POPULAR AND DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY MOHAMED BOUDIAF - M'SILA

FACULTY OF TECHNOLOGY
ELECTRONICS DEPARTMENT
N°:……………………………………………

1985
جامعة محمد بوضياف - المسيلة
Université Mohamed Boudiaf - M'sila

DOMAIN: SCIENCE AND TECHNOLOGY
FILIERE: ELECTRONICS
OPTION: ELECTRONICS OF EMBEDDED SYSTEMS

Dissertation Submitted in partial fulfilment of the requirements
For the Master Academic Degree By:

AZIZI Mohammed
LABZA Nasreddine
MESSAAD Salim

Entitled

Early Enhancing Diabetes Diagnosis using Machine Learning, Deep Learning Models, and Clinical Data Analysis

Diplôme de Master dans le cadre du décret ministériel 1275

Presented on: June 17, 2025, in front of the jury composed of:

Dr. ATTALLAH Bilal	University of Mohamed Boudiaf-M'sila	President
Dr. KHENNOUF Salah	University of Mohamed Boudiaf-M'sila	Supervisor
Dr. BEGHRICHE Tawfiq	University of Mohamed Boudiaf-M'sila	Co-Supervisor
Dr. BENHAMADOUCHE Abdelouahab	University of Mohamed Boudiaf-M'sila	Examiner
Dr. BRIK Youcef	University of Mohamed Boudiaf-M'sila	CATI Representative
	University of Mohamed Boudiaf-M'sila	INCUBATEUR Representative
Dr TOUMI Mabrouk	Doctor	Socio-Economic Partner

Academic year : 2024 / 2025

Acknowledgments

In the Name of Allah, the Most Gracious, the Most Merciful

All praise is due to Allah, the Lord of the Worlds, who has blessed us with the completion of this research. We begin by expressing our deepest gratitude to all those who contributed to its accomplishment and enrichment.

We extend our sincere appreciation to our supervisor, Dr. KHENNOUF Salah, for his invaluable guidance, insightful advice, and relentless efforts in overseeing and supporting this work throughout its various stages.

Our heartfelt thanks also go to Dr. BEGHRICHE Tawfiq for his precious time and meaningful contributions.

We would also like to express our profound gratitude to the esteemed professors of the University of M'Sila, who provided us with academic and intellectual support throughout our educational journey.

A special acknowledgment goes to the members of the jury : Dr. Abdelouahab BENHAMADOUCHE and Dr. BRIK Youcef, for their gracious evaluation of this research and their valuable feedback.

Furthermore, we extend our deepest thanks to all the teaching and administrative staff of the Electronics Department at Mohammed BOUDIAF University in M'Sila for their efforts in fostering a supportive academic environment.

Our final prayer is that all praise be to Allah, the Lord of the Worlds.

Thank ALLAH for finishing this dissertation

Dedication

First and foremost, I express my deepest gratitude to Almighty God for granting me the strength, perseverance, and patience to complete this work, despite the challenges I encountered along the way.

My heartfelt appreciation goes to my dear mother, ***Fadila****, whose boundless love, endless sacrifices, and unwavering prayers have been the foundation of all my achievements. Her strength, kindness, and selflessness continue to inspire me every single day.*

To my beloved father, ***Djedid****, whose wisdom, encouragement, and steadfast belief in me have been a guiding force in my life. His constant support and noble character have shaped my values and aspirations—he is and will always remain my role model and source of strength.*

*To my cherished siblings—****Imen, Wafaa, Karima, Aya****, and* ***Ikhlas****—thank you for being a pillar of love and encouragement, and for providing comfort through every challenge.*

To my dear friends ***Bilal****,* ***AbuBakr****,* ***Salim****,* ***Mondhir****,* ***Mohammed****,* ***Imad*** *,* ***Ali, Salah****,* ***Nassreddin****, for your support and friendship that made this journey brighter..*

Finally, to all the professors of the ***Department of Electronics****, thank you for, your guidance, and the knowledge you have imparted throughout my academic journey*

AZIZI Mohammed

Dedication

*o the one who has always been a warm embrace and a sincere prayer in the depth of night, To the one who planted in my heart the love of excellence and endured the hardships of life to sow hope within me… To my beloved mother **Khadra**, no words can truly express how much I owe you. I dedicate this work to you with pride and gratitude—for you are the source of my strength and the light of my life. To my dear father **Chelali,** A man of great heart, a model of patience and perseverance, thank you for everything you have done to ensure my comfort and education. I offer you this achievement in appreciation and gratitude—you have always been my steadfast support. To my esteemed professors in the Department of Electronics, I extend my deepest thanks and sincere appreciation. Without your efforts and dedication, we would not have taken our first steps with such confidence. You were the guiding lights on my academic journey and a source of true inspiration. To my companions along this path, those who shared every challenge and every moment of joy, who stood by me through doubts and celebrated each step forward— To my dear friends: Ali, Fares, Mouayad, Nour El-Din, Belkacem, Hafida, Mostafa, Mohamed, Abdelrahman, You each hold a special place in my memory and heart. I dedicate this work to all of you, For every word in it carries a part of you, and of the unforgettable moments we lived together.*

LABZA Nasreddine

Dedication

Praise be to God, and prayers and peace be upon the Messenger of God, Mohammed, may God bless him and grant him peace I dedicate this humble work to Dear parents, my mother and father, may God protect them and my full brothers, to classmates Mohammed and Nasr al-Din, brothers abderrahmane, all members of the Suqia Association Moustafa,ahmed,bachier and ismail To my absent brother and my twin, Rabeh, may God have mercy on him And to our brothers in Gaza and Palestine, may God relieve their distress and grant them victory over their enemy

MESSAAD Salim

List of abbreviations

AI: Artificial Intelligence

AUC: Area Under the Curve

CNN: Convolutional Neural Network

CSS: Cascading Style Sheets

CT: Computed Tomography

DL: Deep Learning

DT: Decision Tree

DNA: Deoxyribonucleic Acid

ECG: Electrocardiograms

EEG: Electroencephalograms

EHRs: Electronic Health Records

EMG: Electromyograms

FN: False Negative

FP: False Positive

GDM: Gestational Diabetes Mellitus

HLA: Human Leukocyte Antigen

HPV: Human Papillomavirus

HTML: Hyper Text Markup Language

HTTP: Hypertext Transfer Protocol

ICT: Information and Communication Technologies

IDD: Insulin-Dependent Diabetes

KNN: K-Nearest Neighbors

LSTM: Long Short-Term Memory

Light GBM: Light Gradient Boosting Machine

ML: Machine Learning

MRI: Magnetic Resonance Imaging

NCDs: Non-Communicable Diseases

NIDDM: Non-Insulin-Dependent Diabetes Mellitus

RF: Random Forest

RNA: Ribonucleic Acid

RNN: Recurrent Neural Network

SBN: Strengths-Based Nursing

SHAP: Shapley Additive Explanations

STIs: Sexually Transmitted Infections

SVM: Support Vector Machine

TN: True Negative

TP: True Positive

WHO: World Health Organization

XGBoost: Extreme Gradient Boosting

List of figures

List of tables

Table of Contents

CHAPTER III - SIMULATION AND RESULTS

INTRODUCTION

Introduction

Healthcare remains one of the most vital pillars of modern societies, playing a crucial role in disease prevention, diagnosis, and treatment. It significantly contributes to improving quality of life and extending life expectancy. However, with the rising prevalence of chronic diseases such as diabetes, there is an urgent need for innovative and intelligent tools that can enhance early detection and accurate diagnosis. In this context, artificial intelligence (AI) and machine learning (ML) have emerged as powerful technologies capable of transforming traditional healthcare systems into more proactive, data-driven models.

Among the most pressing public health concerns today is diabetes, a chronic metabolic disease that affects millions of individuals worldwide. Its complications can be severe, ranging from cardiovascular problems to kidney failure, vision loss, and nerve damage. According to the World Health Organization (WHO), diabetes has become one of the leading causes of death globally, particularly in developing countries. This situation highlights the necessity of developing intelligent systems that can detect disease at an early stage and assist both healthcare professionals and patients in monitoring and managing their condition effectively.

The primary aim of this work is to design and develop an intelligent prediction system based on machine learning algorithms to identify early signs of diabetes using clinical data. The proposed solution is supported by a web-based interface that allows users to input medical attributes and receive real-time predictions, making it accessible to both healthcare providers and non-specialist users.

This thesis is divided into three main chapters. The first chapter provides a comprehensive overview of the healthcare system and the medical aspects of diabetes, including its types, symptoms, diagnostic methods, and complications. The second chapter focuses on artificial intelligence and machine learning, presenting key concepts and highlighting their applications in the healthcare field. The third chapter details the implementation of the proposed system — from data preprocessing and model training to performance evaluation and deployment using a user-friendly web interface.

Through this project, we aim to contribute to the integration of AI-powered solutions in healthcare, helping to reduce the impact of diabetes and paving the way for the development of similar systems for other diseases in the future

CHAPTER I

GENERALITIES ABOUT HEALTHCARE AND DIABETES

CHAPTER I

GENERALITIES ABOUT HEALTHCARE AND DIABETES

I.1. Introduction

This chapter gives an overview of healthcare and its value in combating different diseases with the help of automated systems. It discusses the problems developed by chronic illnesses and how technology can aid in improving early detection and management. The chapter underlines, finally, the importance of concentrating on diabetes as one of the most common diseases, pointing out the increasing demand for new methods in its prediction and diagnosis.

I.2. Healthcare

I.2.1. Definition of Healthcare:

Healthcare is a multifaceted concept that encompasses the maintenance and improvement of health through prevention, diagnosis, treatment, and management of illness and injury. It involves a wide range of services provided by professionals across various fields such as medicine, nursing, pharmacy, and allied health professions [1].

I.2.2. Importance of Healthcare:

Healthcare serves as a foundation of contemporary society, significantly contributing to disease prevention via early detection, immunizations, and public health awareness. These preventive services assist in lowering the incidence of chronic conditions and alleviating the burden on healthcare systems. Additionally, a healthy population directly aids economic growth by decreasing absenteeism and enhancing workforce productivity, while simultaneously reducing long-term healthcare costs. Social equity represents a vital aspect of healthcare, as obtaining affordable and high-quality health services lessens inequalities among various social and economic groups, fostering fairness and justice in public health. Moreover, healthcare systems improve overall quality of life by providing services like rehabilitation, mental health assistance, and palliative care, enabling people to lead fulfilling lives despite illness.

Ultimately, by incorporating technological advancements, healthcare provision has become more precise and effective, enhancing diagnoses, treatment results, and the overall experience for patients [2]

I.2.3. Objectives of Healthcare Systems

Healthcare systems are intended to meet particular goals that guarantee the efficient provision of health services and enhance overall health results. These goals may differ depending on the healthcare model and the unique requirements of the community being served. The main fundamentals goals of healthcare systems are [3] [4]:

- Enhancing Health Results: A key goal is to improve health outcomes for both individuals and populations. This entails tackling urgent care requirements and emphasizing population health objectives, even though urgent care frequently takes priority because of pressing demands and media influence.
- Improving Patient Experience: Healthcare systems strive to enhance patient experiences by providing quality care and ensuring patient satisfaction. This is a component of the quadruple aim, which also encompasses enhancing population health, lowering costs, and improving provider experiences.
- Cost Decrease: Effective resource management to lower healthcare expenses is an essential goal. This entails streamlining processes and incorporating digital health solutions to reduce mistakes and enhance efficiency.
- Provider Experience: Enhancing the work atmosphere and experiences of healthcare professionals is crucial for sustaining a motivated and efficient workforce.
- Sustainability and Flexibility: Systems are progressively prioritizing sustainability, incorporating environmental factors into healthcare provision, and adjusting to shifts like climate effects.
- Unified Services and Coordination: Integrated healthcare systems strive to deliver cohesive care across various services, especially for those with complicated chronic illnesses. Leadership and teamwork are essential elements for achieving effective integration.
- Goal-Focused Reforms: Reforms must be issue-based, coherent, and regularly assessed to guarantee they meet the real aims of the healthcare system and adjust to evolving needs.

I.2.4. Challenges in Healthcare

There are several challenges in healthcare systems across the world, which affect the effectiveness, and the quality of services delivered. One of the major problems is the absence of enough number of healthcare professionals, which is even more critical in the rural and hard-to-reach areas, which results in the unequal distribution of health care. This shortage leads to stress and fatigue of the available personnel, which in turn adversely affects the quality-of-service delivery to the patient.

However, there are challenges such as the increased incidence of chronic diseases like diabetes and heart diseases that put a huge burden on the health care systems and hence the need for long term care and funding. Other environmental factors including climate change and pollution also bring about new health threats that compound the existing problems. Financial barriers are still present, and many healthcare systems have the problem of lacking enough funds. This financial pressure often translates to restricted availability of essential services and drugs, especially in the poorer countries.

Moreover, the absence of clear-cut employment regulations and the variability in policies among different healthcare facilities results in poor performance and unequal provision of services. These challenges can only be met through global and coordinated efforts that include more funding of healthcare systems, changes in policies and equitable funding of health care to improve the quality of health care services [5].

I.3. Diseases

I.3.1. Definition of Diseases

A disease is defined as a particular abnormal condition that adversely affects the structure or function of all or part of an organism and is not immediately due to any external injury.

Diseases are often known to be medical conditions that are associated with specific signs and symptoms. They may be caused by external factors such as pathogens or by internal dysfunctions [5].

I.3.2. Classification of Diseases

I.3.2.1. Infectious diseases

Infectious diseases are ailments brought on by harmful microorganisms, including bacteria, viruses, fungi, or parasites. These illnesses can spread, either directly or indirectly, from one person to another. Transmission can take place through different methods, such as direct contact between individuals, tainted food or water, or bites from insects. Infectious illnesses can vary from mild to severe and pose a major issue for public health systems globally [6]

I.3.2.2. Types of infectious diseases

Infectious diseases can be viral, bacterial, parasitic or fungal infections

A) Viral infections

Viral infections result from viruses and can impact different systems in the body. A prevalent category is diseases of the digestive system, where viral infections may result in issues like viral gastroenteritis, leading to symptoms such as diarrhea, vomiting, and abdominal discomfort. Another category consists of skin rashes and lesions, since viruses can induce skin-related issues like chickenpox, which leads to an itchy rash and fever, as well as warts caused by the human papillomavirus (HPV). These infections differ in seriousness and way of spreading but continue to pose a major issue in public health because of their contagious characteristics [6].

B) Bacterial infections

Bacterial infections are caused by bacteria and can result in a wide variety of diseases affecting different parts of the body. One category is pyogenic cocci diseases, which involve infections caused by pus-forming bacteria such as Staphylococcus and Streptococcus, often leading to conditions like skin abscesses, pneumonia, or tonsillitis. Another group includes gram-negative bacterial diseases, which are caused by gram-negative bacteria known for their higher resistance to antibiotics and association with serious infections like urinary tract infections and sepsis. Additionally, zoonotic bacterial diseases are infections transmitted from animals to humans, with

Salmonella being a common example that can cause foodborne illness characterized by gastrointestinal symptoms [6]

C) Fungal infections

Fungal infections can vary from mild surface issues to severe systemic illnesses, based on the extent and severity of the infection. Superficial mycoses impact the outer layers of the body, including the skin, hair, and nails, resulting in ailments like athlete's foot or ringworm. In contrast, opportunistic mycoses mainly affect those with compromised immune systems; one example is candidiasis, which can impact the mouth, throat, or genitals and can become invasive in serious situations. Furthermore, although uncommon, certain diseases can arise from algae, like protothecosis, which has the potential to infect the skin or other tissues, especially in those who are immuno-compromised. These infections illustrate the variety of fungal and algal pathogens and their possible effects on human health. Superficial Mycoses: These are fungal infections affecting the skin, hair, and nails [6].

I.3.2.3. Common infectious diseases

A) Genetic Susceptibility and Pathogenesis

According to research, the genetic composition of the host plays a significant part in determining the vulnerability and course of viral diseases. Genome-wide association studies have revealed 59 risk loci linked to 17 common infections, underscoring the significance of genetic factors on disease response. These loci are present in human leukocyte antigen (HLA) and non-HLA regions, indicating that particular genetic variants may affect immune responses to a range of diseases, such as chickenpox1, shingles, and hepatitis B [7]

B) Infectious Diseases in Vulnerable Populations

Sexually transmitted infections (STIs) and infectious skin diseases are among the infectious diseases that disproportionately affect Indigenous populations, such as Indigenous Australians. In order to overcome these obstacles and enhance health outcomes, Strengths-Based Nursing (SBN) strategies have been put forth. In a similar vein, infectious diseases pose particular difficulties for refugee populations, calling for focused screening and treatment initiatives to control tropical disease burdens [8].

C) Common Infectious Diseases in Children

Infectious diseases can cause serious morbidity and mortality, and children under five are especially susceptible to them. This age group is susceptible to influenza, chickenpox, and measles. Reducing the incidence and impact of these diseases requires effective management and prevention strategies, such as immunization and better hygiene practices [8].

I.3.2.4. Symptoms of infectious diseases

Numerous symptoms can be indicative of infectious diseases, and these symptoms frequently change according to the pathogen causing illness, the body system that is impacted, and the health of the individual. Fever, exhaustion, muscle aches, and gastrointestinal issues like vomiting and diarrhea are typical symptoms. Cough and sore throat are common respiratory symptoms, especially in cases of bacterial and viral illnesses. Rashes and other skin conditions might appear, particularly in cases of measles or chickenpox. More severe cases of some infections may cause neurological symptoms like headache, disorientation, or seizures. It's crucial to remember that the existence and combination of these symptoms can help medical practitioners identify the underlying infectious disease and choose the best course of therapy [9].

I.3.3. Non-Infectious Diseases

Non-communicable diseases (NCDs), another name for non-infectious diseases, are illnesses that cannot be spread from one person to another and are not brought on by infectious agents. They usually last a long time and develop gradually. Diabetes, cancer, chronic respiratory conditions including asthma and chronic obstructive pulmonary disease, and cardiovascular disorders like heart attacks and stroke are the primary categories of noncommunicable diseases (NCDs). Genetic, physiological, environmental, and behavioral variables frequently combine to cause these disorders. The World Health Organization (WHO) reports that noncommunicable diseases

(NCDs) account for 74% of all fatalities worldwide, making them the leading cause of death worldwide [10]

I.3.3.1. Lifestyle-related diseases

Lifestyle-related diseases are among the most significant health challenges in the modern era. They include a range of chronic conditions such as cardiovascular diseases, type 2 diabetes, cancer, obesity, and chronic respiratory diseases. These illnesses are often the result of repeated unhealthy behaviors such as physical inactivity, poor nutrition, smoking, and alcohol consumption. According to the World Health Organization (WHO, 2023), these diseases account for approximately 74% of all global deaths, making their prevention a top public health priority. Adopting a healthy lifestyle — based on balanced nutrition, regular physical activity, and the avoidance of harmful habits — is a key step in reducing the risk of developing these conditions [11].

I.3.3.2. Major non-infectious disease categories

Non-infectious diseases, also known as non-communicable diseases (NCDs), are not caused by pathogens and cannot be transmitted from person to person. The major categories of non-infectious diseases include cardiovascular diseases (such as heart attacks and strokes), cancers, chronic respiratory diseases (such as chronic obstructive pulmonary disease and asthma), and diabetes. These conditions are often linked to genetic, physiological, environmental, and behavioral factors. They tend to be long-lasting and progress slowly. According to the World Health Organization (WHO, 2023), these four major categories account for the majority of deaths from NCDs worldwide, making them a primary concern for global health strategies [11].

I.3.3.3. People at risk

NCDs impact people of all ages, geographical locations, and nationalities. Although these illnesses are frequently linked to older age groups, over 18 million deaths from NCDs happen before the age of 70. More people in this age range die from NCDs than from all other causes combined. It is estimated that 82% of these premature deaths take place in low- and middle-income nations. Youngsters, adults, and the elderly are all susceptible to the risk factors that lead to non-communicable diseases (NCDs), which include poor diets, sedentary lifestyles, exposure to tobacco smoke, hazardous alcohol consumption, and air pollution. Obesity, rising blood lipids, elevated blood pressure, and elevated blood glucose are all symptoms of unhealthy eating habits and inactivity. These are known as metabolic risk factors, and they have the potential to cause cardiovascular disease, which is the NCD that causes the most premature deaths [10]

I.4. Diabetes

Diabetes is a disease that prevents the body from properly utilizing the energy provided by food consumed. It occurs when the pancreas stops producing insulin or when the body becomes resistant to the insulin it produces. Insulin is a hormone that regulates blood sugar levels. Hyperglycemia, or high blood sugar levels, is a common effect of uncontrolled diabetes, which, over time, can lead to serious damage to various body systems, particularly the nerves and blood vessels.

Diabetes is one of the most common diseases, primarily affecting the elderly population worldwide. According to the World Health Organization (WHO), diabetes affects more than 463 million people globally. In Algeria alone, there were approximately 5 million diabetic patients in 2019 [10]

- To fully understand what diabetes is, it is essential to know how our body derives energy from food. (see Figure 1.1)

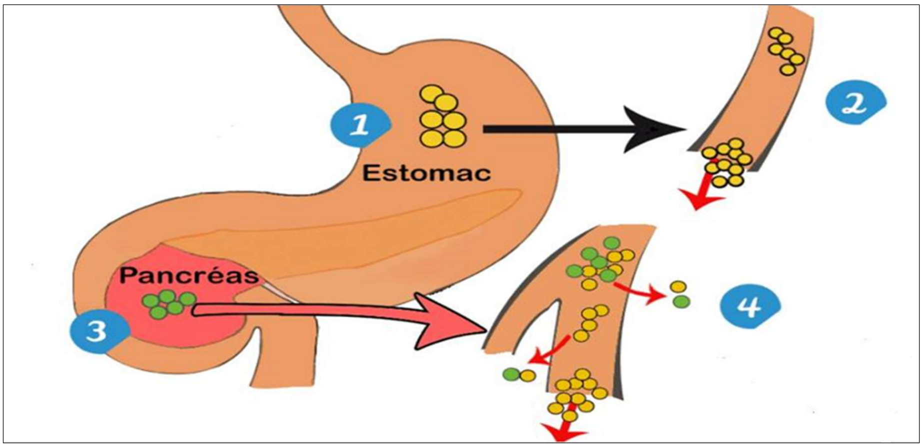

Figure I.1: Illustrative image of diabetes.

Description of figure 1.1:

- The stomach transforms food into glucose (sugar).
- Glucose enters the bloodstream.
- The pancreas no longer produces insulin (type 1 diabetes, 10% of patients) or produces insufficient insulin (type 2 diabetes, 90% of patients). Insulin is the hormone that allows glucose assimilation into the body.
- Glucose, not absorbed by the cells, accumulates in the blood. This excess sugar causes serious discomfort and damages the body

I.4.1. Types of Diabetes

There are mainly two types of diabetes

I.4.1.1. Type 1

Type 1 diabetes, also known as insulin-dependent diabetes, affects about 10% of diabetics. In type 1 diabetes, cells in the pancreas are destroyed, leading to a reduction or even a complete absence of insulin secretion. The pancreas can no longer synthesize enough insulin. As a result, glucose can no longer enter the cells, increasing its blood level and causing hyperglycemia [12]

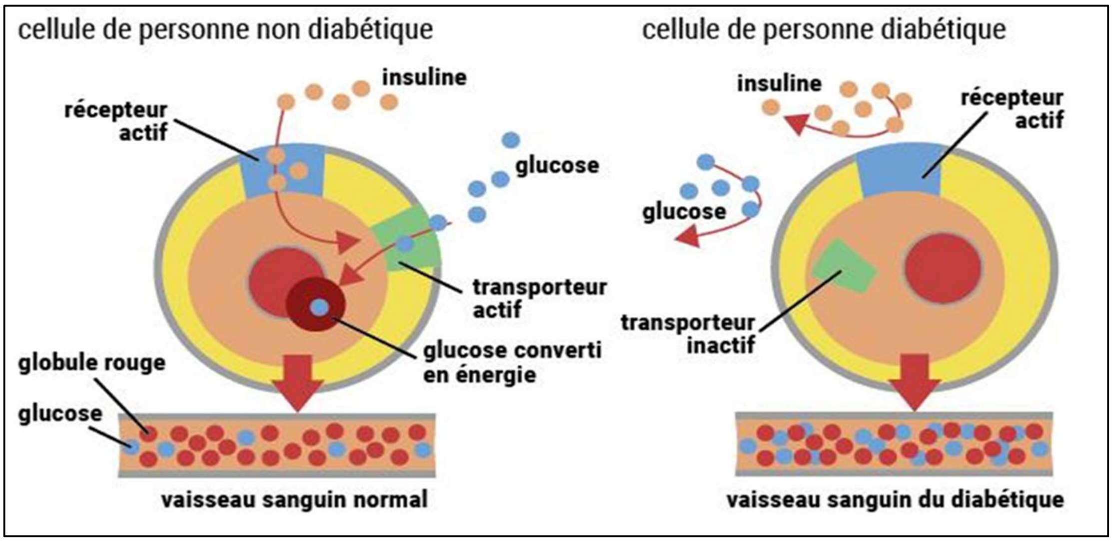

Figure I.2- Insulin resistance and insulin deficiency

A) Symptoms

The symptoms of type 1 diabetes usually appear quickly, unlike type 2 diabetes, which can remain silent for a long time." It starts with frequent urination and thirst greater than usual. This can then lead to weight loss, fatigue, intense hunger, and may complicate ketoacidosis. This serious complication occurs more rapidly in children and can lead to dehydration, vomiting, abdominal pain, and even discomfort. [12]

B) Diagnostic

At the first signs of diabetes, particularly in children, it is essential to consult a doctor quickly to avoid the initial complications. The treating physician can make the diagnosis through a blood test showing hyperglycemia of more than 1.20 g/L. This must be confirmed at least twice for the diagnosis of diabetes to be confirmed [12].

C) Treatment

Type 1 diabetes, being insulin-dependent can only be treated with insulin injections. The number of injections must be adapted throughout the course of the disease, so that the patient always has enough insulin to maintain normal blood sugar levels and a hemoglobin A1c lower than 7%. If this is no longer the case, complications can appear quickly.

Moreover, type 1 diabetes cannot be cured. "It is, therefore, a lifetime treatment for patients," says the dialectologist. "But more and more research is being done on diabetes, as it is a rapidly growing disease worldwide, and innovations have been made," he adds with enthusiasm.

Today, insulin "shots" have been replaced by pens with a single, adapted dose, making the treatment easier. For younger patients, insulin pumps are available that provide continuous insulin delivery for three days. This makes life easier for the younger patients, who may already be very affected by the disease.

Additionally, future innovations are being discussed, such as an artificial pancreas that would eliminate the need for insulin injections, or improved insulin pumps. [12]

I.4.1.2. Type 2

Type 2 diabetes is characterized by insulin resistance, which leads to a significant increase in blood glucose levels, i.e., the amount of sugar in the blood. "This form of diabetes is the most common: 90 to 92% of diabetics are affected by it." Originally, it was considered a diabetes of maturity and aging, as it is promoted by being overweight, obesity, or a sedentary lifestyle. "But it is increasing because there is much more sedentary behavior and obesity worldwide." Although type 2 diabetes can remain silent for years, complications can suddenly arise, causing damage to various organs, sometimes irreversible [14].

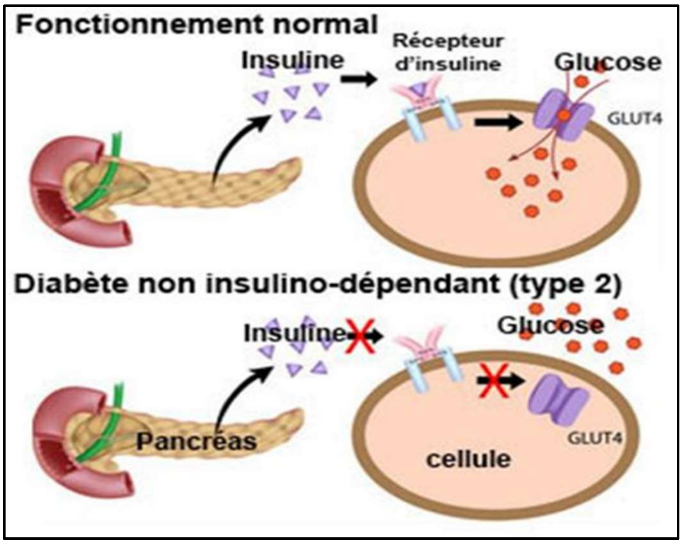

Figure I.3: Non-insulin-dependent diabètes [12]

In a normal person, all the cells in the body function using glucose. This glucose is first supplied by food and then by the liver (outside of meals and during the night). It is transported by blood to the cells that need it. To enter these cells, glucose requires the help of a hormone produced by the pancreas: insulin. Insulin takes up the glucose circulating in the blood (which lowers blood sugar levels) and allows it to enter the cells, where it can be used as fuel. This is normal functioning.

In type 2 diabetes, the cells gradually become resistant to the action of insulin, which can no longer perform its usual role, leading to hyperglycemia. [13]

A) Symptoms

The symptoms of type 2 diabetes generally do not appear at the beginning and may even be absent for years. In fact, it is a disease that can progress for years without causing any symptoms. Moreover, type 2 diabetes often goes unnoticed until complications arise. The signs of type 2 diabetes include fatigue, vision problems, dry mouth, frequent urination, increased hunger or thirst, tingling in the feet, slow-healing infections... These signs appear more gradually. It is important to recognize the subtle signs of diabetes, such as skin infections (abscesses, boils), urinary tract infections, erectile dysfunction, fatigue, shortness of breath... [14].

B) Diagnosis

A blood test in a laboratory measures blood glucose levels; a fasting blood glucose level equal to or greater than 1.26 g/l, confirmed twice, diagnoses diabetes. Other tests may be performed: postprandial blood glucose (2 hours after a meal), glycosuria (presence of sugar in the urine), and glycated hemoglobin. These are rarely used for diagnosing diabetes but can assist in monitoring.

A complete medical examination will look for signs that may indicate the origin of diabetes and the presence of complications: evaluation of overweight (weight, height, fat distribution), blood pressure measurement, heart and vascular auscultation, examination of reflexes and sensitivity in the legs and feet, in particular... [14].

C) Treatment:

The first treatment for type 2 diabetes is diet," assures Dr. Fabrice Bonnet, a researcher. "Limiting carbohydrate intake while regularly engaging in physical activity helps avoid hyperglycemia and thus reduces the need for insulin," he explains. Quitting smoking is also recommended.

Type 2 diabetes can also be treated with medications. There are four classes of drugs: biguanides (Glucophage®, Stagid®...), sulfonylureas (Daonil®, Amarel®...), alpha-glucosidase inhibitors (Diastabol® and Glucor®), and glinides (Novonorm®). These medications work either by directly stimulating insulin secretion, enhancing insulin's action in the body, or slowing sugar absorption. "However, as diabetes progresses, these medications may no longer be sufficient to reduce blood sugar levels, and transitioning to insulin may become necessary," explains the diabetologist.

Today, insulin formulations tailored for type 2 diabetes can limit the number of daily injections, sometimes requiring only one injection per week. "With the global explosion of diabetes, new, less burdensome hormone-based treatments are currently being tested." The therapeutic arsenal for type 2 diabetes is thus undergoing renewal. [14].

In this table we explain the Difference Between Type 1 and Type 2 Diabetes

Table I.1- Difference Between Type 1 and Type 2 Diabetes

Type 1 Diabetes	**Type 2 Diabetes**
Insulin-Dependent Diabetes (IDD)	Non-Insulin-Dependent Diabetes (NIDDM)
Called "lean" diabetes because one of the first symptoms is weight loss, or "juvenile" diabetes because it affects young individuals	Described as "fat diabetes" or maturity-onset diabetes, as it often occurs around the age of fifty in overweight individuals
It accounts for about 10% of cases.	It accounts for approximately 90% of cases
It results from the disappearance of beta cells in the pancreas, leading to a total deficiency of insulin	Two abnormalities are responsible for hyperglycemia: ➢ The pancreas still produces insulin but not enough relative to blood sugar levels, which is called insulinogenic. ➢ Or the insulin does not work effectively, which is known as insulin resistance [15]
The symptoms are usually intense thirst, excessive urination, and rapid weight loss.	Painless, its development can go unnoticed for a long time: it is estimated that an average of 5 to 10 years passes between the onset of the first hyperglycemic episodes and the diagnosis.
The only treatment is insulin administration: either in the form of injections (with a syringe or pen) or with an insulin pump designed to deliver insulin continuously.	It is treated with a diet, along with oral medications if necessary, and possibly insulin after a few years of progression

I.4.2. A Particular Type

Gestational diabetes mellitus (GDM) is defined as glucose intolerance of varying severity with onset or first recognition during pregnancy. It is one of the most common medical complications during pregnancy, affecting approximately 15% of pregnancies worldwide, equating to nearly 18 million births annually. GDM is mainly attributed to hormonal changes during pregnancy that cause insulin resistance and elevated blood glucose levels. If left unmanaged, GDM can lead to significant health risks for both the mother and the baby, including an increased risk of gestational hypertension, preeclampsia, and the need for cesarean delivery. Early detection and effective management of GDM are critical to reducing these risks and ensuring better maternal and fetal outcomes [16].

I.4.3. Relationship of Diabetes with Heredity and Pregnancy

I.4.3.1. Relationship of Diabetes with Heredity

The role of heredity differs depending on whether it is type 1 or type 2 diabetes. When one parent has type 2 diabetes, the risk of transmission to offspring is around 40%, and if both parents are affected, the risk rises to 70%. In type 1 diabetes, the risk is between 4% and 8%, more specifically 8% if the father has diabetes, 4% if it is the mother (but 30% if both parents have it). Therefore, it is useful to create a family tree to identify diabetic family members and understand one's genetic heritage [16].

I.4.3.2. Relationship of Diabetes with Pregnancy

Women with diabetes are at a higher risk of miscarriage or having a baby with congenital malformations (such as heart and kidney defects). This risk increases significantly if blood sugar control is not optimal, especially at the time of conception and during the first 3 months of pregnancy; poorly controlled blood sugar poses many risks for both the mother and the baby. Proper preparation for pregnancy can help reduce the risk of complications and maintain good health throughout the pregnancy, in addition to giving your baby a good start in life.

Blood sugar control is a daily challenge for people with diabetes. It becomes even more critical during pregnancy due to hormonal changes. Most women who manage to control their diabetes well before and during pregnancy will have a happy pregnancy and give birth to a healthy baby. [16]

I.4.4. Complications of Diabetes

Diabetes complications are numerous and can be severe. These complications degrade diabetes and tend to reduce the life expectancy of affected individuals. Most diabetes-related complications can be prevented, reduced, or delayed if diabetes is detected and treated early and properly. The main complications of diabetes include: diabetic retinopathy (which can lead to blindness), cardiovascular complications, diabetic nephropathy leading to kidney failure, diabetic neuropathy, infections, foot and leg ulcers. Acute complications such as hypoglycemic coma can also occur [16].

I.4.5. Prevention of Diabetes

To prevent the onset of diabetes or slow its progression, it is necessary to:

- Engage in regular physical activity
- Control weight and monitor diet
- Regulate blood pressure
- Quit smoking [17]

I.4.6. Diabetes and Diet

Diet plays a crucial role in the progression of diabetes and in the prevention of metabolic risks. So, what eating habits should be adopted? The benefits of a balanced diet on health, particularly on diabetes, are scientifically proven. When you have diabetes, it is important to prioritize foods rich in fiber: fruits (apple, pear), green vegetables, whole grain products (unpearled spelt, barley, crushed durum wheat semolina), legumes (chickpeas, lentils, Chinese peas, or beans), dried fruits, or seeds. Season your dishes with raw fats, a source of essential fatty acids (sunflower, rapeseed, olive, walnut oils...), fresh herbs, or spices (cinnamon, nutmeg, garlic). An interesting and natural alternative to added sugar, sweeteners or dried fruits in moderate amounts can enhance your pastries and cottage cheese [18].

I.5. Conclusion

In this chapter, we explored the critical role of healthcare systems in promoting public health and managing diseases, with a particular focus on diabetes. We emphasized how healthcare serves as the foundation for early detection, prevention, and effective management of both infectious and non-infectious diseases. Special attention was given to diabetes as one of the most prevalent chronic conditions worldwide, highlighting its types, symptoms, diagnosis, treatment, and complications. Furthermore, the chapter discussed the importance of addressing risk factors, adopting preventive strategies, and enhancing healthcare delivery to combat the rising burden of diabetes. Overall, a strong, well-coordinated healthcare system, combined with public awareness and technological innovations, is vital for controlling diabetes and improving the quality of life for affected individuals.

CHAPTER-II

APPLICATION OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

CHAPTER-II

APPLICATION OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

II.1. Introduction

Healthcare is one of the most vital sectors that has greatly benefited from technological advances, particularly the integration of Artificial Intelligence (AI). The continuous evolution of AI techniques, from machine learning to deep learning, has revolutionized the way healthcare services are delivered, enabling earlier diagnoses, personalized treatments, and improved patient monitoring. This chapter provides an overview of the historical development of AI, its key concepts and techniques, and its growing applications in healthcare, especially in supporting medical decision-making and enhancing the quality of care. Understanding these advancements is essential to grasp the profound impact AI has on the healthcare industry today and its promising potential for the future.

II.2. Artificial Intelligence (AI)

II.2.1. Definition of Artificial Intelligence

Artificial Intelligence refers to the ability of non-human systems or artificial entities to carry out tasks, resolve issues, and engage in interactions that exhibit intelligent behavior in practical settings. It encompasses learning, adaptation, communication, and independent decision-making, allowing these systems to replicate human cognitive abilities. This definition highlights that AI involves not just mimicking human intelligence but also utilizing computational efficiency to execute intricate tasks in changing environments [19].

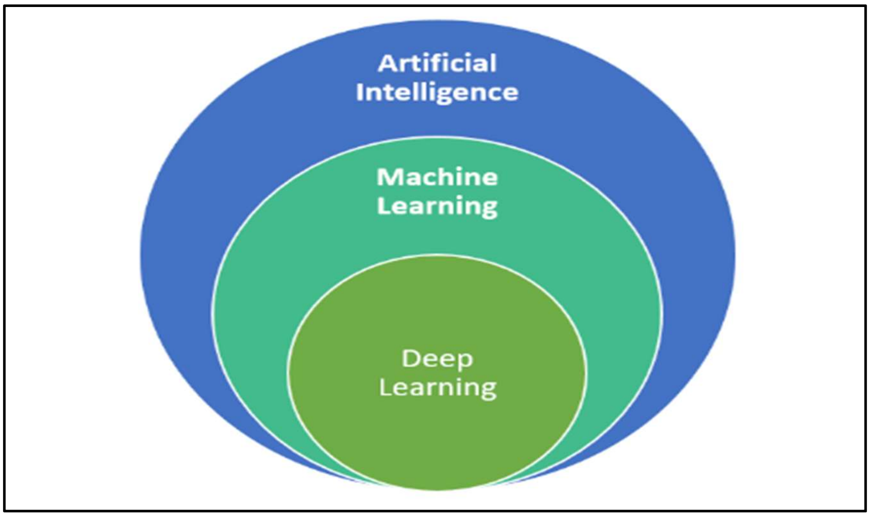

Figure II.1. Structure of AI.

II.2.2. Brief History

From 1943 to 1956 The foundation of AI was laid in 1943 when Warren McCulloch and Walter Pitts introduced the concept of an artificial neuron, proposing a model that could simulate basic neural processes. In 1950, Alan Turing introduced the Turing Test in his seminal paper "Computing Machinery and Intelligence," providing a criterion to evaluate a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. By 1955, Allen Newell and Herbert A. Simon developed the Logic Theorist, considered the first AI program. The formal establishment of AI as a field occurred in 1956 during the Dartmouth Conference, organized by John McCarthy, which brought together researchers to discuss and advance AI research [20].

From 1966 to 1996, Joseph Weizenbaum developed ELIZA, an early natural language processing program that simulated conversation by mimicking a Rogerian psychotherapist. This period also saw the creation of WABOT-1 in 1972, the first full-scale humanoid intelligent robot developed in Japan. A significant milestone was achieved in 1996 when IBM's Deep Blue defeated world chess champion Garry Kasparov, marking the first time a computer beat a reigning world champion in a match under standard chess tournament conditions [20].

From 2006 to 2022 AI began to permeate the business sector around 2006, with companies like Facebook, Twitter, and Netflix integrating AI to enhance user experience and operational efficiency. In 2018, Google introduced Duplex, an AI system capable of conducting natural conversations to carry out real-world tasks over the phone. The release of ChatGPT by OpenAI in November 2022 showcased the advanced capabilities of AI in understanding and generating human-like text, indicating a significant leap forward in AI applications, particularly in education and communication [20].

II.2.3. Fields of AI

Artificial intelligence is constantly present in our daily lives. We observe technological advancements at multiple levels, and AI applications are expanding. We will learn more about some of the key application domains of AI.

II.2.3.1. Health

AI is revolutionizing healthcare by enabling early disease detection, personalized treatments, and improved patient monitoring. For instance, researchers have developed an AI test that predicts which prostate cancer patients will benefit from specific drugs, optimizing treatment plans and reducing unnecessary side effects. Additionally, institutions like AIIMS Patna are incorporating AI-powered diagnostic equipment to enhance accuracy in diagnosing conditions such as cancer and heart diseases [21].

II.2.3.2. Transportation

In transportation, AI enhances traffic management and safety. Brisbane's "Smarter Suburban Corridors" trial employs AI to predict and adapt to traffic patterns, aiming to reduce commuting times by over 20%. Academic research also highlights AI's role in intelligent traffic systems, utilizing techniques like reinforcement learning and neural networks to optimize traffic flow and reduce congestion [22].

II.2.3.3. Tourism

AI significantly improves the tourism industry by enhancing customer service and personalizing experiences. AI-powered chatbots assist travelers with bookings, provide real-time support, and offer personalized recommendations based on user preferences. These chatbots are increasingly used by travel companies to handle inquiries, suggest destinations, and facilitate smoother travel planning [23].

II.2.4. AI Applications in Healthcare

Over the past decade, numerous software applications have been developed in the healthcare sector. These substantial changes have led to significant disruptions in the practices of healthcare system stakeholders. This, in turn, has contributed to facilitating treatment in this field, advancing its development, and reducing medical errors.

AI applications are divided into three sections based on their functionality

II.2.4.1. E-Health

E-health focuses on enhancing communication between patients and healthcare professionals through information and communication technologies (ICT). It aims to enrich symptom evaluation, improve management, and increase patient engagement, thereby ushering in a new era of patient-centered care where patients actively participate in their treatment processes. E-health systems are designed as service provisions that integrate ICT to optimize healthcare delivery [24].

II.2.4.2. Telehealth

Telehealth provides remote accessibility to medical and paramedical care, striving to make healthcare services universally available regardless of geographical barriers. By enabling healthcare delivery at a distance, telehealth expands the reach of medical services and improves access for underserved populations [25]

II.2.4.3. Telemedicine

Telemedicine facilitates remote medical activities, including tele-surgery, tele-diagnosis, tele-guidance, tele-monitoring, and tele-medical training. These applications empower healthcare providers to perform clinical and surgical tasks from a distance, enhancing the capacity and flexibility of healthcare networks [26]

II.3. Machine Learning (ML)

Machine learning is a subfield of artificial intelligence that refers to the use of algorithms capable of "learning" by recognizing patterns and trends from large datasets. Compared to more "traditional" statistical approaches for identifying factors or prognostic scores, such as simple linear or logistic regressions, more "advanced" approaches enable the analysis of highly diverse types of data (structured and unstructured data, text, and images) and very high-dimensional "Big Data" with complex non-linear associations, thanks to increasingly sophisticated methods and algorithms [27].

II.3.1. Types of Machine Learning

Four types of machine learning algorithms are distinguished in **Figure-II.2**: reinforcement learning, supervised learning, unsupervised learning, and semi-supervised learning [28].

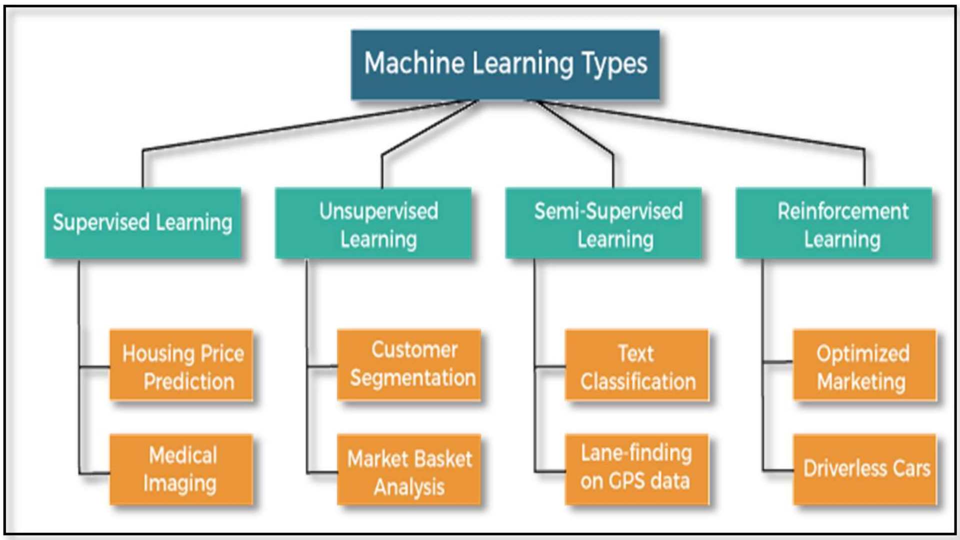

Figure II.2. Types of Machine Learning Techniques

II.3.1.1. Supervised Learning

Supervised learning is applied when the data is in the form of input variables and output target values. The algorithm learns the mapping function from the input to the output. The availability of large scale labeled data samples makes it an expensive approach for tasks where data is scarce. These approaches can be broadly divided into two main categories; Classification (where the output variable is one of some known number of categories) (See figure **Figure II.3**) and Regression (where the output variable is a real or a continuous value)

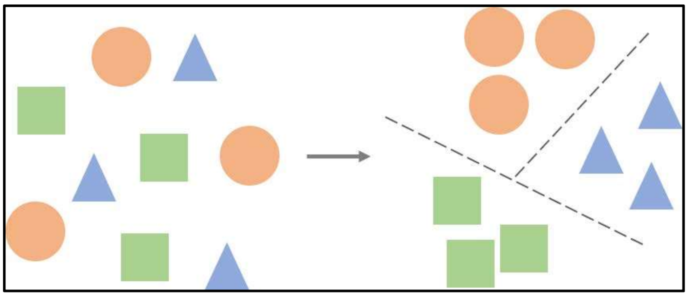

Figure II.3 Example of supervised learning [28]

II.3.1.2. Unsupervised Learning

Unsupervised learning is applied when the data is available only in the form of an input and there is no corresponding output variable. Such algorithms model the underlying patterns in the data to learn more about its characteristics. One of the main types of unsupervised algorithms is clustering. In this technique, inherent groups in the data are discovered and then used to predict output for unseen inputs. An example of this technique would be to predict customer purchasing behavior [28].

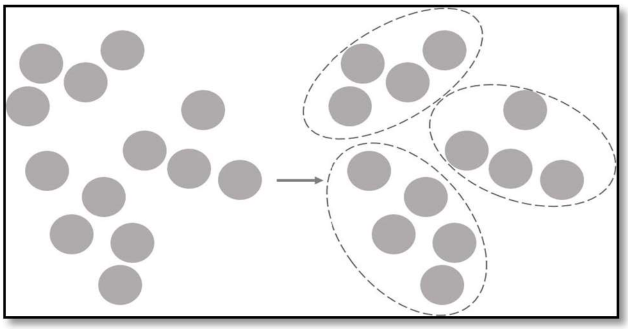

Figure II.4. Example of unsupervised learning [28]

II.3.1.3. Semi-supervised Learning

As the name suggests, this is an intermediate between supervised and unsupervised learning techniques. These algorithms are trained using a combination of labeled and unlabeled data. In a common setting, there is a small amount of labeled data and a very large amount of unlabeled data. A basic procedure involved is that first similar data is clustered using an unsupervised learning algorithm and then existing labeled data is used to label the rest of the unlabeled data.[28]

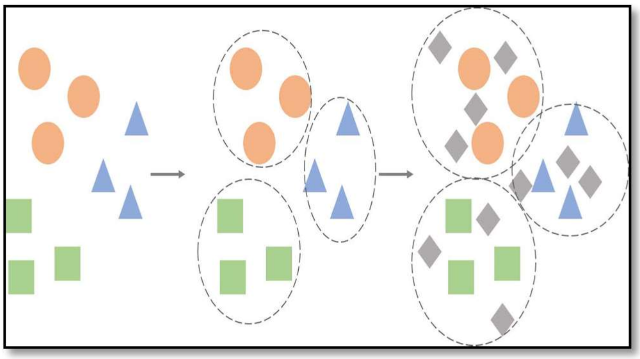

Figure II.5. Overview of semi-supervised learning [28].

II.3.1.4. Reinforcement Learning

Reinforcement learning is applied when the task at hand is to make a sequence of decisions towards a final reward. During the learning process, an artificial agent gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward. Examples include learning agents to play computer games or performing robotics tasks with end goal [28].

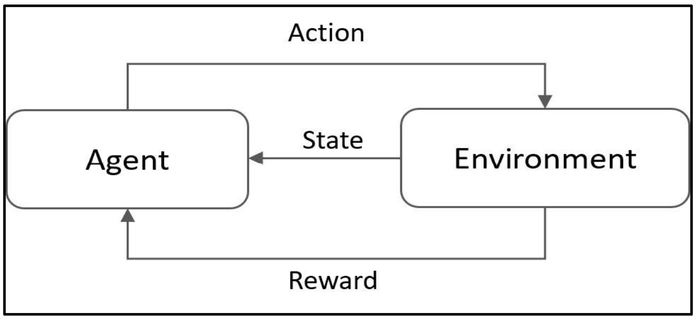

Figure II.6. Overview of reinforcement learning

II.3.2. Some Machine Learning Techniques

II.3.2.1. K-nearest neighbors (KNN)

is an "instance-based learning" or non-generalizing learning, also known as a "lazy learning" algorithm. It does not focus on constructing a general internal model; instead, it stores all instances corresponding to training data in *n*-dimensional space. KNN uses data and classifies new data points based on similarity measures (e.g., Euclidean distance function). Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is quite robust to noisy training data, and accuracy depends on the data quality. The biggest issue with KNN is to choose the optimal number of neighbors to be considered. KNN can be used both for classification as well as regression.[28]

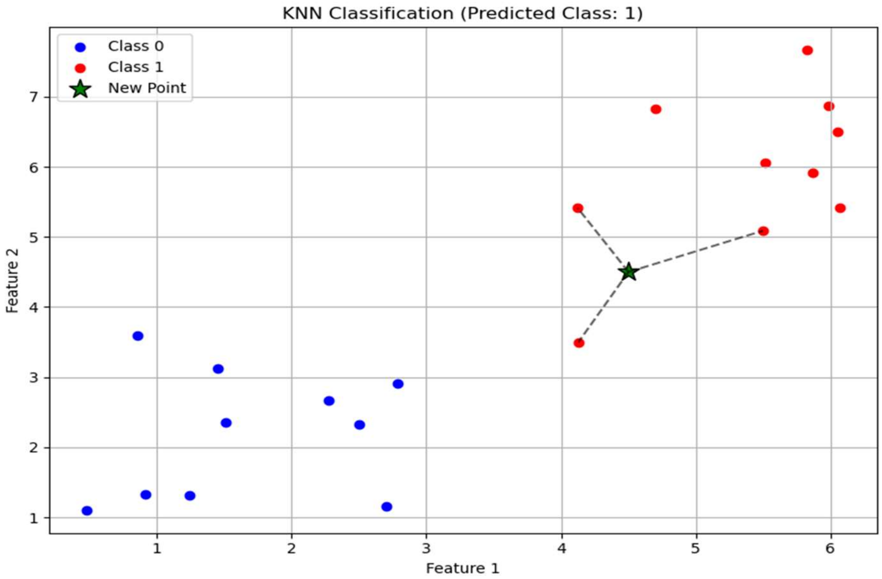

Figure II.7. KNN process.

II.3.2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification and regression tasks. It works by identifying the optimal hyperplane that separates data points of different classes with the maximum margin. SVM is particularly effective in high-dimensional spaces and is known for its robustness in handling both linear and non-linear data using kernel functions. Due to its ability to generalize well on unseen data, SVM has become a popular choice in various fields such as image recognition, bioinformatics, and medical diagnosis. Its performance and mathematical rigor make it a reliable model for both binary and multiclass classification problems [29].

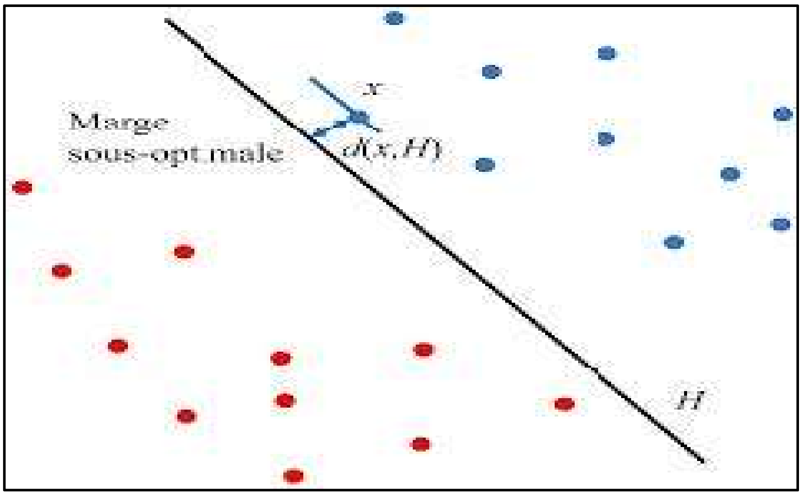

Figure II.8. SVM process

II.3.2.3. Random Forest (RF)

The random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector (Breiman, 1999). The random forest classifier used for this study consists of using randomly selected features or a combination of features at each node to grow a tree. Bagging, a method to generate a training data set by randomly drawing with replacement *N* examples [30].

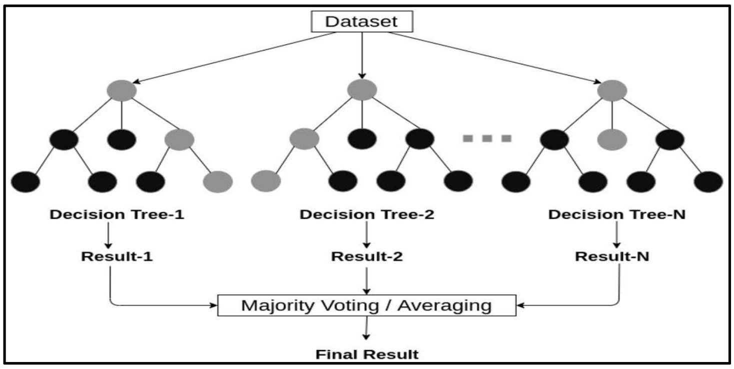

Figure II.9 random forest structure [21].

II.3.2.4. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a scalable and efficient implementation of gradient boosting machines. It has gained prominence in machine learning competitions and real-world applications due to its performance and speed. XGBoost builds an ensemble of decision trees sequentially, where each new tree aims to correct the errors made by the previous ones. The algorithm incorporates a regularized objective function to prevent overfitting and supports parallel processing, making it suitable for large-scale datasets. Its design includes a novel sparsity-aware algorithm and a weighted quantile sketch for approximate tree learning, enhancing its capability to handle sparse data and optimize computational [31].

II.3.2.5. Decision Trees (DT)

Decision Trees are a fundamental component in machine learning, serving as the base learners in ensemble methods like XGBoost. A decision tree splits the data into subsets based on feature values, creating a tree-like structure where each internal node represents a decision rule, and each leaf node represents an outcome. While individual decision trees are easy to interpret and visualize, they can be prone to overfitting, especially with complex datasets. However, when used in ensembles, such as in gradient boosting frameworks, they contribute to models that are both accurate and robust. The integration of decision trees into boosting algorithms like XGBoost leverages their simplicity while mitigating their weaknesses through ensemble learning [31].

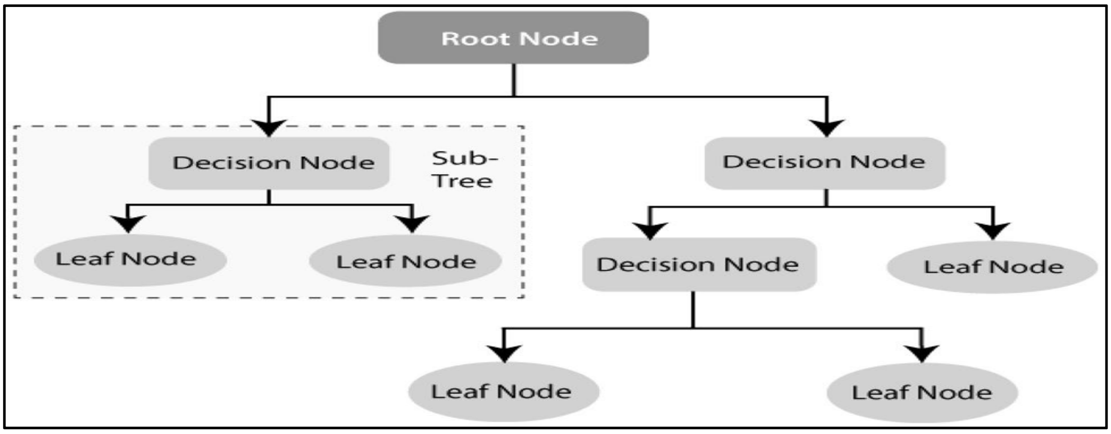

Figure II.10 An example of a decision tree structure

II.4. Deep Learning (DL)

II.4.1. Definition

Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge from experience, there is no need for a human computer operator formally to specify all the knowledge needed by the computer. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones.

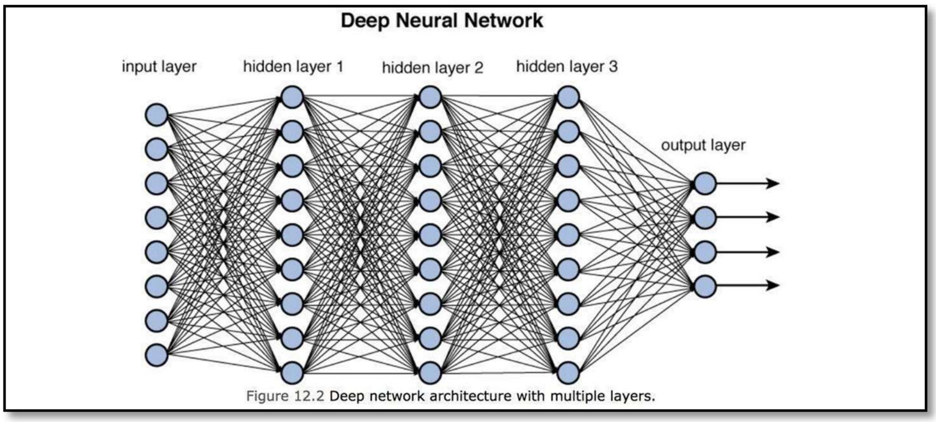

Figure II.11. Deep learning structure.

II.4.2. Deep Learning techniques

II.4.2.1. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning models particularly effective in processing data with a grid-like topology, such as images. They utilize convolutional layers to automatically and adaptively learn spatial hierarchies of features through backpropagation by applying filters to local receptive fields. CNNs have been extensively applied in various domains, including computer vision, speech recognition, and medical imaging, due to their ability to capture spatial and temporal dependencies in data. Their architecture typically comprises convolutional layers, pooling layers, and fully connected layers, enabling them to learn complex patterns with fewer parameters compared to fully connected networks., let us discuss the layers of CNN in detail as shown in **Figure II.12)** [32].

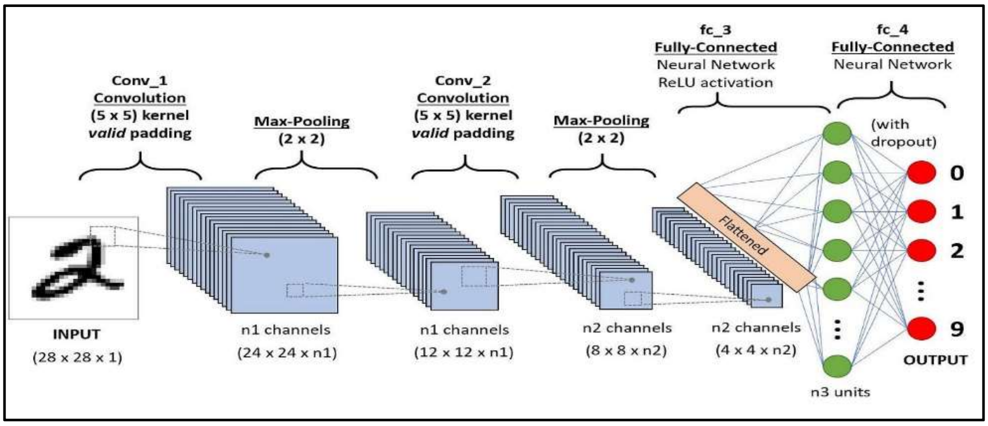

Figure II.12. Layers of CNN

II.4.2.2. Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a type of neural network designed to recognize patterns in sequences of data, such as time series, speech, or text. Unlike feedforward neural networks, RNNs have connections that form directed cycles, allowing them to maintain a 'memory' of previous inputs in their internal state. This feature makes them particularly suited for tasks where context or sequential information is crucial. However, traditional RNNs face challenges like vanishing and exploding gradients, which hinder learning long-term dependencies. To address these issues, advanced variants like Long Short-Term Memory (LSTM) networks have been developed, incorporating gating mechanisms to better capture long-range dependencies in sequential data [32].

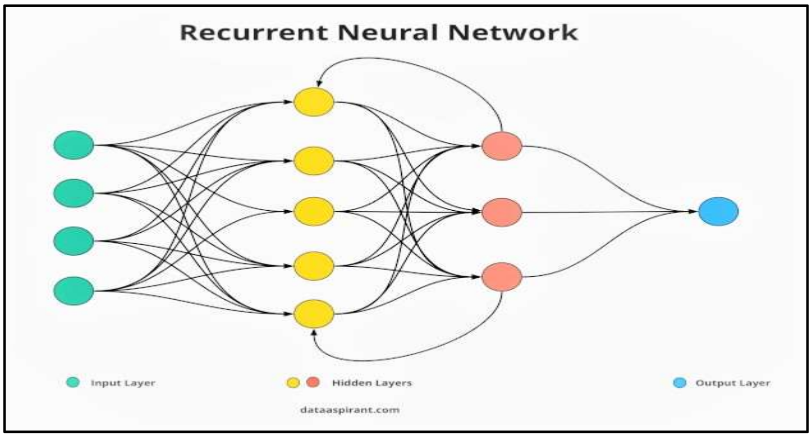

Figure II.13. Recurrent Neural Network architecture

II.5. Biomedical data

Biomedical data encompasses a diverse array of information types, each playing a crucial role in healthcare delivery and research. These data types can be broadly categorized into clinical and administrative data, biological and physiological data, and imaging and textual data [33].

II.5.1. Biomedical data types

II.5.1.1. Clinical and administrative data

Clinical and administrative data consist of both structured and unstructured information collected during patient care and healthcare operations. This includes electronic health records (EHRs), diagnostic codes, treatment plans, laboratory test results, vital signs, and prescription histories. Such data are instrumental in monitoring patient outcomes, managing healthcare resources, and informing policy decisions. They serve as the backbone for hospital management systems, epidemiological studies, and health services research, facilitating the evaluation of care quality and efficiency [33].

II.5.1.2. Biological and Physiological Data

Biological and physiological data pertain to molecular-level and real-time physiological information obtained from the human body. This category encompasses genomic and proteomic data, such as DNA and RNA sequences, gene expression profiles, and protein structures. Additionally, it includes physiological signals like electrocardiograms (ECG), electroencephalograms (EEG), and electromyograms (EMG), as well as metabolomic data and biomarker profiles. These data types are pivotal for advancing personalized medicine, enabling early disease detection, and fostering innovations in biotechnology and pharmacogenomics [33].

II.5.1.3. Imaging and Textual Data

Imaging and textual data involve visual and narrative information utilized in clinical diagnostics and documentation. Medical imaging encompasses modalities like X-rays, magnetic resonance imaging (MRI), computed tomography (CT) scans, and ultrasound, which provide detailed visualizations of internal body structures. Textual data include clinical notes, discharge summaries, and medical literature, often recorded in patient records and used for research purposes. These data types are essential for supporting diagnostic accuracy, facilitating computer-aided detection systems, and advancing fields such as natural language processing and medical knowledge discovery [33].

II.6. Conclusion

Artificial Intelligence has become a powerful tool in healthcare, enabling accurate diagnoses, personalized treatments, and efficient care through technologies like machine learning and deep learning. This chapter reviewed AI's evolution, key concepts, and its applications in areas such as e-health and biomedical imaging. AI holds great promises for improving clinical outcomes and transforming global healthcare systems.

CHAPTER III

SIMULATION AND RESULTS

CHAPTER-III
SUMILATION AND RESULTS

III.1. Introduction

In this final chapter, we provide a detailed technical study, outlining the software environment and tools employed in developing the application. We also offer an in-depth description of the dataset, its characteristics, and the necessary preprocessing steps, including data exploration, cleaning, and the selection of the optimal model. The application section follows where we present the developed graphical interfaces that visualize the system's performance and activities. The chapter concludes with a summary of the results and key insights, along with recommendations for future work.

III.2. Related works

Several recent studies have leveraged machine learning techniques to enhance early-stage diabetes prediction using clinical datasets. A five-year risk prediction model for individuals with prediabetes employed gradient boosting techniques such as LightGBM and XGboost, achieving high AUC scores and was implemented through an interactive clinical interface [34]. Another study developed a deep learning pipeline that integrated data augmentation and convolutional neural networks, achieving 92.31% accuracy on a standard clinical dataset [35]. Gene expression features were used in a separate work to train multiple classifiers, with XGBoost achieving an accuracy of 97%, demonstrating strong diagnostic capability [36]. A model using socio-demographic features and random forest reached 99.36% accuracy, making it particularly useful in settings with limited clinical data access [37]. Ensemble approaches combining AdaBoost with J48 decision trees achieved 99% accuracy when trained on questionnaire-based early-stage diabetes data [38]. Classical machine learning algorithms such as logistic regression, decision trees, and SVM also performed effectively when supported by proper feature selection and data balancing strategies [39]. A deep neural network-based model outperformed several baseline classifiers, achieving 98.1% accuracy while maintaining low training time and computational complexity [40]. Another robust predictive framework improved classification outcomes on imbalanced datasets using optimized ensemble methods [41]. A broad analysis of feature selection and classification strategies found that deep learning and hybrid ensemble techniques consistently outperformed traditional models across diverse datasets [42]. Several of these studies also integrated explainability methods like SHAP to improve interpretability and clinical trust in the prediction models [43].

III.3. Architecture of the Proposed System

The proposed system uses an architecture designed to predict early-stage diabetes risk. It starts with a clinical dataset (Early-Stage Diabetes Prediction dataset), which is processed through a preprocessing module for cleaning and feature selection. The refined data is then passed to a machine learning model to classify the risk. A user-friendly prediction interface allows users to input new data and receive results, while evaluation metrics (accuracy, precision, F1-score) assess the model's performance.

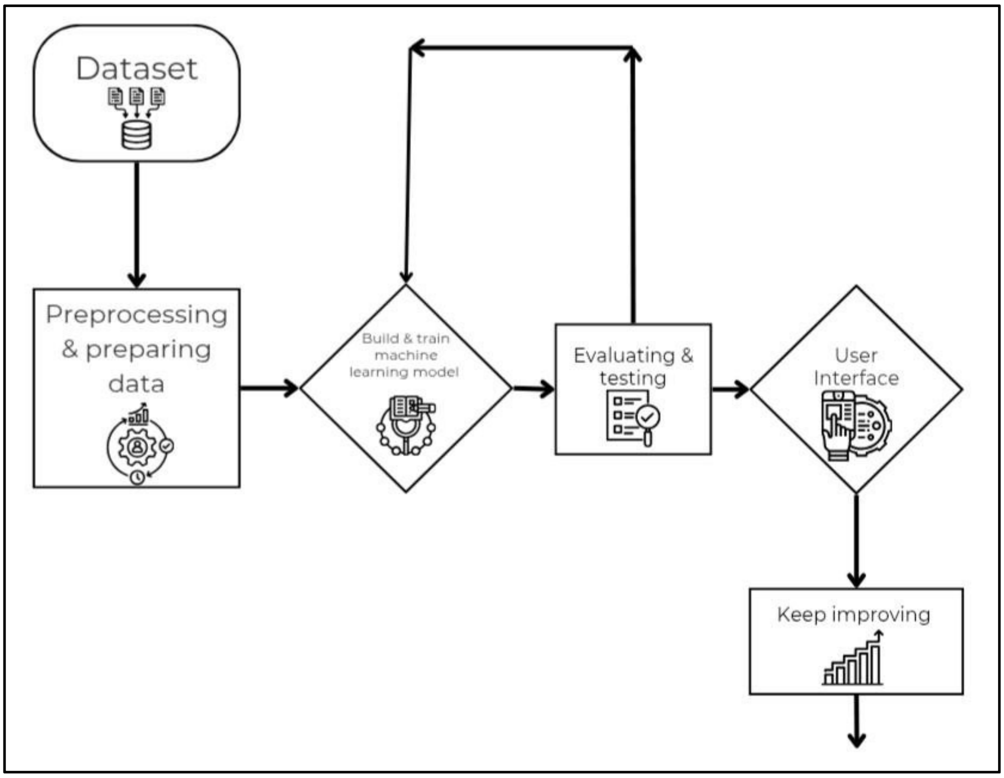

Figure III.1- architecture of the proposed system.

III.4. System Overview

The proposed system follows an architecture where clinical data is fed into a trained machine learning model to predict whether a patient is at risk of developing early-stage diabetes. The proposed system is composed of the following interconnected components:

- **Data Source**: A clinical dataset containing medical attributes related to diabetes risk (in this project we used Early-Stage Diabetes Prediction dataset).
- **Preprocessing Module**: Cleans the data, handles missing values, and applies normalization or feature selection techniques.
- **Machine Learning Model**: A trained algorithm (e.g., XGboost, LightGBM, Gradient Boosting, Bagging Classifier, Extra Trees) used to classify the risk based on input features.
- **Prediction Interface**: A user interface that accepts new data and displays prediction results.
- **Evaluation Metrics**: Used to assess the model's performance (accuracy, precision, F1-score.

III.5. Tools and Technologies Used

The proposed system uses data-driven architecture to predict early-stage diabetes risk. It starts with a clinical dataset (Early-Stage Diabetes Prediction dataset), which is processed through a preprocessing module for cleaning and feature selection. The refined data is then passed to a machine learning mode to classify the risk. A user-friendly prediction interface allows users to input new data and receive results, while evaluation metrics (accuracy, precision, F1-score) assess the model's performance.

III.5.1. Software Environment (Kaggle)

Kaggle is a premier platform for data science and machine learning, offering competitions, datasets, cloud-based notebooks, and free courses. Acquired by Google in 2017, it enables users to solve real-world problems, build portfolios, and collaborate with a global community. With tools like Jupyter notebooks and support for Python and

R, Kaggle is ideal for learners at all levels seeking practical experience and skill development in data science [44].

III.5.2. Python

Python is a widely used high-level programming language valued for its simplicity, readability, and versatility. As an interpreted language, it allows quick execution without prior compilation. Its extensive libraries—like Pandas, NumPy, and Scikit-learn—make it especially popular in data analysis, machine learning, and web development, supporting efficient and scalable application development across many fields [45].

III.5.3. Flask

Flask is a lightweight Python web framework used in this project to build a user-friendly interface for interacting with the machine learning model. It supports handling HTTP requests, rendering HTML templates, and managing user input, making it ideal for quick and flexible web app development. Flask's simplicity enabled rapid prototyping, while its extensibility allowed seamless integration with tools like Matplotlib and Pandas for dynamic result visualization. Its use of the Jinja2 templating engine also helped create interactive and informative web pages [46].

III.5.4. HTML & CSS

HTML and CSS were used together to build and style the web interface of the application. HTML structured the content by creating elements like headings, forms, tables, and visualizations to display the machine learning model's results in a clear and accessible way. CSS enhanced the layout and appearance by customizing fonts, colors, spacing, and responsiveness, making the interface visually appealing and easy to navigate across different devices. Together, HTML and CSS provided a user-friendly and interactive front-end for effective presentation and interaction with the model's predictions [47].

III.6. Implementation Stages

The implementation of a machine learning project involves several key stages. It begins with data preprocessing, where raw data is cleaned, transformed, and prepared for analysis by handling missing values, encoding categorical variables. Next is model building, where an appropriate machine learning algorithm is selected and structured based on the problem type. This is followed by model training, where the model learns patterns from the training data to make accurate predictions. Finally, model evaluation is performed using suitable metrics to assess the model's performance and ensure it generalizes well to new unseen data.

III.6.1. Dataset Description

This dataset was gathered through direct questionnaires administered to patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, and later reviewed and approved by a medical professional. Its main goal is to predict whether a patient has diabetes based on a set of relevant clinical attributes derived from their medical history and symptoms. The dataset consists of 520 records, each representing a patient's data with 16 diagnostic features and a binary target variable (Class). Unlike the Pima Indian dataset, this collection includes both male and female patients, aged between 20 and 65 years. The attributes reflect a mix of demographic information, physiological symptoms, and clinical indicators related to diabetes.

Table III.1- Definition of variables and features.

Features	**Description**	**Range**
Age	Patient's age in year	20-65 years
Sex	Biological sex	1: Male, 2: Female
Polyuria	Excessive urination	1: Yes, 2: No
Polydipsia	Excessive thirst	1: Yes, 2: No
Sudden weight loss	Unintended rapid weight loss	1: Yes, 2: No
Weakness	Persistent fatigue or weakness	1: Yes, 2: No
Polyphagia	Excessive hunger	1: Yes, 2: No

Genital thrush	Yeast infection in genital area	1: Yes, 2: No
Visual blurring	Blurred vision	1: Yes, 2: No
Itching	Persistent skin itching	1: Yes, 2: No
Irritability	Increased irritability	1: Yes, 2: No
Delayed healing	Slow wound recovery	1: Yes, 2: No
Partial paresis	Muscle weakness	1: Yes, 2: No
Muscle stiffness	Reduced muscle flexibility	1: Yes, 2: No
Alopecia	Hair loss	1: Yes, 2: No
Obesity	Presence of obesity	1: Yes, 2: No
Class	Target variable	1: Positive, 2: Negative

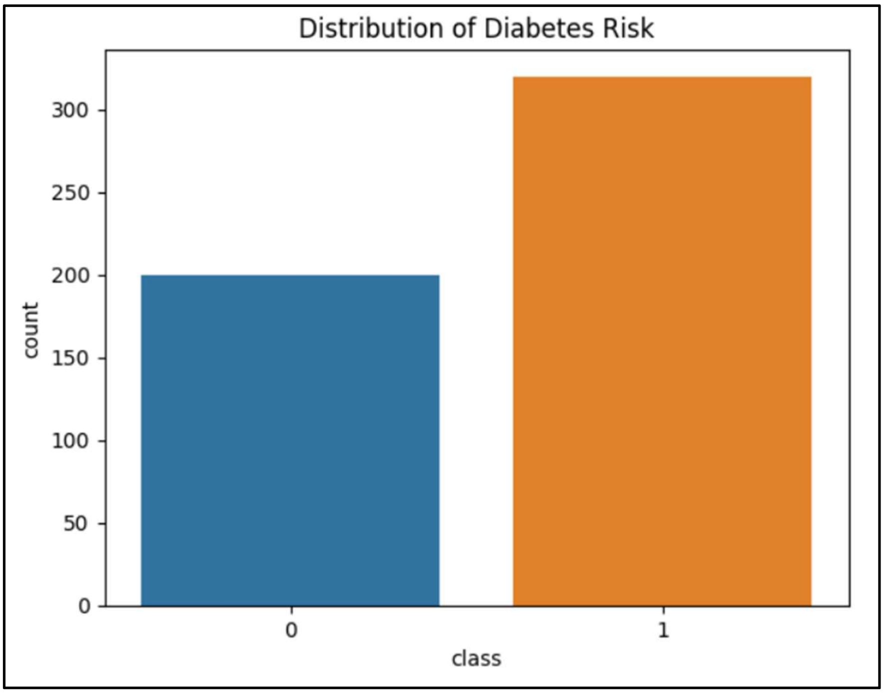

Figure III.2- Distribution of dataset's target.

III.6.2. Data Preprocessing

As evident from the output, the dataset contains no missing values across any of its features—an excellent starting point for analysis. A clean dataset like this allows us to proceed directly to modeling without the need for imputing missing data or removing incomplete records. For diabetes prediction, this is particularly crucial, as gaps in medical data could compromise model reliability.

```
# Check for missing values
df.isnull().sum()
```

```
Age                  0
Gender               0
Polyuria             0
Polydipsia           0
sudden weight loss   0
weakness             0
Polyphagia           0
Genital thrush       0
visual blurring      0
Itching              0
Irritability         0
delayed healing      0
partial paresis      0
muscle stiffness     0
Alopecia             0
Obesity              0
class                0
dtype: int64
```

Figure III.3-Checking missing values.

III.6.3. Model Building

Building a machine learning model is a critical step in the implementation process. It involves selecting the appropriate type of model based on the nature of the problem, such as classification, regression, or clustering. In our case, since the goal is to predict whether a person is diabetic or not, the problem is clearly a classification task. Therefore, we focused on using powerful classification algorithms such as Extra Trees, Bagging Classifier, XGBoost, LightGBM, and Gradient Boosting, which are well-suited for handling structured tabular data and can provide high accuracy and robustness. After selecting the appropriate algorithms, the models were initialized and configured with suitable parameters. Careful consideration was given to how the models handle input features and how they balance learning capacity with the risk of overfitting or underfitting. This stage sets the foundation for effective training and accurate predictions in later phases.

III.6.4. Selection of Algorithms

algorithms was carefully selected. The aim was to include algorithms from different methodological families to ensure a broad exploration of modeling techniques suitable for classification tasks. The selection criteria were based on algorithm popularity, proven effectiveness in prior research, availability in standard machine learning libraries, and their ability to handle different data characteristics such as linearity, non-linearity, noise, and dimensionality. The selected algorithms are:

- **Gradient Boosting** – Gradient Boosting is a powerful ensemble machine learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees. It works by sequentially training models, where each new model tries to correct the errors of the previous ones. The goal is to minimize a loss function using gradient descent.

- **Extra Trees** – short for **Extremely Randomized Trees**, is an ensemble machine learning algorithm based on decision trees. It is like the Random Forest algorithm but introduces more randomness during the tree-building process to reduce variance and improve generalization.

- **Bagging Classifier** – The Bagging Classifier (Bootstrap Aggregating) is an ensemble method that builds multiple models on random subsets of the training data and combines their predictions using majority voting. By reducing variance and overfitting, it enhances the stability and accuracy of classification models, especially when using decision trees as base learners.

- **XGBoost** – is a powerful and efficient implementation of the gradient boosting algorithm. It builds decision trees sequentially, where each new tree corrects the errors of the previous ones. Known for its speed and high accuracy, XGBoost includes regularization techniques to prevent overfitting, making it a popular choice for structured data and classification tasks.

- **LightGBM:** is a high-performance gradient boosting framework designed for speed and efficiency. It uses histogram-based algorithms and grows trees leaf-wise rather than level-wise, which allows it to handle large datasets with lower memory usage. LightGBM is widely used for classification tasks due to its fast training and high accuracy.

- This selection ensures a balanced experimentation setup that can reveal the strengths and weaknesses of different algorithms across varying conditions.

III.6.5. Model Training

After selecting the appropriate algorithms, the next step involved training each model using the prepared dataset. The training process was carried out using a supervised learning approach, where the models learned to map input features to the binary output label indicating the presence or absence of early-stage diabetes.

To ensure robust evaluation and prevent overfitting, the dataset was split into two subsets:

- **Training Set (80%)**: Used for model fitting.
- **Testing Set (20%)**: Reserved for evaluating the model's performance on unseen data.

The ***train_test_split*** function from the Scikit-learn library was used to perform the split in a stratified manner to preserve class distribution.

Each model was trained using default hyperparameters initially. The models were then evaluated using the testing set to compare baseline performance. The training process included the following steps:

- Fitting the model to the training data using the ***fit()*** method.
- Evaluating model performance using accuracy, precision, and F1-score
- Storing results for comparison and visualization.

This phase provided the foundation for the next step, which involved optimizing the models through hyperparameter tuning to enhance prediction performance.

III.6.6. Model Evaluation

Once the models were trained, their performance was evaluated using a variety of classification metrics to determine their effectiveness in predicting early-stage diabetes.

Evaluation was conducted on the testing set, which contained unseen data not used during training. This step is crucial to assess the generalization capability of the models. The confusion matrix (as shown in **Figure III.4**) is a widely used tool for summarizing the performance of classification models. It provides a clear overview of how well the model distinguishes between different classes. In classification tasks, each prediction falls into one of four categories. If the actual class is positive and the model correctly predicts it as positive, it is labeled a True Positive (TP). If the model incorrectly predicts a negative outcome for a true positive case, it is a False Negative (FN). Conversely, when the actual class is negative and the model predicts it correctly, it is a True Negative (TN). However, if the model incorrectly predicts a positive outcome for a negative case, this results in a False Positive (FP).

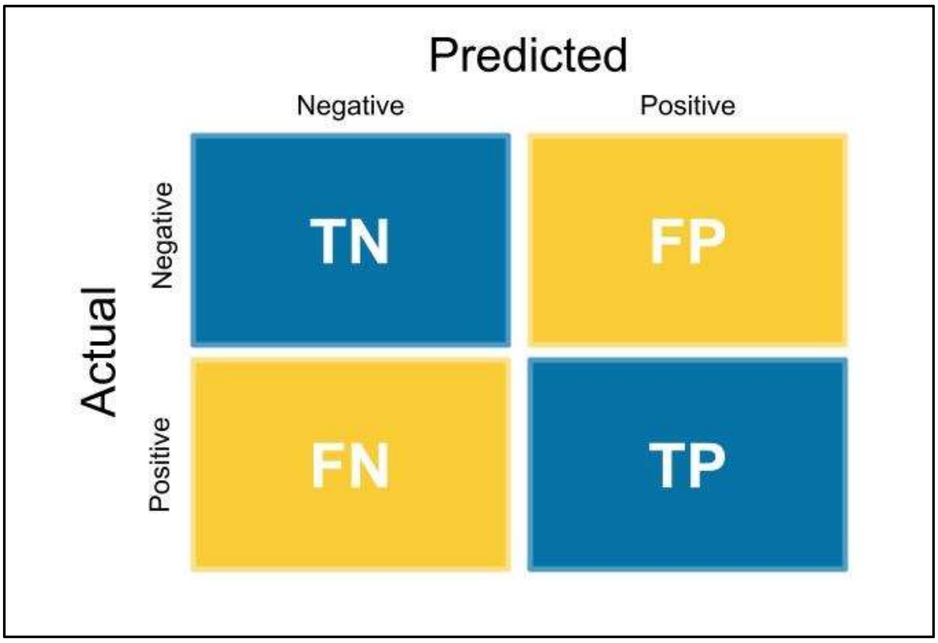

Figure III.4- A confusion matrix.

Using the confusion matrix allows for a straightforward assessment of the model's accuracy and reliability. Based on this matrix, several evaluation metrics are calculated to analyze the performance of the proposed model. Among the most important evaluation metrics used in this study are Accuracy, F1-core, and Precision.

- **Accuracy:** It measures the overall correctness of the model by calculating the ratio of correctly predicted observations (both positive and negatives) to the total number of predictions. While it provides a general sense of performance, it may not always be reliable when dealing with imbalanced datasets [48].

$$Accuracy = \frac{TP + TN}{\mathrm{TP + TN + FP + FN}} \quad (1)$$

- **Precision** focuses specifically on the positive class and represents the proportion of true positive predictions out of all instances predicted as positive. High precision indicates that the model makes few false positive errors, which is especially important in medical applications like diabetes prediction [49].

$$\mathrm{Precision} = \frac{TP}{\mathrm{TP + FP}} \quad (2)$$

- **F1-Score** is the harmonic mean of precision and recall. It provides a balanced measure that takes both false positives and false negatives into account, making it particularly useful when the class distribution is uneven or when both precision and recall are important for decision-making [50].

$$\mathrm{F1 - score} = \frac{2 * \mathrm{TP}}{2 * TP + FN + FP} \quad (3)$$

III.7. Experimental work and Results

This section presents the experiments conducted, including the implementation of various machine learning algorithms and ensemble techniques. It also provides a detailed comparison of the obtained results using multiple evaluation metrics to assess model performance.

III.7.1. Experiment 1: Using Machine Learning Algorithms

In this experiment, five machine learning algorithms were employed to evaluate and compare their performance on the selected dataset. These algorithms are XGBoost, Extra Trees, LightGBM, Gradient Boosting, and Bagging Classifier. They were chosen to represent a variety of ensemble and boosting techniques, allowing for a comprehensive assessment of their accuracy and classification efficiency. This selection ensures a robust comparison across different ensemble learning strategies. The results obtained after implementing and evaluating these algorithms are presented in Table III.2.

Table III.2 - Results were obtained after implementation and evaluation.

Model	Performance (%)		
	Accuracy	**F1- Score**	**Precision**
Extra Trees	99.04	99.29	100.0
Bagging Classifier	99.04	99.29	100.0
XGboost	98.08	98.57	100.0
LightGBM	98.08	98.57	100.0
Gradient Boosting	97.12	97.84	100.0

III.7.2. Experiment 2: Using Ensemble Learning

In this experiment, we explore the effectiveness of an ensemble learning approach by combining multiple high-performing tree-based classifiers. The goal is to improve the classification performance observed in the previous experiment by leveraging the collective strengths of several advanced models, including Extra Trees, Bagging Classifier, XGBoost, LightGBM, and Gradient Boosting. These models are known for their robustness, efficiency, and high accuracy in a wide range of classification problems.

We employ a Voting Classifier, which aggregates the individual predictions from different models using a majority voting strategy. This method is intended to reduce individual model biases and variances, and to produce a more stable and accurate final prediction. Through this experiment, we aim to assess whether this ensemble configuration can outperform each of the individual models and provide better generalization on unseen data. **Table III.3** demonstrated that the ensemble model outperformed the individual classifiers in most evaluation metrics. This indicates that combining multiple strong learners through ensemble voting contributed to improved overall performance and robustness.

Table III.3- Results of individual models and voting ensemble.

Performance (%)			
Model	**Accuracy**	**F1- Score**	**Precision**
Soft Voting	**100.0**	**100.0**	**100.0**
Hard Voting	**100.0**	**100.0**	**100.0**

the confusion matrix of ensemble learning machine mentioned give us

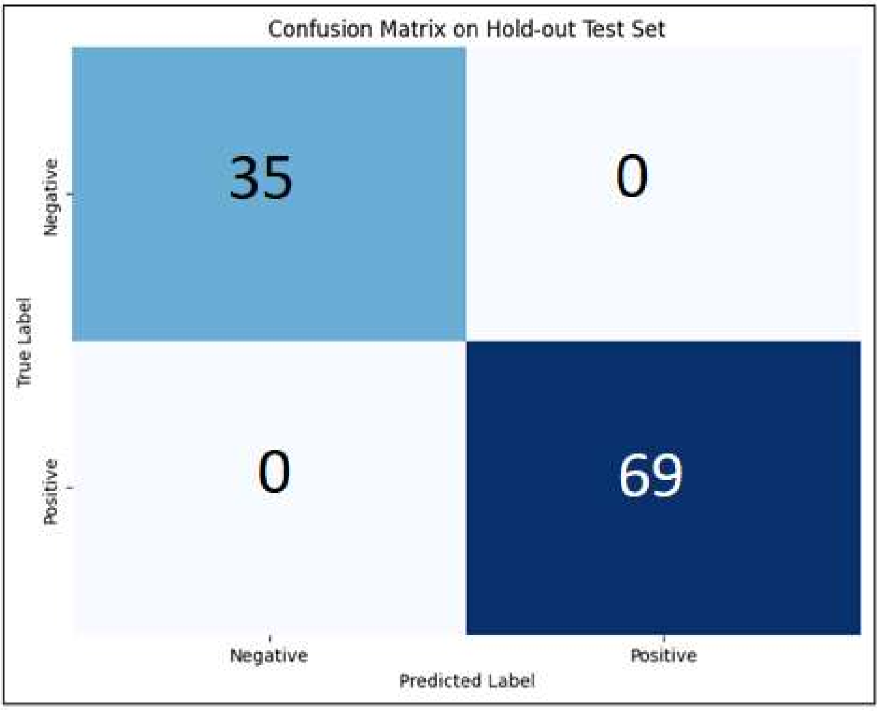

Figure III.5- Confusion matrix for the soft voting method.

III.8. Platform Deployment

To make the prediction model accessible and user-friendly, a web application was developed using the Flask framework. Flask is a lightweight and flexible Python web framework that allows quick deployment of machine learning models through simple APIs and user interfaces. The deployment process involved the following key steps:

- **Model Saving**: The final trained model was serialized using the *joblib* library to allow easy loading without the need for retraining.
- **Web Application**: A web interface was created using Flask, where users can input clinical features such as gender, age, and other diabetes-related symptoms. The application then processes this input and displays the prediction result (e.g., likely diabetic or not diabetic).
- **Prediction Pipeline Integration**: All necessary preprocessing steps (e.g., data scaling, encoding) were incorporated within the Flask backend to ensure consistency between training and inference.
- **User Experience Design**: The web interface was designed to be simple and intuitive, making it accessible for non-technical users, such as medical staff or patients.
- **Testing and Validation**: The deployed app was tested with different input scenarios to ensure stability, correctness, and real-time responsiveness.

This deployment makes it possible to use the trained model in real-world scenarios, potentially assisting in early diagnosis and awareness about diabetes risk.

III.8.1. Home Page of the Web Application

This interface displays the input section of the Diabetes Risk Assessment system. It allows the user to enter their biological information and select relevant symptoms. This data is then used by the model to predict the likelihood of early-stage diabetes.

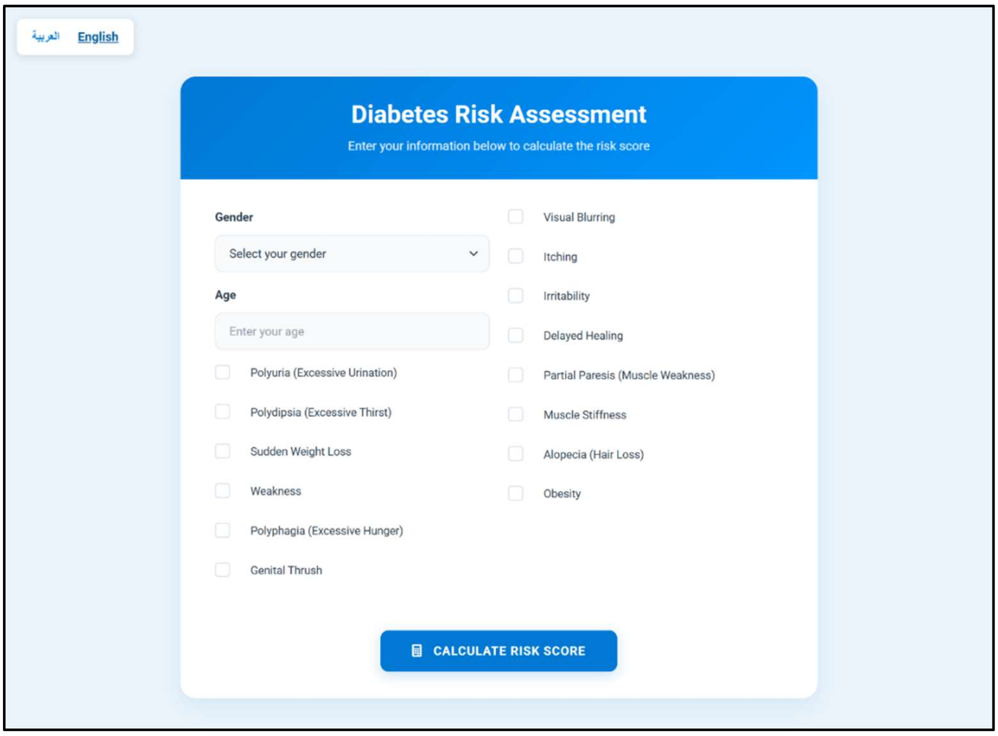

Figure III.6 – Diabetes prediction system interface.

III.8.2. Prediction Output

Figure III.7 shows the result page of the Diabetes Risk Assessment system. After processing the user's input data, the system displays a clear prediction outcome (positive or negative) along with a corresponding risk level. The result is visually supported by a color-coded scale, helping users easily understand the degree of risk and take appropriate action if needed.

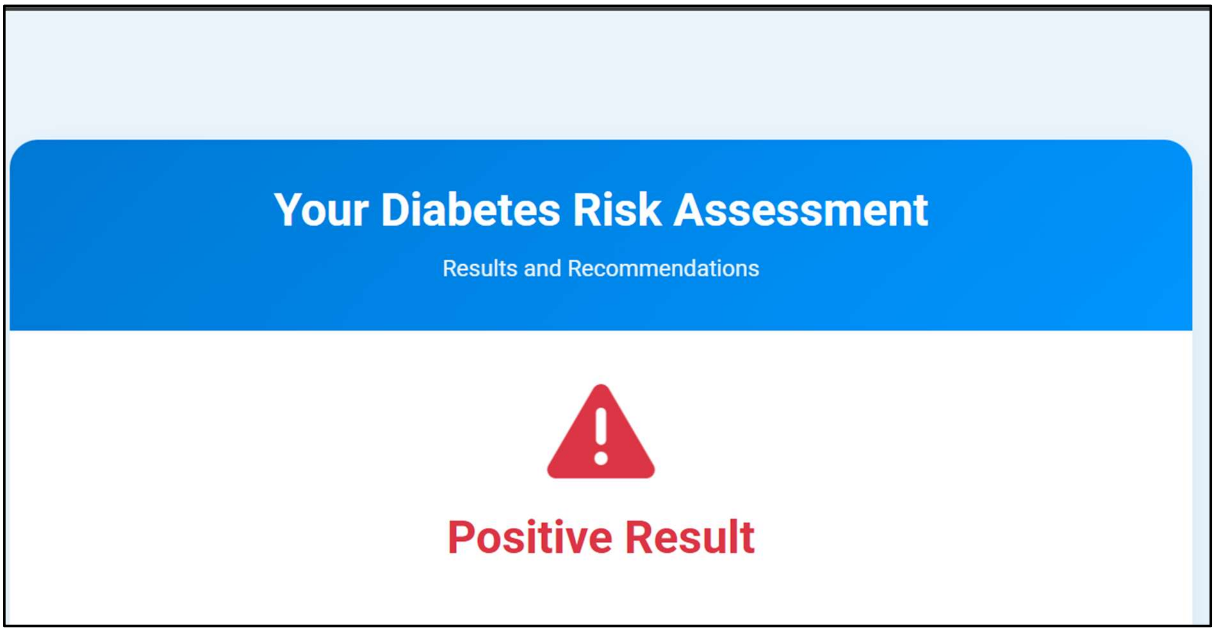

Figure III.7- Prediction result displayed to the user.

III.8.3. The recommendations section

This section of the system provides personalized recommended actions based on the predicted diabetes risk. It suggests lifestyle modifications such as maintaining a healthy diet, engaging in regular physical activity, managing weight, and reducing stress. A disclaimer is also included to clarify that the assessment is informational and not a substitute for professional medical advice.

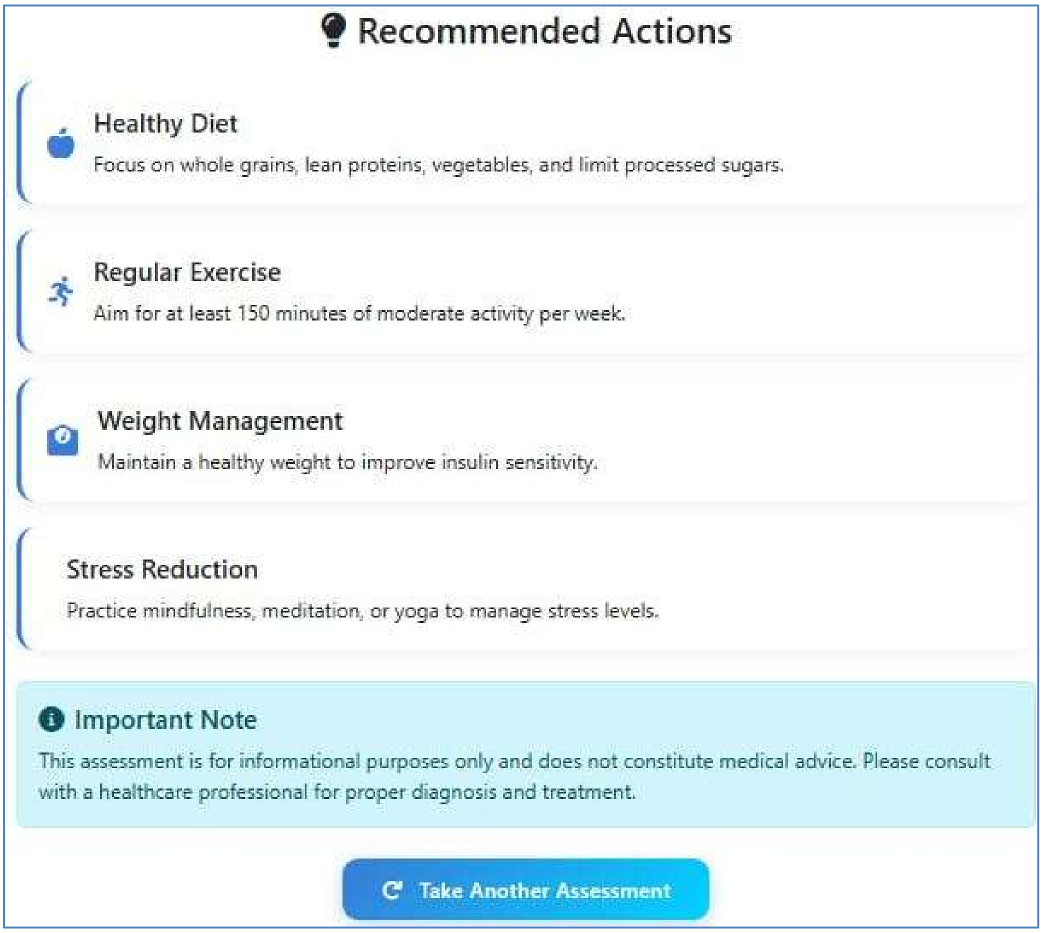

Figure III.8 – recommended actions section.

III.9. Comparison with the State-of-the Art Methods

Compared to recent state-of-the-art approaches for diabetes prediction, the proposed ensemble-based model demonstrates superior performance across all evaluation metrics. While prior studies reported high accuracies such as 99.36% with Random Forest on socio-demographic features [37], 99% using AdaBoost and J48 [38], and 98.1% using a deep neural network [40], our Voting Classifier ensemble achieved 100% accuracy, precision, and F1-score. This ensemble integrated five strong learners: Extra Trees, Bagging Classifier, XGBoost, LightGBM, and Gradient Boosting. Individually, the base classifiers performed well. The ensemble effectively aggregated their strengths using majority voting, resulting in perfect classification performance on the Early-Stage Diabetes Risk Prediction Dataset. Unlike some studies that relied on deep learning architecture or complex feature engineering, our model-maintained simplicity by utilizing only categorical encoding, benefiting from the dataset's high quality and clinical relevance. Although no interpretability methods such as SHAP or feature attribution tools were employed, the model's design and performance metrics demonstrated its potential for highly reliable diabetes prediction in early stages.

Table III.4- Comparison with the State-of-the Art Methods

Ref	**Datasets**	**Methods**	**Performance (%)**		
			Accuracy	**F1-Score**	**Precision**
[37]	Early-stage dataset	Random Forest	99.36	100	99.24
[38]	Early-stage dataset	AdaBoost	99.00	100	99.00
[40]	Early-stage dataset	deep neural network	98.1	100	98.22

III.10. Conclusion

In this chapter, we presented the full development cycle of an intelligent system for early-stage diabetes prediction using clinical data. The process included data collection, preprocessing, model selection and training, performance evaluation, and finally, deployment via a web-based interface for user interaction. Various machine learning algorithms were tested and compared to identify the most suitable models for this classification task. Notably, the use of ensemble learning techniques allowed us to achieve perfect performance with the accuracy 100% among all tested models, highlighting their strength in combining multiple weak learners to produce more robust and reliable predictions. The final system not only delivers accurate predictions but also provides an intuitive interface that makes it accessible for both healthcare professionals and individuals. These results demonstrate the promising role of artificial intelligence in preventive healthcare and support the adoption of such intelligent tools in real-world medical applications.

CONCLUSION

CONCLUSION

In this work, we have proposed a comprehensive intelligent system for the early detection and prediction of diabetes using both clinical data and machine learning techniques. The growing global burden of diabetes and its severe complications call for efficient and accessible diagnosis tools. Traditional diagnostic methods often suffer from delays, limited scalability, and the need for expert involvement. However, the integration of artificial intelligence, especially machine learning, has shown great potential in overcoming these limitations.

To address this, we have developed a diabetes prediction system using a variety of machine learning models including XGBoost ‹Extra Trees ‹LightGBM ‹Gradient Boosting ‹Bagging Classifier .... Each model was evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-score to ensure robust and fair comparison.

The dataset used in this work was the Early-Stage Diabetes Risk Prediction dataset, which includes clinical symptoms and patient demographics. Comprehensive preprocessing steps were applied, including data cleaning, transformation, and visualization to enhance model performance.

For deployment, a user-friendly web application was implemented using Flask to make the prediction system accessible to non-technical users such as healthcare providers or patients. The system allows users to input relevant clinical data and instantly receive a prediction regarding their diabetes risk.

Among all tested models, Ensemble Learning provided the highest prediction accuracy, making it the best candidate for real-world implementation.

- Collaboration with healthcare professionals to validate the system's predictions against clinical assessments and ensure alignment with real medical practices.
- Patient education and awareness: Developing a simplified interface or mobile application aimed at educating patients about diabetes risk factors and prevention strategies based on their inputs

REFERENCES

[1] Perumal, S., & Dhanasekaran, S. (2023). A Study on Role of Health Care Management Streams. Shanlax International Journal of Economics. https://doi.org/10.34293/economics.v11i3.6197.

[2] Parker, D. (2024). The essential role of healthcare services in modern society. Journal of Health Care Communications, 9(4), Article 39. https://www.primescholars.com/articles/the-essential-role-of-healthcare-services-in-modern-society-130625.html

[3] Abelson, J., Allin, S., Grignon, M., Pasic, D., & Walli-Attaei, M. (2017). Uncomfortable trade-offs: Canadian policy makers' perspectives on setting objectives for their health systems.. Health policy, 121 1, 9-16 . https://doi.org/10.1016/j.healthpol.2016.10.007

[4] Woods, L., Eden, R., Green, D., Pearce, A., Donovan, R., McNeil, K., & Sullivan, C. (2024). Impact of digital health on the quadruple aims of healthcare: A correlational and longitudinal study (Digimat Study). International journal of medical informatics, 189, 105528 . https://doi.org/10.1016/j.ijmedinf.2024.105528.

[5] Wikipedia contributors. (2024, April 14). Disease. Wikipedia. https://en.wikipedia.org/wiki/Disease

[6] Lichtenberg, F. (1992). Pathology of Infectious Diseases. Histopathology, 21, 293-296. https://doi.org/10.1111/J.1365-2559.1992.TB00394.X.

[7] Tian, C., Hromatka, B., Kiefer, A., Eriksson, N., Noble, S., Tung, J., & Hinds, D. (2016). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nature Communications, 8. https://doi.org/10.1038/s41467-017-00257-5.

[8] Pottie, K., & Girard, V. (2021). Common Infectious Diseases.. Primary care, 48 1, 45-55 . https://doi.org/10.1016/j.pop.2020.11.002.

[9] Mayo Clinic. (n.d.). Infectious diseases - Symptoms & causes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173

[10] World Health Organization. (2024, April 14). Noncommunicable diseases. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[11] World Health Organization. (2023). *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[12] nutrident. Le Diabète. https://nutrident.fr/diabete/

[13] Diabete-type-1-2-definition-causes-symptomes-traitement-regime-complications/#diabete-alimentation

[14] TOOMED. Tous savoir sur le matériel médical et la santé. https://www.toomed.com/blog/le-diabete-quest-ce-que-cest/?cn-reloaded=1

[15] Doctissimo. Santé. Disponible sur :

https://www.doctissimo.fr/html/sante/encyclopedie/sa_1290_diab_02.htm

[16] Zhu, Y., & Zhang, C. (2016). Gestational diabetes mellitus: Prevalence, risk factors, maternal and fetal outcomes, and screening and diagnosis. Nature Reviews Endocrinology, 12(11), 639–649. https://doi.org/10.1038/nrendo.2016.98

[17] Federation français de Cardiologie.Réduire le risque cardio-vasculaire LE DIABETE. [en ligne]. Disponible sur : https ://www.fedecardio.org/Je-minforme/ Reduire-le-risque-cardio-vasculaire/le-diabete

[18] Docteurclic. Diabète non insulino-dépendant, diabète de type 2 ou DNID. Disponible sur : https://www.docteurclic.com/maladie/diabete-non-insulino-dependant-type-2-DNID.aspx

[19] Kreps, S. E., & Kriner, D. L. (2023). A scholarly definition of artificial intelligence (AI). *Journal of Information Technology & Politics*, 20(1), 1–15. https://doi.org/10.1080/10584609.2023.2290497

[20] Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5–14. https://doi.org/10.1177/0008125619864925

A Rouyar, É Alvarez Andres, et al. Intelligence artificielle en radiothérapie:

radiomique, pathomique, et prédiction de la survie et de la réponse aux traitements. Cancer/Radiotherapies', 25(6-7) :630–637, 2021

[21] The Guardian. (2025, May 30). *New AI test can predict which men will benefit from prostate cancer drug*. https://www.theguardian.com/society/2025/may/30/new-ai-test-can-predict-which-men-will-benefit-from-prostate-cancer-drug

[22] Saxena, R. R. (2024). *Artificial Intelligence in Traffic Systems*. arXiv. https://arxiv.org/abs/2412.12046

[23] Iancu, I., & Iancu, B. (2023). *Artificial Intelligence in Tourism Through Chatbot Support in the Booking Process—An Experimental Investigation*. *Tourism and Hospitality*, 6(1), 36. https://www.mdpi.com/2673-5768/6/1/36

[24] Sharma, A., Bhatia, M., & Chawla, A. (2022). Role of e-health in modern healthcare: A comprehensive review. *Journal of Healthcare Engineering*,

2022, Article ID 7592486. https://doi.org/10.1155/2022/7592486

[25] Bashshur, R. L., Doarn, C. R., Frenk, J. M., Kvedar, J. C., & Woolliscroft, J. O. (2020). Telemedicine and the COVID-19 pandemic, lessons for the future. *Telemedicine and e-Health*, 26(5), 571-573. https://doi.org/10.1089/tmj.2020.29040.rb

[26] Tuckson, R. V., Edmunds, M., & Hodgkins, M. L. (2017). Telehealth. *The New England Journal of Medicine*, 377(16), 1585-1592. https://doi.org/10.1056/NEJMsr1503323

[27] R Sun, M Lerousseau, T Henry, A Carré, A Leroy, T Estienne, S Niyoteka, S Bockel,.

[28] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21

[29] Evgeniou, T., & Pontil, M. (2001). Support vector machines: Theory and applications. In Lecture Notes in Computer Science (Vol. 2049, pp. 249–257)

[30] Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222

[31] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://arxiv.org/abs/1603.0275

[32] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

[33] . Elzeheiry, H. A., Barakat, S., & Rezk, A. (2022). Different Scales of

Medical Data Classification Based on Machine Learning Techniques: A Comparative Study. *Applied Sciences, 12*(2), 919. https://doi.org/10.3390/app12020919

[34] Park, J., Lee, S., Kim, H., & Choi, J. (2025). Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes. *JMIR*, 1, e73190. https://www.jmir.org/2025/1/e73190

[35] García-Ordás, M. T., & Hidalgo, J. (2024). Diabetes Detection Using Deep Learning Techniques with Oversampling and Feature Augmentation. *arXiv*. https://arxiv.org/abs/2402.02188

[36] Roy, A., Banerjee, S., & Das, P. (2024). Leveraging Gene Expression Data and Explainable Machine Learning for Enhanced Early Detection. *arXiv*. https://arxiv.org/abs/2411.14471

[37] Arif, M., Alhussain, T., & Hossain, M. S. (2024). Machine Learning-Based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics. *Algorithms*, 16(11), 503. https://www.mdpi.com/1999-4893/16/11/503

[38] Palabaş, M. (2023). Early-Stage Diabetes Risk Prediction Using Machine Learning Techniques Based on Ensemble Approach. *Estubtdc Journal*, 8(2), 123-135. https://dergipark.org.tr/en/pub/estubtdc/issue/86402/1320922

[39] Kumari, P., Singh, K., & Sharma, V. (2023). Forecast the Early Stage of Diabetes Mellitus Using Machine Learning. In *Advances in Computational Intelligence* (pp. 523-532). Springer. https://link.springer.com/chapter/10.1007/978-981-99-3485-0_45

[40] Beghriche, T., Djerioui, M., Brik, Y., & Belhaouari, S. B. (2022). An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. *ResearchGate*.

https://www.researchgate.net/publication/357141281

[41] Ahmed, S., Khan, R., & Islam, S. (2024). Robust Predictive Framework for Diabetes Classification Using Optimized Machine Learning on Imbalanced Datasets. *Frontiers in Artificial Intelligence*. https://www.frontiersin.org/articles/10.3389/frai.2024.1499530/full

[42] Aziz, S., Anwar, S., & Aslam, N. (2024). Analyzing Classification and Feature Selection Strategies for Diabetes Prediction. *Frontiers in Artificial Intelligence*. https://www.frontiersin.org/articles/10.3389/frai.2024.1421751/full

[43] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://arxiv.org/abs/1705.07874

[44] . GeeksforGeeks. (2023, February 5). Run R Markdown in Kaggle. Retrieved from https://www.geeksforgeeks.org/run-r-markdown-in-kaggle/

[45] GeeksforGeeks. (2023, January 31). Python – The New Generation Language. https://www.geeksforgeeks.org/python-the-new-generation-language/

[46] Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python* (2nd ed.). O'Reilly Media.

[47] Duckett, J. (2014). *HTML and CSS: Design and Build Websites*. John Wiley & Sons.

[48] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R* (Springer Texts in Statistics). Springer. https://doi.org/10.1007/978-1-4614-7138-7

[49] Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of

Machine Learning Technologies, 2(1), 37–63. https://arxiv.org/abs/2010.16061

[50] Sokolova, M., & Lapalme, G. (2009). *A Systematic Analysis of Performance Measures for Classification Tasks*. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

ملخـــــص

يُعد التشخيص المبكر لمرض السكري عاملًا أساسيًا للتدخل الفعّال وتحسين نتائج العلاج. تهدف هذه الدراسة إلى تطوير إطار تشخيصي يعزز دقة الكشف المبكر من خلال دمج تقنيات التعلم الآلي والتعلم العميق مع تحليل البيانات السريرية. تم استخدام مجموعة بيانات واقعية تشمل مؤشرات مثل مستوى الجلوكوز، مؤشر كتلة الجسم، العمر، وضغط الدم، لتقييم أداء نماذج متنوعة كأشجار القرار، SVM، الغابات العشوائية، والشبكات العصبية العميقة. شملت المنهجية خطوات معالجة مسبقة للبيانات واختيار الميزات المهمة. أظهرت النتائج أن الجمع بين النماذج العميقة والتحليل السريري يُحسّن الأداء بشكل ملحوظ مقارنة بالأساليب التقليدية، مما يؤكد دور الذكاء الاصطناعي في دعم التشخيص المبكر والدقيق لمرض السكري.الدقيق والمبكر في

الكلمات المفتاحية : تشخيص السكري، التعلم الآلي، التعلم العميق، تحليل البيانات السريرية

Abstract

Early diagnosis of diabetes is essential for timely intervention and improved patient outcomes. This study proposes an enhanced diagnostic framework that integrates machine learning (ML), deep learning (DL), and clinical data analysis. Using a real-world dataset containing clinical features such as glucose level, BMI, age, and blood pressure, we evaluate the performance of ML models (e.g., decision trees, SVM, random forests) and DL models (e.g., deep neural networks). Data preprocessing and feature selection techniques are applied to improve accuracy and reduce noise. Experimental results show that deep learning models, especially when combined with clinical insights, outperform traditional approaches in early diabetes prediction. This research highlights the potential of AI-based systems to support accurate and early clinical decision-making in diabetes care.

Keywords: Diabetes Diagnosis, Machine Learning, Deep Learning, Clinical Data Analysis

Résumé

Le diagnostic précoce du diabète est essentiel pour une prise en charge rapide et une amélioration significative des résultats cliniques. Cette étude présente un cadre diagnostique avancé combinant apprentissage automatique (ML), apprentissage profond (DL) et analyse des données cliniques. Un ensemble de données réelles, contenant des variables telles que le taux de glucose, l'IMC, l'âge et la pression artérielle, est utilisé pour entraîner des modèles ML (arbres de décision, SVM, forêts aléatoires) et DL (réseaux neuronaux profonds). Des techniques de prétraitement et de sélection des caractéristiques sont appliquées pour améliorer la précision et réduire le bruit. Les résultats expérimentaux indiquent que les modèles DL, notamment lorsqu'ils sont enrichis de données cliniques, surpassent les méthodes classiques. L'étude met en lumière le potentiel de l'intelligence artificielle pour soutenir le dépistage précoce et précis du diabète en milieu médical..

Mots-clés : Diagnostic du diabète, Apprentissage automatique, Apprentissage profond, Analyse des données cliniques.