 0,971, and 0.937, respectively.

Our methodology involved the integration of these models with respective optimization techniques, aiming to enhance feature selection and improve classification performance. The Enhanced Harmony Search (EHHO) algorithm was employed to select relevant features, while the Long Short-Term Memory (LSTM) network was used for classification in the EHHO-LSTM model. Similarly, the Grey Wolf Optimizer (GWO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) were coupled with LSTM, and the ACO was also combined with the Gated Recurrent Unit (GRU) network.

The ACO-LSTM model exhibited the highest accuracy, achieving 97.1%, indicating its superior performance in predicting diabetes. This model's success can be attributed to ACO's efficient exploration and exploitation capabilities, which enhance the LSTM's ability to learn from sequential data effectively. In comparison, the standalone LSTM model achieved a commendable accuracy of 0.95, underscoring the robustness of LSTM in handling time-series data.

The PSO-LSTM model also demonstrated significant accuracy, reaching 0.938. PSO's capability to optimize the weights and biases of the LSTM network contributed to its high performance. The ACO-GRU model followed closely with an accuracy of 0.937, showcasing the effectiveness of GRU in handling sequential data, although slightly lagging behind the ACO-LSTM combination.

The GWO-LSTM and EHHO-LSTM models achieved accuracies of 0.911 and 0.9055, respectively. While these models performed well, their lower accuracies compared to ACO-LSTM suggest that the choice of optimization algorithm plays a crucial role in enhancing model performance. The comparative analysis indicates that ACO, when integrated with LSTM, offers the best performance, likely due to its efficient pheromone-based search mechanism that optimizes feature selection and network parameters effectively.

In conclusion, the ACO-LSTM model stands out as the most accurate and reliable model for diabetes prediction among the ones tested in this study. This model's superior performance underscores the potential of hybrid approaches in addressing complex machine-learning challenges in medical diagnostics. The integration of swarm intelligence techniques with deep learning models not only enhances predictive accuracy but also aids in identifying critical factors influencing disease onset. These findings hold significant promise for early detection and intervention, ultimately improving healthcare outcomes. Future work should explore the application of these hybrid models to a broader range of medical conditions, leveraging their potential to revolutionize disease prediction and patient management.

Chapter 6

General conclusion and perspectives

6.1 Introduction

This thesis is an in-depth exploration of the essential area of predicting diabetes mellitus. Our primary goal in this study is to develop a new and enhanced approach for predicting the likelihood of individuals developing this condition. Given the substantial influence of diabetes on general health and quality of life, it is crucial to accurately forecast and identify it as soon as possible.

Our main goal is to improve diabetes prediction through thorough exploration of different methods and techniques in this field. Although many strategies have been considered, we have started a quest to discover an innovative approach that could surpass current methods.

In our efforts to accomplish this ambitious goal, our study explored a broad range of methods, focusing especially on deep learning and swarm intelligence strategies. These methodologies were chosen based on their established effectiveness and wide use in various areas such as medical research and healthcare. Using data-driven algorithms, our goal was to create a model that could effectively pinpoint people who are at risk of diabetes mellitus, enabling early intervention and better health results.

Following a thorough evaluation of the possible choices, we consciously decided to utilize deep learning and swarm intelligence models as the foundation of our strategy. By doing this, we have supported the scientific community's agreement on the successful use of deep learning algorithms in predicting medical conditions. Out of these algorithms, a range of LSTM-based models, combined with optimization methods, have been identified as the best options for our study. We examined how well LSTM, EHHO-LSTM, GWO-LSTM, PSO-LSTM, ACO-LSTM, and ACO-GRU models performed to identify the best combination.

6.2 Methodology

Our thesis is organized in a logical sequence of chapters, each serving a distinct purpose:

6.2.1 General introduction and motivation

The first chapter establishes the groundwork for our research. It explains our motivations, outlines the objectives guiding our investigation, details our selected research methodology, and provides an overview of the dissertation's components.

6.2.2 Background and Related Concepts

The second chapter acts as an intellectual framework for our study. In this section, we carefully define and clarify the key concepts essential to our research. These include the different types of diabetes, the prevalence of diabetes in Algeria, a detailed exploration of machine learning and its various forms, an introduction to swarm intelligence, and a brief review of well-known swarm intelligence algorithms.

6.2.3 The State of the Art

The third chapter marks a crucial point in our research journey. Here, we conduct an extensive review of existing research on diabetes prediction. This involves a thorough analysis and synthesis of previous studies. To aid comparative analysis, we summarize these studies into a detailed table, covering essential elements such as the study's title, authors, datasets used, proposed methodologies, resulting model performance metrics, and the advantages of each approach. Following this, we engage in a detailed discussion and critique of these diverse studies, extracting valuable insights that guide the development of our novel prediction approach.

6.2.4 Proposed Approach for Diabetes Prediction

In In the fourth chapter, we transition from reviewing existing research to presenting our own research methodology. We outline the detailed steps involved in preparing our dataset, explain the workings of various LSTM-based models, and explore the complexities of the respective optimization algorithms. Importantly, we describe our strategy for seamlessly integrating these methodologies to enhance predictive accuracy.

6.2.5 Experimental Setup and Evaluation

The fifth chapter represents the practical realization of our research efforts. In this section, we provide a detailed account of our approach's implementation, focusing specifically on using the Python programming language as our implementation tool. Additionally, we rigorously evaluate our models, using a set of performance metrics to determine their effectiveness in predicting diabetes.

6.2.6 General conclusion and perspectives

The final chapter serves as the culmination of our dissertation. Here, we summarize the essence of our research, highlighting our achievements and their implications. Furthermore, we look to the future, outlining potential areas for further research and innovation in the field of diabetes prediction.

6.3 Limits

Despite our exhaustive efforts, some limitations became apparent during our research. While the integration of various optimization algorithms with LSTM and GRU models yielded substantial improvements, certain constraints persisted. Notably, the ACO-LSTM model, which achieved the highest accuracy of 97,1, demonstrated the challenges inherent in balancing model complexity and

computational efficiency. Furthermore, the marginal improvement of 0.02% observed in LSTM models highlights the intricacies of fine-tuning hyperparameters and the potential diminishing returns of increasingly complex models.

Additionally, our aspiration to develop an interactive application to democratize the benefits of our prediction methodology faced formidable challenges. The complexities of integrating deep learning models, real-time data feeds, and user-friendly interfaces proved to be more intricate than anticipated. As a result, we did, develop an application using Strapi, but it has not been deployed yet.

6.4 Perspectives

As we advance in our research endeavors, our focus centers on an extensive exploration of advanced prediction techniques within the domains of deep learning and swarm intelligence. Our approach involves a rigorous investigation of cutting-edge methodologies, aimed at pushing the boundaries of achievable predictive analytics.

Furthermore, aligned with our strategic vision, we aim to materialize our innovative approach by developing a user-friendly application that transcends platform limitations. This application will be meticulously designed for accessibility by a diverse user base, including both web and mobile users. This commitment to inclusivity is integral to our overarching mission of fostering widespread adoption and empowerment of predictive technologies making them accessible to people from all walks of life.

6.5 Final Conclusion

After evaluating the performance of our six models, it is evident that the ACO-LSTM model is the superior choice, achieving the highest accuracy of 97.1%. This model's remarkable performance underscores the potential of hybrid approaches that integrate swarm intelligence with deep learning techniques. The consistent high performance of the PSO-LSTM and ACO-GRU models further validates the efficacy of combining optimization algorithms with advanced neural networks. As we move forward, we are committed to refining these models, addressing their limitations, and expanding their applications to a broader range of medical conditions, thereby contributing to the advancement of predictive healthcare technologies.

Bibliography

[1] News Medical. Insulin's role in the human body. https://www.news-medical.net/health/Insulins-role-in-the-human-body.aspx, March 30 2021. Accessed: 2023-05-04.

[2] Sarah Toy. Diabetes screening should start at 35, u.s. panel recommends. *The Wall Street Journal*, 2021. Last Updated: Aug. 24, 2021 11:00 am ET, Accessed: 2023-05-05.

[3] GeeksforGeeks. Introduction to deep learning. https://www.geeksforgeeks.org/introduction-deep-learning/, 2024. Last updated: May 26, 2024, Accessed: May 5, 2023.

[4] Géza Katona. The operation of the ant colony algorithm. ResearchGate, 2019. Last updated: January 2019, Accessed: 2023-05-05.

[5] S Geneva. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a who consultation. part 1: Diagnosis and classification of diabetes mellitus. *World Health Organisation*, 1999.

[6] International Diabetes Federation. Diabetes facts figures, 2024. Accessed: 2024-05-20.

[7] MH Shaeena and Rupesh Kumar Mani. Diabetes and nanotechnology–a recent advance in treatment of diabetes. *Journal of University of Shanghai for Science and Technology*, 23(11):445–453, 2021.

[8] Clifford J Bailey and Caroline Day. Traditional plant medicines as treatments for diabetes. *Diabetes care*, 12(8):553–564, 1989.

[9] Alpesh Goyal, Yashdeep Gupta, Rajiv Singla, Sanjay Kalra, and Nikhil Tandon. American diabetes association "standards of medical care—2020 for gestational diabetes mellitus": a critical appraisal. *Diabetes Therapy*, 11:1639–1644, 2020.

[10] American Diabetes Association. 2. classification and diagnosis of diabetes: standards of medical care in diabetes—2021. *Diabetes care*, 44(Supplement_1):S15–S33, 2021.

[11] American Diabetes Association Professional Practice Committee and American Diabetes Association Professional Practice Committee:. 2. classification and diagnosis of diabetes: Standards of medical care in diabetes—2022. *Diabetes care*, 45(Supplement_1):S17–S38, 2022.

[12] John M Lachin and David M Nathan. Understanding metabolic memory: the prolonged influence of glycemia during the diabetes control and complications trial (dcct) on future risks of complications during the study of the epidemiology of diabetes interventions and complications (edic). *Diabetes Care*, 44(10):2216–2224, 2021.

[13] MD Bethesda. National institute of diabetes and digestive and kidney diseases 2009. *US Renal Data System: USRDS 2009 Annual Data Report*, 2009.

[14] Jeremy T Warshauer, Jeffrey A Bluestone, and Mark S Anderson. New frontiers in the treatment of type 1 diabetes. *Cell metabolism*, 31(1):46–61, 2020.

[15] H. El Bouhissi, R. E. Al-Qutaish, A. Ziane, K. Amroun, N. Yaya, and M. Lachi. Towards diabetes mellitus prediction based on machine-learning. In *2023 International Conference on Smart Computing and Application (ICSCA)*, pages 1–6. IEEE, February 2023.

[16] Juan José Marín-Peñalver, Iciar Martín-Timón, Cristina Sevillano-Collantes, and Francisco Javier del Cañizo-Gómez. Update on the treatment of type 2 diabetes mellitus. *World journal of diabetes*, 7(17):354, 2016.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] Bibi Aamirah Shafaa Emambocus, Muhammed Basheer Jasser, and Angela Amphawan. A survey on the optimization of artificial neural networks using swarm intelligence algorithms. *IEEE Access*, 11:1280–1294, 2023.

[20] Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm intelligence: from natural to artificial systems*. Oxford university press, 1999.

[21] Amrita Chakraborty and Arpan Kumar Kar. Swarm intelligence: A review of algorithms. *Nature-inspired computing and optimization: Theory and applications*, pages 475–494, 2017.

[22] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. ieee, 1995.

[23] H. El Bouhissi, A. Ziane, L. Rahmani, M. Medbal, and M. Kostiuk. Rf-pso: An optimized approach for diabetes prediction. *CEUR Workshop Proceedings*, 3513:227–238, 2023.

[24] M Dorigo and T Stutzle. Ant colony optimization. mit press, cambridge, ma. 2004.

[25] Qasem Al-Tashi, Helmi Rais, and Said Jadid Abdulkadir. Hybrid swarm intelligence algorithms with ensemble machine learning for medical diagnosis. In *2018 4th international conference on computer and information sciences (ICCOINS)*, pages 1–6. IEEE, 2018.

[26] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. Grey wolf optimizer. *Advances in engineering software*, 69:46–61, 2014.

[27] Safial Islam Ayon and Md Milon Islam. Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, 13(2):21, 2019.

[28] Safial Islam Ayon and Md Milon Islam. Diabetes prediction: a deep learning approach. *International Journal of Information Engineering and Electronic Business*, 13(2):21, 2019.

[29] Haneen Qteat and Mohammed Awad. Using hybrid model of particle swarm optimization and multi-layer perceptron neural networks for classification of diabetes. *International Journal of Intelligent Engineering & Systems*, 14(3), 2021.

[30] KJ Rani. Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 6:294–305, 2020.

[31] Md Maniruzzaman, Md Jahanur Rahman, Benojir Ahammed, and Md Menhazul Abedin. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8:1–14, 2020.

[32] Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti, Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, 12(2):295–302, 2018.

[33] Sandip Kumar Singh Modak and Vijay Kumar Jha. Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications*, 83(13):38523–38549, 2024.

[34] Khondokar Oliullah, Mahedi Hasan Rasel, Md Manzurul Islam, Md Reazul Islam, Md Anwar Hussen Wadud, and Md Whaiduzzaman. A stacked ensemble machine learning approach for the prediction of diabetes. *Journal of Diabetes & Metabolic Disorders*, pages 1–15, 2023.

[35] Aditya Gupta, Ishwari Singh Rajput, Gunjan, Vibha Jain, and Soni Chaurasia. Nsga-ii-xgb: Meta-heuristic feature selection with xgboost framework for diabetes prediction. *Concurrency and Computation: Practice and Experience*, 34(21):e7123, 2022.

[36] Newton Lee. Interview with bill kinder: January 13, 2005. *Computers in Entertainment (CIE)*, 3(1):4–4, 2005.

[37] Huma Naz and Sachin Ahuja. Deep learning approach for diabetes prediction using pima indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19:391–403, 2020.

[38] Surabhi Kaul and Yogesh Kumar. Artificial intelligence-based learning techniques for diabetes prediction: challenges and systematic review. *SN Computer Science*, 1(6):322, 2020.

[39] Md Kamrul Hasan, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, 8:76516–76531, 2020.

[40] Palanigurupackiam Nagaraj and P Deepalakshmi. Diabetes prediction using enhanced svm and deep neural network learning techniques: An algorithmic approach for early screening of diabetes. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 16(4):1–20, 2021.

[41] Chetan Nimba Aher and Ajay Kumar Jena. Improved invasive weed bird swarm optimization algorithm (iwbsoa) enabled hybrid deep learning classifier for diabetic prediction. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3929–3945, 2023.

[42] Dinesh Chellappan and Harikumar Rajaguru. Detection of diabetes through microarray genes with enhancement of classifiers performance. *Diagnostics*, 13(16):2654, 2023.

[43] Yassine Ayat, Wiame Benzekri, Ali El Moussati, Ismail Mir, Mohammed Benzaouia, and Abdelaziz El Aouni. Novel diabetes classification approach based on cnn-lstm: enhanced performance and accuracy. *Diagnostyka*, 25, 2024.

[44] M. Ganesan, N. Sivakumar, M. Thirumaran, and R. Saravanan. Optimal deep learning based data classification model for type-2 diabetes mellitus diagnosis and prediction system. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(3):278–284, January 2020.

[45] Suja A Alex, NZ Jhanjhi, Mamoona Humayun, Ashraf Osman Ibrahim, and Anas W Abulfaraj. Deep lstm model for diabetes prediction with class balancing by smote. *Electronics*, 11(17):2737, 2022.

[46] P. Bhandari. How to find outliers | 4 ways with examples explanation. *IEEE Xplore*, November 2021. Accessed: 2024-05-18.

[47] Python Software Foundation. About python. `https://www.python.org/doc/essays/blurb`. Accessed: 2024-05-18.

[48] Numpy. `https://numpy.org/`. Accessed: 2024-05-15.

[49] Pandas documentation version 2.1.0. `https://pandas.pydata.org/docs/`. Accessed: 2024-05-15.

[50] scikit-learn: machine learning in python – scikit-learn 1.3.0 documentation. `https://scikit-learn.org`. Accessed: 2024-05-15.

[51] Keras. `https://datascientest.com/keras`. Updated: 2021-06-18, Accessed: 2024-05-15.

[52] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, and Greg S. Tensorflow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org/`, 2015. Accessed: 2024-05-15.

[53] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[54] Pauli Virtanen, Ralf Gommers, and Travis E. Oliphant. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[55] Scikeras Contributors. Scikeras: Keras api wrapper for scikit-learn. `https://github.com/adriangb/scikeras`, 2024. Accessed: 2024-05-15.

[56] Michael Waskom and the seaborn development team. Seaborn: Statistical data visualization. `https://seaborn.pydata.org/`, 2021. Accessed: 2024-05-15.

[57] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[58] IDBC. Tutorial: How to read a roc curve and interpret its auc?, 2022. Accessed: 2024-05-30.

[59] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006.

[60] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[61] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.