Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Abderrahmane Mira Bejaia


Faculty of exact sciences
Computer science department

To obtain an academic master's degree

Option: Advanced Information Systems

# Theme

# Prediction of gestational diabetes

***Presented by:***

MEDBAL Meriem & RAHMANI Lamia

***Evaluated by:***

| | | | |
|---|---|---|---|
| *President* | Dr. ALOUI Soraya | M.C.A | U. A/Mira Bejaïa |
| *Examiner* | Dr. MOKTFI Mohand | M.C.B | U. A/Mira Bejaïa |
| *Supervisor* | Dr. EL BOUHISSI Houda Wife BRAHAMI | M.C.A | U. A/Mira Bejaïa |
| *Co-supervisor* | Mr. ZIANE Amine | PHD student | U. A/Mira Bejaïa |

**Promotion** 2022 – 2023

**In the name of Allah, the merciful**

# *Thanks*

*We begin by expressing our heartfelt gratitude to Allah, the Merciful, for His boundless blessings, unwavering assistance, and His responsiveness to our prayers.*

*We would like to express our deepest appreciation goes out to our parents, whose love, nurturing care, moral and financial support, as well as their countless sacrifices on our behalf, have been the pillars of our journey.*

*We extend our profound thanks to all our family members, friends, comrades, and those who lent us their helping hand during the arduous journey of completing this thesis.*

*Special recognition is reserved for our supervisor,* ***Mrs. El Bouhissi****, whose remarkable patience, unwavering support, invaluable guidance, and constant availability have been instrumental in our academic pursuits. We express our sincerest gratitude for her unwavering commitment to our growth and success.*

*We thank our heartfelt gratitude to all the teachers who have supported us, from our early years in primary school through to our university education. Your dedication and guidance have been instrumental in shaping our academic journey, and for that, we are profoundly thankful.*

# *Dedications*

*I wish to express my gratitude to the Almighty Allah for bestowing upon me the courage and fortitude to attain this academic milestone and for guiding me to successfully complete my thesis.*

*This work is dedicated with profound love and reverence. It is dedicated to my cherished mother, and also in loving memory of my dear father, who has departed from this world but remains dearly missed. May God bless his soul and grant him a place in His vast paradise.*

*Additionally, I dedicate this work to my beloved and only sister Feriel, and extend my dedication to encompass my entire family, friends, my pair Meriem with whom I had the pleasure to work, and to my fellow section mates.*

***Lamia***

# *Dedications*

*I dedicate the fruit of my efforts to the one who carried me, protected me, and gave me life, my dear mother, who ensured my education with her patience and sacrifices for my success.*

*To my beloved father, who supported me in my academic journey from my very first steps to school. I would also like to express my dedication to my dear grandparents, who supported me with their prayers. May God grant them long lives.*

*I would like to express my deep gratitude to my family for their unwavering support throughout my life. My parents, grandparents, and sisters 'Abir','Lina' and my little 'Nada', you have always encouraged me to pursue my academic dreams, and I couldn't have done this without you.*

*A heartfelt thank you to my thesis advisor EL Bouhissi for their guidance, expertise, and patience. Your mentorship was instrumental in shaping this research.*

*I want to extend my appreciation to my thesis partner Lamia, for our exceptional collaboration. Our teamwork was vital to the success of this thesis.*

*I'm also thankful to all my professors and mentors who have shared their knowledge and insights with me over the years.*

***Meriem***

# *Abstract*

Diabetes is a persistent medical condition that arises from a malfunction in the pancreas, leading to elevated blood sugar levels and potentially affecting various bodily functions. Over time, this condition can inflict damage upon the heart, blood vessels, eyes, kidneys, nerves, and other vital organs. To mitigate these complications, it is imperative to develop a reliable diagnostic system that can identify diabetic patients based on their medical information. In the pursuit of this goal, various machine learning algorithms have been explored for the prediction of diabetes. These algorithms play a crucial role in early disease detection and the prevention of associated health issues. Building upon our previous research focused on predicting gestational diabetes using the Random Forest algorithm, the current study takes a step further. Here, we employ a swarm intelligence strategy to discern the optimal set of features for training the Random Forest algorithm, with the overarching aim of enhancing its predictive performance. The efficacy of this proposed approach was rigorously assessed, and the results yielded promising insights. Notably, combining the Random Forest algorithm with Particle Swarm Optimization led to a marked improvement with an accuracy of 99%. This innovative fusion of algorithms showcases significant potential for advancing the field of diabetes diagnosis and risk assessment.

**Keywords :** Diabetes, Prediction, Machine learning, Optimization, Random Forest, Particle Swarm Optimization, Dataset.

# *Résumé*

Le diabète est une maladie persistante qui résulte d'un dysfonctionnement du pancréas, entraînant une élévation du taux de sucre dans le sang et pouvant affecter diverses fonctions corporelles. Avec le temps, cette maladie peut endommager le cœur, les vaisseaux sanguins, les yeux, les reins, les nerfs et d'autres organes vitaux. Pour atténuer ces complications, il est impératif de mettre au point un système de diagnostic fiable capable d'identifier les patients diabétiques sur la base de leurs informations médicales. Dans la poursuite de cet objectif, divers

algorithmes d'apprentissage automatique ont été explorés pour la prédiction du diabète. Ces algorithmes jouent un rôle crucial dans la détection précoce de la maladie et la prévention des problèmes de santé associés. S'appuyant sur nos recherches antérieures axées sur la prédiction du diabète gestationnel à l'aide de l'algorithme de forêt d'arbres aléatoires, la présente étude va plus loin. Nous utilisons ici une stratégie d'intelligence en essaim pour discerner l'ensemble optimal de caractéristiques pour l'entraînement de l'algorithme de la forêt aléatoire, dans le but principal d'améliorer ses performances prédictives. L'efficacité de l'approche proposée a été

rigoureusement évaluée et les résultats ont donné des indications prometteuses. Notamment, la combinaison de l'algorithme de forêt d'arbres aléatoires et de l'optimisation par essaims particulaires a permis une nette amélioration, avec une précision de 99%. Cette fusion innovante d'algorithmes présente un potentiel significatif pour faire progresser le domaine du diagnostic du diabète et de l'évaluation des risques.

**Mots clés :** Diabète, Prédiction, Apprentissage automatique, Optimisation, Forêt d'arbres aléatoires, Optimisation par essaims particulaires, Ensemble de données.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# *List of Abbreviations*

**ACC** Accuracy

**ACO** Ant Colonny Optimization

**ANN** Artificial Neural Networks

**BMI** Body Mass Index

**CSV** Comma Separated Values

**DNN** Deep Neural Networks

**DT** Decision Tree

**FP** False Negative

**FN** False Positive

**GDM** Gestational Diabetes Mellitus

**GUI** Graphical User Interface

**IQR** Inter Quantile Range

**KNN** K-Nearest Neighbours

**LR** Logistic Regression

**ML** Machine learning

**NB** Naïve Bayes

**NUMPY** Numerical Python

**PANDAS** Python Data Analysis

**PSO** Particle Swarm Optimization

**SKLEARN** Scikit-learn

**SNS** Seaborn

**SVM** Support Vector Machine

**T1D** Type 1 Diabetes

**T2D** Type 2 Diabetes

**TN** True Negative

**TP** True Positive

**WHO** World Health Organization

# Chapter 1

# General Introduction

When the pancreas does not create enough insulin or when the body is unable to utilise the insulin that is produced, diabetes develops. Blood sugar levels are controlled by the hormone insulin. Uncontrolled diabetes frequently results in hyperglycaemia, also referred to as high blood sugar, which over time causes significant harm to a number of organs in the body, including the neurons and blood vessels. Since a decade ago, there has been a sharp rise in the number of persons who have diabetes. The biggest contributing factor to the rise in diabetes is the current human lifestyle, which includes bad eating habits, an unhealthy way of life, stress, and a lack of exercise. The majority of diabetics reside in low- and middle-income nations. Throughout the last decades, diabetes has been progressively rising in both incidence and prevalence.

In the last ten years, the frequency of diabetes has sharply climbed in Algeria, a poor nation in North Africa. A significant fraction of people either have diabetes or are at danger of developing it because of the many risk factors present in the community, such as obesity, inactivity, and hereditary vulnerability to the disease.

Gestational diabetes, is characterised by hyperglycaemia that appears for the first time during pregnancy. Like other types of diabetes, gestational diabetes affects how your cells use sugar (glucose), it's a condition of carbohydrate tolerance that produces hyperglycaemia of varying severity, beginning or being observed for the first time during pregnancy, irrespective of the need for treatment, and perhaps evolving in the postpartum.

The intricacy of this disease's treatment, however, increases along with the number of people who have it. Researchers have created a number of techniques to help with early illness diagnosis and to learn more about the factors that greatly affect the occurrence of infections.

To support lowering the risk of infection, several machine-learning-based solutions have been put into place. Several areas of our daily lives benefit from risk prediction. Knowing the most important aspects to control diabetes early on gives you an advantage.

Gestational diabetes is a very dangerous condition that affects both the mother and the baby, it increases the risk of macrosomia (large babies) and birth complications, higher likelihood of developing type 2 diabetes later in life and potential effects on the baby's long-term health.

Due to the fact that gestational diabetes is hard to diagnose, doctors and healthcare workers for most of the time have a hard time detecting it. For that, our goal is to create an application that can make it easier to detect this type diabetes.

Our application would be used easily, if someone wants to know if they can be at risk of getting diabetes, all they have to do is insert data about themself such the age, glucose level, blood pressure...etc. Then according to those personal pieces of information, the application will show them if they could be at risk of getting the disease or not.

This application is meant to help pregnant women to be able to know if they're prone to diabetes, in order to improve their habits or visit a doctor before the disease is left undetected

and untreated for a long time.

The major contributions of this work are:

- Study and explanation and of some of the most important works that have be done on the prediction of diabetes mellitus.
- Analyse of the Frankfurt Hospital diabetes dataset.
- Approach based on a famous machine learning method and an optimization algorithm to enhance the efficacy of the ML model.

The goal of this work is to help women getting a prior prevention of their possibility to be affected by gestational diabetes and assist diabetologists and gynaecologists in their medical monitoring, and in order to create the application we wanted we have organised our work into the follow phases :

- **Research:** first we made research on gestational diabetes, it's causes and symptoms, most reliable diabetes datasets.
- **Analyse of research:** this phase was about studying the results of our research and making conclusions about the main items that we should focus on.
- **Detection of the problem:** the problem is the complexity of the diagnosing gestational diabetes, it's important to early prevent any women of the risk of getting it.
- **Finding the solution:** we made large research on the different ways to create an approach to forecast gestational diabetes, we choose machine learning and optimization techniques.
- **Implementation:** this step contained the coding of our approach and the realisation of the application.
- **Evaluation:** this is the last step where we tested our model and to calculate it's efficacity in detecting the type of diabetes which we're concerned about.

The chapters of our thesis that come after this one are:

**chapter 2:** This chapter contain the notions and the subjects that concern our study, diabetes and it's types, diabetes in Algeria, machine learning and swarm intelligence.

**chapter 3:** This chapter discuss various works made on the prediction of diabetes, we explain the approaches of the studies and their results by developing a state of art.

**chapter 4:** This chapter focuses on our proposed approach for predicting gestational diabetes. We provide a detailed exposition of the methodology, its techniques, followed by a comprehensive evaluation of the method's effectiveness, this section represents our contributions in the domain of prediction of gestational diabetes.

**chapter 5:** This chapter encompasses both the experimental phase and the evaluation of our proposed prediction model. It includes a detailed description of the dataset, information about the hardware and software environment, and a comprehensive explanation of the phases involved in implementing our approach.

**chapter 6:** In this concluding chapter, we present our final thoughts on the thesis and delve into our perspectives regarding future work.

# Chapter 2

# Fundamental concepts

## 2.1 Introduction

Diabetes is a chronic disease that develops when the body either cannot use the insulin that the pancreas makes properly or does not create enough of it, a hormone called insulin controls blood sugar. To prevent the risk to get any type of diabetes, medical scientists had constantly searched for efficient ways to early diagnosis this health condition to better control it, it has been concluded that medicine cannot successfully achieve this goal alone but with the help of other sciences especially computer and information science.

Information technology played an important role in diabetes prediction through its various methods and techniques. Machine learning, neural networks, and deep learning are the most predominant concepts for the problem of diabetes prognosis. These technological advances have made it possible to fully leverage health data to identify individuals at risk and implement more effective prevention and management strategies.

In the contemporary landscape of personalized medicine and the data-driven revolution, diabetes prediction represents a fascinating example of how technology and data science can be put to the service of human health. In light of this perspective, persistent efforts in the domain of diabetes prediction assume an indispensable role in the ongoing battle against this pernicious ailment.

In this chapter, we will define the notions and concepts that are involved in our study, we will discover diabetes and its varieties, diabetes in our country, machine learning and swarm intelligence.

## 2.2 Diabetes

### 2.2.1 Definition

Diabetes is a chronic disease that affects people of all ages, is a silent killer disease that is very prevalent in recent years, with the worldwide number of individuals living with diabetes currently at 537 million [1]. This figure is projected to rise to 643 million people by 2030 [2], indicating that one out of every ten adults in the future may have diabetes. In Algeria approximately 15% of the population aged 18 and over, nearly 2.8 million individuals, are currently diagnosed with diabetes, according to the Ministry of Health. Without preventive measures, this number is expected to reach 5 million by the year 2030 [3].

Diabetes occurs when blood sugar levels become excessively high, primarily due to an absolute deficiency of insulin, a hormone responsible for regulating blood sugar levels. Insulin is produced by the $\beta$ cells within the islets of Langerhans in the pancreas [4]. When insulin

function is compromised or when there's insufficient insulin production, it leads to elevated blood glucose levels, a condition associated with a range of health complications. The journey of glucose begins with the breakdown of carbohydrates we consume. Both glucose and amino acids are directly absorbed into the bloodstream, resulting in an increase in blood glucose levels, which can pose significant health risks. In response to this heightened blood sugar, the pancreas detects the imbalance and initiates the release of insulin into the bloodstream, a process mediated by beta cells located in the islets of Langerhans [5]. Once in the bloodstream, insulin facilitates the entry of glucose into body cells, where it undergoes metabolism, ultimately reducing blood glucose levels [6]. Insulin plays a crucial role in enabling glucose uptake by various cell types in the body. Liver cells, for instance, store glucose in the form of glycogen and regulate glycogen hydration. Additionally, muscles store glucose as glycogen, which serves as an exclusive energy source for muscle cells during physical activity. These cells possess membrane receptors for insulin, which are pivotal in maintaining glucose homeostasis. For a visual overview of the normal utilization of glucose, refer to figure 2.1.

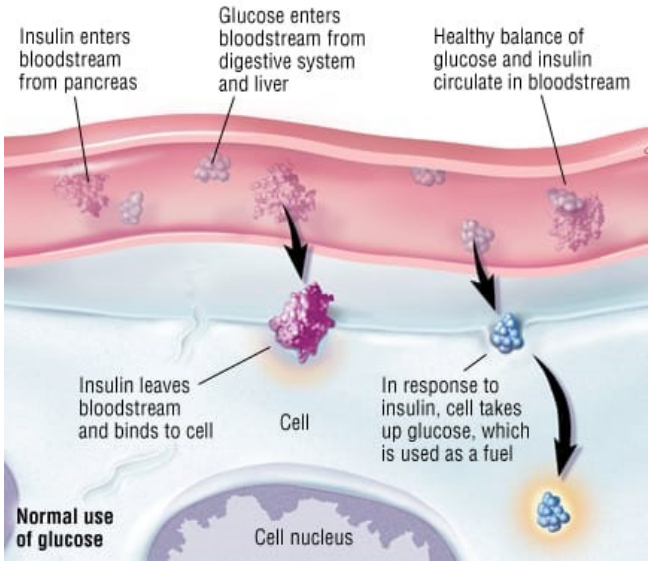



Figure 2.1: The normal use of glucose [5].

### 2.2.2 Classification of diabetes

There are mainly three types of diabetes [7]:

#### Diabetes type 1

Type 1 diabetes, also known as insulin-dependent diabetes or juvenile-onset diabetes, is a chronic autoimmune disease in which the body's immune system attacks and kills the beta cells that produce insulin in the pancreas. The immune system's immunological-mediated destruction of beta cells stops the synthesis of insulin.

The precise cause of Type 1 diabetes remains unknown, but it is primarily attributed to a malfunction in the body's immune system. Type 1 diabetes is classified as an autoimmune disease, characterized by the destruction of insulin-producing beta cells located in the pancreas. The onset of diabetes occurs when more than 90% of these beta cells are destroyed. In this process, the body's immune cells attack and dismantle the insulin-producing pancreatic cells [8].

There are certain subtypes of Type 1 diabetes for which the underlying causes remain unknown. Some of these patients do not exhibit autoimmune markers, yet they display persistent insulin production and a predisposition to ketoacidosis. This subtype is most commonly observed among individuals of African or Asian heritage. It is characterized by episodic ketoacidosis, with varying degrees of insulin insufficiency between these episodes. While there is no conclusive scientific evidence of B-cell autoimmunity associated with this type of diabetes, it demonstrates a strong genetic predisposition. The necessity for insulin replacement therapy in affected patients may vary, and it may not always be required.

The insulin-dependent diabetes (I.D.D.) is linked to factors that increase the risk of developing type 1 diabetes, among them heredity factors and environmental factors such as virus, Exposure to toxic chemicals, and diseases affecting the pancreas may indirectly cause type 1 diabetes.

### Diabetes type 2

Type 2 diabetes, previously referred to as adult-onset diabetes or non-insulin-dependent diabetes, primarily affects individuals with insulin resistance and often involves a relative, rather than absolute, insulin deficiency. This type of diabetes accounts for 90–95% of all cases.

A significant proportion of individuals with Type 2 diabetes are overweight or obese, contributing to varying degrees of insulin resistance. It's important to note that even individuals who are not conventionally categorized as obese may have a higher percentage of body fat concentrated in the abdominal region.

Unlike Type 1 diabetes, ketoacidosis is rare in Type 2 diabetes, and the condition often goes undiagnosed for years due to its gradual onset and mild initial symptoms. People with Type 2 diabetes are at an increased risk of both macro vascular and microvascular complications. Risk factors include age, obesity, inactivity, hypertension, dyslipidaemia, and a history of gestational diabetes in women.

### Gestational diabetes

Gestational diabetes is a condition characterized by elevated blood glucose levels that develop for the first time during pregnancy. It affects how cells use sugar (glucose) and is diagnosed in approximately 3-10% of pregnancies globally, typically between the 24th and 28th week of pregnancy [9].

According to the World Health Organization (WHO), gestational diabetes is defined as a disorder of carbohydrate tolerance that causes variable hyperglycaemia, starting or being noticed for the first-time during pregnancy, independent of the need for treatment, and the condition can persist postpartum.

Gestational diabetes poses increased risks for both the mother and the baby, but these complications can be minimized and reduced by adequate management, it is essential for pregnant women, particularly those with risk factors like age over 35, obesity, or a family history of diabetes, to undergo diagnosis and work closely with healthcare providers to ensure a healthy pregnancy.

## 2.3 Diabetes in Algeria

In North Africa and the Middle East, the number of patients climbed by 733,000 and 2,164,000, respectively, between 1990 and 2017, amounting to an increase of more than 200%. Algeria is an impoverished country in North Africa, and during the past ten years, the prevalence of diabetes there has rapidly increased. Due to the number of risk factors in the community, such

as obesity, inactivity, and genetic susceptibility to developing diabetes, a sizable proportion of individuals either have diabetes or are at risk of having it. It was revealed that Algeria had the highest rate of diabetes type 1 in North Africa and the Middle East, with estimates that if the disease is not effectively managed, over half of the population will be affected by 2030. Still, as diabetes is linked to a wide range of co-morbidities brought on by macro and microscopic vascular diseases, the lack of awareness among healthcare policy leaders in Algeria might create problematic situations.

A high incidence and prevalence of diabetes are present in Algeria. According to the eighth edition of the International Disability Federation Disability Atlas, Algeria ranks seventeenth among all countries in terms of the prevalence of children under the age of 15 who have T1D (n = 20,100). In this age range, Algeria also has the highest incidence of T1D in the Middle East and North Africa (MENA) area, involving 3100 children in 2019.

The global prevalence of T2D in 2014 was 12.3% (7.4 -18.8) for men and 12.6 % (7.7 -18.9) for women, according to the World Health Organization. In terms of risk factors for diabetes, Algeria had a prevalence of 33.6 for diabetes and 26.4 for physical inactivity and obesity (BMI 30) among individuals aged 18 and older in 2016.These numbers are likely to rise as a result of bad lifestyles and poor nutritional habits. Additionally, Algeria's age-standardized mortality rate (ASM) for no communicable diseases—including diabetes—is comparatively high at 430.7 per 10,000 people. 8390 people died from diabetes in 2016, and the mortality rate was 2.12 per 10,000 people, according to the WHO [10].

## 2.4 Machine learning

### 2.4.1 Definition

Since the beginning of time, humans have used a variety of tools to carry out various tasks more easily. The inventiveness of the human mind led to the creation of various machines that made life easier for people by enabling them to meet needs such as travel, industry, and computing. Machine learning is one such invention. Arthur Samuel defines machine learning as : "the scientific discipline that enables computers to learn without having to be specifically programmed ". For his checkers-playing program, Arthur Samuel became well-known. To educate machines how to handle data more effectively, machine learning (ML) is used. Sometimes, despite analysing the data, we are unable to decipher the information it contains. In that situation, machine learning is used. The need for machine learning algorithms has grown because of the number of datasets [11].

### 2.4.2 Types of Machine learning

We distinguish different types of machine learning [11] as described in figure 2.2:

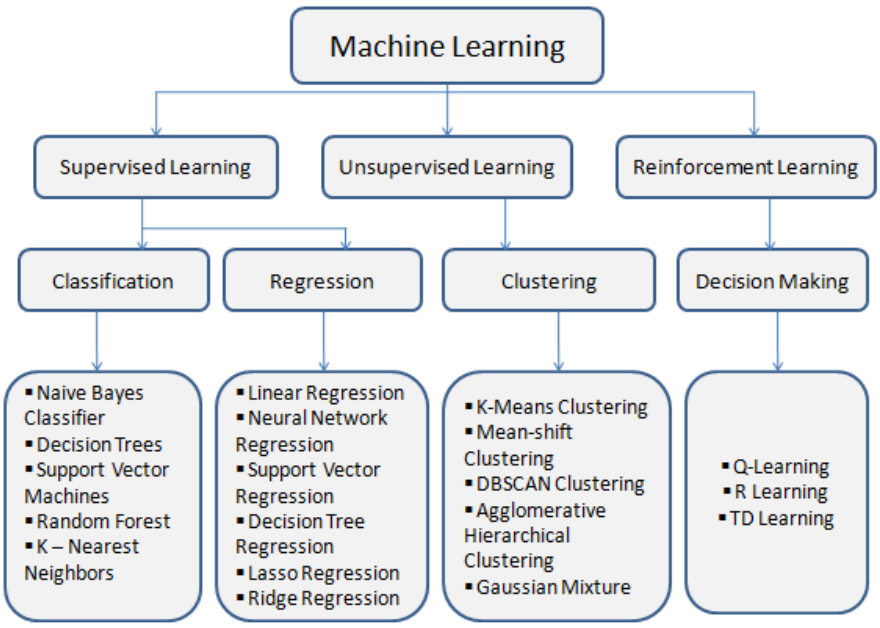



Figure 2.2: Machine Learning Types
[12].

### Supervised learning

Learning a function that translates a parameter to an outcome using sample input-output pairs is a machine learning challenge. It uses labelled training data made up of a collection of training examples to infer a function. Algorithms that require outside help are those that fall under the category of supervised machine learning. Train and test datasets are created from the input dataset. The output variable in the train dataset must be predicted or categorized.

### Unsupervised Learning

These are referred to as unsupervised learning because, in contrast to the supervised learning described above, there are no right answers and no teachers. The algorithms are left to find and display the intriguing pattern in the data on their own. Few features are learned from the data by the unsupervised learning algorithms. When new data is implemented, it recognizes the class of the data using the previously learnt features. It is mostly utilized for reducing characteristics and grouping.

### Semi Supervised Learning

Combining both supervised and unsupervised methods of machine learning is known as semi-supervised machine learning. It may be useful in fields of data analysis and machine learning where there is existing unlabelled data and obtaining labelled data requires a laborious process. With increasingly widespread supervised machine learning techniques, an algorithm is trained on a "labelled" dataset where each record contains the outcome data. Below is a discussion of the sum of Semi-Supervised learning algorithms.

### Reinforcement Learning

The field of machine learning known as reinforcement learning looks at how software agents should behave in each environment to maximize a theoretical concept of cumulative reward.

Along with supervised learning and unsupervised learning, reinforcement learning is one of the three fundamental machine learning paradigms.

### Multitask Learning

A branch of machine learning called "multi-task learning" uses the similarities between several tasks to attempt to solve many different problems at once. This can increase the effectiveness of learning and serve as a regularize. Formally, Multitask Learning (MTL) can help to enhance the learning of a specific model by using the information included in all of the n tasks if one has n tasks (conventional methods of deep learning aim to solve just 1 task via 1 specific model). These n tasks or a portion of them are connected to one another but not identical.

## 2.5 Optimization

Because optimization permeates everything, it is a fundamental concept with numerous applications a variety of uses. We constantly strive to optimize anything in engineering and business applications, whether it's to reduce costs and energy use or to increase the revenue, product, performance, and effectiveness. Any available resource must be used to its full potential, which necessitates a change in mindset in scientific reasoning, as the majority of real-world applications have much more complex elements and settings to modify the system's behaviour. Computer simulations are widely used in modern engineering design. Due to these new challenges, optimization faces. As structures and systems get more sophisticated and there is a greater demand for accuracy, the simulation process takes longer and takes more time. The optimization algorithm, an effective numerical simulator, and a realistic representation of the physical processes we aim to describe and optimize are all interconnected parts of the optimization process for any optimization problem. This is frequently a labour intensive operation, and the computing expenses are frequently very expensive. Once we have a suitable model, the optimization methods used for search and the mathematical solution utilized for simulation decide the overall computing costs [13]. In simulation-driven optimization and modelling, three fundamental challenges include the efficiency of optimization algorithms, the accuracy and efficiency of numerical simulators, and the selection of appropriate algorithms for specific problems. Despite their importance, there are no universal laws or guidelines for addressing these issues. Efficient optimizers are essential but may vary in effectiveness depending on factors like internal structure, data requirements, and implementation details. The choice of optimization algorithm depends on the problem type, desired solution quality, available computing resources, and other factors. Efficient evaluation of objectives in optimization is time-consuming, often requiring extensive computations. Approaches like approximation techniques and lower-fidelity models are used to reduce assessment time while still achieving accurate results [13], the figure 2.3 gives an overview of the optimization:

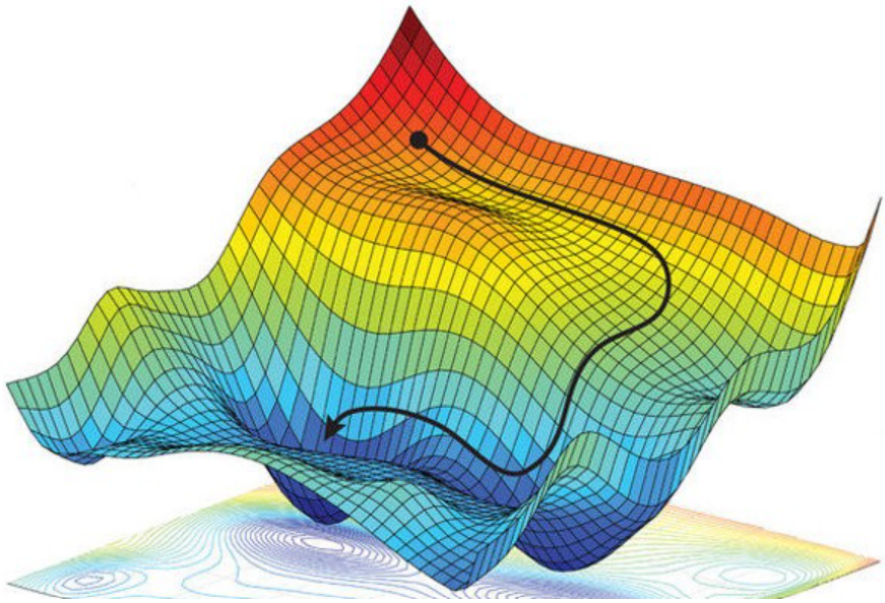


Figure 2.3: Optimization
[14].

## 2.6 Swarm Intelligence

Swarm intelligence [15] is an important concept in artificial intelligence and computer science with emergent properties. The essential idea of swarm intelligence algorithms is to employ many simple agents applying almost no rule which in turn leads to an emergent global behaviour.

To put it in a simple way, swarm intelligence can be described as the collective behaviour emerged from social insects working under very few rules. Self-organization is the main theme with limited restrictions from interactions among agents. Many famous examples of swarm intelligence come from the world of animals, such as birds flock, fish school and bug's swarm. The social interactions among individual agent help them to adapt to the environment more efficiently since more information are gathered from the whole swarm. This paper aims to introduce several well-known and interesting algorithms based on metaheuristic derived from nature and their applications in problem solving.

Swarm intelligence algorithms are inspired by various populations of biological organisms in nature. They usually replicate certain characteristics of specific organisms as they interact among themselves and their environment for achieving certain tasks intelligently by making use of simple organized steps. Swarm intelligence algorithms consist of a population of artificial search agents that interact similarly to a specific group of biological organisms in a search space. These simple interactions allow the algorithm to look for the optimal solution for a problem in a heuristic manner. Hence, they can solve a myriad of optimization problems by providing either optimal or near-optimal solutions in a reasonable time.

Swarm intelligence algorithms can be employed to solve different types of optimization problems including continuous, discrete, or multi-objective optimization problems. Hence, they have numerous applications in a variety of domains. For example, they can be used in water resources engineering, in wireless networks, in cloud-based Internet of Things, in optical systems, in recommender systems, in anomaly detection systems, and in supply chain management. They can also be used for clustering, for feature selection and for solving the traveling salesman problem. Moreover, they have several applications in optimal designs, electrical engineering, networking, mechanical engineering, machine learning, resource allocation, and digital image processing.

Recently, SI and ML have attracted close attention of researchers and have also been applied

successfully in many fields (e.g., engineering, transportation, commerce, industry and so on).

However, there are still huge space for improvement about SI algorithms and ML algorithms. On one hand, SI algorithms (e.g., ant colonies, bird flocking and so on) are often limited by weak points of computation time and local solution for large and complex problems. On other hand, ML algorithms are often limited by weak points of data and parameters. The figure 2.4 shows the principle of swarm AI:

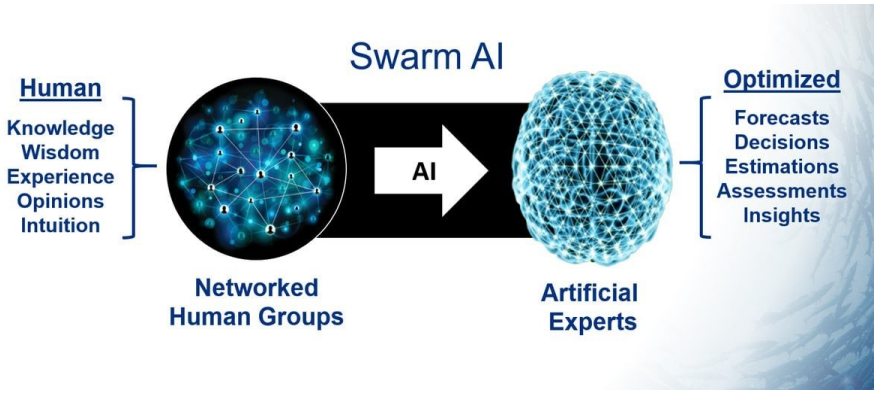



Figure 2.4: Swarm artificial intelligence [16].

To model the broad behaviours arisen from a swarm, we introduce some examples of swarm intelligence algorithms:

### 2.6.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a nature-inspired optimization algorithm that is based on the social behaviour of birds and fish. It was introduced by Dr James Kennedy and Dr Russell Eberhart in the mid-1990s. PSO is a population-based algorithm where a group of potential solutions, known as particles, iteratively search for the optimal solution to an optimization problem.

In PSO, each particle represents a potential solution in the search space. These particles move through the search space with the aim of finding the best solution. The movement of each particle is influenced by its own best-known position (local best) and the best-known position among all particles in the population (global best). The algorithm simulates the social behaviour of particles as they adjust their positions based on their own experience and the experience of the entire swarm. The core idea behind PSO is that particles adjust their positions iteratively, moving towards the global best-known position in the search space, which represents the optimal solution. PSO is particularly effective for optimization problems that involve continuous and high-dimensional search spaces [17] The figure 2.5 demonstrates an example of a particle moving to find it's best position:

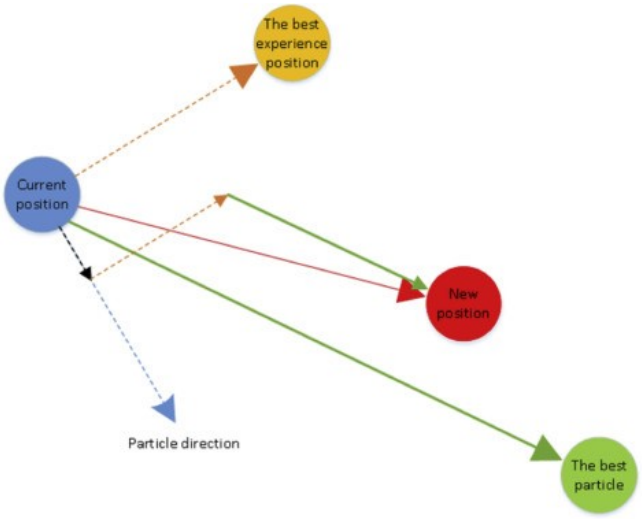



Figure 2.5: Motions of particles in PSO [18].

### 2.6.2 Elephant herd Optimization

Elephant Herding Optimization (EHO) is a metaheuristic optimization algorithm inspired by the collective behaviour of elephants. In EHO, a clan operator is used to update the distances between elephants within each clan relative to the position of a lead elephant, known as the matriarch. EHO has demonstrated its superiority over several cutting-edge metaheuristic algorithms in numerous benchmark problems and across various application domains. Elephants, being highly social animals, organize themselves into family structures consisting of females and their young calves. A clan of elephants is headed by a matriarch and includes multiple elephants. Female elephants prefer to reside with their family members, while male elephants tend to live separately and gradually become independent before eventually leaving their families. The overall population of all elephants is depicted in Figure 2.6. EHO, as proposed by Wang et al. in 2015, draws its foundation from a thorough study of the natural herding behaviour of elephants and incorporates several key assumptions into its methodology [19].

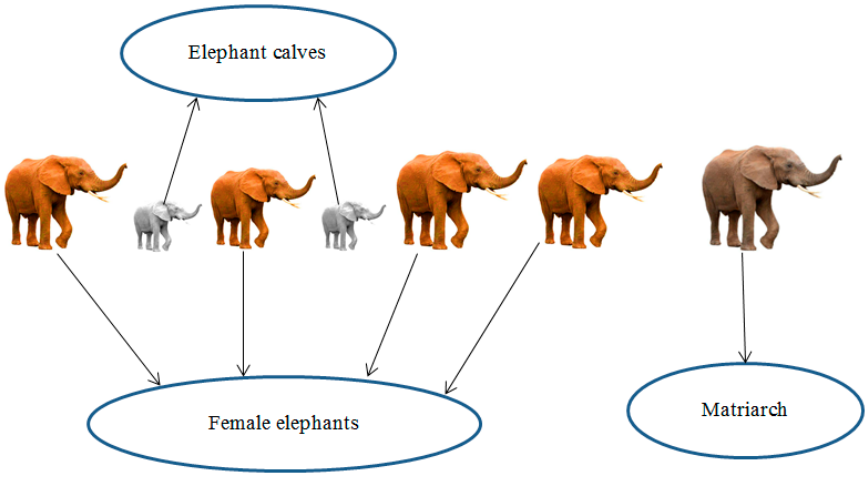



Figure 2.6: Population of elephants [19].

### 2.6.3 Grey wolf Optimization

Grey Wolf Optimization (GWO) is a nature-inspired metaheuristic optimization algorithm that draws its inspiration from the social hierarchy and hunting behaviour of grey wolves, particularly their alpha, beta, omega, and subordinate roles within a pack. It was introduced by Mir Jalili et al. in their paper titled "Grey Wolf Optimizer" published in the journal Advances in Engineering Software in 2014 [20].

The algorithm uses the alpha, beta, and omega positions to guide the search and adapt the population over successive iterations. Grey Wolf Optimization has been applied to a wide range of optimization problems and has demonstrated competitive performance compared to other optimization techniques. The suggested approach draws its inspiration from the way grey wolves in packs organize themselves. Alpha wolves, who value discipline and order more than physical strength, take the lead. Beta wolves assist the alphas, while omega wolves play a role in reducing tension within the group. Subordinate wolves, encompassing scouts, sentinels, elders, hunters, and caretakers, handle tasks such as territorial defence and ensuring safety [20], the figure 2.7 represents the hunting behaviour of wolves:


Figure 2.7: Hunting behaviour of grey wolves: (A) chasing, approaching, and tracking prey (B-D) pursuing, harassing, and encircling (E) stationary situation and attack [20].

### 2.6.4 Ant Colony Optimization

ACO is a metaheuristic algorithm inspired by the foraging behaviour of ants, it was introduced by Marco Dorigo in the early 1990s [21] and is part of the broader field of swarm intelligence. ACO is used to solve combinatorial optimization problems, where the goal is to find the best solution from a finite set of possibilities, it is used to find approximate solutions to combinatorial optimization problems by simulating the way ants find the shortest path between their nest and a food source. ACO algorithms utilize pheromone information and heuristic knowledge to guide a population of artificial ants in their search for optimal or near-optimal solutions.

The most recognized example of swarm intelligence in real world is the ants. To search for food, ants will start out from their colony and move randomly in all directions. Once an ant find food, it returns to colony and leave a trail of chemical substances called pheromone along the path. Other ants can then detect pheromone and follow the same path. The interesting poin.t is that how often is the path visit by ants is determined by the concentration of pheromone along the path. Since pheromone will naturally evaporate over time, the length of the path is

also a factor. Therefore, under all these considerations, a shorter path will be favoured because ants following that path keep adding pheromone which makes the concentration strong enough to against evaporation. As a result, the shortest path from colony to foods emerges [22].The Ant Colony Optimization (ACO) algorithm consists of five key steps: Initialization process involves setting up the population of ants with random values, defining fitness values based on the optimization goal, and determining the maximum number of iterations. In the Evaluation process, ants are randomly assigned to features, and they build solutions while aiming to minimize the mean square error of the classifier; ants failing to improve over successive steps exit. The Construction process relies on a constructive heuristic to probabilistically assemble solutions from a finite set of components. In the Update process, pheromone concentrations are adjusted on nodes, with all ants depositing pheromone on the graph and the best ant contributing more. Finally, the Decision process aims to identify the best ant globally across all iterations and record its selected consequent combination [23], figure 2.8 represents the ACO algorithm steps:

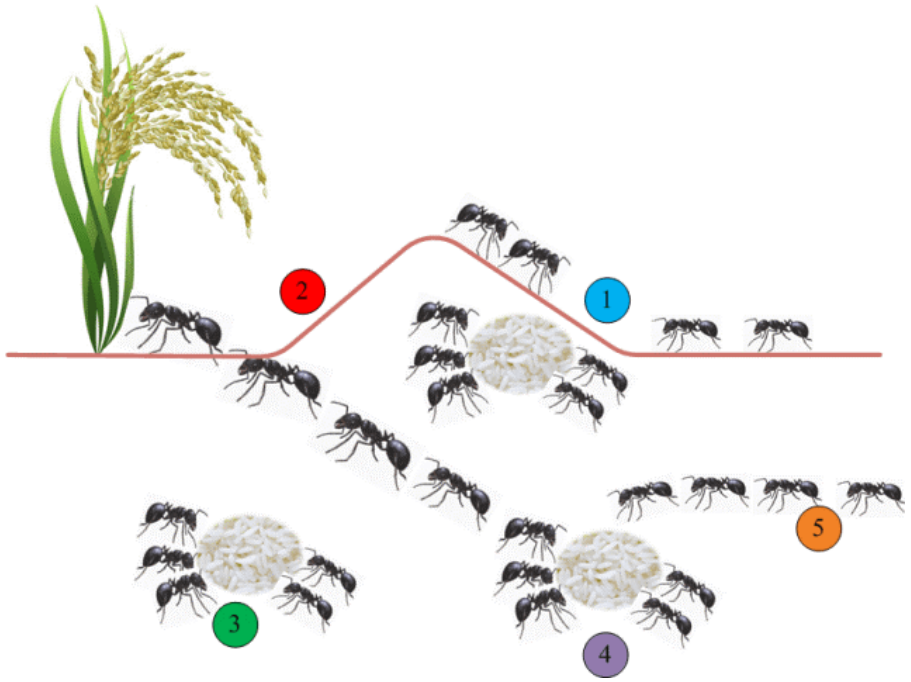



Figure 2.8: Ant Colonny Optimization [23].

## 2.7 Conclusion

This chapter has provided a comprehensive overview of general concepts related to diabetes and its various types, diabetes in Algeria, definition of machine learning and swarm intelligence as some of the most famous swarm algorithms. Machine learning and swarm intelligence, two innovative fields that hold great promise in advancing our understanding of diabetes, improving its diagnosis and treatment, and ultimately enhancing the quality of life for those affected by this pervasive disease. In the next chapter we will delve into the extensive body of research and studies that have been conducted in the realm of diabetes prediction. Diabetes is a global health concern, and its early detection and prediction play a pivotal role in effective management and prevention of complications. To gain a comprehensive understanding of the current state of the field, we review and analyse various approaches, methodologies, and technologies employed in predicting diabetes.

# Chapter 3

# State of the art

## 3.1 Introduction

The world is currently in the midst of a global AI-driven transformation that is permeating various industries, this technological surge is particularly catalysing a healthcare revolution, specifically in the realm of chronic disease management and treatment. The continuous advancements in AI are providing doctors with powerful tools for diagnosing and treating patients while enabling researchers to develop intelligent systems aimed at enhancing patient care and predicting diseases, with a specific focus on diabetes. The focus of diabetes prediction research is critical to reducing the risk of complications and implementing preventive measures. Identifying individuals at higher risk of developing diabetes through early detection and intervention holds the promise of more effective disease management and better patient outcomes. As we navigate this paradigm shift, there is a burgeoning synergy between AI and healthcare.

Notably, Diabetes prediction has been a major area of focus for researchers, who have conducted numerous studies in this domain employing a variety of Machine Learning techniques and comparing respective accuracy rates. In this chapter, we will highlight some of the key research works related to diabetes prediction and discuss the methods used in each study.

## 3.2 Related works

Artificial intelligence refers to the use of methods, sets of rules, and algorithms to solve complex problems and make smart decisions. In the context of diabetes classification, techniques have been employed such as data mining, machine learning (ML), and artificial neural networks (ANN) to analyse large datasets and develop models that can predict the onset of diabetes, Specifically, when it comes to figuring out who might get diabetes. The success and the remarkable utilization of these computer techniques in the classification of diabetes has led to a huge number of publications in journals and conference proceedings. Therefore, in this study, we explore some of the significant contributions that have been made by different researchers in this field.

**Rahimloo et al.** [24], combine statistical models and neural networks (a hybrid neural network and logistic regression model) to create a new compound that has at least error and maximum reliability. The accuracy and efficiency of the method were investigated, and acceptable results compared to the neural network and logistic regression methods were obtained. This research examined the relationship between complications in diabetic patients and their properties such as blood glucose, triglycerides, cholesterol, haemoglobin, and body mass index to predict complications based on their symptoms. The dataset used in this article is prepared and documented to diagnose diabetes from the Association of Diabetics in the city of Urmia.

It contains 180 samples with 8 features. Out of these 180 samples, 60 have type 1 diabetes, 60 have type 2 diabetes, and 60 are healthy individuals. As we said, the proposed model with inputs and outputs is used to accurately predict diabetes, with fewer errors. The results of logistic regression were applied to the neural network and had a significant impact on its function. In the results section, the researchers discussed the results of using logistic regression, as well as the results of combined neural network with logistic regression. These methods were compared and analysed. First, the researchers applied logistic regression with the properties already said with an estimated coefficients value and P-value of persons and with non-diabetes and prediabetes and showed the importance of each variable in predicting both diabetes and prediabetes within the logistic regression model. Then, they also showed the results of combining statistical models and neural network, and how to reduce error in this new compound. The final results showed that the combined neural network yield better results with the error function is equal to 0,0002.

**In Alehegn et al.** [25], the authors analysed study's primary objectives are to identify high-accuracy algorithms and predict the diabetic disease ultimately, decide on the optimal methodology to early diabetes disease prediction, for that they proposed to build a hybrid model based on different machine learning algorithms to increase the accuracy of the prediction of diabetes. The approach of the authors was to combine four algorithms: Radom Forest, J48, KNN and Naïve Bayes, with WEKA hybrid system, since individual classification algorithms do not provide results for predictions, it was preferable to combine individual classifier predictions to provide a single result by integrating different classifiers into one, an ensemble technique overcomes the issue or limitation associated with separate classifiers and boosts accuracy. The study has been done on the Pima Indian diabetes database, the first step they have done was the data pre-processing, then they split the dataset into 90% of training set and 10% of testing set, the evaluation of the model was based on the accuracy, F-measure, Recall and precision, they tested the four algorithms individually and the hybrid model. The combination of the four algorithms provided a better accuracy than each algorithm individually.

**Zou et al.** [26] used machine learning techniques to predict diabetes mellitus. They utilized two datasets, the first one is hospital physical examination data from Luzhou, China, containing 14 attributes. And the second one is the Pima Indians diabetic data, which contains 8 attributes. The authors used three models as classifier which are decision tree, random forest, and neural networks. The terms used in this study to evaluate the quality of the classification models are sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews correlation coefficient (MCC). The authors discuss two commonly used validation methods in machine learning: the hold-out method and the k-fold cross validation method. The hold-out method involves dividing the dataset into training and testing sets, where the training set is used to train the model and the testing set is used to evaluate the model's performance. On the other hand, the k-fold cross validation involves dividing the dataset into k equal-sized folds, where the model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, and the final performance is the average of all the tests. They highlighted that the choice of validation method depends on the problem and the size of the data, so they used the hold-out method to verity the methods' universal applicability and the five-fold cross validation method in their study to ensure that the entire dataset is used in training and testing. To reduce the dimensionality of the data they used Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR). The results showed that the random forest model had the highest accuracy (ACC = 0.8084) when all the attributes were used.

Another approach proposed by **Alam et al.** [27] which used relevant features to predict diabetes mellitus and described the link between the various traits. The model of prediction is based an artificial neural network and machine learning algorithms. The test of the model has been done on the National Institute of Diabetes and Digestive and Kidney Diseases (publicly

available at: UCI ML Repository, after cleaning the dataset, it included some zero (0) values for glucose, blood pressure, skin thickness, insulin, and BMI. Consequently, the median value of that property was used to replace all zero values. The methodology of this work was composed of three steps, first one was the data preprocessing, this step included data cleaning, data reduction and data transformation, the second step was the association rule mining, in which data mining techniques were used to generate rules and meaningful information. The branch of association rule mining that identifies patterns and frequently used objects in a dataset is crucial. There are two sections to it: Establish the frequent item set first, then create rules second. The final step was the modelling, the scientists proposed three models: artificial neural network (ANN), random forest and k-means clustering. First, they implemented the ANN, the random forest algorithm and finally the k-means clustering, the accuracy that was predicted with the help of the confusion matrix and the AUROC curve were assessed to evaluate each model. The accuracy of the ANN approach was 74.7%, that of the Random Forest method was 75.7%, and that of the K-means clustering method was 73.6%.

**Mujumdar et al.** [28], focus on developing a predictive model for diabetes prediction utilizing machine learning algorithms and data mining approaches. The authors proposed a model that included five modules, first the dataset collection, data Pre-processing, Clustering, building the model and the evaluation of the model. The study was done on the PIMA dataset, they implemented k-means clustering algorithm on two attributes of the dataset that are age and Glucose, to classify each patient into a diabetic or non-diabetic class. The machine learning model was built with Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbours, Gaussian Naïve Bayes, bagging algorithm and Gradient Boost Classifier, then created a pipeline for the algorithms that gave the highest accuracy and then evaluated the pipelines. Finally, in order to evaluate the prediction results, they assessed the accuracy of the classifications, the confusion matrix, and the f1-score. The evaluation showed that Logistic Regression had the highest accuracy of 96%, but using the pipeline Ada Boost algorithm was the best model with an accuracy of 98.8%. With two separate datasets, the researchers have seen comparisons of the accuracy of machine learning algorithms. It is evident that using this dataset instead of the previous dataset allows the algorithm to predict diabetes more accurately and precisely.

Another method proposed by **Nadeem et al.** [29], which highlighted the potential benefits of using machine learning algorithms to propose solutions, which are a significant health concern. The proposed solution aims to support the improvement of survival in diabetic patients by providing patients with information on individual patient treatment options. By informing patients on the optimal treatment on an individual patient basis. These researchers first viewed the latest articles (2003-2018), they proposed a fusion-based approach. They used two datasets in the training and testing of the proposed fusion-based machine learning architecture. The first dataset is derived from the National Health and Nutrition Examination Survey (NHANES), and the second is acquired from the online repository. The fused dataset has 10,627 records with 8 features with an age distribution between 21 and 77 years to predict and manage diabetes after merging data with the Data In-Data Out (DAI-DAO) method, which involves combining data from multiple sources to improve data quality. after that they pre-process the data by normalizing and standardizing it, split it into train and test set using 5 fold-cross validations, and applying the SVM and artificial neural network algorithms. The algorithms and the fusion of SVM-ANN's performances were compared according to accuracy, specificity, sensitivity, precision, miss rate, false positive ratio, and false negative ratio, the fusion of SVM-ANN had the highest performance with an accuracy of 94.67%.

**Al Yousef et al.** [30] concluded that BN is suitable for predicting the probability of diabetes in patients with an AUC of 0.71 and 0.75 and an accuracy of 63% and 66%, between

six machine learning algorithms with and without Synthetic Monitoring Oversampling Techniques (SMOTE). These algorithms are: K-nearest neighbours (KNN), the random tree forest (RTF), the support vector machine (SVM), the naïve Bayesian classifier (BC), the Bayesian network (BN), and logistic regression (LR). All machine learning algorithms were applied to the dataset, which was collected from 5 hospitals in different cities using National Guard Health Affairs databases and comprised 38 attributes of 21431 patients between 2015 and 2019. This dataset was first collected, where the attributes were selected, and prepared by cleaning and preprocessing the data (the missing values were replaced by the mean of the relevant value attribute). The authors used as a sampling method under-sampling and oversampling to balance the data using Synthetic Monitory Oversampling Technique (SMOTE). This approach allowed for a fair comparison of the performance of machine learning algorithms based on evaluation metrics of accuracy, precision, recall, and F1-score, with and without SMOTE. After building the model using data mining, it was evaluated using the hold out method by splitting the dataset into two parts: the training data and the test data, with respectively 70% and 30% of the data. In this work, the main parameter of evaluation between algorithms is accuracy.

**Edeh et al.** [31], had as goal to build a model for an early prediction of diabetes with multiple machine learning models, therefore they choose four supervised machine learning algorithms: Random Forest (RF), Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM) and an unsupervised algorithm: K-means clustering to classify the patients into categories. The experiments were performed on the PIMA Indian Diabetes (PIDD) provided by the UCI machine learning repository and the database provided of the hospital of Frankfurt in Germany. This study included four steps, first they proceeded to clean the data in which they implemented the k-means classifier on the two databases to clean the columns where there is data that are missing and replace it with the cluster, then they split the PIDD dataset into 80% for training and 20% for testing and the German dataset into 70% for training and 30% for testing, the second step was the evaluation method, they tested each one of the ML algorithms described above, finally the experimental results they evaluated each one of the four algorithms on the datasets, they calculated the accuracy, the precision, f1-measure and recall of each algorithm to evaluate and compare them. The results of the evaluation process showed that RF performed the best on the Frankfurt dataset with 97.6% of accuracy, in other words, 582 cases are accurately identified out of 600 variables that were used to evaluate this model on the other hand SVM was the best model on the PIDD dataset with an accuracy of 83.1%, that is 127 cases are accurately identified among the 153 variables that were used to evaluate this model. Due to its great accuracy, they decided that the SVM model was the most suitable one for their database. From the experiment's findings, they deduced that the values of accuracy for the two data sets with all measurements are satisfactory with a disturbance, though changing the number of clusters and the random setup of cluster centres can have an impact on the accuracy values. This study's major goal was to assist diabetologists in developing an accurate treatment regimen to treat their patients with diabetes. Because of the work's high accuracy, rapid disease diagnosis, and rapid treatment, a digital healthcare system for people with diabetes may be created as a result. In a larger sense: creation of an Algerian diabetic patient database, deep learning-based diabetes diagnosis. developed a method to help people determine if they have diabetes with the help of an Android app.

**Qin et al.** [32] compared the efficacy of five machine-learning models, namely CATBoost, XGBoost, Random Forest, Logistic Regression, and Support Vector Machine, in predicting diabetes using lifestyle data from the NHANES database. The study explored the utilization of machine learning techniques to predict diabetes based on lifestyle types. The researchers employed the power of machine learning to precisely identify a patient's lifestyle type. First the researchers developed a machine learning model to predict the lifestyle type of a patient, then they utilized this prediction to predict diabetes. They evaluated the model's performance on a

large dataset of diabetes patients, revealing that the model was able to predict diabetes using lifestyle type. The study utilized data from 17,833 individuals. It also utilized various techniques for preprocessing and evaluating the machine-learning models. After excluding missing values with four categories of feature variables, preprocessing using Synthetic Minority Over-Sampling Technique Nominal Continuous which is effectively operates on datasets containing both numerical and categorical features, and using the AIC forward propagation algorithm to screen training data for machine models, the model's performance was evaluated by using accuracy, sensitivity, specificity, precision, the F1 score, and the ROC curve. Among the five models, CATBoost was found to be the best performing model, achieving an accuracy of 82.1% and an AUC of 0.83. Predicting diabetes patients in this article was most influenced by the dietary intake levels of energy, carbohydrate, and fat.

**Ahamed et al.** [33] used various machine learning techniques to predict type-2 diabetic mellitus illness, their goal was to create a predictive model that can accurately predict whether a person has diabetes. The following sections make up the different parts of this study: The relevant works in DM are shown in the section relevant Works. The theoretical underpinnings of the different algorithms are outlined in Section Theoretical Approaches of the Classifiers. The construction and application of the classifiers are decided in Section Results and Discussion. The study's conclusions and upcoming projects were discussed within the section Conclusion and Upcoming Work. The chosen classifiers are logistic decision trees, ExtraTrees, random forest, regression, XGBoost, gradient boosting, and the light gradient boosting machine (LGBM). Initially gathered and kept in the database are the data required for the investigation. For use, the PIMA dataset is acquired from the UCI Repository. Following that, the dataset is pre-processed using various strategies for exploratory data analysis. The dataset is separated into "training data" and "testing data." The best algorithm that works and has the highest accuracy is chosen as the best prediction model for predicting DM disease after the various algorithms listed have been compared. After evaluating each algorithm, the LGBM algorithm showed the best performance with an accuracy of 95,20 %. By providing a greater accuracy in comparison to the other algorithms and taking into account the remarks made here, the researchers determined that the "LGBM algorithm" worked best for the dataset used, they suggested, in the future, that different datasets may be taken and compared with various classifiers in order to determine which algorithm can yield the greatest results, an advanced LGBM algorithm can be employed, and the forecast accuracy percentage could be raised by further adjusting the parameters used in LGBM.

By utilizing a variety of machine learning techniques, the authors **Kale et al**. in [34] did early diabetes prediction in a human body or patient for a higher degree of accuracy. approaches for machine learning better prediction outcomes can be achieved by building models from patient-collected datasets. The objective of this work is to create a system which, by combining the findings of several machine learning approaches, can accurately conduct early diabetes prediction for a patient. The algorithms are Gradient Boosting (GB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest neighbours (KNN) and Logistic Regression (LR). The emergence of different symptoms has been used to detect the presence of disease. The methodologies, metrics, and features that are employed affect the outcome of the prediction. A Disease Influence Measure (DIM) based diabetic prediction has been provided as a step toward diabetes prediction. The approach performs a preliminary processing on the input data set, removing the noisy records. The method calculates disease influence measure (DIM) in the second step using the characteristics of the input data point. The technique does diabetic diagnosis depending on the DIM value. Various disease prediction methods have been taken into consideration, and their effectiveness has been compared. The analysis's findings have been provided in-depth for development. The project work reveals that the model is capable of accurately predicting diabetes with an accuracy of 95% or higher.

## 3.3 Analysis and comparison

Previously, we have presented the main prediction approaches in the field of healthcare. In the table 3.1, we will carry out a comparative study of the approaches proposed above according to the following 7 factors:

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

| Approach | Dataset | Used techniques | Software tools | Evaluation of performances | Target | Advantages |
|---|---|---|---|---|---|---|
| Rahimloo et al [24]. | prepared and documented from the Association of Diabetics in the city of Urmia. | Artificial neural network (ANN), Logistic regression (LR) statistical model. The fusion of ANN and LR. | / | The error function of artificial neural network: ANN= 0.18808. The error function of the artificial neural network combined with logistic regression: (ANN- LR) = 0.00025753. | / | • Handles both categorical and continuous predictor variables.<br>• Simplicity and robustness in performance evaluation.<br>• Enhances model robustness.<br>• Compensates for weaknesses in either method.<br>• Ensures accurate and effective data analysis. |
| Alehegn et al. [25] | Pima Dataset. | RF, KNN, Naïve Bayes, J48 | Hybrid System Weka, Java tools | / | Type 1 diabetes, type 2 diabetes | The study evaluated various data mining methods and their applications, revealing that different data sets enhance machine learning techniques. A model combining multiple algorithms demonstrated better precision than a model using each individually. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Zou et al. [26] | The hospital physical examination data in Luzhou, China. Pima Indian dataset. | Decision tree. Random forest. Neural network five-fold cross validation principal component analysis (PCA). Minimum redundancy maximum relevance (mRMR) | / | With Luzhou dataset: RF= 0.8084 J48= 0.7853 ANN= 0.7841 With Pima Indians dataset: RF= 0.7604 J48= 0.7275 ANN= 0.7667 | predict the presence of diabetes mellitus in individuals | • Performs multiple experiments, enhancing study reliability.<br>• Enables training and testing on all dataset samples, reducing variance risk.<br>• Strengthens study validity by using real-world data.<br>• Reduces the impact of missing and abnormal values on results, improving data quality.<br>• Utilizes an independent test set for validation. |
| Alam et al. [27] | National Institute of Diabetes and Digestive and Kidney Diseases Database. | Artificial Neural Network (ANN), RF, K-means clustering | / | ANN=75.5% RF=74.7% K-means=73.6% | Type 2 diabetes | This model's implementations outcomes have demonstrated that ANN performs better than the ML models. The findings using association rule mining demonstrate a potent relationship between BMI and glucose levels and diabetes. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mujumdar et al. [28] | Pima Dataset | RF, DT, Extra Tree Classifier, Support Vector Classifier (SVC), Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm (LDA), Logistic Regression, KNN, Gaussian, Naïve bayes, Bagging algorithm, Gradient Boost Classifier | / | **Results on the dataset used on the paper:** Logistic Regression=96% LDA=94% Gaussian NB=93% Gradient Boost Classifier=93% AdaBoost=93% RF=91% Extra Trees=91% Bagging=90% KNN= 90DT= 86% Perceptron= 76% SVC=60% **Results on the Pima dataset:** LDA=77% AdaBoost= 77% Gradient Boost Classifier= 77% Extra Trees=76% Logistic Regression= 76% Bagging=75% DT= 74% RF=72% KNN= 72% SVC=68% Gaussian NB=67% Perceptron= 67% | Type 1 diabetes, type 2 diabetes, type 3 diabetes (gestational diabetes) | The benefit of this research is that the authors used multiple ML algorithms which permits to evaluate each one of them, also the authors used a pipeline on the algorithms which performed the best to evaluate them and conclude which one gave the best results. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | **Results after using pipeline on the algorithms that gave the highest accuracy:** AdaBoost Classifier= 98.8% Gradient Boost Classifier= 98.1% RF= 98.1% Logistic Regression= 97.5% Extra Trees Classifier= 96.3% LDA=95% | | |
| Nadeem et al. [29] | was obtained from the National Health and Nutrition Examination Survey (NHANES). Pima Indian diabetes dataset. | principal component analysis (PCA). Artificial neural networks (ANN). support vector machines (SVM). Fusion of SVM and ANN. Cross-validation for performance evaluation. | / | SVM=88.30% ANN=93.63% SVM-ANN=94.67% | / | • Adapts to new data without explicit programming.<br>• Maintains accuracy as data evolves by learning from new training samples.<br>• Especially effective with complex datasets.<br>• The proposed approach Improves the accuracy and offers an effective and accurate method for predicting diabetes onset. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Al Yousef et al. [30] | Collected from 5 hospitals using National Guard Health Affairs databases (NGHA databases) | K-nearest neighbours (KNN) The random tree forest (RTF) Support vector machine (SVM) The naïve Bayesian classifier (BC) Bayesian network (BN) Logistic regression (LR) Classification and regression tree (CART) Root mean square error (RMSE) River operating characteristic (ROC) Synthetic minority oversampling technique (SMOTE) | / | **without using the synthetic minority oversampling technique (SMOTE):** RF= 53% SVM= 60% LR= 62% BC= 63% BN= 63% KNN: K=1: Accuracy=55% K=10: Accuracy=62% K=50: Accuracy=60% **With using the synthetic minority oversampling technique (SMOTE):** RF= 28% SVM= 61% LR= 56% BC= 59% BN= 66% KNN: K=1: Accuracy=53% K=10: Accuracy=58% K=50: Accuracy=60% | Type 2 diabetes | • Utilizes a large dataset.<br>• Handles data balance effectively<br>• Includes all three groups: diabetic, prediabetic, and non-diabetic patients.<br>• Ensures no at-risk patient category is excluded from the study. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Edeh et al. [31] | Pima Dataset Hospital Frankfurt dataset | RF (Random Forest), SVM (Support Vector Machin), NB (Naive Bayes), DT (Decision Tree) | / | Frankfurt database: RF=98% DT=97% SVM=78% NB=77% Pima Indian Database: SVM=83% RF=80% NB=78% DT=70% | Type 2 diabetes | The four algorithms aim to enhance future observation categorization and reduce classification mistakes. Compared to other type 2 diabetes prediction models, these models are more accurate. The k-means algorithm was successfully used to improve data quality, as demonstrated in the research tables. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Qin et al. [32] | The National Health and Nutrition Examination Survey (NHANES) dataset. | CATBoost. XGBoost. Random Forest (RF). Logistic Regression (LR). Support Vector Machine (SVM). | / | CATBoost = 82.1% XGBoost = 70.8% RF = 78.4% LR = 68.9% SVM = 67% | Predict type 2 diabetes | • Utilizes the Boosting algorithm to mitigate gradient bias and prediction offset.<br>• Imbalanced Class Data Handling by using the SMOTE-NC method within the machine-learning model.<br>• Highlights the significance of feature selection.<br>• Reduces classification model complexity.<br>• Optimizes the diabetes prediction model. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Ahamed et al. [33] | Pima Indian dataset. | Logistic Regression (LR), The XGBoost (XGB), Gradient Boosting (GB), Decision Trees (DT), Extra Trees, RF, Light Gradient Boosting Machine (LGBM) | / | LGBM=95.2% RF=94.8% Extra Trees=94.6% DT=94.4% GB= 94.1% XGB= 83.3% LR= 75.20% | Diabetes mellitus type 2 | The researchers predicted diabetes using some of the most famous ML algorithms that are DT, RF and LR but the most positive side of this study is that they brought to light several extensions of decision tree algorithm: XGBOOST, GB, ET and LGBM, to conclude which one is the most efficient for an early prognostic of DT2. |
| Kale et al. [34] | | RF, SVM, DT, K-Nearest Neighbour (KNN), Logistic Regression (LR), Gradient Boosting (GB) | / | / | Type 1 diabetes, type 2 diabetes, type 3 diabetes (gestational diabetes) | This study was to create a system that, by fusing the findings of the algorithms mentioned before, each algorithm's accuracy was calculated along with the model's accuracy, the model for predicting diabetes was chosen from those with good accuracy. Also a diabetes prediction based on the Disease Influence Measure (DIM) has been made. |

Table 3.1: Table of the state of art

After studying the works presented in the section above, we found out that there are various artificial intelligence concepts to use for building a system of early prediction of diabetes. Presently, the predominant approaches involve the utilization of machine learning algorithms, artificial neural networks, and deep learning approaches. The presented studies demonstrate various methods and techniques for the diabetes prediction, the effectiveness of these predictive model can be significantly influenced by the choice of algorithms and data preprocessing techniques. Through various methodologies and evaluation metric, these studies show the efficacy of machine learning algorithms in predicting and managing diabetes, this ongoing research reflects a dynamic field, where the focus remains on harnessing artificial intelligence techniques to enhance accuracy and achieve early diabetes prediction.

As researchers continue to ameliorate these models, the capacity for predicting diabetes early and activating health care management becomes increasingly possible. These innovations contribute to better health results and the potential for lowering healthcare expenses.

Machine learning (ML) is a form of artificial intelligence (AI) that focuses on the development of algorithms and statistical models, allowing computer systems to improve their performance. Many research studies have leveraged these algorithms, including: the study [26] which conducted an extensive examination using a large dataset of hospital physical examination data and Pima Indians data. The results showed that random forest achieved the highest accuracy among the tested algorithms with an accuracy of 80.84%. The study emphasizes the importance of data preprocessing and feature engineering techniques based on data size and specific research problems for improving the accuracy of machine learning models. The study has the benefit of using a large sample of over 7,000 Chinese adults, making the results more generalizable. However, the study could have two limitations. The first is that the study is limited to the Chinese population (a single geographic region) which might not be applicable to other populations. The second is that it did not consider other risk factors for diabetes that could influence the results, such as genetic predisposition or lifestyle choices.

Another study [30] also employed machine learning techniques. This study encompassed several key steps, including data preprocessing, feature selection or extraction, model training, and evaluation with different metrics. The results showed that BN was the best classifier with an accuracy of 63% and 66% with and without SMOTE. However, it's important to note some limitations of the study. Firstly, missing values were a significant concern. Dealing with missing data is a crucial aspect of data preprocessing, as it can impact the analyses and the performance of machine learning models. Some algorithms can be sensitive to missing values, which can affect the accuracy of the model. In this study, the presence of missing data had a noticeable impact on accuracy. Additionally, Bayesian networks and random forest algorithms can be complex and demand a comprehensive understanding, which can lead to challenges when explaining the results to medical professionals.

Furthermore, these studies [25], [31], and [32] also aimed to predict and evaluate predictive models for diabetes mellitus using machine learning techniques. The articles present a comprehensive overview of how machine learning techniques can be used to predict diabetes mellitus. The authors employed commonly used machine learning (ML) algorithms such as Random Forest (RF), Decision Tree (DT), Linear Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), J48 algorithm, K-means clustering, and K Nearest Neighbour (KNN). Their models share similarities in approach, involving the prediction of diabetes based on their respective datasets using each algorithm individually. This approach allows for the evaluation and comparison of algorithm performance, ultimately selecting the best-performing model with

the highest accuracy. These approaches have for advantages that it uses trusted and approved ML methods, the realization of the model is not expensive and doesn't need very sophisticated software tools.

The articles [28], [32], and [34], all explored a variety of machine learning techniques and boosting methods in their respective studies: In their publication [33], the authors used a standard machine learning algorithms and boosting techniques such as Logistic Regression (LR), Random Forest (RF), Decision Trees (DT), XGBoost (XGB), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), and Extra Trees (ET). These boosting algorithms enhance the predictive capabilities of the models.

The authors in [32], on the other hand, conducted a comparative analysis of five machine learning models: CATBoost, XGBoost, Random Forest, Logistic Regression, and Support Vector Machine (SVM) for prediction diabetes on lifestyle data. This study found that CATBoost was the best model with an accuracy of 82.1%. in addition, it allows for better comprehension of how lifestyle factors can impact the risk of developing diabetes. Despite its effectiveness, it's worth noting that these methods, while highly accurate, are based on complex algorithms, potentially making their application may consume the time, rendering them impractical.

The study of the authors [28] employed a range of machine learning algorithms, including Random Forest (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), Linear Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB). They also incorporated boosting algorithms like: Ada boost, Gradient Boosting (GB), and an ameliorated algorithm of Naïve Bayes known as Gaussian NB. Additionally, they brought to light less applied ML methods in diabetes prediction: Linear Discriminant Analysis algorithm (LDA), Bagging and perceptron, to optimize their results, the authors created a pipeline that combined the six best-performing methods.

Whereas, the studies [25] and [34] didn't mention the results of the evaluation of their model which makes their model not valid yet in the domain of Machine learning prediction of diabetes. By analysing these studies, in terms of ML algorithms, it's shown that RF, SVM, DT and LR all performed efficiently but the boosting techniques were more successful, the extra trees are more effective than the traditional decision tree model. The accuracies of each algorithm vary from a dataset to another so their performances depend on the quality and the type of the data.

In some works, the authors universally incorporated artificial neural networks (ANNs) into their models. Particularly, the authors in [27] compared the ANN with RF and K-means, the ANN method had the highest accuracy and they assumed that an ANN model performed better than any ML technique. Additionally, in the work [24] the authors utilized a blend of Artificial Neural Network (ANN) and Logistic Regression (LR), this study demonstrated that the combined model yielded the most favourable outcomes. Among its advantages, ANNs are powerful and flexible algorithms which have the ability to learn, adapt and adjust to new situations. On the other hand, logistic regression is a widely used predictive algorithm is known for its ease to implement and interpret, as well as its speed in classifying unknown records. The combined utilization of these approaches offers unique advantages. It capitalizes on the accuracy and flexibility of ANNs, complemented by the simplicity of logistic regression which is a widely adopted predictive tool known for its ease of implementation and interpretation.

Coupled with, the work [29] which proposed an innovative machine learning architecture based on the combination of SVM and artificial neural network. This combination improved the prediction accuracy compared to the individual algorithms, reaching a high level of accuracy of 94.67%. The study focused on the importance of feature selection in improving model performance and also emphasized on increasing the survival rate of diabetic patients.

After an extensive review of related works, which included various models. Building upon the earlier research [7], which aimed to predict early gestational diabetes by comparing the

Deep Neural Deep Neural Network (DNN), Support Vector Machine (SVM), and Random Forest (RF) classifiers, that showed that Random Forest (RF) stood out with the highest accuracy rate, reaching 96%. We have opted to employ the Random Forest (RF) classifier as a pivotal element in our model. as we strive to further enhance it by employing the Particle Swarm Optimization (PSO) algorithm to select the optimal features for gestational diabetes prediction. Our primary objective is to enhance the accuracy of diabetes prediction, thereby advancing the existing body of knowledge in this field.

## 3.4 Conclusion

In this chapter, we have conducted a state-of-the-art review where we presented some of the most influential works in diabetes prediction that employed concepts and techniques of data mining, machine learning, deep learning, and artificial neural networks. We thoroughly examined each article in these works, analysing the approaches proposed by the authors, evaluating the models' results, summarizing the conclusions drawn by the researchers, and exploring their perspectives for future research. Each model exhibited its set of advantages and disadvantages; some utilized classic machine learning algorithms like RF and SVM, while others highlighted less well-known algorithms. Some even proposed original deep learning approaches. Each of these approaches made a valuable contribution to the early forecasting of diabetes mellitus. Additionally, the research targets varied, encompassing different types of diabetes. Some studies focused solely on one type, such as Type 2 diabetes, Type 1 diabetes, or gestational diabetes, while others addressed multiple types, covering all three of them.
In the next chapter, we will present our approach for early diabetes prediction in detailed.

# Chapter 4

# Contributions

## 4.1 Introduction

Every year, millions of people around the world are diagnosed with diabetes, a chronic disease that can be life-changing. Treating diabetes often begins with the critical step of predicting who is at risk of developing the disease. This need for prediction becomes even more important when considering the potential consequences of undiagnosed or poorly managed diabetes, such as serious health complications.

In this study, we propose an innovative diabetes prediction method that exploits the power of PSO and RF algorithms. The aim is to identify people at risk of diabetes at an early stage, allowing for timely intervention and personalized healthcare, not only benefits at-risk individuals, but has far-reaching implications for public health. It empowers individuals, reduces disease burden, lowers healthcare costs, and informs evidence-based public health policies, ultimately leading to healthier communities and a best in the fight against diabetes.

The challenge lies in the complexity of diabetes prediction, which involves analysing a multitude of variables, including medical history, lifestyle factors, and genetic predisposition, identifying the most relevant features (variables) within the dataset, choosing appropriate machine learning models for diabetes prediction. Our approach seeks to address this by combining advanced machine learning techniques with optimization algorithms like PSO, to enhance the accuracy of our predictions.

Our research aims to provide the people with a reliable aid by developing a diabetes prediction system that considers various factors, including medical data, lifestyle information, and genetic markers. This system will serve as a valuable tool to assist individuals, clinicians, and healthcare providers in making informed decisions and implementing preventive measures.

Our current study represents a logical extension of our previous work [7]. In this study, we aimed to identify the best features to improve gestational diabetes prediction by improving the random forest (RF) algorithm using the particle swarm optimization (PSO) algorithm. Through this research project, we hope to have a significant impact on the field of diabetes prediction. By striving to improve the accuracy of our predictions, facilitate early interventions, and ultimately contribute to better health outcomes for individuals at risk of diabetes.

This chapter delves into the conceptual intricacies of our approach, detailing the utilization of PSO and RF algorithms for diabetes prediction. We will explore how these methods optimize feature selection, enhance model performance, and contribute to more accurate risk assessments.

## 4.2 Proposed approach

In this section, we describe in detail our approach to developing our prediction system. Our system relates to the health field, with the objective to assist individuals, clinicians, and health-care providers in making well-informed decisions by developing an accurate and efficient system based on individual's health profiles, using Particle Swarm Optimization (PSO) algorithm for feature selection and the Random Forest (RF) algorithm for prediction. The overall architecture of the RF-PSO proposal is presented in Figure 4.1 and involves four main steps.

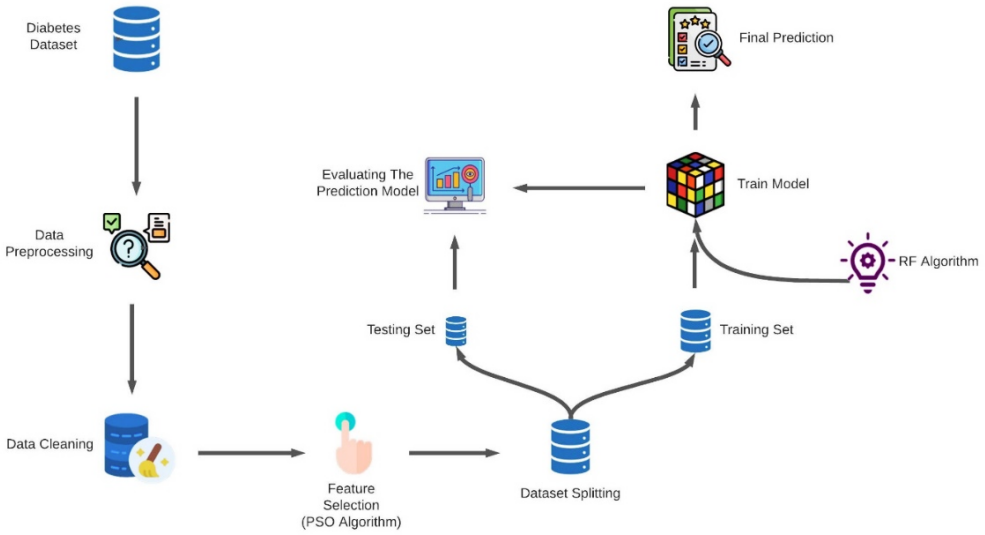



Figure 4.1: Proposed approach.

The first step "Data collection" is responsible to gather a comprehensive information in a dataset that includes diverse health-related features such as pregnancies, age, BMI, insulin, and family history of diabetes. The second step "Data Preprocessing" consists of performing data preprocessing by identifying missing values in our dataset. Most datasets represent missing values as NaN, which can be replaced with the mean or the median of the feature, handling outliers by first detecting them, either through the calculation of the Z-Score for each data point and identifying those beyond a certain threshold, or by identifying data points outside the IQR boundaries. Once identified, outliers are either removed or transformed, normalizing or standardizing the feature data to ensure consistent scales. Finally, ensuring data quality by addressing any noise present in the dataset. The third step "Feature selection" involves using PSO algorithm: this step aims to Implement the PSO algorithm to search for the best feature subset, uses a population of particles (solutions) that move through the feature space to find the optimal subset. And finally, the fourth step "Prediction": After selecting our optimal features, we use RF to make predictions. These predictions are first tested on a dataset to verify the efficiency of the proposal, and next on new, unseen data. Users can input the various characteristics, and the model will predict whether the person is likely to have diabetes or not in future.

Here's an algorithmic which describes the sequence of the entire process.

**Algorithm 1** Diabetes Prediction Process.

**Input:** Comprehensive dataset with diverse health-related features.

**Output:** Prediction result for each individual (diabetic or not).

**Begin**

1. Collect a comprehensive dataset containing diverse health-related features (e.g., pregnancies, age, BMI, insulin, family history of diabetes) for a set of individuals.
2. Identify and handle missing values:
   - Locate missing values in the dataset if not found go to 5.
   - Replace missing values with either the mean or median of the respective feature.
3. Detect and handle outliers:
   - Identify outliers using methods like Z-Score or Interquartile Range (IQR) if not found go to 4
   - Decide whether to remove outliers or transform them.
4. Normalize or standardize the feature data.
5. Applied PSO algorithm for feature selection mentioned below in **algorithm 3**.
6. Allow users to input the characteristics mentioned (e.g., pregnancies, age, BMI, insulin, family history of diabetes).
7. Use the trained model to predict whether the individual is likely to have diabetes or not.
8. Provide the prediction result to the user (e.g., "you have diabetes" or " you don't have diabetes").

**End.**

In the following, we will present in detailed the steps of the proposed approach.

### 4.2.1 Data collection

Data collection is an integral initial step in diabetes prediction and management, serving as a cornerstone for accurate model predictions. The calibre and comprehensiveness of the collected medical data serve as fundamental factors influencing the precision of predictive models.

In this study, the dataset used, is publicly available and sourced from Kaggle. It consists of several predictive medical variables obtained from a hospital in Frankfurt, Germany [35]. This dataset includes several various medical predictive variables, and encompasses 2000 individuals, categorized into diabetic and non-diabetic groups, each defined by nine distinct attributes. The dataset has a size of 12.0 kB and two distinct diabetes classes:

- Class 0: Represents healthy individuals.
- Class 1: Represents diabetic individual

Importantly, we augmented this dataset with an additional 200 records, each adhering to the same attribute set. These supplementary records were collected confidentially from the Internal Medicine Department of Khellil Amrane Hospital in Bejaia City [36], respecting patient data confidentiality and obtaining necessary permissions throughout the process.

Comprehensively, data collection for diabetes prediction, encompassing clinical and non-clinical factors associated with diabetes risk. plays a pivotal role in building accurate prediction models. The quality and quantity of collected data play a crucial role in the accuracy of the diabetes prediction model. In essence, the more complete and representative of the data is the target population, the more reliable the model is in its predictions. Here are some key aspects of data collection for predicting diabetes:

1. **Clinical Data:**a wide range of details related to a person's health, medical history, and physical condition. Common clinical data points include:
    - Blood glucose levels.
    - Haemoglobin A1c levels.
    - Body mass index (BMI).
    - Blood pressure.
    - Cholesterol levels.
    - Family history of diabetes.
    - Previous diagnoses of diabetes or prediabetes.
    - Medication history.
    - Age and gender.
2. **Lifestyle Factors:** refer to a set of behaviours, habits, and choices that individuals make in their daily lives that may influence diabetes risk. These factors can include:
    - Diet and nutrition patterns.
    - Physical activity levels.
    - Smoking status.
    - Alcohol consumption.
    - Stress levels.
3. **Biometric Measurements:** Biometric data, such as Blood Glucose Levels, Body Mass Index (BMI), Heart Rate, Cholesterol Levels, and Skinfold Thickness.
4. **Genetic and Family History:**are significant factors in understanding an individual's risk of developing diabetes, particularly type 2 diabetes. Including family history of diabetes and specific genetic markers, can be relevant for assessing an individual's predisposition to diabetes.
5. **Demographic Information:** Data on age, gender.
6. **Environmental Factors:**Environmental data, such as virus, exposure to toxic chemicals, as environmental factors can contribute to diabetes risk.

### 4.2.2 Data preprocessing

Data preprocessing is an important step after collecting the data, that involves cleaning, transforming the data into a format that is more easily and effectively processed in machine learning, and integrating raw data to prepare it for analysis. The goal of data preprocessing is to enhance data quality and to make it more suitable for the specific data mining task. When we collect data from real world it consists of redundant data, missing values, and outliers that will always end up with a model that would not be affective for prediction and data analysis. By using data preprocessing, we can able to remove all these issues.

Data preprocessing includes various phases:

- **Data exploration and visualization**
  Data exploration is the first phase of data analysis. It allows us to explore and visualize data. It is the process of examining and analysing a set of data to understand its structure, patterns, and characteristics that can be useful for decision making or for creating predictive models. Data exploration is a crucial step in the data analysis process because it also helps identify anomalies or outliers within the data that can inform subsequent data cleaning.

  The process of data exploration is multifaceted and involves several key steps. It begins with data collection and preparation, where raw data is gathered and organized into a format suitable for analysis. Once the data is ready, exploratory data analysis (EDA) techniques are applied. EDA involves generating summary statistics, visualizing data distributions, and identifying potential relationships between variables. As well as, data visualization is an essential part of data analysis and reporting results, it plays a vital role in data exploration. It represents data through use of common graphics, such as charts, plots, and infographics that are commonly used to visualize data distributions and relationships. These visual representations help analysts grasp the underlying patterns and variations in the data.

  These visual displays of information represent data graphically in order to extract information, detect trends and communicate complex data relationships. It facilitates information extraction and communication of data by allowing human brain to understand information in an easy, clear, understandable and smart way. This tool helps to identify correlations, outliers and anomalies in data.

- **Data cleaning**
  When building datasets, there are many chances for data to be duplicated or mislabelled. Data cleaning is the process used to rectify or eliminate inaccurate, corrupted, improperly formatted, duplicate, or incomplete data within a dataset. This process is used to ensure the integrity and quality of the data, making it suitable for reliable analysis and modelling.

  Data cleaning involves a series of steps, such as identifying and correcting errors, handling missing values, outliers and removing duplicate entries. Moreover, data cleaning involves ensuring data consistency and standardization. This may include converting units of measurement, resolving inconsistencies in data formatting.

Among the techniques commonly used in data preprocessing are:

→ **Remove duplicates values**
Removing duplicates from datasets is an important data preprocessing step in data analysis and machine learning, it is critical to ensure data accuracy, improve analysis results and model performance, and maintain data quality. As duplicate datasets may distort statistical analysis results. Eliminating duplicates allows us to get more accurate statistics.

→ **Remove null values**
This technique consists of replacing missing values with values estimated or calculated from other available data. Imputation methods include mean, median, mode, or even more advanced methods.

→ **Visualizing and checking outliers**
Visualizing and checking for outliers is an important step in data analysis. It helps to understand the distribution of the data and identify potential anomalies or data points that deviate from the majority of the dataset. Outliers are data points that significantly differ from the majority of the data, they are often found in variables with skewed distributions, and they can have a negative impact on machine learning and statistical modelling performance. Therefore, detecting outliers is crucial for accurate and reliable data analysis.

There are several commonly used visualization techniques such as: Scatter Plot and box plot. To visualize our distribution and summary statistics, we create a box plot using the Seaborn or Matplotlib library. Box plots are a valuable tool for visualizing the spread of data, as well as for identifying potential outliers. Box plots show the five-number summary of a data set: the minimum score, the first (lower) quartile, the median, the third (upper) quartile, and the maximum score.

Here are the elements that make up a boxplot:

- **Box**: The main rectangular box represents the interquartile range (IQR), which encompasses the middle 50% of the data.
- **Median**: This is the midpoint of the dataset. 50% of the data points will be below this value and 50% of the data will be above this value.
- **First quartile**: this is the 25th percentile point. The values of 25% of the data points are less than this value and 75% are greater than this value.
- **Third quartile**: this is the 75th percentile point. The values of 75% of the data points are lower than this value.
- **Interquartile range (IQR)**: these are the points between the 25th and 75th percentile.
- **Whiskers**: are the lines that extend outward from the box the minimum and maximum values within a certain range.

  **Minimum (Lower Whisker)**: This is the lowest value in the dataset excluding outliers.

  **Maximum (Upper Whisker)**: this is the point of the maximum value excluding outliers.

The outlier points fall outside the range of the whiskers and are typically displayed individually as data points or small dots.
The figure 4.2 shows the detecting of outliers using IQR.

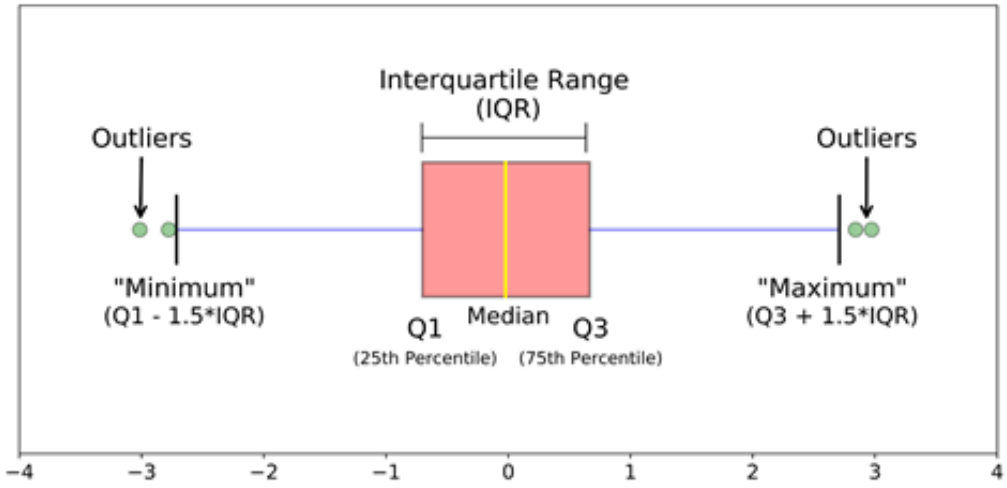



Figure 4.2: IQR to detect Outliners
[37].

**Detecting outliers using the Inter Quantile Range(IQR)**

- Sort the dataset in ascending order.
- Calculate the first quartile (Q1) and third quartile (Q3) of the dataset.
- Calculate the Interquartile Range (IQR): IQR=Q3-Q1.
- Determine the lower bound (LB) and upper bound (UB) for potential outliers: LB =Q1 – (1.5 * IQR), UB = Q3 + (1.5 * IQR).
- Use the bounds to highlight any outliers, all values that fall outside the bounds below the lower bound and above the upper bound and mark them as outliers.
- Remove or modify the outliers as needed to clean the dataset.

Here's the algorithm used in our case:

**Algorithm 2** Outlier Detection and Handling for a list or array of numerical values.

**Input:** A list or array of numerical values.

**Output:**A list of outliers.

**Begin**

1. Sort the dataset in ascending order.
2. Calculate the first quartile (Q1) and third quartile (Q3) values.
   - Q1 is the median of the lower half of the dataset.
   - Q3 is the median of the upper half of the dataset.
3. Calculate the Interquartile Range (IQR) as:
   - IQR = Q3 - Q1
4. Determine the lower bound and upper bound for potential outliers:
   - Lower Bound = Q1 - (1.5 * IQR)
   - Upper Bound = Q3 + (1.5 * IQR)
5. Iterate through the list and identify values that fall below the lower bound or above the upper bound. These values are considered outliers.
6. Store the outliers in a list or array.
7. Return the list of outliers as the output.

**End.**

In the following we will explain our algorithm using an example, to demonstrate the outlier detection and handling algorithm, let's use a list with outliers.

**Example 1:** A list with Outliers
Suppose we have a dataset of ages for a group of people:
Age= 20, 22, 150, 25, 37, 26, 40, 28, 30, 32, 35, 62].

1. Arrange the dataset in ascending order:
   Age = [20, 22, 25, 26, 28, 30, 32, 35, 37, 40, 62, 150].
2. Calculate Q1 and Q3:
   - Q1 = 26
   - Q3 = 37
3. Calculate the Interquartile Range (IQR):
   - IQR = Q3 - Q1 = 37 - 26 = 11
4. Determine the lower and upper bounds for potential outliers:
   - Lower Bound = Q1 - 1.5 * IQR = 26 - 1.5 * 11 = 9.5
   - Upper Bound = Q3 + 1.5 * IQR = 37 + 1.5 * 11 = 53.5
5. Identify outliers: In this case, the values 62 and 150 fall outside the range of 9.5 to 53.5. Therefore, 62 and 150 are the outliers.

6. Outliers = [62,150]
7. Once the outliers are detected, we can either choose to remove them or handle them.

→ **Correlation between all the features**
Correlation matrix is an essential tool in statistics and data analysis, used to examine the relationship that is frequently encountered between two variables in all human or applied sciences. The correlation coefficient measures the intensity of the co-variation between the two variables. Generally, there are two types of correlation tests: parametric tests, such as the Pearson test, and non-parametric tests, such as the Spearman test. The Pearson coefficient, denoted as "r," is calculated using the formula:

$$r = \frac{\mathrm{Cov}(X, Y)}{(\sigma_X \cdot \sigma_Y)} \tag{1}$$

where Cov(X, Y) is the covariance between variables X and Y, and $\sigma$X and $\sigma$Y are their respective standard deviations. The closer the absolute value of the correlation coefficient "r" is to 1, the more correlated the variables are.

→ **Data scaling and normalization**
Data scaling, is a fundamental preprocessing phase, it involves transforming the numerical attributes or features of a dataset to a specific range or distribution, this transforms all data points so that they fall between 0 and 1. The goal of data scaling is to ensure that all features contribute equally to the analysis, preventing some features from dominating others due to differences in their scales. This can improve the accuracy of machine learning algorithms.

### 4.2.3 Feature selection

The feature selection process is performed using the Particle Swarm Optimization (PSO) algorithm, chosen for its ability to search for optimal solutions in complex spaces. In this context, PSO is utilized to pinpoint and choose the best features from the dataset. The reduced subset includes important features relevant to the dataset. Next, the RF algorithm is used for classification to achieve better prediction accuracy. Swarm intelligence is a distributed solution to complex problems which intend to solve complicated problems by interactions between simple agents and their environment, is used to search for optimal solutions in complex problem spaces, is a computational optimization algorithm that draws inspiration from the collective behaviour observed in natural systems, like the social behaviour of bird flocking or fish schooling. It is particularly useful for solving complex optimization problems. In the context of PSO,as shown in **figure 4.3**, each member of the swarm is called a "particle." In the standard PSO process, after the initialization of a random particle population. During each iteration, particles select attributes, separate data into training and test sets, train on the training data, classify using the test data, and store evaluation metrics. Each particle then updates its velocity and position during each iteration based on its individual experience (pbest) and the collective best experience of all particles (gbest), as it is described in the equation 2. At the end of each iteration, the performance of all particles is evaluated using a predefined cost function. This process continues until a stopping criterion is met.

$$v(t+1) = w \cdot v(t) + c_1 \cdot r_1 \cdot (p_{\mathrm{best}}(t) - p(t)) + c_2 \cdot r_2 \cdot (g_{\mathrm{best}}(t) - p(t)) \tag{2}$$

Where:

- i ranges from 1 to N , where N represents the number of particles in the swarm population.

- v[t] is i-the velocity vector of the th particle at the t-th iteration.
- p[t] ] represents the current position of the i-th particle at the t-th iteration.
- pbest[t] is the previous best position of the i-th particle at the t-th iteration.
- gbest[t] is the previous best position among all particles in the swarm at the t-th iteration [38].
- The parameters w, cog, and cos are positive acceleration coefficients that control the balance between local and global search efforts.

‘

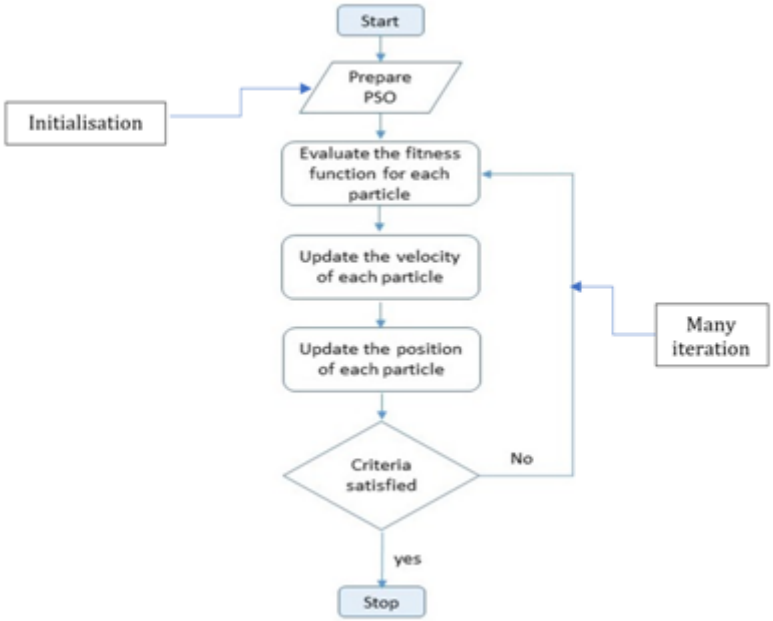



Figure 4.3: PSO flowchart

The common proposed feature selection algorithm is given below:

**Algorithm 3** Feature Selection with PSO.

**Input:** K, M, max-iterations, X.

**Output:** Selected features.

**Begin**

1. Initialize Population of K particles with random binary vectors of length N (the number of features
2. Initialize global-best-particle to None and global-best-fitness to 0.
3. Set PSO parameters: w, c1, c2.
4. FOR iteration = 1 to max-iterations.
   a. FOR i = 1 to K
   i. Create a subset of features based on the binary vector of the current particle.
   ii. Split data into Training and Validation sets.
   iii. Train a machine learning classifier (Random Forest) on the Training data.
   iv. Classify using Test data to make predictions.
   v. Calculate the fitness (e.g., accuracy) of the model's predictions.
   vi. Update global-best-particle and global-best-fitness if the fitness is better.
   vii. Update the velocity and position of each feature in the binary vector.

   b. NEXT i
   c. End of the loop for particles
5. End of the loop for iterations.
6. Select features based on the binary vector of the global-best-particle.

**End.**

This algorithm illustrates the iterative process of selecting the best subset of features using PSO. It also outlines the continuous adjustment of particle positions and velocities to improve feature selection.

### 4.2.4 Prediction process

The main objective of this phase is to predict whether a patient is diabetic or not using the RF classifier, which is a robust machine learning algorithm known for its effectiveness in both regression and classification tasks.

The RF algorithm is widely employed in the field of medicine due to its versatility and applications in various healthcare scenarios, including identifying disease trends and assessing disease risks.

Random Forest is a supervised learning algorithm that constructs an ensemble of decision trees. These decision trees are trained using a technique known as "bagging," which involves growing trees on randomly selected subsets of the data. In a Random Forest, each tree in

the ensemble generates predictions, and the final prediction is determined by a majority vote for classification tasks or the average value for regression tasks. This ensemble approach is attributed to Leo Breiman [39].

During this step, we utilize the RF classifier on the dataset while incorporating a feature selection process powered by the PSO. The aim is to classify each patient as either diabetic or non-diabetic.

The purpose of the **Algorithm 5**, is to create an ensemble of Decision Trees using feature selection results obtained through PSO and then use these trees to make predictions for diabetic classification.

The predictions made by Random Forest can fall into two categories:

1. **Regression:** When the output is a mean value, the Random Forest is used for regression tasks, aiming to predict continuous numerical values.

2. **Classification:** When the output is the mode among the trees predictions, the Random Forest is applied to classification problems, where the goal is to assign data points to specific classes.

The RF classifier operates in two distinct phases:

1. **Phase 1:** RF Ensemble Construction
   In the first phase, we construct the Random Forest ensemble by combining a total of N Decision Trees (DT).

2. **Phase 2:** Prediction Generation
   The second phase involves generating predictions for each decision tree within the ensemble.

Generally, the process follows this structure:

1. **Input Parameters:** The dataset (X, y), the number of decision trees to create (K), and a set of new data points (NewData).

2. **Decision Tree Creation:** The algorithm selects K data points from the original dataset (X, y) and creates K decision trees using these selected data points.

   (a) **Prediction Process:**
   For each new data point, the algorithm iterates through each decision tree and makes predictions. The predictions from all decision trees are collected.

   (b) **Majority Voting:**
   The algorithm determines the class with the majority votes among the predictions from the decision trees for each new data point.

   (c) **Output:**
   The algorithm returns the predicted class for each new data point.

At the conclusion of these phases, a decision is made by selecting the class with the highest number of votes (majority voting) among the predictions generated by the individual decision trees.

Here is Algorithm 4, which employs the same structure.

**Algorithm 4** Prediction Process with Random Forest.

**Input:**Dataset (X, y), K, NewData.

**Output:**Predicted class for each new data point.

**Begin**

1. Select K data points.
2. Initialize an empty list to store the decision trees: DecisionTrees = []
3. For i = 1 to K:
   (a) Create a decision tree (DT) using the selected data points (Each decision tree is constructed independently).
   (b) Add the DT to the DecisionTrees list.
4. For each new data point in NewData:
   (a) Initialize an empty list to store predictions for the data point: Predictions = []
   (b) For each decision tree DT in DecisionTrees:
      i. Make a prediction using DT for the new data point.
      ii. Add the prediction to the Predictions list.
   (c) Determine the class with the majority votes in the Predictions list.
   (d) Assign the new data point to the class with the majority votes.
5. Return the predicted class for each new data point.

**End.**

**Algorithm 5** Random Forest with PSO Feature Selection.

**Input:** Dataset (X, y), N, PSO Feature Selection Results

**Output:** Predictions for Diabetic Classification

**Begin**

1. Construct an empty ensemble of N Decision Trees (DT).
2. For each DT in the ensemble:
   (a) Train the DT using the selected features obtained from PSO.
   (b) Generate predictions for each patient in the dataset.
3. Combine the predictions from all DTs through majority voting.
4. Output the final classification predictions for each patient.

**End.**

## 4.3 Conclusion

In this chapter, we have introduced our approach to diabetes prediction, utilizing advanced machine learning technique RF, specifically harnessing the power of the PSO. Our system encompasses various critical stages, including data collection, feature selection through PSO, model training with RF, and result interpretation. We have introduced the PSO algorithm as a means of optimizing feature selection, enhancing the model's ability to discern important factors in diabetes prediction. Additionally, we have elucidated the mechanics of the Random Forest algorithm, demonstrating how it efficiently processes data and generates accurate predictions.

Our diabetes prediction system represents a significant advancement in the field of Health-Care. Its primary objective is to advance early detection and intervention for individuals at risk of diabetes, leading to improved overall public health outcomes.

In the next chapter, we will take our research to the next level by implementing and rigorously evaluating our approach in the context of diabetes prediction. Furthermore, we will present the tools and development environment employed in our study.

In conclusion, our diabetes prediction system stands as a vital tool in the pursuit of proactive healthcare management. By applying cutting-edge technology to predict and prevent diabetes, our aim is to empower both individuals and healthcare providers with the information they need to make informed decisions and, in turn, alleviate the burden of this widespread chronic disease.

# Chapter 5

# Experiment and evaluation

## 5.1 Introduction

The overarching goal of this project is to create an advanced predictive model tailored for the early detection of gestational diabetes (GDM). To achieve this objective, we have chosen to leverage the Random Forest (RF) algorithm, a highly regarded ensemble learning method renowned for its exceptional accuracy and resilience in handling complex classification tasks.

Furthermore, we recognize the critical importance of optimizing our model's performance, and to that end, we have integrated Particle Swarm Optimization (PSO) into the development process. PSO, a well-established optimization technique, will enable us to fine-tune the hyper-parameters of the predictive model. This optimization will enhance its predictive capabilities and, in turn, increase its effectiveness in identifying cases of gestational diabetes.

The creation of a Gestational Diabetes Prediction System that combines the prowess of RF and the optimization capabilities of PSO carries profound implications for the field of prenatal care. It has the potential to substantially enhance the quality of care provided to pregnant women at risk of developing GDM. By harnessing the synergistic power of cutting-edge machine learning techniques and advanced optimization methods, our objective is to equip healthcare professionals with a valuable and precise tool.

This tool will facilitate early identification and intervention, contributing significantly to the overall well-being of both expectant mothers and their unborn children during the crucial period of pregnancy. Ultimately, our project aspires to make a positive impact on maternal and foetal health outcomes, offering a promising avenue for improving the healthcare landscape surrounding gestational diabetes.

## 5.2 Dataset description

### 5.2.1 Definition

Oxford Dictionary defines a dataset as "a collection of data that is treated as a single unit by a computer". A dataset is a structured collection of data that is organized and presented in a way that makes it suitable for analysis, research, or other purposes. It consists of individual data points, which can be in the form of numbers, text, images, or other types of information, and these data points are typically organized into rows and columns, with each row representing a single observation or record, and each column representing a specific attribute or variable [40].

The Frankfurt Hospital Diabetes Dataset [35] is a collection of data related to diabetes patients. This dataset is commonly used for various data analysis and machine learning tasks to predict diabetes or understand factors associated with diabetes. Researchers and data scientists

often use this dataset to build predictive models for diabetes diagnosis or to conduct exploratory data analysis to better understand the relationships between these features and diabetes. It's downloaded from Kaggle and consists of several predictive variables and is in CSV format because it is more convenient for Python to handle this type of file in the field. The dataset has a size of 62.06 kB and includes 2000 diabetic and non-diabetic patients.

Below is a description of the columns or features typically found in this dataset:

- **Glucose:** This column represents the plasma glucose concentration at 2 hours during an oral glucose tolerance test. It measures the level of glucose in the patient's blood after a specific period.
- **Pregnancies:** The "Pregnancies" column indicates the number of times the patient has been pregnant. This information is relevant as pregnancy can affect a person's risk of developing diabetes.
- **Blood Pressure:** This column represents the diastolic blood pressure of the patients, typically measured in millimetres of mercury (mm Hg). High blood pressure can be a risk factor for diabetes.
- **Skin Thickness:** The "Skin Thickness" column measures the thickness of the triceps skinfold, typically in millimetres (mm). Skinfold thickness can be an indicator of body fat, which is related to diabetes risk.
- **Insulin:** This column represents the serum insulin level measured 2 hours (mu U/ml). after a specific event. Insulin is a hormone that regulates blood sugar levels, and its measurement can provide insights into diabetes.
- **BMI (Body Mass Index):** The BMI, or Body Mass Index, is calculated using the patient's weight (in kilograms kg) divided by the square of their height (in meters m). It is a measure of body fat and is often used to assess obesity, which is a risk factor for diabetes.
- **Diabetes Pedigree Function:** This column typically contains a numerical value that represents a genetic factor or a function related to diabetes. It may indicate the genetic predisposition of a patient to develop diabetes.
- **Age:** The "Age" column represents the age of the patients in years. Age is an important factor in assessing diabetes risk, as the likelihood of developing diabetes tends to increase with age.
- **Outcome:** The "Outcome" column is the target variable or the label. It is a binary variable, where:
  - 0 usually indicates that the patient does not have diabetes.
  - 1 usually indicates that the patient has diabetes.

As it's mentioned in the previous chapter, the dataset was expanded by 200 additional records (patient cases), and the data for the individuals was collected using the same standards. During student internships at the hospital, these additional entries were obtained under strict confidentially from the Internal Medicine Department of Khellil Amrane Hospital in Bejaia City, Algeria [36]. The figure 5.1. shows an overview of the dataset.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 2 | 138 | 62 | 35 | 0 | 33.6 | 0.127 | 47 | 1 |
| **1** | 0 | 84 | 82 | 31 | 125 | 38.2 | 0.233 | 23 | 0 |
| **2** | 0 | 145 | 0 | 0 | 0 | 44.2 | 0.630 | 31 | 1 |
| **3** | 0 | 135 | 68 | 42 | 250 | 42.3 | 0.365 | 24 | 1 |
| **4** | 1 | 139 | 62 | 41 | 480 | 40.7 | 0.536 | 21 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1995** | 2 | 75 | 64 | 24 | 55 | 29.7 | 0.370 | 33 | 0 |
| **1996** | 8 | 179 | 72 | 42 | 130 | 32.7 | 0.719 | 36 | 1 |
| **1997** | 6 | 85 | 78 | 0 | 0 | 31.2 | 0.382 | 42 | 0 |
| **1998** | 0 | 129 | 110 | 46 | 130 | 67.1 | 0.319 | 26 | 1 |
| **1999** | 2 | 81 | 72 | 15 | 76 | 30.1 | 0.547 | 25 | 0 |

2000 rows × 9 columns

Figure 5.1: Overview of the dataset.

### 5.2.2 Statistical summary of the Data Frame

This summary shown in the figure 5.2 gives a quick overview of the dataset. We use pandas describe method to view some basic statistical details like: mean, std, min. . . , pandas information shows column data types (feature), number of non-zero values and memory usage. In our Data Frame, the maximum number of a pregnancies in our data frame that woman has is 17 pregnancies and she's a diabetic.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 |
| **mean** | 3.703500 | 121.182500 | 69.145500 | 20.935000 | 80.254000 | 32.193000 | 0.470930 | 33.090500 | 0.342000 |
| **std** | 3.306063 | 32.068636 | 19.188315 | 16.103243 | 111.180534 | 8.149901 | 0.323553 | 11.786423 | 0.474498 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| **25%** | 1.000000 | 99.000000 | 63.500000 | 0.000000 | 0.000000 | 27.375000 | 0.244000 | 24.000000 | 0.000000 |
| **50%** | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 40.000000 | 32.300000 | 0.376000 | 29.000000 | 0.000000 |
| **75%** | 6.000000 | 141.000000 | 80.000000 | 32.000000 | 130.000000 | 36.800000 | 0.624000 | 40.000000 | 1.000000 |
| **max** | 17.000000 | 199.000000 | 122.000000 | 110.000000 | 744.000000 | 80.600000 | 2.420000 | 81.000000 | 1.000000 |

Figure 5.2: Statistical summary of the Data Frame.

In the above table, we can see that the min value of Glucose, Blood Pressure, Scantiness, Insulin, BMI is zero, which is not common that these values be zero and thus indicates missing values.

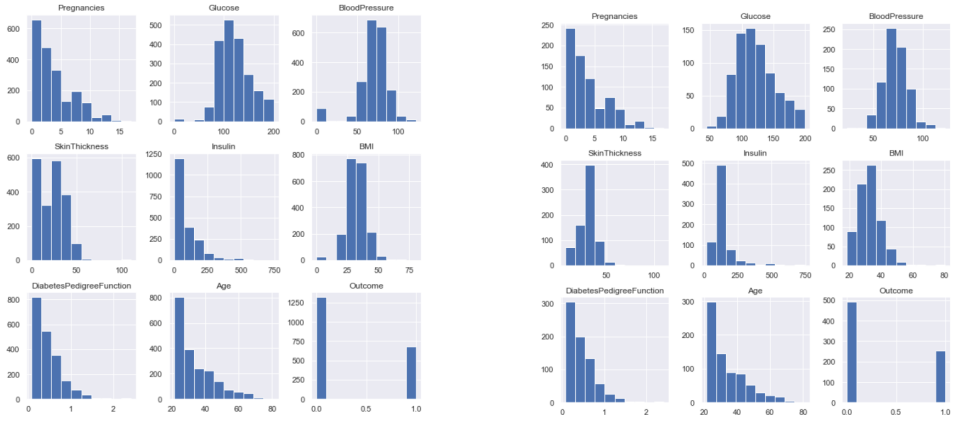
### 5.2.3 Plotting the data distribution plots

#### Histogram

Figure 5.3 represents the plots that show the frequency distribution of all the columns before and after cleaning missing values and duplicates, we observe that:

- In the histogram plot of blood pressure, it is evident that before data cleaning, there were approximately 100 individuals whose recorded blood pressure was at a value of 0.
- 230 women are around 0 or 2 times of pregnancy; 50 people are pregnant more than 5 times.

- In the glucose histogram, without replacing values we can see that the glucose was 100 for around 520 people and become 160 people with the same glucose as we have modified the data frame the distribution has changed.




(a) Before (b) After

Figure 5.3: Frequency distribution of all the columns before and after cleaning missing values.

### The balance of the data

The graph below shows that the number of diabetics is almost half the number of non- diabetics. The figure 5.4 shows that age and pregnancies, skin thickness and BMI are well correlated.

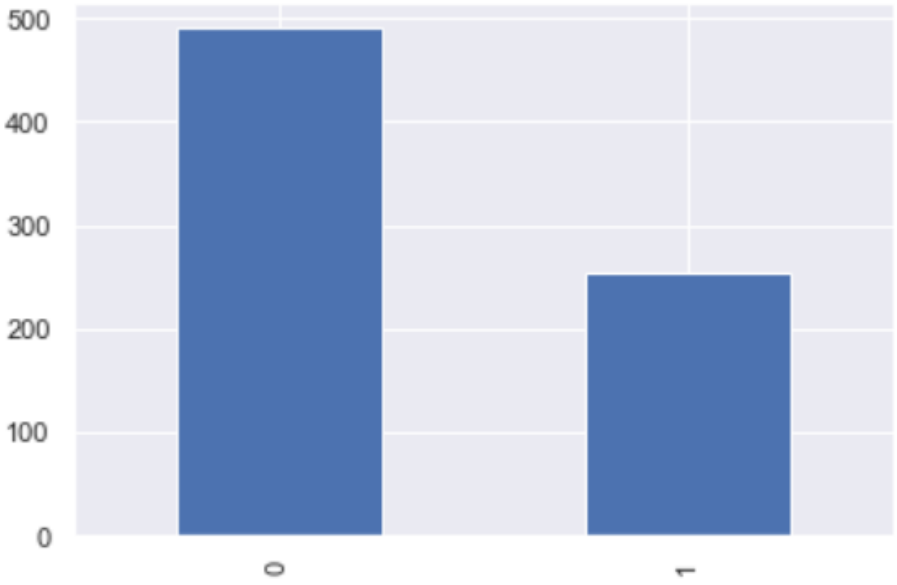



Figure 5.4: The balance of the data.

## 5.2.4 Data cleaning

### NaN values and outliers removal

As explained in Chapter Four, it was essential to preprocess the database before utilizing it. To achieve this, we removed the NaN values (null values) and the outliers.

```
Pregnancies                 0
Glucose                     5
BloodPressure              34
SkinThickness             215
Insulin                   359
BMI                        10
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

(a) Before

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

(b) After

Figure 5.5: Number of NaN values before and after cleaning.

Here in figure 5.5, we can see that the number of NaN values in glucose was 5, in blood pressure was 34. To clean these missing values, we will replace it with the mean. In results clearly see that there are no null values present in any of the features.

Following the process of removing the outliers from the dataset, it was observed that the dataset went from initially containing 2500 rows, each representing a patient, to a reduced dataset with 1813 rows, each still corresponding to an individual patient. This reduction in the number of rows occurred because the outliers, which are data points significantly different from the majority, were excluded from the dataset during the cleaning process.

### Correlation Matrix

As defined in the chapter 4, a correlation matrix is a table that displays the correlation coefficients between many variables. Each cell in the table shows the correlation between two variables. Correlation coefficients quantify the strength and direction of a linear relationship between two variables. Figure 5.6 represents the correlation matrix of the variables of our dataset, it shows that age and pregnancies, skin thickness and BMI are well correlated.

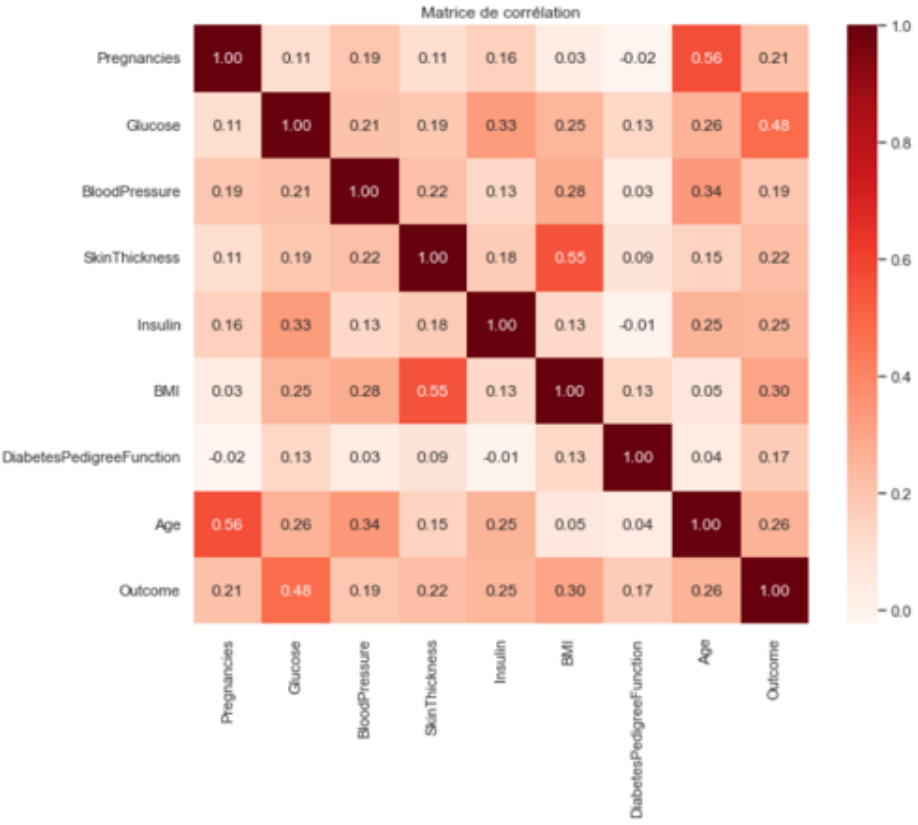



Figure 5.6: Correlation matrix.

## 5.3 Development environment

### 5.3.1 Hardware environment

- **Personal computer 1**

  **Machine type:** DELL Latitude 5310 2-in-1.
  **Processor:** 10th Gen Intel® Core(TM) i5-10310U CPU @ 1.70GHz, 2.21 GHz.
  **Random access memory (RAM):** 16,0 Go.
  **Exploitation system:** 64 bits, processor x64.
  **Operating system:**Windows 10 Professional.

- **Personal computer 2**

  **Machine type:** DELL Vostro 15 3510.
  **Processor:** 11th Gen Intel® Core(TM) i7-1165G7 @ 2.80GHz, 2803 MHz
  **Random access memory (RAM):** 16,0 Go.
  **Exploitation system:** 64 bits, processor x64.
  **Operating system:**Windows 10 Professional.

### 5.3.2 Software environment

Python, created by Guido van Rossum in 1991, is a widely-used programming language known for being open source, high-level, and suitable for various purposes. Python supports different

programming paradigms like object-oriented, imperative, functional, and procedural, and it's especially favoured in the field of machine learning. It was among the pioneering languages to offer libraries and tools for machine learning [38]. In our implementation, we rely on several frequently used libraries, including:

**NumPy:** a Python library providing data structures and functions for scientific and numerical computations, offering advanced mathematical and statistical operations. NumPy is extensively employed in domains such as machine learning and data science due to its high performance and user-friendly nature [41].

**Pandas:**s an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [42].

**Seaborn:**abbreviated as sns, seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is particularly useful for visualizing complex datasets and statistical relationships in a concise and aesthetically pleasing manner [43].

**Scikit-learn:** often abbreviated as sklearn, is a popular open-source machine learning library for Python. It is a versatile and user-friendly library that provides tools for various aspects of machine learning, including data preprocessing, feature selection, model training, and evaluation. Scikit-learn is built on top of other widely used Python libraries like NumPy, SciPy, and matplotlib, making it an integral part of the Python machine learning ecosystem, it is widely used in both academia and industry for tasks such as classification, regression, clustering, and more. It plays a crucial role in making machine learning accessible to a broader audience due to its simplicity and ease of use [44].

**Matplotlib:**: a well-known Python library for crafting data visualizations and graphs. Matplotlib empowers users to tailor every aspect of their graphs, such as axes, labels, legends, colours, and line styles. This adaptability makes it a versatile tool for visually representing data in diverse fields, including data science, research, and data visualization [45].

**PySwarms:** PySwarms stands as a versatile and robust research toolkit designed specifically for the realm of particle swarm optimization (PSO) within the Python ecosystem. This library serves as a valuable resource for a wide spectrum of users, ranging from swarm intelligence researchers to practitioners and students, all of whom seek a sophisticated and expressive platform to harness the power of PSO for their problem-solving needs. With PySwarms, users are granted access to an elevated, high-level, and declarative interface, simplifying the implementation of PSO methodologies for their unique challenges. The library empowers individuals to embark on fundamental optimization tasks using PSO while also facilitating seamless interaction with various swarm optimization techniques. In essence, it transcends the boundaries of conventional PSO libraries, offering a comprehensive and adaptable environment to explore the full potential of swarm intelligence for a myriad of problem domains [46].

**Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It is widely used for data analysis, scientific computing, machine learning, and data visualization tasks. Jupyter supports a wide range of programming languages, including Python, R, Scala, etc [47].

**Django:** Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source [48].

**CSS:** Cascading Style Sheets, is a stylesheet language used to describe the presentation of a document written in HTML or XML (including XML dialects such as SVG, MathML or XHTML). CSS describes how elements should be rendered on screen, on paper, in speech, or on other media [49].

**HTML:** HyperText Markup Language, it is a standard markup language for web page creation. It allows the creation and structure of sections, paragraphs, and links using HTML elements (the building blocks of a web page) such as tags and attributes [50].

**Bootstrap:** it is a free, open source front-end development framework for the creation of websites and web apps. Designed to enable responsive development of mobile-first websites, Bootstrap provides a collection of syntax for template designs. As a framework, Bootstrap includes the basics for responsive web development, so developers only need to insert the code into a pre-defined grid system. The Bootstrap framework is built on Hypertext Markup Language (HTML), cascading style sheets (CSS) and JavaScript. Web developers using Bootstrap can build websites much faster without spending time worrying about basic commands and functions [51].

## 5.4 Implementation

In alignment with the comprehensive insights provided earlier in this document, we take pride in introducing a sophisticated web application meticulously constructed using the powerful Django framework. Our application serves a vital niche within the healthcare domain, focusing primarily on addressing the needs of medical specialists, particularly gynaecologists and endocrinologists. Its core purpose revolves around equipping these esteemed professionals with a versatile tool, tailor-made to facilitate the input of essential patient information. This, in turn, empowers them to conduct comprehensive assessments and prognostic evaluations pertaining to the risk of gestational diabetes development during pregnancy. This web application seamlessly accommodates the unique requirements of the medical practitioners it serves. It not only provides a user-friendly interface but also facilitates the input of personalized patient data with the utmost convenience and security, in the following titles, we will offer a detailed description of our application's graphical user interfaces (GUI).

### 5.4.1 Home Interface

Within Figure 5.7, we showcase the primary interface of the application. In this GUI, a conspicuous and engaging element takes centre stage: the "Here we go" interactive button. This button, akin to a gateway, affords the user, who assumes the persona of a medical practitioner, the opportunity to progress effortlessly to the subsequent interface, it's specifically dedicated to the task of predicting gestational diabetes, and it becomes accessible with a mere, straightforward click of the button.


Figure 5.7: Home Interface.

### 5.4.2 Diabetes Prediction Interface

Within the scope of Figure 5.8, we delve into the specialized GUI tailored for the prediction of gestational diabetes. This interface stands as a pivotal tool for healthcare professionals, particularly doctors, who are entrusted with the critical task of assessing their patients' susceptibility to this condition. At the heart of the interface lies an intricately designed interactive form, thoughtfully engineered to facilitate the seamless input of patient data. This form serves as a conduit through which doctors can convey the essential information necessary to perform an accurate prediction of diabetes risk. These data fields, each meticulously labelled and organized, are prominently displayed on the interface, ensuring utmost clarity and user-friendliness. It is worth noting that the parameters and data elements required for this prediction process have been comprehensively defined and elucidated in prior sections. This deliberate approach aims to equip healthcare professionals with a clear and concise understanding of the data inputs necessary to inform precise gestational diabetes predictions, thereby empowering them to make informed clinical decisions.


Figure 5.8: Diabetes Prediction Inteface.

## 5.5 Evaluation

### 5.5.1 Evaluation metrics

#### Accuracy

Accuracy [52] is a commonly used metric in machine learning and statistics to assess the overall performance of a classification model. It measures the proportion of correct predictions made

by the model out of all the predictions it made. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Here's a breakdown of the components involved in this formula:

1. **True Positives (TP):**These are the instances that were correctly predicted as positive by the model.
2. **True Negatives (TN):** These are the instances that were correctly predicted as negative by the model.
3. **False Positives (FP)** These are the instances that were incorrectly predicted as positive by the model.
4. **False Negatives (FN)**These are the instances that were incorrectly predicted as negative by the model.

Accuracy provides an overall assessment of how well a classification model is performing in terms of making both positive and negative predictions correctly.

### Precision

Precision [52] is a fundamental metric used in various fields, including statistics, machine learning, and information retrieval, to evaluate the performance of classification models, particularly in binary classification problems. Precision measures the accuracy of positive predictions made by a model and is defined as the ratio of true positive predictions to the total number of positive predictions made by the model, or more formally:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Let's dissect the elements included in this equation:

1. **True Positives (TP):** These are the instances that were correctly predicted as positive by the model. In other words, they are the cases where the model correctly identified positive instances.
2. **False Positives (FP):** These are the instances that were incorrectly predicted as positive by the model when they were actually negative. In other words, they are the cases where the model made a positive prediction, but it was incorrect.

Precision focuses on the quality of the positive predictions made by a model. It quantifies the ability of the model to avoid making false positive predictions. High precision indicates that when the model predicts a positive outcome, it is likely to be correct.

### F1score

The F1 score [52] is a widely used metric in machine learning and statistics, especially in binary classification tasks. It combines both precision and recall into a single measure and is particularly useful when dealing with imbalanced datasets where one class significantly outnumbers the other. The F1 score is defined as the harmonic mean of precision and recall and is given by the following formula:

$$\text{F1 score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

**Recall (Sensitivity or True Positive Rate):** Recall measures the model's ability to correctly identify all relevant instances in the dataset. It's calculated as the ratio of true positives to the total number of actual positive instances, or more formally:

$$\text{Precision} = \frac{TP}{TP + FN} \tag{4}$$

### 5.5.2 Evaluation of the proposed model

In the previous study of El Bouhissi et al. [7], it was demonstrated that the Random Forest classifier algorithm outperformed the other approaches. Nevertheless, to enhance the Random Forest model's performance, we opted to employ a swarm intelligence algorithm. Specifically, we selected the Particle Swarm Optimizer, as elaborated upon in the preceding chapter. Combining Random Forest with Particle Swarm Optimization (PSO) can offer several benefits in various applications, particularly in machine learning and optimization tasks. Here are some of the advantages:

- Improved Model Performance: PSO can optimize the hyperparameters of the Random Forest algorithm, such as the number of trees, tree depth, and the number of features considered at each split. This optimization process can lead to a Random Forest model that performs better in terms of accuracy and generalization.
- Feature Selection: PSO can be used to select the most relevant features for the Random Forest model. This can result in a more compact and interpretable model that retains the most important information while reducing noise and overfitting.
- Reduced Overfitting: By fine-tuning the hyperparameters using PSO, you can prevent overfitting, which is a common issue in machine learning. This ensures that the Random Forest model generalizes well to new, unseen data.
- Faster Model Training: Optimizing the Random Forest hyperparameters with PSO can lead to a more efficient model, reducing training time while maintaining or even improving performance. This can be especially beneficial when working with large datasets.
- Automated Hyperparameter Tuning: PSO automates the process of hyperparameter tuning, making it easier for machine learning practitioners to find optimal settings for Random Forest without the need for extensive manual experimentation.

To evaluate the effectiveness of our RF-PSO (Random Forest combined with Particle Swarm Optimization) model, we conducted a thorough analysis using the Jupyter environment. Our evaluation process involved using the Frankfurt Hospital Diabetes Dataset, which we divided into two distinct sets: one consisting of 80% of the data for training and the other containing 20% for testing.

The evaluation consisted of two main phases:

1. **Initial Prediction with Random Forest (RF):** In the first phase, we employed the Random Forest algorithm to make predictions on the dataset independently. Random Forest is a powerful machine learning algorithm known for its robustness and accuracy in various applications.
2. **Enhancement with Particle Swarm Optimization (PSO):** In the second phase, we took the predictive capabilities of Random Forest a step further by implementing Particle Swarm Optimization (PSO). PSO is a swarm intelligence optimization algorithm that can optimize the hyperparameters of the Random Forest model to achieve improved performance.

We then carefully compared the outcomes of these two approaches: RF alone and RF with the addition of PSO. The objective was to demonstrate and quantify the superiority of the RF-PSO combination over using Random Forest on its own.

To measure the performance of these models, we chose three widely used evaluation metrics which we defined above, those metrics are Accuracy, Precision and F1score.

The results of our analysis, including the accuracy, precision, and F1-score, are presented in the table 5.1. This table allows for a clear and concise comparison between the performance of Random Forest on its own and the enhanced RF-PSO model.

| Approach | Accuracy % | Precision % | F1score % |
|---|---|---|---|
| Random Forest | 97 | 96 | 96 |
| Random Forest + Particle Swarm Optimizer | 99 | 98 | 98 |

Table 5.1: Comparison between RF and RF+PSO

Upon a thorough examination of the table provided earlier, we draw the following conclusion: the integration of the Particle Swarm Optimization (PSO) algorithm into our model development process has led to a notable enhancement in the performance of the Random Forest (RF) algorithm. This enhancement is reflected in the results, which exhibit a 0.2% improvement. While this 0.2% improvement might seem marginal at first glance, it carries substantial implications for our primary goal, which is the accurate prediction of diabetes cases. In the context of medical diagnosis, even small increments in prediction accuracy can be of immense significance. A 0.2% increase in predictive performance can translate into more precise early identification of individuals at risk of gestational diabetes.

This enhanced accuracy can potentially result in earlier interventions and more effective management of diabetes during pregnancy, ultimately leading to improved maternal and the health of the developing baby outcomes. Therefore, while the numerical difference may appear modest, it wields a profound impact on the real-world application of our model, making it a valuable and impactful tool for healthcare professionals in their efforts to enhance prenatal care.

### 5.5.3 Prediction using the RF-PSO model

Figure 5.9 serves as an illustrative representation of the model's function and the outcomes it produces through its execution. In this scenario, we engaged the model with a specific dataset pertaining to a patient's medical case, where the individual was known to be diabetic. In the code illustrated here, x_final represents the best selected features by PSO and rf_classifier_final is the RF model which has been optimized by PSO. The remarkable aspect of this portrayal lies in the model's ability to make an accurate prediction, corroborating the patient's actual condition of being diabetic. This alignment between the model's prediction and the real-world diagnosis exemplifies the model's efficacy in healthcare applications, where it plays a pivotal role in assisting medical professionals in making critical decisions and enhancing patient care through data-driven insights.

```python
# Fonction interactive pour la prédiction
def predict_diabetes():
    print("Entrez les détails du patient pour la prédiction du diabète :")
    pregnancies = int(input("Nombre de grossesses : "))
    glucose = float(input("Niveau de glucose : "))
    blood_pressure = float(input("Pression artérielle : "))
    skin_thickness = float(input("Épaisseur du pli cutané : "))
    insulin = float(input("Insuline : "))
    bmi = float(input("Indice de masse corporelle (BMI) : "))
    age = int(input("Âge : "))
    DiabetesPedigreeFunction = float(input("DiabetesPedigreeFunction : "))
    # Créez un tableau NumPy avec les données du patient
    patient_data = X_final

    # Utilisez le modèle RF+PSO pour prédire le diabète
    prediction = rf_classifier_final.predict(patient_data)

    if prediction[0] == 1:
        print("Le patient est diabétique.")
    else:
        print("Le patient n'est pas diabétique.")

# Appelez la fonction de prédiction
predict_diabetes()
```

```
Entrez les détails du patient pour la prédiction du diabète :
Nombre de grossesses : 0
Niveau de glucose : 135
Pression artérielle : 68
Épaisseur du pli cutané : 42
Insuline : 250
Indice de masse corporelle (BMI) : 42.3
Âge : 24
DiabetesPedigreeFunction : 0.365
Le patient est diabétique.
```

Figure 5.9: Prediction of diabetes with RF-PSO

## 5.6 Conclusion

In this chapter, we have conducted an extensive review of the existing literature pertaining to the diagnosis of diabetes. Additionally, we have devised a novel hybrid model that combines a swarm intelligence technique with a machine-learning algorithm, achieving the highest level of accuracy recorded thus far.

In this particular study, our methodology involved employing Particle Swarm Optimization (PSO) for feature selection and utilizing the Random Forest (RF) algorithm for data classification. Our primary goal was to attain the highest level of accuracy in the early prediction of diabetes. The collaborative utilization of RF and PSO has demonstrated its potential to effectively address a wide spectrum of machine-learning challenges.

To evaluate the performance of our model we exploited a dataset with various patient characteristics, such as glucose levels, blood pressure, BMI, and age, to make accurate predictions about the likelihood of gestational diabetes in pregnant individuals.

We developed a software tool in Python to facilitate the execution of experiments conducted under two distinct scenarios. These scenarios were designed to evaluate the Enhanced Harmony Search Optimization (EHO) algorithm's impact on prediction speed, quality, and accuracy. In the first scenario, we exclusively relied on the RF classifier for diabetes prediction, while in the second scenario, we leveraged the RF-PSO model. Our findings consistently demonstrated that the RF-PSO model outperformed the RF-only approach, resulting 0.2 % increase in clas-

sification accuracy when compared to alternative algorithms. The integration of RF and PSO not only enhances the predictive accuracy but also aids in identifying the most influential factors contributing to the condition. This achievement holds great promise for early detection and intervention, ultimately improving the healthcare outcomes for pregnant women at risk of gestational diabetes.

Based on these highly promising findings, we assert that our approach possesses the potential for broader applications in the diagnosis of diverse diseases across various domains. The integration of machine learning within the realm of medical diagnostics is poised to become increasingly indispensable in the future, particularly in the context of disease prediction, where it can significantly expedite the process of diagnosis and patient triage.

In the upcoming chapter, we will delve into our ultimate conclusions drawn from this project and outline our prospective endeavours for the future.

# Chapter 6

# General conclusion

This dissertation serves as an extensive investigation into the crucial field of forecasting gestational diabetes. Our main focus in this research revolves around creating a fresh and improved method for anticipating the chances of expectant mothers developing this condition. Considering the significant impact of gestational diabetes on both maternal health and the well-being of the foetus, the importance of precise and early prediction cannot be overstated. Our overarching objective is to contribute to the enhancement of gestational diabetes prediction, a goal that necessitates a profound investigation into various methodologies and techniques within this domain. While numerous strategies have been explored, we embarked on a journey to uncover a pioneering approach that could outperform existing methods. In our quest to achieve this ambitious aim, our research canvassed a wide spectrum of approaches, with a particular emphasis on machine learning and deep learning techniques. The selection of these methodologies was underpinned by their proven efficacy and widespread application in diverse fields, including medical research and healthcare. By harnessing the power of data-driven algorithms, we aspired to develop a model capable of accurately identifying pregnant women at risk of gestational diabetes, thus facilitating early intervention and improved health outcomes for both mother and child. After a meticulous review of the available options, we made a deliberate choice to employ a machine learning model as the cornerstone of our approach. In doing so, we aligned with the consensus of the scientific community, which has consistently demonstrated the effectiveness of machine learning algorithms in the prediction of medical conditions. Among these algorithms, the Random Forest classifier emerged as the optimal choice for our research. Renowned for its robustness and remarkable predictive accuracy, the Random Forest classifier has established itself as a potent tool in the domain of diabetes prediction, making it an ideal candidate for our study. However, we were not content with merely adopting a well-established technique. Our research also sought to push the boundaries of performance within the framework of the Random Forest classifier. To this end, we delved into the realm of swarm intelligence, a collective problem-solving approach inspired by the behaviour of social insects like bees and ants. Swarm intelligence comprises a plethora of algorithms designed to optimize the outcomes of machine learning techniques. In our research, we honed in on the Particle Swarm Optimization (PSO) algorithm as the linchpin of our optimization strategy. PSO, with its capacity to select and prioritize the most influential features for prediction accuracy, held the potential to elevate the performance of the Random Forest classifier to unprecedented heights. Furthermore, we have developed a straightforward web application that comprises two distinct graphical user interfaces (GUIs), tailored primarily for medical professionals, with a specific focus on specialists in gynaecology and diabetology. Within this application, we have seamlessly integrated our robust prediction model. Its primary purpose is to empower healthcare practitioners, particularly doctors, to harness their patients' personal information for the precise prediction of gestational diabetes. This marks a significant advancement in the field, enhancing the accuracy and effi-

ciency of diagnostic processes for these specialists. The structure of our dissertation unfolds in a logical progression of chapters, each serving a distinct purpose:

**Introduction and motivation:** The inaugural chapter lays the foundation for our research. It articulates our motivations, delineates the objectives that steer our inquiry, expounds upon our chosen research methodology, and previews the components that comprise our dissertation.

**Fundamental concepts:** The second chapter serves as an intellectual scaffold for our study. Here, we meticulously define and elucidate the key concepts central to our research. These encompass the various classifications of diabetes, the prevalence of diabetes within the Algerian context, an in-depth exploration of machine learning and its various iterations, an introduction to swarm intelligence, and a brief survey of popular swarm intelligence algorithms.

**State of art:** The third chapter represents a pivotal juncture in our research journey. Here, we embark on a comprehensive survey of the existing body of research dedicated to diabetes prediction. This entails a meticulous analysis and synthesis of prior studies. To facilitate comparative analysis, we distil the essence of these studies into a comprehensive table, encompassing crucial elements such as the study's title, authors, datasets employed, proposed methodologies, resultant model performance metrics, and the perceived advantages of each approach. Subsequently, we engage in a nuanced discussion and critique of these diverse studies, distilling valuable insights that inform the development of our innovative prediction approach.

**Contributions:** In the fourth chapter, we pivot from the review of existing research to the exposition of our own research methodology. We delineate the intricate steps involved in the preparation of our dataset, elucidate the inner workings of the Random Forest classifier, and delve into the intricacies of the Particle Swarm Optimization algorithm. Crucially, we elucidate our strategy for the seamless integration of these two methodologies to amplify predictive accuracy.

**Implementation and Evaluation:** The fifth chapter represents the practical culmination of our research efforts. Here, we detail the nuts and bolts of our approach's implementation, with a specific focus on the utilization of the Python programming language as our implementation medium. Furthermore, we subject our model to rigorous evaluation, leveraging a suite of performance metrics to ascertain its effectiveness in predicting gestational diabetes.

**General conclusion and perspectives:** The concluding chapter serves as the capstone of our dissertation. Within its pages, we encapsulate the essence of our research, summarizing our achievements and their implications. Furthermore, we cast a discerning gaze towards the future, delineating potential avenues for further research and innovation within the realm of gestational diabetes prediction.

It is with a sense of disappointment that we report our endeavours to enhance the RF (Random Forest) algorithm have not met the lofty objectives we had set for ourselves. The incorporation of the Particle Swarm Optimization (PSO) technique, while holding promise, has yielded only a modest improvement, amounting to a mere 0.2% enhancement in the performance of the RF model. This outcome, while not insignificant, falls short of our initial expectations. Our team has invested considerable time and resources into fine-tuning the RF algorithm, exploring various avenues for optimization, and rigorously testing different approaches. Despite our collective dedication and diligence, the results have not reflected the magnitude of improvement we had

aimed for. In conclusion, our journey to improve the RF algorithm and deliver an interactive application has been marked by challenges and less-than-ideal outcomes, but we are resolute in our determination to overcome these obstacles and ultimately provide a more robust and accessible predictive solution to our users. As we chart our course into the future, our overarching goals extend to a multifaceted approach that encompasses an exhaustive exploration of advanced prediction techniques within the expansive domains of machine learning and deep learning. We are dedicated to the rigorous investigation of state-of-the-art methodologies, continually pushing the boundaries of what's achievable in predictive analytics. In tandem with our pursuit of cutting-edge predictive paradigms, we are committed to unravelling the intricacies of optimization algorithms, including but not limited to the Grey Wolf Optimization (GWO) and Elephant Herd Optimization (EHO) algorithms. Our objective is to harness the potential of these algorithms, leveraging their unique strengths to enhance the predictive accuracy and efficiency of our systems. Additionally, we have aspirations to refine and enhance our application, with the ultimate goal of making it accessible to a broader audience by launching it on the Android platform. Our commitment to ongoing improvement remains unwavering as we strive to provide an even more comprehensive and versatile tool for healthcare professionals.

# Bibliography

[1] A. Rocha, S. I. Lopes, and C. Abreu. A cost-effective infrared thermographic system for diabetic foot screening. In *18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 106–111, Thessaloniki, Greece, 2022.

[2] D. J. Magliano. IDF DIABETES ATLAS. NCBI Bookshelf.

[3] M. Lounis. Assessment of diabetes related-knowledge and practice among algerian university students: a cross-sectional study. *Research Square*, 2023.

[4] M. S. Rahman, K. S. Hossain, S. Das, S. Kundu, E. O. Adegoke, M. A. Rahman, M. A. Hannan, M. J. Uddin, and M. Pang. Role of insulin in health and disease: An update. *International Journal of Molecular Sciences*, 22(12):6403, 2021.

[5] Kindred Hospitals. Pathophysiology of diabetes mellitus, November 07 2013.

[6] P. Rorsman and F. M. Ashcroft. Pancreatic $\beta$-cell electrical activity and insulin secretion: Of mice and men. *Physiological Reviews*, 98(1):117–214, 2018.

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

[8] D. J. Klinke. Extent of beta cell destruction is important but insufficient to predict the onset of type 1 diabetes mellitus. *PLoS ONE*, 3(1):e1374, 2008.

[9] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Type 2 diabetes, 2023. Accessed on: March 4, 2023.

[10] M. A. Rais, A. K. Awad, S. Swed, H. T. Ali, and R. Kashyap. An alarming trend concerning diabetes mellitus in algeria. *International Journal of Surgery*, 106, 2022.

[11] B. Mahesh. Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, 9(1), 2020.

[12] Analytics vidhya | learn everything about data science, artificial intelligence and web 3.0. `https://www.analyticsvidhya.com`. Accessed 15-07-2023.

[13] X. Yang. *Optimization algorithms*, pages 13–31. 2014.

[14] A. Kumar. Convex optimization explained: Concepts  examples. Accessed on 03/07/2023.

[15] B. A. S. Emambocus, M. B. Jasser, and A. Amphawan. A survey on the optimization of artificial neural networks using swarm intelligence algorithms. *IEEE Access*, 11:1280–1294, 2023.

[16] nesta. Collective intelligence grants 1.0. Accessed on 10/09/2023.

[17] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948, Perth, WA, Australia, 1995.

[18] A. Banerjee, D. Singh, S. Sahana, and I. Nath. Impacts of metaheuristic and swarm intelligence approach in optimization. In *Cognitive Big Data Intelligence with a Metaheuristic Approach*, pages 77–99. 2022.

[19] J. Li, H. Lei, A. H. Alavi, and G. G. Wang. Elephant herding optimization: Variants, hybrids, and applications. *Mathematics*, 8(9):14–15, 2020.

[20] S. Mirjalili, S. M. Mirjalili, and A. Lewis. Grey wolf optimizer. *Advances in Engineering Software*, 69:46–61, March 2014.

[21] M. Dorigo, M. Birattari, and T. Stützle. Ant colony optimization: Artificial ants as a computational intelligence technique. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[22] Q. Al-Tashi, H. Rais, and S. J. Abdulkadir. Hybrid swarm intelligence algorithms with ensemble machine learning for medical diagnosis. In *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, pages 1–6, 2018.

[23] Ai techniques ant colony optimization (aco) algorithm to solve numerical optimization problem. `https://transpireonline.blog/`. Accessed 16-08-2023.

[24] P. Rahimloo and A. Jafarian. Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them. *Bulletin de la Société Royale des Sciences de Liège*, 85:1148–1164, 2016.

[25] M. Alehegn and R. Joshi. Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology (IRJET)*, 4(10):10, 2017.

[26] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang. Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9:515, 2018.

[27] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 2019.

[28] A. Mujumdar and V. Vaidehi. Diabetes prediction using machine learning algorithms. *Elsevier*, 165(c):292–299, 2019.

[29] M. W. Nadeem, H. G. Goh, V. Ponnusamy, I. Andonovic, M. A. Khan, and M. Hussain. A fusion-based machine learning approach for the prediction of the onset of diabetes. *Healthcare (Basel, Switzerland)*, 9(10):1393, 2021.

[30] M. Z. Al Yousef, A. F. Yasky, R. Al Shammary, and M. S. Ferwana. Early prediction of diabetes by applying data mining techniques: A retrospective cohort study. *Medicine*, 101(29):29588, 2022.

[31] M. O. Edeh, O. I. Khalaf, C. A. Tavera, S. Tayeb, S. Ghouali, G. M. Abdulsahib, N. E. Richard-Nnabu, and A. Louni. A classification algorithm-based hybrid diabetes prediction model. *Frontiers in Public Health*, 10, 2022.

[32] Y. Qin, J. Wu, W. Xiao, K. Wang, A. Huang, B. Liu, J. Yu, C. Li, F. Yu, and Z. Ren. Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, 19(22):15027, 2022.

[33] B. S. Ahamed, M. S. Arya, and A. O. Nancy. Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques. *Frontiers in Computer Science*, 4, 2022.

[34] S. Kale, P. Rahane, M. Ghumare, and S. Patil. Diabetes prediction using different machine learning approaches. *International Journal of Scientific Development and Research (IJSDR)*, 7(5):531–534, May 2022.

[35] Frankfurt hospital. Retrieved July 2023.

[36] Diabetics data, khelil amran hospital, bejaia, algeria. Collected July 2023.

[37] P. Bhandari. How to find outliers | 4 ways with examples & explanation. IEEE Xplore, November 30 2021. Revised on June 21, 2023.

[38] A. Kumar, S. Pant, M. Ram, and S. Singh. On solving complex reliability optimization problem using multi-objective particle swarm optimization. In *Elsevier eBooks*, pages 115–131. 2017.

[39] C. C. Olisah, L. N. Smith, and M. L. Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220:106773, 2022.

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[41] Numpy. `https://numpy.org/`. Accessed 14-09-2023.

[42] Pandas documentation version 2.1.0. `https://pandas.pydata.org/docs/`. Accessed 14-09-2023.

[43] Seaborn : statistical data visualization. `https://seaborn.pydata.org/`. Accessed 14-09-2023.

[44] scikit-learn: machine learning in python x2014 scikit-learn 1.3.0 documentation. `https://scikit-learn.org`. Accessed 14-09-2023.

[45] Python plotting matplotlib 3.4.3 documentation. `https://matplotlib.org`. Accessed 14-09-2023.

[46] Pyswarms documentation. `https://pyswarms.readthedocs.io/en/latest/`. Accessed 14-09-2023.

[47] Jupyter. : https://jupyter.org/. Accessed 10-09-2023.

[48] Meet django. `https://www.djangoproject.com/`. Accessed 11-09-2023.

[49] Mdn web docs-guides, css. `https://developer.mozilla.org/`. Accessed 11-09-2023.

[50] Hostinger tutorials, what is html. `https://www.hostinger.com/tutorials/what-is-html`. Accessed 11-09-2023.

[51] Techtarget, boostrap. `https://www.techtarget.com/whatis/definition/bootstrap`. Accessed 11-09-2023.

[52] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

# *Abstract*

Diabetes is a persistent medical condition that arises from a malfunction in the pancreas, leading to elevated blood sugar levels and potentially affecting various bodily functions. Over time, this condition can inflict damage upon the heart, blood vessels, eyes, kidneys, nerves, and other vital organs. To mitigate these complications, it is imperative to develop a reliable diagnostic system that can identify diabetic patients based on their medical information. In the pursuit of this goal, various machine learning algorithms have been explored for the prediction of diabetes. These algorithms play a crucial role in early disease detection and the prevention of associated health issues. Building upon our previous research focused on predicting gestational diabetes using the Random Forest algorithm, the current study takes a step further. Here, we employ a swarm intelligence strategy to discern the optimal set of features for training the Random Forest algorithm, with the overarching aim of enhancing its predictive performance. The efficacy of this proposed approach was rigorously assessed, and the results yielded promising insights. Notably, combining the Random Forest algorithm with Particle Swarm Optimization led to a marked improvement with an accuracy of 99%. This innovative fusion of algorithms showcases significant potential for advancing the field of diabetes diagnosis and risk assessment.


**Keywords :** Diabetes, Prediction, Machine learning, Optimization, Random Forest, Particle Swarm Optimization, Dataset.

# *Résumé*

Le diabète est une maladie persistante qui résulte d'un dysfonctionnement du pancréas, entraînant une élévation du taux de sucre dans le sang et pouvant affecter diverses fonctions corporelles. Avec le temps, cette maladie peut endommager le cœur, les vaisseaux sanguins, les yeux, les reins, les nerfs et d'autres organes vitaux. Pour atténuer ces complications, il est impératif de mettre au point un système de diagnostic fiable capable d'identifier les patients diabétiques sur la base de leurs informations médicales. Dans la poursuite de cet objectif, divers

algorithmes d'apprentissage automatique ont été explorés pour la prédiction du diabète. Ces algorithmes jouent un rôle crucial dans la détection précoce de la maladie et la prévention des problèmes de santé associés. S'appuyant sur nos recherches antérieures axées sur la prédiction du diabète gestationnel à l'aide de l'algorithme de forêt d'arbres aléatoires, la présente étude va plus loin. Nous utilisons ici une stratégie d'intelligence en essaim pour discerner l'ensemble optimal de caractéristiques pour l'entraînement de l'algorithme de la forêt aléatoire, dans le but principal d'améliorer ses performances prédictives. L'efficacité de l'approche proposée a été

rigoureusement évaluée et les résultats ont donné des indications prometteuses. Notamment, la combinaison de l'algorithme de forêt d'arbres aléatoires et de l'optimisation par essaims particulaires a permis une nette amélioration, avec une précision de 99%. Cette fusion innovante d'algorithmes présente un potentiel significatif pour faire progresser le domaine du diagnostic du diabète et de l'évaluation des risques.

**Mots clés :** Diabète, Prédiction, Apprentissage automatique, Optimisation, Forêt d'arbres aléatoires, Optimisation par essaims particulaires, Ensemble de données.