



UNIVERSITÀ POLITECNICA DELLE MARCHE  
FACOLTÀ DI INGEGNERIA

---

Master's Degree in Biomedical Engineering

**DUAL APPROACH FOR THE CLASSIFICATION  
OF TYPE 2 DIABETES USING CONTINUOUS  
GLUCOSE MONITORING DATA**

Advisor:

Dott. Micaela Morettini

Candidate:

Elisea Creato

Co-Advisors:

Prof. Laura Burattini

Dott. Andrea Tura

Academic year 2019 – 2020



# ABSTRACT

With the worrying increase of diabetes cases all over the world and the importance of early diagnosis for a successful management of the disease, the use of machine learning models is emerging as a powerful tool to support the diagnostic process in a reliable and efficient manner. Continuous Glucose Monitoring (CGM) data represents a rich and relatively unexplored source of information that it is worth exploiting for the classification of the metabolic state of healthy and diabetic subjects.

This thesis proposes a dual approach for the classification of subjects as type 2 diabetic or normoglycemic, based on CGM data.

The first approach used CGM recordings spanning across multiple days, belonging to a population of 57 subjects, of which 38 normoglycemic and 14 prediabetic/diabetic. From an initial set of 13 features extracted from CGM data, five candidate sets of predictors of different sizes were outputted by a wrapper type feature selection algorithm based on the maximization of the classification sensitivity. Among these sets, the one with 6 predictors was finally selected and PCA was then applied to it. A logistic regression model, fed with the selected features has been validated using 5-fold cross validation, and the model so obtained was called  $M_{6+PCA}$ . This model showed mean accuracy and sensitivity of  $0.81\pm 0.10$  and  $0.85\pm 0.18$  respectively, with a mean Area Under the Receiver Operating Characteristic Curve equal to 0.88.

The second approach used CGM data recorded during a standardized meal test of 3 h, that is the postprandial glycemic response. From each of the 151 CGM recordings, of which 105 belonged to healthy subjects and 46 to prediabetic/diabetic subjects, 19 features were extracted. Feature selection was carried out in two steps, namely: a filter type feature selection (first step) based on the scatterplot matrix and point biserial correlation and a wrapper type feature selection (second step) based on the deviance of the fit. A logistic regression model, fed either with the features selected after the filter type or after the wrapper type feature selection, has been validated using a 5-fold cross-validation. The resulting classification models, indicated as  $M_{FT}$  (14 features) and  $M_{FT+WT}$  (9 features), showed mean accuracy of  $0.82\pm 0.10$  vs.  $0.84\pm 0.05$ , and mean sensitivity of  $0.82\pm 0.13$  vs.  $0.85\pm 0.13$ , with Area Under the Receiver Operating Characteristic Curve respectively equal to 0.94 and 0.93.

In conclusion, both approaches represent a suitable solution to discriminate healthy subjects from prediabetic/diabetic ones using CGM data.

# INDEX

<b>Introduction.....</b>	<b>I</b>
<b>1. Glucose metabolism.....</b>	<b>1</b>
1.1 Anatomy and physiology of the pancreas .....	1
1.2 Blood glucose regulation.....	2
<b>2. Diabetes mellitus.....</b>	<b>6</b>
2.1 Type 1 diabetes.....	6
2.2 Type 2 diabetes.....	7
2.2.1 Prediabetes .....	8
2.3 Other types of diabetes .....	9
2.4 Clinical manifestation and diagnosis.....	10
2.4.1 Long-term complications of diabetes.....	13
2.5 Treatment and prevention.....	14
<b>3. Continuous Glucose Monitoring .....</b>	<b>16</b>
3.1 Interstitial glucose and blood glucose .....	18
<b>4. Experimental premises.....</b>	<b>19</b>
4.1 Clinical decision support systems .....	19
4.2 Classification.....	22
4.2.1 Generalization is the key.....	23
4.2.2 The role of features in ML models .....	24
4.2.3 Curse of dimensionality and overfitting .....	26
4.2.4 Performance metrics .....	30
4.2.5 The three components of learning.....	33
4.2.6 Logistic regression .....	34
<b>5. Daily CGM classification .....</b>	<b>37</b>
5.1 Methods.....	37

5.1.1	Dataset.....	37
5.1.2	Classification problem .....	40
5.1.3	Feature extraction and selection.....	40
5.1.4	Model validation and performance measures .....	49
5.2	Results and discussion.....	49
<b>6.</b>	<b>Postprandial glycemc response classification .....</b>	<b>56</b>
6.1	Methods.....	56
6.1.1	Dataset.....	56
6.1.2	Classification problem .....	59
6.1.3	Feature extraction and selection.....	59
6.1.4	Model validation and performance measures .....	61
6.2	Results and discussion.....	62
	<b>Conclusion .....</b>	<b>III</b>
	<b>Bibliography .....</b>	<b>V</b>

## INTRODUCTION

Diabetes is one of the most widespread diseases of these times, with approximately 463 million people in the world who currently suffer from one form of this disease. The World Health Organization estimated that every year diabetes causes about 1.6 million deaths worldwide, while other 3.7 million people die due to complications of diabetes and high blood glucose. The prevalence of diabetes is rising at an alarming rate, especially in low- and middle-income countries, with a projected increase to 700 million by 2045.

The most common form of diabetes, whose proportion is increasing in most countries, is type 2 diabetes, which is often associated to unhealthy lifestyle and diet, physical inactivity and obesity. To contrast the risk of developing type 2 diabetes, besides changes to lifestyle and diet, it is very important to keep blood glucose levels under control by regular check-ups. Indeed, the International Diabetes Federation estimated that about 50% of people with diabetes are not diagnosed, and untreated diabetes may lead to serious health complications. For this reason, prevention and early detection of type 2 diabetes plays a key role in the management of the disease.

The long-established method to diagnose diabetes relies on single time-point measurements derived from blood tests, which however represent the overall metabolic state of the subject.

In the last decades, a new technology based on frequent sampling of glucose is emerging as a promising tool for an optimized and easier management of diabetes, especially for type 1, namely Continuous Glucose Monitoring (CGM) systems. These systems consist in a sensing part placed subcutaneously (needle), which reads interstitial glucose values usually every 5 minutes and wirelessly sends them to either a device that stores data for visualization and analysis or to an insulin pump that delivers insulin based on glucose readings.

The use of these devices for research and diagnostic purposes has recently been brought to attention for its enormous unexplored potential. Indeed, the information carried inside the glucose dynamics can reveal important insights of the metabolic state of the subject which would not be possible to extrapolate from sporadic sampling of blood glucose.

A powerful tool to discover hidden patterns inside data is offered by machine learning algorithms, whose unique inductive capabilities can tackle problems in several domains, including healthcare. When the mathematical models of machine learning are used to make predictions, classify, recognize patterns and in general help the diagnostic process, we can talk

about clinical decision support systems. The first step in the development of a proper decision support system is to find models for a given type of data, capable of achieving accurate and reliable results.

The purpose of this thesis is to use machine learning to recognize whether a continuous glucose recording belongs to a diabetic subject or to a normoglycemic one, that is classification; to tackle this problem, two approaches have been taken: the first one was based on the classification of daily CGM recordings, that are recordings spanning across multiple days, while the second one was based on the classification of CGM recordings during a standardized meal test, that is the postprandial glycemc response.

# 1. GLUCOSE METABOLISM

In humans and other complex organisms many regulatory processes occur simultaneously and continuously in order to keep the internal environment as stable as possible. This tendency to maintain a relatively constant internal environment is called homeostasis (from the Greek *homeo*, similar + *stasis*, state of standing) and it is absolutely crucial to an organism's survival. Homeostatic regulation occurs at many levels as a response to potentially dangerous changes in the external or internal environment, and they usually involve negative feedback control mechanisms which tend to minimize those changes and restore the internal balance. [1]

Homeostatic regulation of the blood glucose concentration is just one of the numerous control mechanisms existent in human body, such as the thermoregulation, or the control of blood pressure and pH, and it will be the focus of this chapter.

Glucose is the primary energy source for human body, and it is central for the physiologic brain functioning. The high energy demand of neurons requires a continuous delivery of glucose from the blood (about 5.6 mg of glucose per 100 g of brain tissue per minute), therefore a tight regulation of blood glucose levels is crucial for the individual's survival [2].

In physiologic conditions, basal plasma glucose concentration ranges between 70 and 110 mg/dL, with an average of 80 to 90 mg/dL [1, 3, 4].

The most important hormones involved in blood glucose control are insulin and glucagon and they are both secreted by the pancreas [3, 4].

## 1.1 Anatomy and physiology of the pancreas

The pancreas is a large gland (about 15 cm long) located between the inferior part of the stomach and the proximal portion of the small intestine (duodenum). Its head lies within the curvature of the duodenum and the body extends laterally toward the spleen, where it ends with a narrow and rounded tail (Figure 1). The pancreas is mainly made up of exocrine cells as they represent 99% of its volume. These cells, called *pancreatic acini*, secrete about 1 liter of digestive enzymes per day into the duodenum through a network of pancreatic ducts, making them the main actors of digestion in the small intestine. [1]

The endocrine cells of the pancreas, which are grouped in clusters known as *pancreatic islets* or *islets of Langerhans* scattered around exocrine cells, secrete hormones into the bloodstream that are vital for our survival. Despite constituting only about 1% of the pancreatic volume,



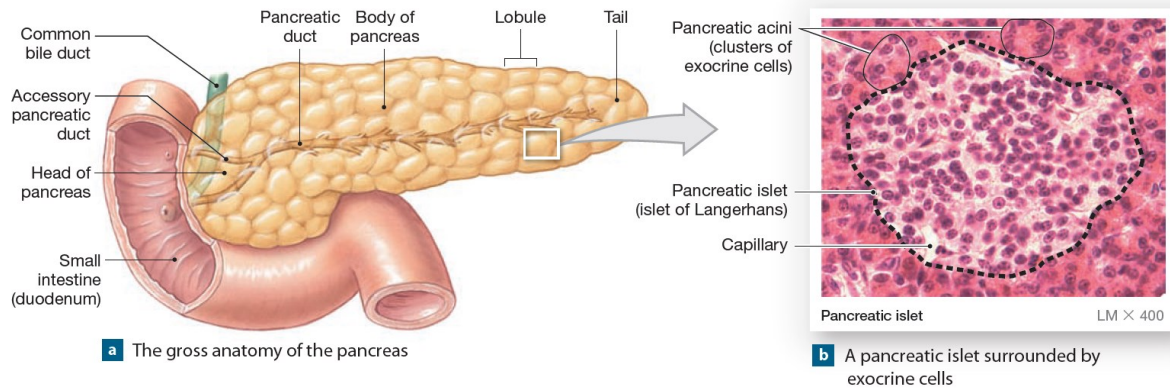


FIGURE 1 – SCHEMATIC REPRESENTATION OF THE GROSS ANATOMY OF THE PANCREAS WITH FOCUS ON PANCREATIC ISLET [1]

there are approximately 2 million of islets in a healthy pancreas, and each islet contains on average 2500 cells of four types [1]:

- Alpha ( $\alpha$ ) cells which secrete *glucagon*, a hormone responsible for blood glucose level increase [1].
- Beta ( $\beta$ ) cells which produce *insulin*, a hormone that lowers blood glucose level [1].
- Delta ( $\delta$ ) cells which produce *somatostatin*, a hormone which inhibits insulin and glucagon secretion and slows nutrient absorption and enzyme secretion rates [1].
- Pancreatic polypeptide cells (PP cells) which secrete the hormone pancreatic polypeptide (PP) that inhibits gallbladder contraction and regulates the production of pancreatic enzymes [1].

## 1.2 Blood glucose regulation

The maintenance of blood glucose level within the normal range is ensured by the combined secretion of insulin and glucagon (Figure 2). The action of these regulating hormones are opposite: insulin is secreted by pancreatic  $\beta$  cells when blood glucose concentration increases (hyperglycemia), and it promotes the uptake and use of glucose by target cells, thus lowering blood glucose concentration; glucagon is secreted by pancreatic  $\alpha$  cell in response to a low blood glucose concentration (hypoglycemia), and it stimulates the mobilization of energy reserves in order to restore an adequate blood glucose concentration. [1]

In physiologic condition, the counter-regulatory hormones keep blood glucose levels inside the normality range after large inputs associated with food intake and following large usage rates

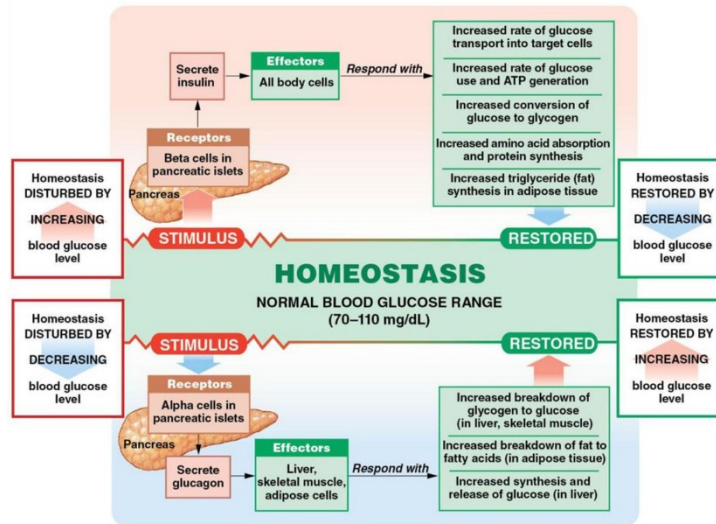


FIGURE 2 – HOMEOSTATIC REGULATION OF THE BLOOD GLUCOSE CONCENTRATION [1]

related to physical exercise. This is possible thanks to the ability of human body to store excess glucose in different forms, and to use it between meals, during exercises or prolonged fasting [5].

Most of the energy reserve (75%) is stored as triglycerides in adipose tissue, which constitutes up to 30% of body weight in healthy individuals and it can reach 80% in obese ones. Fat is a very efficient energy storage, but very little is synthesized from glucose (10 – 12 g per day) [3]. Proteins account for almost 25% of the energy storage, but they are used as a source of energy just in extreme cases of starvation [3]. Carbohydrates are stored in form of a glucose polymer called glycogen, and although representing less than 1% of the total energy reserve, it is essential for the CNS metabolism and to support intense muscle work [3]. About 25% of the glycogen store is localized in the liver and it can be broken down through glycogenolysis and made available to other tissues in form of glucose [3]. The remainder glycogen is stored in muscle fibers and it can't be used anywhere else because muscular tissue is not provided of glucose-6-phosphatase, an enzyme necessary for glucose release into the bloodstream [3]. Glycogen can be synthesized from glucose, galactose and fructose introduced through the diet, or de novo in the liver and to a lesser extent in the kidneys, through a process called gluconeogenesis, which derives glucose from non-carbohydrate precursors [3, 5].

When a food intake causes blood glucose level to rise, insulin secretion increases while glucagon secretion decreases. The increase of plasma insulin concentration above basal promotes glucose uptake and use by insulin dependent cells, whose plasma membrane contain

insulin receptors. When insulin reaches the plasma membrane of a target cell through the bloodstream, it binds to its receptor inducing a cascade of reactions that causes the mobilization of glucose transporters to the plasma membrane, where they facilitate glucose entry from the bloodstream to the cytoplasm of the target cell [1, 3].

Once glucose is inside the cell, a part of it is used to produce ATP and sustain the cell functioning, while the rest is stored in form of glycogen and triglycerides to be used when blood is poor of glucose [1]. Glycogen, obtained through gluconeogenesis, is mainly stored in the liver, while triglycerides are obtained by lipogenesis and stored in adipose tissue.

This is what happens to *insulin dependent* cells, which are the majority of the cells of our body. There are other cells defined *insulin independent*, which do not have insulin receptors on their plasma membranes and they can accept and use glucose without insulin mediation. Insulin independent cells are those in the brain, in the kidneys, in the inner layer of the digestive tract and red blood cells [1]. As insulin enhances glucose uptake and use, the level of glucose in the blood returns in the normal range and consequently insulin secretion rate is restored to its basal value [5].

Between meals or during physical exercise, when glucose concentration falls below basal, insulin secretion decreases while glucagon secretion increases, mobilizing energy reserves. When glucagon is released into the bloodstream by pancreatic  $\alpha$  cells, its main action is on the liver, where it binds to a hepatic plasma membrane glycoprotein receptor, triggering a cascade of reactions that stimulates glycogen breakdown (glycogenolysis). The glucose released by glycogenolysis is then used by skeletal muscle fibers or poured into the bloodstream by liver cells. After 12 to 15 hours of fasting or during intense physical work, hepatic glycogen reserves are depleted (in fact glycogen is a short-term energy storage), and gluconeogenesis becomes the source of blood glucose. Gluconeogenesis mainly occurs in the liver, where hepatic cells synthesize glucose from pyruvate, lactate, glycerol and glucogenic amino acids which were released into the bloodstream by adipose tissue and muscles. Finally, the product of hepatic glucose production (gluconeogenesis) is released into the circulation. Gluconeogenesis is also supported by the breakdown of triglycerides (lipolysis) that releases fatty acids into the bloodstream for use by other tissues [1, 3, 5].

The combined action of insulin and glucagon during glucose deficiency results in the reduction of glucose use and the release of more glucose into the bloodstream (endogenous production).

When glucose levels return within the normal range, also insulin and glucagon basal secretion rates are restored [1].

These two antagonist hormones, insulin and glucagon, are secreted without endocrine or nervous stimuli because  $\alpha$  and  $\beta$  cells are highly sensitive to changes in blood glucose concentration and can directly respond to them by enhancing or reducing hormonal secretions. However, insulin production is also influenced by autonomic activity: the brain has insulin receptors in multiple regions, including the hypothalamus, cerebellum, and hippocampus, and through sympathetic and parasympathetic innervations of the pancreas, it enhances or inhibits insulin secretion [1, 4].

## 2. DIABETES MELLITUS

The previous chapter described how blood glucose homeostasis is maintained in a healthy individual through a number of regulating mechanisms involving various hormones and organs. However, for several reasons, the regulation of blood glucose level might fail leading to a pathological condition known as *diabetes mellitus* or simply *diabetes*.

Diabetes is a chronic metabolic condition characterized by hyperglycemia due to impaired insulin production or usage. In 2019, the worldwide prevalence of diabetes was estimated to be 463 million people (9.3% of the population), and it is expected to increase to 578 million (10.2%) by 2030 and to 700 million (10.9%) by 2045 [6]. The higher prevalence observed in urban areas and high-income countries (10.8% and 10.4% respectively) compared with rural areas and low-income countries (7.2% and 4.0% respectively) [6] suggests a positive correlation of the pathology with the lifestyle [7].

Besides lifestyle, there are multiple factors that can trigger diabetes onset, and according to the cause, we can classify the disease in different types, and the most common ones are type 1 and type 2 diabetes.

### 2.1 Type 1 diabetes

Type 1 diabetes (T1D), once known as juvenile or insulin-dependent diabetes, represents about 5-10% of all cases [5, 8], and it usually develops in children and young individuals [1].

In most cases, T1D is caused by the autoimmune destruction of pancreatic beta cells, which usually leads to absolute insulin deficiency [8]. The preclinical phase of the disease is distinguished by the presence of autoimmune markers such as islet cell autoantibodies or insulin autoantibodies, which however are not thought to directly cause the disease [8, 9]. The destruction of  $\beta$ -cells, which may occur at different rates (usually faster in infants and children and slower in adults), is related to genetic predispositions and environmental factors, still not fully clear [8]. The most prominent genetic factor points back to the human leukocyte antigen (HLA); however, at least other 47 non-HLA genetic factors are associated with the risk of T1D [9]. The main putative environmental factors include maternal factors (e.g. gestational infections, high maternal age), viral infections (e.g. mumps virus or rotavirus), dietary factors (e.g. bovine milk or short breastfeeding, cereals, vitamin D deficiency), high birth weight and growth rate, psychologic stress (e.g. stress during pregnancy) and toxic substances (e.g. alloxan, streptozocin, vacor) [9].

A small fraction of T1D patients exhibits a permanent reduction in insulin secretion due to  $\beta$ -cell function loss but without evidence of  $\beta$ -cell autoimmunity. This form of T1D is defined idiopathic, and it is highly inheritable and not associated with HLA. [9]

Whether autoimmune or idiopathic type 1 diabetes, the onset of the disease is characterized by the loss of pancreatic  $\beta$ -cell secretory function. The absolute lack of insulin production inhibits glucose uptake by insulin-dependent tissues and the body reacts to cell starvation by augmenting glucagon secretion [9]. Increased levels of glucagon stimulate hepatic glucose production through glycogenolysis and gluconeogenesis which in turn causes glucose intolerance and finally hyperglycemia [3]. Furthermore, the state of fasting caused by the absence of insulin triggers the lipolysis of adipose tissue into free fatty acids, which are converted in the liver into ketone bodies through beta oxidation [9]. Ketone bodies represent an alternative source of energy for human body, and beside increasing glycemia, they turn the blood acidic (Diabetic Ketoacidosis or DKA) [10].

Diabetes ketoacidosis mainly occurs in T1D patients, but it may also appear in some patients with type 2 diabetes. It is a potentially life-threatening condition, usually accompanied by dehydration, deep and sighing respirations, sweet-smelling fetor on the breath caused by the ketone bodies, clouded consciousness, abdominal pain and vomiting [9]. Without medical intervention it can lead to coma and death [3, 9].

## **2.2 Type 2 diabetes**

Type 2 diabetes (T2D), also known as non-insulin-dependent diabetes, represents about 90-95% of all cases [5, 8] and it is usually associated with obesity, sedentary lifestyle and elderly age; however, with the increasing incidence of the disease, it is now less uncommon in adolescent individuals [9].

During the first stages of the disease, the production of insulin is not altered, but there is an impairment in insulin-mediated blood glucose uptake by target tissues. The reduced responsiveness of tissues to insulin is commonly referred to as *insulin resistance*, a metabolic syndrome that precedes the disease onset and whose causes are mainly attributed to obesity, physical inactivity and genetic predisposition [5, 9, 11, 12]. Insulin resistance is a synonym for *low insulin sensitivity*.

At first, in order to compensate for the lowered insulin sensitivity of tissues, the pancreas increases insulin secretion (hyperinsulinemia) enabling blood glucose to enter the cells and

glycemia to stay within the normal range [11, 12]. Over time, the pancreas loses the ability to keep up with the augmented demand for insulin as its  $\beta$  cells get smaller (hypotrophy) and eventually die (hypoplasia) [11]. Dysfunctionality of  $\beta$  cells soon leads to decreased insulin production and overt hyperglycemia, hence type 2 diabetes onset [9, 11]. It is interesting to remark that insulin resistance is furtherly worsened by chronic hyperglycemia, which itself causes insulin resistance [9]; for this reason, insulin resistance is present throughout the progression from preclinical phase to overt T2D [12]. However, insulin resistance doesn't always progress into diabetes because in some subjects (about two third of obese, insulin-resistant individuals)  $\beta$  cells are able to compensate for the decreased insulin effectiveness [12].

Unlike T1D, in T2D a small amount of insulin is generally still produced by the pancreas so the insulin-glucagon balance is maintained and diabetic ketoacidosis is usually not developed [11]. Nevertheless, a condition typically involved with T2D is Hyperosmolar Hyperglycemic State (HHS): in case of hyperglycemia, because glucose molecules can't passively diffuse across cell membrane, the water inside the cells diffuses by osmosis into the bloodstream inducing an increase in urination and finally causing total body dehydration [11].

The development of insulin resistance and eventually chronic hyperglycemia are often associated with an inflammatory state, whose main cause is attributed to visceral obesity [9].

High energetic diets, especially rich in glucose and free fatty acids, induce stress in pancreatic islets and insulin-dependent tissues such as adipose tissue, causing the local production and release of cytokines and chemokines, two molecules involved with the immune system. As a consequence, inflammation is induced in adipose tissue and elsewhere such as in the liver and in the islets, leading to insulin resistance [9, 12].

### **2.2.1 Prediabetes**

Type 2 diabetes is a chronic disease which can only be treated through drugs and changes in lifestyle and diet, but it can't be reversed. However, during the first stages of the disease and when glucose levels are higher than normal, but still not so high to be diagnosed as diabetic, we are in a potentially reversible metabolic condition known as *prediabetes*.

In 2017, the number of people in US with prediabetes was estimated to 84 million, and without intervention, 37% to 70% of them were expected to develop diabetes within the next 4 years [7]. Lifestyle interventions aiming to minimize the probability of the progression of prediabetes

in diabetes include weight loss, physical exercise and personalized diet. However, diagnosing prediabetes might not be straightforward, especially when the subject shows normal glycemia.

## 2.3 Other types of diabetes

There are other cases in which individuals experience hyperglycemia but they don't fit in neither type 1 or type 2 diabetes category. In these cases, subjects experience specific types of diabetes, caused by certain circumstances or substances.

- **Gestational diabetes**

Pregnant women may experience hyperglycemia during the second or third trimester of pregnancy [8, 11] and although it may remit after delivery, it indicates a high risk for future type 2 diabetes outbreak [9]. The risks involved with gestational diabetes include fetal death or prematurity, congenital malformations, neonatal hypoglycemia, jaundice, macrosomia and increased risk of developing childhood obesity [9].

Although ultimately unknown, the exact cause of gestational diabetes is thought to be related to hormones secreted during pregnancy that could interfere on insulin receptor hence on insulin action [11].

- **Drug-induced diabetes**

Many commonly used drugs interfere with glucose homeostasis, causing or worsening pre-existing hyperglycemia. Drugs that induce hyperglycemia in healthy patients usually (but not always) lead to reversible and not insulin-dependent diabetes, while for diabetic patients it is fundamental to investigate the diabetogenic properties of drugs to prevent deterioration in glycemic control. The underlying mechanisms through which drugs effect glucose homeostasis are of two types: reduction of insulin biosynthesis or secretion, or reduction of tissue sensitivity to insulin [9].

Among the most commonly prescribed drugs, glucocorticoids have by far the greatest impact on glucose homeostasis, exacerbating hyperglycemia in T2D patients and, given in high doses, causing significant increases in blood glucose concentrations in normoglycemic individuals; however, hyperglycemia is generally reversible once glucocorticoids is quit, regardless the duration of drug treatment [9].



Oral contraceptive pills, especially those with high estrogen content or progesterone levonorgestrel, have been proven to cause hyperglycemia with a probability of developing impaired glucose tolerance of 35% or even greater in women with a history of gestational diabetes. In contrast to the older high-estrogen pills, the current low-dosage contraceptive pills show only a minor risk of development of T2D; moreover, hyperglycemia induced by hormonal contraceptives is usually reversible on withdrawal of the pill [9].

Other examples of chemicals affecting glucose homeostasis include thiazide diuretics,  $\beta$  - adrenoceptor antagonists and agonist, HIV protease inhibitors, antirejection drugs and antidepressants [9].

- **Monogenic diabetes**

Monogenic diabetes has an incidence of 1-2% and it is due to inheritance of one or more mutations in a single gene that cause  $\beta$ -cell dysfunction or, rarely, insulin resistance. It is often initially misdiagnosed as type 1 or type 2 diabetes. [9]

- **Endocrine disorders and pancreatic disease related diabetes**

Endocrine disorders may cause diabetes when hormones that are antagonist to insulin (inhibiting insulin secretion and/or action) are excessively secreted. This is the case, for instance, of acromegaly, Cushing syndrome, polycystic ovarian syndrome or glucagonoma and somatostatinoma (rare islet cell tumors) [9].

Pancreatic disease account for less than 0.5% of all cases of diabetes, and examples of them include: acute and chronic pancreatitis, tropical calcific pancreatitis, pancreatic carcinoma, cystic fibrosis [9].

## **2.4 Clinical manifestation and diagnosis**

In both type 1 and type 2 diabetes, a variety of genetic and environmental factors can lead to a progressive loss of  $\beta$ -cell mass and/or function, that ultimately manifests as hyperglycemia. Regardless the type of diabetes, once hyperglycemia arises, patients are alike to develop the same chronic complications, although rates of progression may differ [8].

Clinical symptoms of uncontrolled diabetes include *polyphagia*, *glycosuria*, *polyuria* and *polydipsia*. Both type 1 and type 2 diabetic patients show high blood glucose concentrations due to insulin deficiency or resistance, and because glucose can't enter insulin-dependent cells

without insulin, these cells starve for energy; in response, adipose and muscle tissues break down to provide nutrients to the cells but causing weight loss and sense of hunger (polyphagia). When blood glucose concentration is too high, kidneys fail to reabsorb excess glucose so it gets excreted through urine (glycosuria). At the same time, since glucose is osmotically active, water tends to dilute urine resulting in increased urination (polyuria), which leads to dehydration, hence increased sense of thirst (polydipsia) [11].

Diabetic ketoacidosis is a serious complication of type 1 diabetes, occasionally present also in type 2 diabetes, while hyperosmolar hyperglycemic syndrome is a dangerous condition that can occur only in type 2 diabetes; both DKA and HHS have been described in paragraphs 2.1 and 2.2 respectively. Other non-specific symptoms include tiredness, general malaise, blurred vision and repeated or persistent skin infections. [9]

Unfortunately, in many cases diabetes may remain asymptomatic for years and at the time of diagnosis, the patient is likely to present long-term complication of diabetes. It is estimated that about 50.1% of people living with diabetes are not aware of their condition [6]; for this reason, screening tests are very important, especially for individuals at high-risk of diabetes [9].

There are several ways to diagnose diabetes, and they usually rely on blood tests. The most common tests to detect diabetes are based on glycated hemoglobin (HbA1c or A1c), Fasting Plasma Glucose (FPG), plasma glucose 2 hours after a 75-g oral glucose tolerance test (2-h PG after OGTT) and random (or casual) plasma glucose test.

- **Glycated hemoglobin**

Hemoglobin, the protein within red blood cells which transports oxygen throughout the body, becomes *glycated* when glucose chemically attaches to it, hence the name *glycated hemoglobin* [13]. Most monosaccharides, including glucose, spontaneously bind to hemoglobin through glycation; normal blood glucose concentrations generate a normal amount of glycated hemoglobin, and as blood glucose levels rise, the fraction of hemoglobin combined with glucose increases proportionally [14]. Since the average lifespan of red blood cells is about 4 months, the amount of glycated hemoglobin found in the blood reflects the average glycemia over the past 2 to 3 months prior to the test [13, 14].

- **Fasting Plasma Glucose**

Fasting plasma glucose (FPG) test measures the level of glucose in the blood after a fasting

of at least 8 hours, hence it is generally performed before breakfast [15].

- **Oral Glucose Tolerance Test**

During an Oral Glucose Tolerance Test (OGTT), a standard dose of 75g of glucose diluted in water is orally ingested and the glucose level is sampled at specific time intervals. The level of plasma glucose 2h after the OGTT is used as a diagnostic metric for diabetes.

- **Random plasma glucose test**

A random plasma glucose test is taken to measure blood glucose level at any time of the day, for instance in case of severe diabetes symptoms.

With respect to FPG and 2hPG after OGTT, HbA1C test is more convenient as it doesn't require fasting and it is representative of the average blood glucose levels in the past 2-3 months rather than being a single-time-point measure [7, 8, 9]. However, HbA1C test has a greater cost, and it has limitations on the age, ethnicity and hemoglobinopathies of the tested person [8].

The American Diabetes Association (ADA) provides the reference values of HbA1c, FPG and 2h-PG after OGTT to diagnose diabetes and prediabetes (Table 1) [15].

Unless there are clear clinical manifestation of diabetes, for instance with classic symptoms of hyperglycemia or random plasma glucose greater than 200 mg/dL, the same test is repeated for two consecutive times, on different blood samples [8]. The diagnosis of diabetes is confirmed when two different tests or the same test repeated twice, show both positive results [8].

TABLE 1 - ADA GUIDELINES FOR DIABETES DIAGNOSIS

	<i>HbA1c (%)</i>	<i>FBG (mg/dL)</i>	<i>2-h PG after OGTT (mg/dL)</i>
<i>Healthy</i>	< 5.7	< 100	< 140
<i>Prediabetic</i>	≥ 5.7 and < 6.5	≥ 100 and < 126	≥140 and ≤ 199
<i>Diabetic</i>	≥ 6.5	≥ 126	≥ 200

### **2.4.1 Long-term complications of diabetes**

Untreated diabetes leads to a series of severe health problems that ultimately result in death. Diabetes management (see paragraph 2.5) focuses on reducing the risk of long-term complications through glycemic control and screening [9].

Chronic side effects of diabetes are generally divided in microvascular (diabetic nephropathy, neuropathy, and retinopathy) and macrovascular (coronary artery disease, peripheral arterial disease, and stroke) subgroups [9]. The risk of developing vascular complications related to diabetes increases in individuals exposed to longer periods of hyperglycemia [9]. Besides hyperglycemia, hypertension is an important risk factor for vascular damages [9].

At the level of small blood vessels, diabetes causes hyaline arteriosclerosis, which makes arteriole hard and inflexible, and it thickens the basement membrane of capillaries, hindering the transfer of oxygen into the cells thus leading to hypoxia [11].

- **Diabetic nephropathy**

Diabetic nephropathy is an important complication of diabetes which involves the kidneys and it's the primary cause of renal failure in the US [9, 16]. The nephron, the functional unit of the kidney, contains a net of capillaries involved with blood filtration; alterations to these small blood vessels, due to uncontrolled glycemia, leads to a series of consequences such as widespread capillary occlusion, microaneurysm formation, mesangial nodule formation and increased intraglomerular pressure that ultimately result in a progressive decrease in glomerular filtration rate (GFR) and eventually in the end-stage renal disease (ESRD) [9, 11, 16]. The hallmarks of diabetic nephropathy is the microalbuminuria (increased excretion of albumin in the urine) that eventually worsen in proteinuria [9, 16].

- **Diabetic neuropathy**

Diabetic neuropathy affects about 50% of diabetic patients and it is characterized by nerves damages in different parts of the body, but most commonly in legs and feet [17]. According to the nerves involved, there are different types of neuropathies: peripheral, autonomic, focal and proximal neuropathies, and more than one type can be present at the same time [17]. Peripheral neuropathy is the most common type and it affects the lower limbs and feet first, followed by the upper limbs and hands; it leads to decreased sensations, numbness, sharp pain and eventually foot problems, ulcers and infections [9, 11, 17].

- **Diabetic retinopathy**

Diabetic retinopathy is probably the most common microvascular complication of diabetes, responsible for about 10,000 new cases of blindness every year in the US [16]. Diabetic retinopathy is caused by damages to the capillaries that nourish the retina, including hemorrhages, microaneurysms and occlusions that eventually cause injuries to the retina with subsequent vision impairment or loss [9].

In large blood vessels, diabetes causes arterial wall damages, hence atherosclerosis [11], which leads to arterial walls narrowing throughout the body [16]. Atherosclerosis is directly linked to cardiovascular disease (CVD), which is the primary cause of death in people with diabetes [16].

## **2.5 Treatment and prevention**

Until now, it is unfortunately impossible to cure diabetes, but it can be effectively treated, especially if diagnosed on time. Diabetes treatment's goal is to reproduce the physiologic insulin profile and keep blood glucose level within the normal range without incurring significant risk of hypoglycemia [9].

People with T1D are completely dependent on exogenous insulin because of their irreversible pancreatic  $\beta$ -cells function loss; the therapy for T1D not only needs to ensure adequate amounts of insulin after each meal, but it also need to provide basal insulin for the background control of glucose between meals [9]. The usual insulin therapy involves fingerstick testing of glycemia and subcutaneous insulin administration usually in the abdomen, thigh and deltoid [9]. In the last decades, the use of continuous subcutaneous insulin pumps and continuous glucose monitoring for T1D management has led to an improved metabolic control and increased lifestyle flexibility [9]. Indeed, these tools allow for a continuous monitoring of blood glucose level and a tight and automated control of insulin delivery, resulting in a closed-loop system known as *artificial pancreas*, which mimics the physiologic pancreatic insulin secretion, hence the name [9]. Continuous Glucose Monitoring is further discussed in Chapter 3.

People with T2D, as their pancreatic  $\beta$ -cells may still be producing some insulin, might not require continuous insulin therapy as T1D patients do; they are usually treated with oral hypoglycemic agents, which improve insulin sensitivity and safeguard  $\beta$ -cells function [18],

along with lifestyle and dietary modifications. However, more than 50% of T2D require insulin injections at some point of their lives [9].

Unfortunately, type 1 diabetes cannot currently be prevented, while there are some preventive actions that can significantly lower the risk of developing type 2 diabetes or at least prevent the side effects associated with any type of diabetes [19]. Preventive interventions include regular physical exercise, healthy and balanced diet, avoiding smoking, controlling blood pressure and adipose tissue, especially in the abdomen and of course screening tests [9, 19].

### 3. CONTINUOUS GLUCOSE MONITORING

Early diagnosis and strict monitoring and treatment of diabetes are the key elements for a successful management of the disease, thus a significant risk reduction of long-term complications. Continuous glucose monitoring (CGM) systems not only dramatically facilitate diabetes monitoring but they also have the potential to be used for diabetes detection [20].

The sensing part of a CGM system is a small sensor placed subcutaneously, usually in the arm or in the belly (Figure 3), which periodically measures the interstitial glucose level (glucose within the fluid surrounding the cells) [18]. The sensor samples glucose level usually every 5 minutes and wirelessly sends it to a monitor or to a smartphone/tablet. Data can also be sent to an integrated system consisting of an insulin pump and a control unit which modulates insulin release based on the data received from the CGM sensor [18]. For this reason, CGM system are particularly used for T1D management.

Monitoring glucose in real time helps the patient (or the automated system) to make more informed decisions about the insulin therapy, taking into account also food intakes and physical activity [18]. Indeed, CGM systems provide additional information about the dynamic of blood glucose level, in contrast with the traditional fingerstick tests which give accurate but isolated measures of glycemia.

Once the sensor is placed, it can be worn for up to 7 days, depending on the model, and it continuously (day and night) records and sends data to a given device, e.g. an alarming system that alerts the patient or a second person when glycemia is too high or too low, or an insulin pump which automatically adjusts insulin bolus [18].

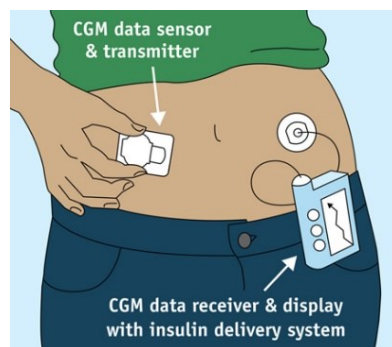


FIGURE 3 – CONTINUOUS GLUCOSE MONITORING SYSTEM ILLUSTRATION. PHOTO COURTESY: U.S. FOOD AND DRUG ADMINISTRATION

Some drawbacks of CGM systems are:

- Lower accuracy of the results with respect to standard blood glucose meters. In fact, it is still necessary to check the validity of CGM glucose readings against finger-stick glucose test twice a day [18].
- Most CGM models can't be used to make treatment decisions alone, which should also rely on a finger-stick glucose test [18].
- Higher price with respect to a standard glucose meter [18].

On the other hand, a number of remarkable advantages are associated with the use of CGM, especially when integrated with an insulin pump:

- Real time monitoring of glycemia with detailed information on glucose patterns and dynamic [9, 20].
- Improved metabolic control and quality of life in patients with type 1 diabetes [9].
- Increased lifestyle flexibility [9].
- Reduced the risk for both hypoglycemia and hyperglycemia, often associated with a decrease in HbA1c [9].
- Promising diagnostic tool [20].

The use of CGM for diagnostic purposes is still developing but it appears to be a promising application. Current methods to diagnose diabetes (or prediabetes) rely on single-time-point measurements or on average measures of glycemia, and they do not consider how blood glucose fluctuates over time. However, multiple studies have demonstrated that information hidden inside glucose dynamic are critical for assessing the metabolic status of a person. Particularly important is the quantification of the glycemic variability, which is thought to be an independent risk factor for diabetes development. Another important risk factor is the postprandial hyperglycemia, which induces oxidative stress, hypercoagulability, endothelial dysfunction, and inflammation, hence it is considered a sign of glucose dysregulation. Through the use of a CGM system, it is possible to spot glycemic variability in individuals otherwise considered nondiabetic by standard measures and, through diet and lifestyle modifications, planning an intervention aimed to minimize glucose variability hence reduce the risk of diabetes onset. Indeed, without intervention, 37% to 70% of prediabetic patients are expected to develop diabetes within 4 years, with increased incidence of diabetes complications. [7, 20]



### 3.1 Interstitial glucose and blood glucose

The main difference between a standard blood glucose (BG) test and a CGM acquisition is the site of glucose sampling: while the first measures the glucose concentration directly in the blood, the latter senses the glucose concentration in the interstitial fluid (ISF), (Figure 4).

It follows from this that BG and ISF glucose differ in a number of aspects: while blood distributes glucose throughout the whole body, the ISF only transfers glucose to cells; BG measurements provide limited insights about the dynamic of glucose fluctuations, while ISF glucose measurements, due to the significantly higher sampling rate, provide information about dynamic changes and patterns which can be deployed to optimize diabetes care; on the other hand, BG measurements are more reliable than ISF ones [21]. Another important difference is that the BG concentration represents an overall measure of the total currently available glucose in the bloodstream, while ISF glucose concentration is a marker of the local conditions in terms of glucose diffusing from the blood to the cells nearby the needle sensor; this means that the sampling volume of BG is significantly larger than ISF one [21].

Despite these differences, most of the time BG and ISF measurements can be used interchangeably for treatment plan; however, the difference between the two measurements increases when the dynamic changes very quickly [21].

For the sake of simplicity, in this work we will not make distinction between BG and ISF glucose measurements.

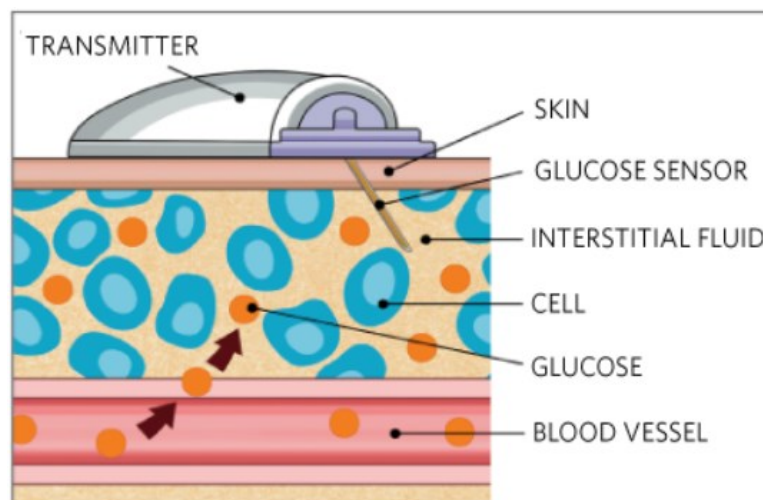


FIGURE 4 - CGM SENSOR

## 4. EXPERIMENTAL PREMISES

Before going into the details of the experimental part of this thesis, it seems appropriate to introduce the reader to the definition of clinical decision support systems and see how these promising tools may assist clinicians in the decision-making process. Then, the theoretical background of classification with some technical and practical aspects of learning will be outlined.

### 4.1 Clinical decision support systems

In the last decades, we have assisted to a massive increase of clinical data availability due to considerable improvements of the sensors' electronics, computational power and storage capacities of machines. Alongside with data availability, remarkable advances in computer science and information technology have allowed the development of computer systems that use medical data to support the clinical decision process. These systems are generally referred to as *clinical decision support systems* (CDSSs), and they are never meant to replace the role of the medical doctor, but rather to help him/her with clinical decision-making tasks, e.g. the diagnosis of a certain disease, the programming of a treatment plan etc..

A main distinction between CDSSs is made between knowledge-based and data-driven systems (Figure 5): in knowledge-based systems, an artificial intelligence (AI) inference engine takes decisions on the basis of the scientific knowledge, usually encoded in the form of rules, while data driven systems use machine learning (ML) and deep learning (DL) models to find hidden patterns inside data.

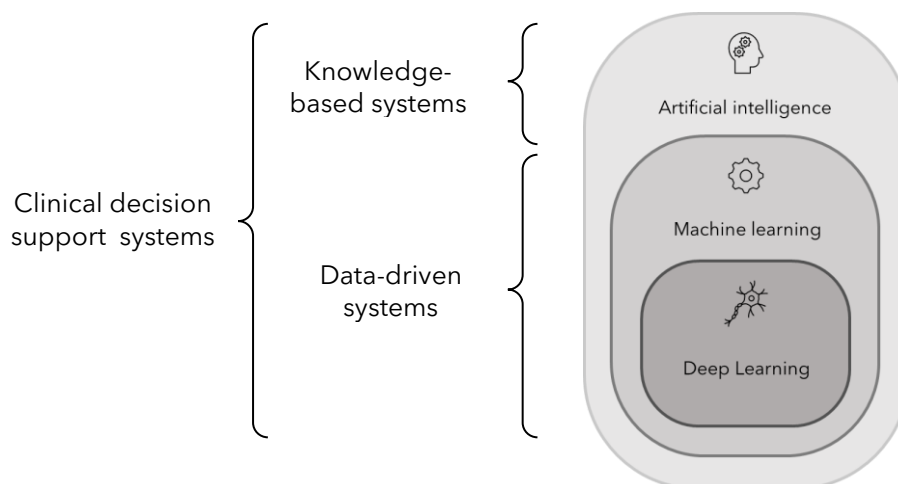


FIGURE 5 - ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, DEEP LEARNING HYERARCHY AND CLINICAL DECISION SUPPORT SYSTEMS TYPES

AI is a field of computer science which aims to build systems capable of mimic human cognitive activities such as learning, reasoning, and perception to perform complex tasks that would otherwise require human intelligence. Although there is a debate about considering ML as a subfield of AI or rather seeing just the “intelligent” part of ML as a subset of AI, for the sake of simplicity we’ll stick to the first theory.

ML uses mathematical models to automatically learn programs from large amount of data, so that once the model has been trained to perform a certain task, it can be used to predict the outcome given unseen input data. Basically, what ML aims to do is to generalize beyond the examples [22], and build models capable of tackling the task for which the systems has been trained, e.g. spam/fraud detection, credit scoring, recommender systems, web search etc. DL is a subfield of ML whose models mimic human brain structure, hence called artificial neural networks, in order to independently learn and make intelligent decisions [23]. With respect to ML, DL models require less human intervention to learn and they can handle larger and unstructured data; in particular, deep neural networks learn the features while training, whereas feature extraction in ML models needs to be done manually (Figure 6) [23]. The peculiarity of DL to deal with large and unstructured data, makes it especially suitable for images, which are given as input to special neural networks called convolutional neural networks (CNN).

In the healthcare domain, examples of knowledge-based systems are: diagnostic systems based on IF-THEN rules, risk assessment and patient monitoring systems based on fuzzy logic and diagnostic and treatment support systems based on Bayesian belief network.

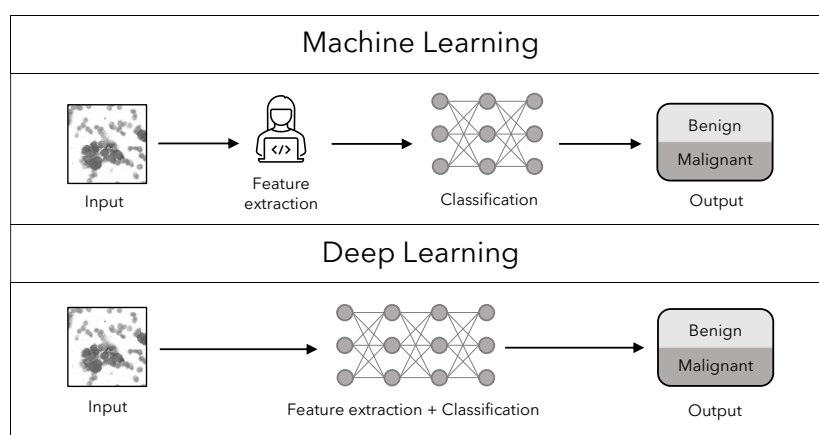


FIGURE 6 - MACHINE LEARNING VS DEEP LEARNING. IMAGE CLASSIFICATION (BREAST CANCER WISCONSIN DATASET)

Healthcare data driven systems include ML models for anomalous pattern recognition from physiological signals, disease diagnosis/prediction, cancer detection from medical images, medical image enhancement etc. In contrast with knowledge based models, which are known as “white box”, data driven systems are also called “black box” (Figure 7) because they provide a mathematical relationship between input and output data, but they often hide the insights of the underlying process, hence lacking in interpretability [24]. This is particularly true for DL, where much of the learning pipeline is automated, so the models are probably very good in terms of accuracy, but difficult to interpret and understand [23].

Especially in domains such life sciences, uninterpretable results are considered unreliable, and most medical doctors do not lean on them for diagnoses, but they rather use them as post-diagnostic tools, to suggest patterns to look into in more depth. [24]

It is worth exploiting the extraordinary capability of data science models to learn patterns from large data, combining it with the solid scientific knowledge; actual systems that try to integrate these two paradigms are known as “grey box” or “theory-guided data science” (TGDS) [24]. However, some data-driven health information systems still allow some degree of freedom for knowledge integration, e.g. one can encode the scientific knowledge dependent on the particular task under study through manual feature extraction (Figure 6).

Regardless the type of CDSS, the ultimate goal of these tools is to improve the quality of care by increasing its efficacy, e.g. earlier detection of abnormalities, and efficiency, e.g. cost and time reduction. Another remarkable advantage of using these systems is that they would enable health democratization by standardizing medical care and by reaching isolated areas through telemedicine. Last but not least, CDSSs provide paradigm shift toward precision medicine hence care customization because complex data analysis can reveal differences among the patients with a given pathology that standard medical approaches might miss [25].

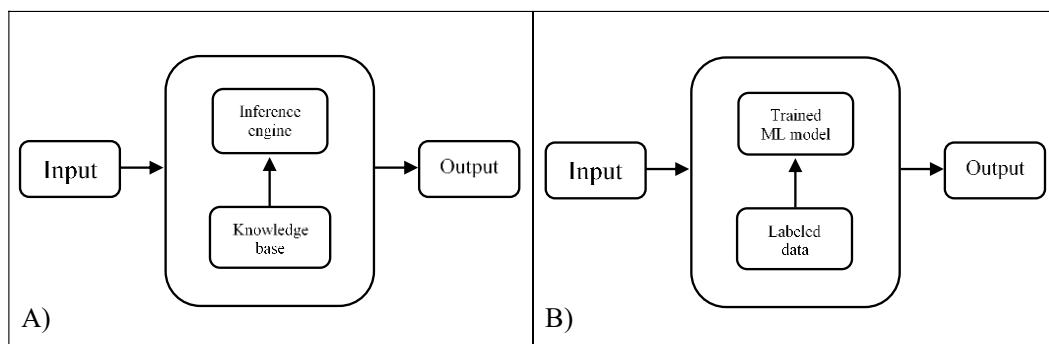


FIGURE 7 - A) KNOWLEDGE-BASED OR WHITE-BOX SYSTEM. B) DATA-DRIVEN OR BLACK-BOX SYSTEM

## 4.2 Classification

Classification is a supervised learning paradigm of ML (Figure 8) used for pattern recognition, which aims to predict qualitative outputs, i.e. descriptive labels rather than numerical values. A classification problem's purpose is to identify to which of a set of categories (or classes) an observation belongs and the choice is based on a training dataset that contains labeled observations, i.e. instances whose category membership is known. An example would be assigning a diagnosis to a given patient based on the values of a predetermined set of characteristics (otherwise called features or predictors), e.g. heart rate, weight, gender, presence/absence of certain symptoms etc.. The algorithm implementing the classification is called *classifier*, and it is usually a mathematical function that maps input data to a category.

The steps of a classification algorithm are schematized in Figure 9. Given a training dataset whose elements are labeled, a set of features is extracted from the data and fed to the algorithm. The algorithm takes as input the values of the features for each element of the dataset with the respective labels (thus supervised learning) and it trains itself to distinguish elements of different classes, based on the values assumed by the features; in other words, it learns how to recognize to which class an element belongs. After the classifier has been trained, when it receives a new unlabeled input, it will extract the same features extracted during the training, and based on what it has learned, it will label the new input with one of the possible classes.

The word “learning” might seem inappropriate for a machine, but these systems are actually miming the human learning process; for example, an infant needs to see many cats and dogs before being able to recognize and distinguish them, and the underlying process that enables him to learn is exactly the same: at first the infant sees many examples of cats and dogs around and, as his parents will tell him which is a dog and which a cat (label), he will start to unconsciously select characteristics from the dogs and cats (feature extraction), that will enable him to distinguish the two species (training).

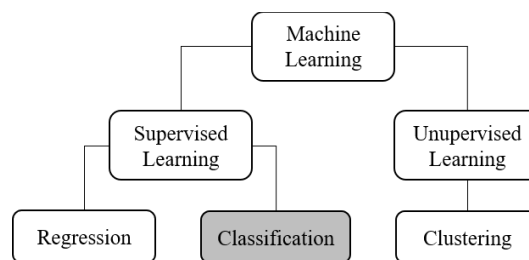


FIGURE 8 - SUPERVISED AND UNSUPERVISED LEARNING PARADIGMS

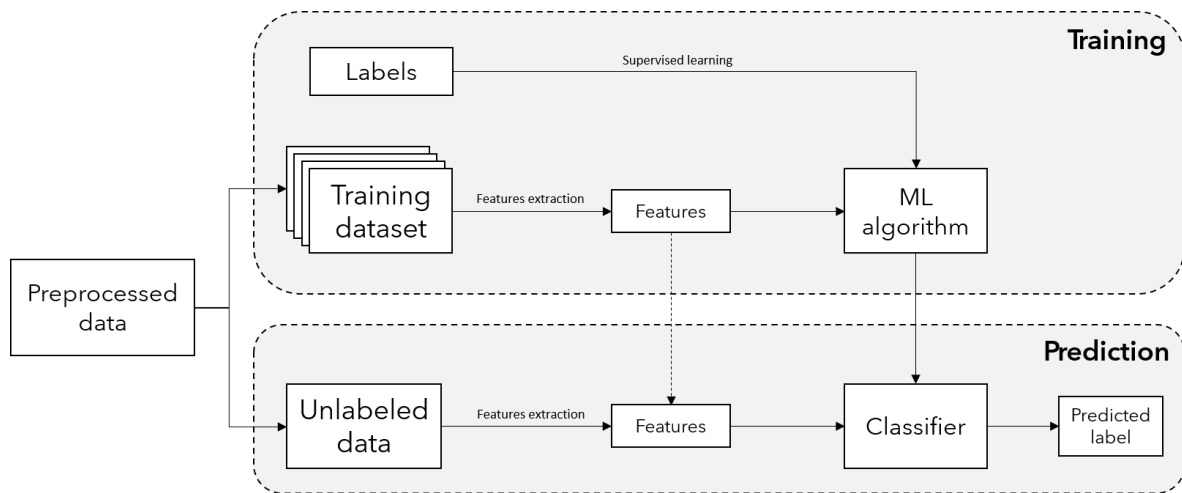


FIGURE 9 - CLASSIFICATION ALGORITHM: TRAINING AND PREDICTION

Once the infant has seen enough labeled examples of these pets, he will be able to autonomously recognize which is a dog and which a cat, based on the same features through which he has learned to distinguish them, e.g. size, shape, coat, sound, behavior, etc. (prediction). Clearly, the processes behind human learning are much more complex and barely known, but this analogy gives us an idea of what a classifier does. What a classifier actually “learns” are parameters or weights associated with features.

The next section will be dedicated to the discussion of some important concepts about ML models in general with focus on classification. The theoretical background in this domain is so wide that an extensive discussion would be out of the scope of this work. For this reason, only the main implications will be described, prevalently using an intuitive approach.

#### 4.2.1 Generalization is the key

The fundamental goal of ML is to generalize beyond the examples of the training set [26], namely inducing the law starting from observations, and this is possible only if we have a consistent amount of data [22].

Returning to the above example, the infant might not be able to correctly classify whether a given animal is a dog or a cat until he has seen enough examples of both species; to put it very simple, if the parents showed their son only 1 grey cat, 1 black kitten, 1 black dog and 1 grey puppy, and assuming the features to be the color and the size, then if the infant sees a large black cat, then his prediction might not be better than random guessing (Figure 10).

Alongside with data abundance, it is very important to have enough examples for all classes, ideally the same amount (balanced dataset). The reason is intuitive: a classification model can't learn properly how to separate two classes if one of them is scarce in the training set, hence poorly represented. In the example of the infant, if he sees plenty of dogs and only few cats during the training phase, he will be surely very good at recognizing dogs, but it is also likely that he will confound a cat for a dog.

Data availability and balanced classes are undoubtedly necessary for generalization; however, data alone is not sufficient [22], but we also need some means to avoid sticking too much on training examples, see paragraph 4.2.3.

## 4.2.2 The role of features in ML models

In standard ML, feature extraction is done manually, hence it is crucial to choose the appropriate features for the classification task under consideration, because the performances of the classifier are highly dependent on them.

Indeed, extracting the “right” features from data means to extrapolate information which are relevant for the purpose of the specific classification task.

Sometimes, the choice of the features might not be straightforward, and some features that one had believed to be relevant may not be so, and irrelevant features in the model may act as confounders and degrade classification accuracy. Moreover, there might be pairs of highly correlated features, which basically carry almost the same information, hence it would be pointless keeping both. For these reasons, the so called *feature selection* should follow feature extraction. To get an idea of what would be an irrelevant feature, we can think about trying to guess the gender of a person by the color of his/her eyes, which is totally irrelevant for gender classification. In addition to the issue of high correlation and irrelevance, reducing the number of features results in a simpler model, which implies both a better interpretability and a lower computational cost.

A further reason to perform feature selection is discussed in section 4.2.3.

There are three categories of feature selection:

- **Filter type feature selection** → with filter type feature selection, the features are ranked on the base of their characteristics and on the relevance to the response variable. This type of feature selection is performed before training the model, hence it is uncorrelated to the training algorithm [27].

- **Wrapper type feature selection** → wrapper type feature selection is based on the classification performances, hence it is carried out during training. From an initial set/subset of features, the algorithm starts to remove/add features based on a certain selection criterion, which measures the change in model performance resulting from removing/adding a feature. The algorithm stops when the stopping criteria is satisfied [27].
- **Embedded type feature selection** → the embedded type feature selection incorporates feature selection in its learning process, hence it learns feature relevance [27].

Going back to the example of the infant, in that case the process of feature extraction is not so simple as in standard ML, but above all, it is autonomous, e.g. the infant will not actively select any features to look for in the animal.

DL models resemble this type of mechanism because the features are directly learned by the neural network and not fed to it as input; this implies the fact that the features learned by the network are optimal for the specific classification task, but on the other hand, they are difficult to interpret (Figure 6).

In domains such as healthcare, the importance of the features is even greater because they can have a physiological meaning which is worth interpreting.

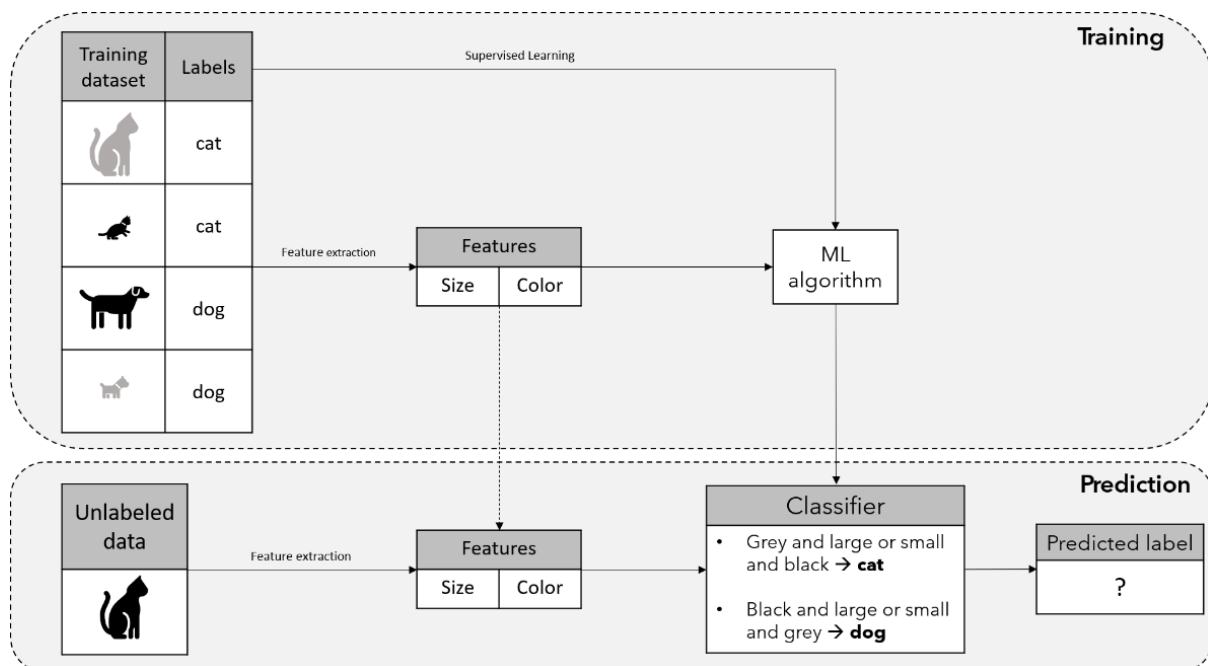


FIGURE 10 – INTUITIVE EXAMPLE OF A CLASSIFICATION WITH A SMALL TRAINING DATASET



### 4.2.3 Curse of dimensionality and overfitting

Another good reason to select features is to avoid the so called *curse of dimensionality*, which often occurs when dealing with high-dimensional data; we might indeed think that by adding more features we can better characterize the different classes within the dataset, but this is not always the case [28]: if we keep increasing the dimensionality without increasing the number of training samples, the classifier's performance will increase until the optimal number of features is reached, and after that point the performance will decrease very fast [28]. In fact, as we increase the dimensionality of the problem, the density of the training samples decreases exponentially, becoming sparser and sparser; due to this sparsity, it becomes very easy to find an hyperplane which separates the classes perfectly. However, the same hyperplane is not likely to be good for separating new unseen data, hence we failed to generalize [22] or, in other words, we *overfitted* the model to the training data [28].

We can better figure out this phenomenon by a visual example (Figure 10):

Let's say that we want to train a linear classifier to separate two classes (red and blue) based on 20 training instances (10 per class); by using only one feature (Figure 10, A), we can't separate the two classes perfectly, so we try to add another feature (Figure 10, B), but it still won't result in a linearly separable classification because it doesn't exist a line that can perfectly separate red from blue. In three-dimensional feature space (Figure 10, C) there exists a plane which perfectly separates the classes, this means that a linear combination of the three features allows to obtain a perfect classification for our 20 samples. This example seems to suggest that we should increase the number of features until we obtain a perfect classification on the training set; however, we can notice that the density of the samples decreases exponentially as the dimensionality of the feature space increases, and when data are sparse we don't have enough information from the data to choose the best hyperplane for that problem, hence we end up having an hyperplane which is perfect for the training set, but probably very bad for another unseen set of data. In this example, we should chose the model with two features and accept some misclassified samples rather than going for the model with 3 features and risking to lose in generalization. Indeed, generalizing correctly becomes exponentially harder as the dimensionality increases [22].

How can we avoid losing in generalization (or overfitting)?

The data available to build a classification model are assumed to be representative of the population from which they have been drawn, and the larger is the dataset, the more it resemble

the real population. This means that by increasing the number of training samples the model can generalize better and avoid overfitting (Figure 14, left). As already discussed in this section, also the number of features, hence the model complexity, plays a role in overfitting (Figure 14, right).

The way to keep track of the generalization error is to use a test set and eventually a validation set (Figure 12). A validation set is simply a small fraction of the original dataset that is used to evaluate the performance of the classifier during training, while we keep adjusting hyperparameters and model specifications, whereas a test set is a small portion of the original dataset used for the final estimation of the model performances once the training is completed. A valid alternative to keep overfitting under control is cross validation (Figure 13): since holding out data reduces the size of the training set, it is possible to randomly divide the original dataset in  $k$  sets (or folds), e.g. 5-fold cross validation, and for  $k$  times validating on one set the model that was trained in the  $k-1$  complementary sets [22]. The performance of the classification are calculated on the validation set for each fold and finally averaged so that we have a general idea of how good/bad the model performs on unseen data.

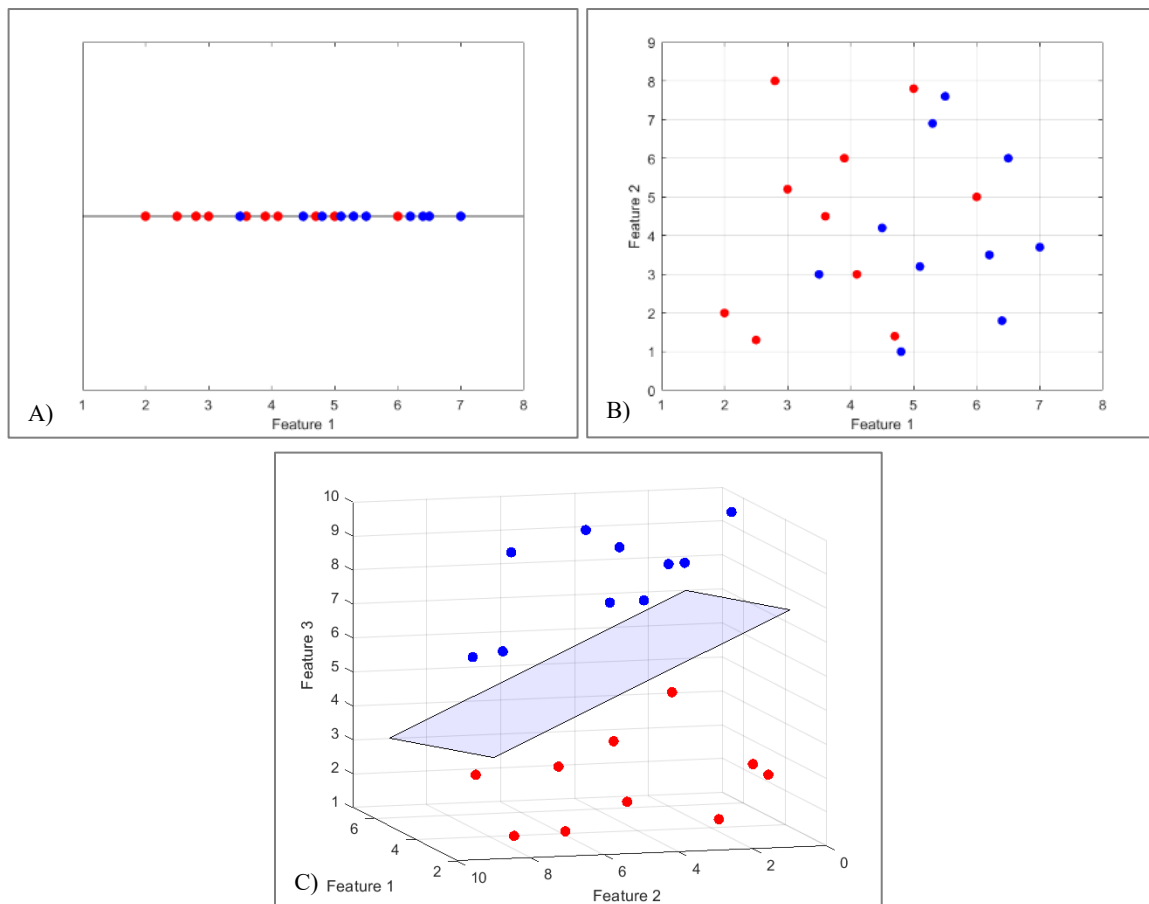


FIGURE 11 - CURSE OF DIMENSIONALITY VISUALLY EXPLAINED

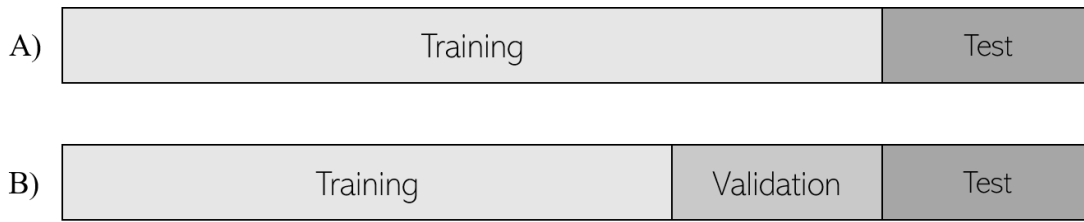


FIGURE 12 - DATASET SPLIT INTO TRAINING, VALIDATION AND TEST SETS

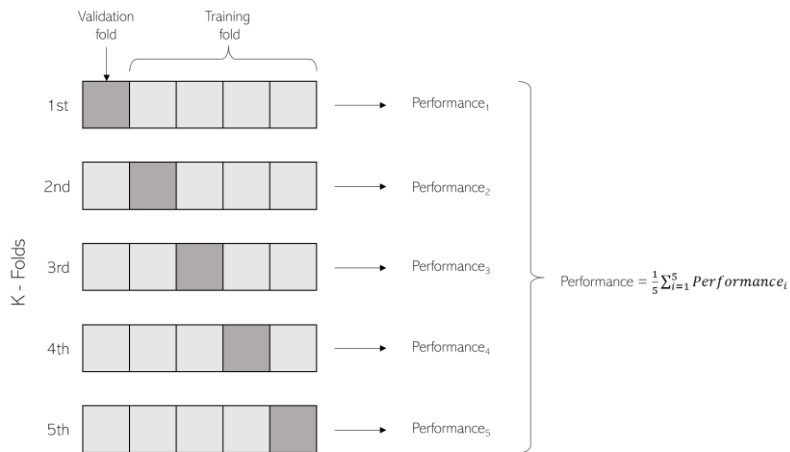


FIGURE 13 - K-FOLD CROSS VALIDATION (K = 5)

The other face of overfitting is underfitting (Figure 14, right) and it occurs when we poorly fit data to the training set, and the performance of the classification are scarce in both the training and test set. For instance, this could be the case of the classification red/blue above, when we only selected 1 feature (Figure 11, A).

Other methods applied to avoid oversampling include regularization and dimensionality reduction. Dimensionality reduction aims to minimize the feature space while maximizing the information content; as opposed to feature selection, this method doesn't just select a subset of features among the extracted ones, but it actually transforms the data from a high-dimensional into a lower-dimensional space [29]. One of the most widely used method for dimensionality reduction is the Principal Component Analysis (PCA), which transforms the features by projecting them onto a lower dimensional space. The idea behind PCA is to find a new basis to re-express the features in such a way that the transformed variables are linear functions of the original ones, that maximize the variance and that are uncorrelated with each other [30].

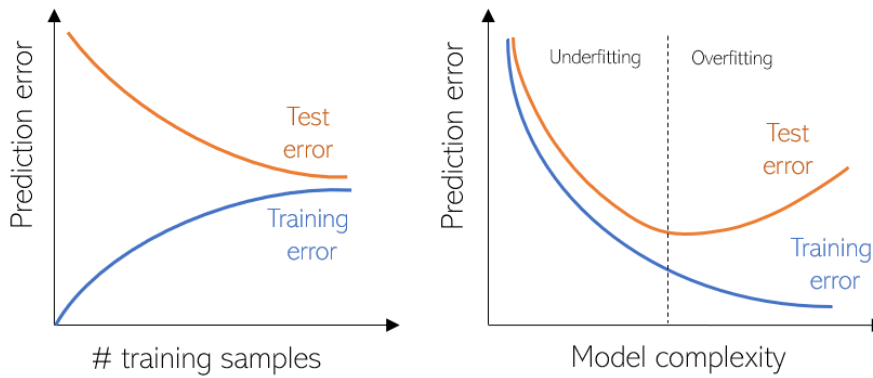


FIGURE 14 - TRAINING VS TEST ERROR WITH VARYING SAMPLE SIZE AND MODEL COMPLEXITY

Figure 15 shows an intuitive graphical representation of the transformation applied by PCA: from the original space of features ( $x_1, x_2$ ), PCA finds a new set of axis ( $z_1, z_2$ ), such that the data are mainly spread along one axis, which is the dimension where the variance of the data is maximum ( $z_1$ ); in this way, the variance along the other axis ( $z_2$ ) is so small that we can make rid of this dimension without losing too much information. In this example, the principal component that we retained would be  $z_1$ . In higher dimensional spaces, we can either decide how many components we want to keep, or keep a number of components such that an arbitrary percentage of the total variance is retained, e.g. 90%.

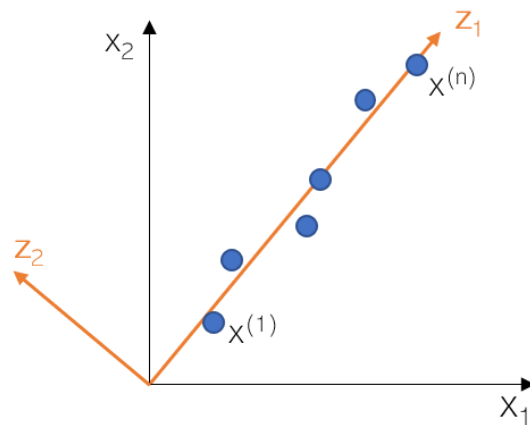


FIGURE 15 – VISUAL REPRESENTATION OF PCA TRANSFORMATION

#### 4.2.4 Performance metrics

Once we trained our model, we need to test its performances on unseen data (the test set). By feeding our test set to the model, we can find the predicted label for each sample, e.g. healthy (or negative) and cancer (or positive), and compare it with the expected true label. The comparison leads to 4 possible outcomes, collected in the *confusion matrix* (Figure 16):

- **True negative (TN)** → the sample is correctly classified as negative, e.g. predicted label = healthy, true label = healthy;
- **True positive (TP)** → the sample is correctly classified as positive, e.g. predicted label = cancer, true label = cancer;
- **False negative (FN)** → the sample is wrongly classified as negative, e.g. predicted label = healthy, true label = cancer;
- **False positive (FP)** → the sample is wrongly classified as positive, e.g. predicted label = cancer, true label = healthy;

		Predicted		Σ
		-	+	
Actual	-	TN	FP	TN+FP
	+	FN	TP	FN+TP
	Σ	TN+FN	FP+TP	1

FIGURE 16 - CONFUSION MATRIX FOR BINARY CLASSIFICATION. "+" IS THE POSITIVE CLASS AND "-" IS THE NEGATIVE CLASS

From the confusion matrix, it is possible to derive a number of statistical measures which indicate how well the classifier correctly identifies or excludes a condition on the test set:

- **Accuracy (ACC)** =  $\frac{TP+TN}{TP+TN+FP+FN}$

It's the ratio between the number of correct classifications among the total number of classified samples.

- **Sensitivity (SE) or True Positive Rate (TPR)** =  $\frac{TP}{TP+FN}$

It's the ratio between the true positives and the total number of actually positive samples. E.g. it's the probability of a positive prediction, given that the subject has the disease.

- **Specificity (SP) or True Negative Rate (TNR)**  $= \frac{TN}{TN+FP}$

It's the ratio between the true negatives and the total number of actually negative samples. E.g. it's the probability of a negative prediction, given that the subject is healthy.

- **Precision or Positive Predictive Value (PPV)**  $= \frac{TP}{TP+FP}$

It's the ratio between the true positives and the total number of samples classified as positive. E.g. it's the probability that a subject with a positive prediction truly has the disease.

- **Negative Predictive Value (NPV)**  $= \frac{TN}{TN+FN}$

It's the ratio between the true negatives and the total number of samples classified as negative. E.g. it's the probability that a subject with a negative prediction is truly healthy.

- **False Positive Rate (FPR)**  $= 1 - TNR = \frac{FP}{FP+TN}$

It's the ratio between the false positives and the total number of actually negative samples.

- **False Negative Rate (FNR)**  $= 1 - TPR = \frac{FN}{FN+TP}$

It's the ratio between the false negatives and the total number of actually positive samples.

- **Balanced accuracy (bACC)**  $= \frac{TPR+TNR}{2}$

Accuracy might not be a suitable metric for unbalanced datasets, i.e. different class sizes. In these cases, we can use balanced accuracy, which is the average between true positives normalized by the number of positive samples (TPR) and true negatives normalized by number of negative samples (TNR).

- **F1 score, F score or F measure ( $F_1$ )**  $= 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$

It's a measure of a test's accuracy and it is calculated as the harmonic mean of precision and sensitivity.

Taking a step back, the predictions of the classifier are usually not labels but probabilities, i.e. probabilistic classification; namely, for each sample, the output vector contains the probabilities of the sample to belong to each class (the sum of these probabilities is always one). In general, the label is assigned by determining the class corresponding to the maximum

value in the output vector of probabilities, that is, in binary classification, setting the decision threshold to 0.5, e.g. if the probability of a given sample to belong to the healthy class is 0.6 ( $>0.5$ ), the sample will be labeled as “healthy”, otherwise as “cancer”.

However, sometimes we might want to investigate how the performance of the classification would change with decision threshold different from 0.5, e.g. if we fix the decision threshold to 0.7, a probability of 0.6 to belong to the healthy class, would lead to a “cancer” label. There are indeed some applications the decision threshold can be choose strategically, to favour a certain classification outcome, e.g. in diagnostic application, a false negative (predicting no disease when the subject is sick) is way worse than a false positive (predicting a disease when the subject is healthy), therefore, we can play with the decision threshold to discourage false negatives.

For this purpose, it is useful to plot the Receiver Operating Characteristic (ROC) Curve, which represents the TPR as function of the FPR, for different values of decision threshold. In fact, we want to find a trade-off between the TPR and FPR, such that changing the threshold of classification will change the balance of predictions towards improving the TPR at the expense of FPR, or vice versa. Figure 17 represents ROC curves of a generic classifier (solid line), a perfect classifier (dash-dotted line) and a random classifier (dashed line). The shaded area is the Area Under the Curve (AUC), and it is a general indicator of the classifier performance in terms of how well the classifier can separate the classes: the closer to 1, the better. Finally, the Equal Error Rate (EER), is the point on the ROC curve corresponding to an equal probability of miss-classifying a positive or negative sample. By minimizing the EER, one can optimize the trade-off between FPR and TPR.

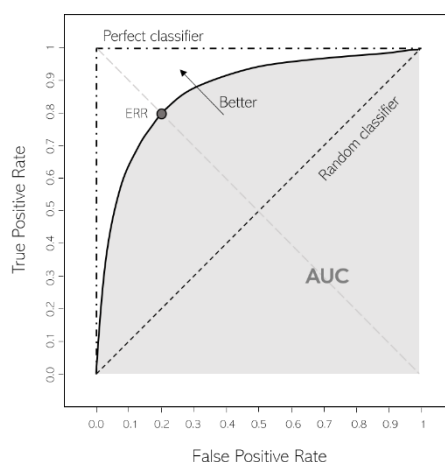


FIGURE 17 - RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

### 4.2.5 The three components of learning

Despite the diversity of application's domains and the impressive variety of algorithms available in machine learning, the fundamental components of learning are always the same three: representation, evaluation and optimization [22].

Let's consider a training set of  $n$  samples  $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$ , where  $x^{(k)}$  is the input variable of the  $k$ -th example and  $y^{(k)}$  is the corresponding output variable, which in the case of binary classification assumes discrete values, e.g.  $y \in \{0, 1\}$  or  $y \in \{-1, 1\}$ . Each example input  $x^{(i)}$  is represented by its feature vector  $\mathbf{x}^{(i)} = (x_0^{(i)}, x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)})^T$ , where  $x_0^{(i)} = 1$  and it's called the bias term. Given these premises, we want to find:

- A **representation** of our data [22], that is the mathematical model used to approximate the unknown relationship between input and output data [26]. This model, called the ***hypothesis function***, actually maps the input variables to the output through a mathematical relationship [31]. The hypothesis function depends on the input variables and on parameters (or weights)  $\mathbf{w} = (w_0, w_1, w_2, \dots, w_m)^T$ , that we want to estimate.
- An **evaluation** function (commonly referred to as cost or loss function), which is the function that the algorithm uses internally to evaluate the classifier [22] and quantify how close (or far) are the predicted outputs to the real ones. Note that the evaluation function used internally by the algorithm may differ from the external one, i.e. performance metrics [31] (see paragraph 4.2.4).
- An **optimization** technique to apply to the loss function in order to improve the performance of the classifier [22]. This corresponds to finding the best parameters  $\mathbf{w}$  that lead to the lowest error between the real output and the predicted one [31].

Common examples of these three components are reported in Figure 18.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

FIGURE 18 - THE COMPONENTS OF ANY LEARNING ALGORITHM [22]



## 4.2.6 Logistic regression

ONE OF THE MOST SIMPLE AND COMMON MODELS FOR CLASSIFICATION IS LOGISTIC REGRESSION, WHICH OWES ITS NAME TO THE *LOGISTIC FUNCTION* (EQ. (4) AND

Figure 19), and in its basic form is used for binary problems. Despite its name, logistic regression is not used for regression, hence it does not predict a continuous outcome, but the category of membership of the dependent variable. However, logistic regression and linear regression are somehow related as the first makes use of the latter to build its prediction.

Linear models are characterized by the linear combination of the input variables (features and weights) in the form  $s = \mathbf{w}^T \cdot \mathbf{x}$ , where  $\mathbf{w}$  is the vector of weights and  $\mathbf{x}$  is the matrix of features [32]. In linear regression (Eq. (1)), this signal is directly used as output, indeed  $s$  belongs to the real numbers and it is unbounded [32]. Linear classification (Eq. (2)), uses a hard threshold on the signal  $s$  to produce  $\pm 1$  as output [32]. Logistic regression (Eq. (3)) is in-between these two cases because it applies the logistic function to the signal  $s$  and it produces an output which smoothly varies between 0 and 1. In practice, it yields a probability, which will be later interpreted and turned into a class membership according to the decision threshold [32].

$$h_{\mathbf{w}}(x) = \mathbf{w}^T \cdot \mathbf{x} \quad (1)$$

$$h_{\mathbf{w}}(x) = \text{sign}(\mathbf{w}^T \cdot \mathbf{x}) \quad (2)$$

$$h_{\mathbf{w}}(x) = \theta(\mathbf{w}^T \cdot \mathbf{x}) \quad (3)$$

$$\text{where } \theta(s) = \frac{1}{1+e^{-s}} \quad \text{is the logistic function} \quad (4)$$

hence the explicit form of hypothesis function for logistic regression is:

$$h_{\mathbf{w}}(x) = \frac{1}{1 + e^{-\mathbf{w}^T \cdot \mathbf{x}}} \quad (5)$$

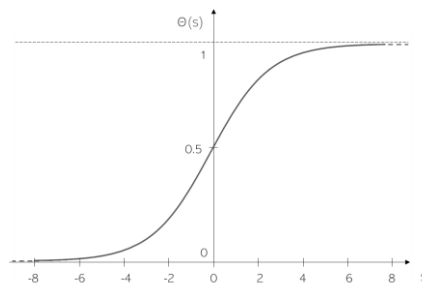


FIGURE 19 - LOGISTIC FUNCTION, ALSO CALLED SIGMOID FUNCTION DUE TO ITS S-SHAPE

Since the predicted output of logistic regression is a probability, i.e.  $\Pr(y_{pred} = 1 | \mathbf{x}; \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T \cdot \mathbf{x}}}$ , we want to find a cost function which modulates the error associated with each sample with its output probability. In other words, the outcome of logistic regression is not simply the class membership, i.e. 0 or 1, that would mean either a correct or wrong prediction, but it is rather a measure of how close/far the prediction is from the real output, and this measure is in fact the probability of that sample to belong to one class (or the other).

For example, given the true output of sample  $i$ -th is equal to class 1, i.e.  $y_{true}^{(i)} = 1$ , we find 2 sets of parameters,  $\mathbf{w}_1$  and  $\mathbf{w}_2$ , such that  $\Pr(y_{pred}^{(i)} = 1 | \mathbf{x}; \mathbf{w}_1) = 0.4$  and  $\Pr(y_{pred}^{(i)} = 1 | \mathbf{x}; \mathbf{w}_2) = 0.1$ .

Assuming the decision threshold being equal to 0.5, both models misclassify this sample. However, the first one is closer to the real solution than the latter, and this needs to be taken into account when quantifying the misclassification error. For this reason, we want to find a cost function that when the true output is one would associate a null error to a predicted probability that tends to 1 and an infinite error to a predicted probability that tends to 0, and vice versa for a true output equal to 0.

The mapping between the interval  $[0, 1]$  (probability) to the interval  $[0, \infty]$  (error) can be achieved with the logarithmic function (Figure 20), hence the cost function of logistic regression, called log-likelihood, takes the form of Eq. (6). Note that the log of a number between 0 and one is always minor or equal to 0, hence the value inside the square brackets is always negative (or equal to 0 in case of 100% correct classifications). It is possible to either leave the function negative or to multiply it by -1 and reverse the overall sign to positive as in Eq. (6).

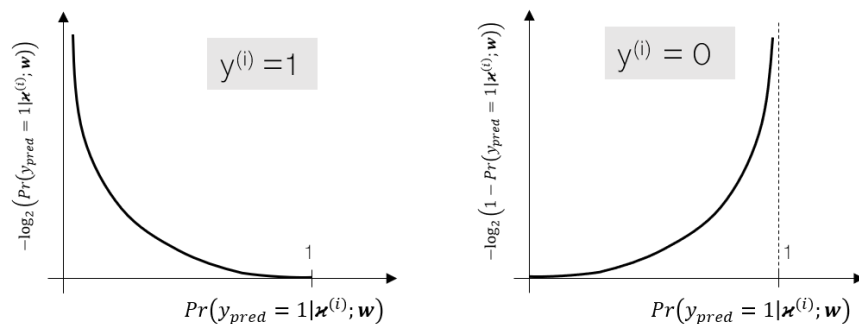


FIGURE 20 - VISUAL REPRESENTATION OF COST FUNCTION FOR LOGISTIC REGRESSION

$$J(\mathbf{w}) = -\frac{1}{n} \left[ \sum_{i=1}^n y^{(i)} \log_2 \left( \Pr(y_{pred} = 1 | \mathbf{x}^{(i)}; \mathbf{w}) \right) + (1 - y^{(i)}) \log_2 \left( 1 - \Pr(y_{pred} = 1 | \mathbf{x}^{(i)}; \mathbf{w}) \right) \right] \quad (6)$$

Now that we have the hypothesis function and the cost function, it's time to find the best  $\mathbf{w}$ , that are the values of  $\mathbf{w}$  that would make the cost function as close as possible to zero (log-likelihood maximization for negative  $J(\mathbf{w})$  and log-likelihood minimization for positive  $J(\mathbf{w})$ ). The cost function minimization (or maximization) in logistic regression is usually performed by running the gradient descent algorithm.

Without going into details, gradient descent is an iterative optimization method which seeks for a local minimum of a differentiable function by moving in the direction of the steepest descent.

TABLE 2 - THE THREE COMPONENTS OF LOGISTIC REGRESSION

Hypothesis function	Cost function	Optimizer
$h_{\mathbf{w}}(x) = \frac{1}{1 + e^{-\mathbf{w}^T \cdot x}}$	Log-likelihood	Gradient descent

## 5. DAILY CGM CLASSIFICATION

In paragraph 2.4, we have seen that diabetes diagnosis usually relies on physiological parameters measured by means of blood tests. Under the assumption that dysfunctions in glucose metabolism reflect in the glucose dynamics, we want to build a system to discriminate a normoglycemic from a non-normoglycemic individual based on CGM data. For this, we will combine the learning ability of ML models with clinical knowledge, encoded in the form of features of the systems.

In this chapter we will build a classification system based on CGM data recorded across days, while in the next chapter we'll use CGM data of short duration, recorded during a standardized meal test, that is a postprandial glycemc response.

### 5.1 Methods

Data visualization, pre-processing and manipulation have been entirely handled in MATLAB<sup>®</sup> R2020b environment.

#### 5.1.1 Dataset

The dataset used in this chapter was found from a collection of links to publicly available CGM datasets [33], and it goes under the name © 2018 Hall et al. [7]. The article related to this dataset is open access and distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>).

The dataset contains CGM recordings of 57 healthy adult participants without prior diagnosis of diabetes, aged between 25 and 76 years old (median 51), of which 32 are females and 25 males. CGM data were recorded in the normal environment of the participants for at least 7 days, using Dexcom G4 CGM devices, which records interstitial glucose every 5 minutes. Participants were instructed to calibrate monitors once to twice a day using glucose meters. They were blinded to the results of monitoring, hence their dietary habits were not influenced by the glucose recordings [7]. Although all participants were supposed to be healthy, during screening tests, 14 of them met criteria for having prediabetes and 5 T2D, while the remainder were normoglycemic (Table 3) [7]. The diagnosis has been made on the basis of ADA guideline reported in Table 1.

Table 3 resumes the participant's characteristics as mean  $\pm$  standard deviation.

TABLE 3 - PARTICIPANTS CHARACTERISTICS

Participants Characteristics	All	Healthy	Prediabetes	T2D
<b>n</b>	57	38	14	5
<b>Age</b> [years]	49 ± 14	45 ± 13	57 ± 14	57 ± 7
<b>BMI</b> [kg/m <sup>2</sup> ]	26.68 ± 4.70	25.44 ± 3.88	29.82 ± 5.27	27.22 ± 5.53
<b>SSPG</b> [mg/dL]	129.76 ± 71.28	122.64 ± 63.38	119.67 ± 77.70	187.8 ± 89.92
<b>FBI</b> [mIU/L]	7.75 ± 4.44	6.68 ± 4.15	8.93 ± 4.29	11.8 ± 4.32
<b>FBG</b> [mg/dL]	93.21 ± 12.61	87.30 ± 5.07	100.43 ± 11.02	116.8 ± 18.93
<b>2h-OGTT</b> [mg/dL]	124.76 ± 41.49	106.05 ± 21.59	150.85 ± 35.40	213 ± 46.50
<b>HbA1C</b> [%]	5.41 ± 0.42	5.20 ± 0.20	5.68 ± 0.31	6.14 ± 0.63
<b>hsCRP</b> [mg/L]	1.84 ± 2.36	1.51 ± 1.80	2.78 ± 3.47	1.7 ± 2.22
<b>Tri/HDL</b> [-]	1.76 ± 1.61	1.43 ± 1.05	1.86 ± 1.26	3.94 ± 3.71

BMI = body mass index; SSPG = steady state plasma glucose concentration; FBI = fasting blood insulin concentration; FBG = fasting blood glucose concentration; 2h-OGTT = 2-hour plasma glucose after oral glucose tolerance test; HbA1c = glycated hemoglobin; hsCRP = high-sensitivity c-reactive protein; Tri/HDL = triglyceride to high-density lipoprotein ratio, an approximation of insulin resistance. Values are expressed as mean ± standard deviation.

Data inspection revealed that the length of the subjects' recordings varied from 7 to 17 days, being sometimes discontinuous, with gaps of minutes/hours/days/months. Filling the missing values was not needed for our purposes, however, it was necessary to separate the recordings from day to day for each subject in the dataset. Figure 21 represents the CGM recording of an healthy subject across 7 days; we can see gaps of few hours or minutes as well as days where the recording lasted only few hours. Figure 22 shows the same recording of Figure 21, structured day by day and time aligned. The CGM recording over the 24h time will be called daily CGM.

Besides adjustments to the structure of the CGM recordings, pre-processing included also the manual labeling of the 57 instances of the dataset based on the values of HbA1c, FBG and 2h-OGTT (see Table 1) with two classes: "control" (38 samples) for healthy subjects and "highRisk" (19 samples), for diabetic or prediabetic subjects.

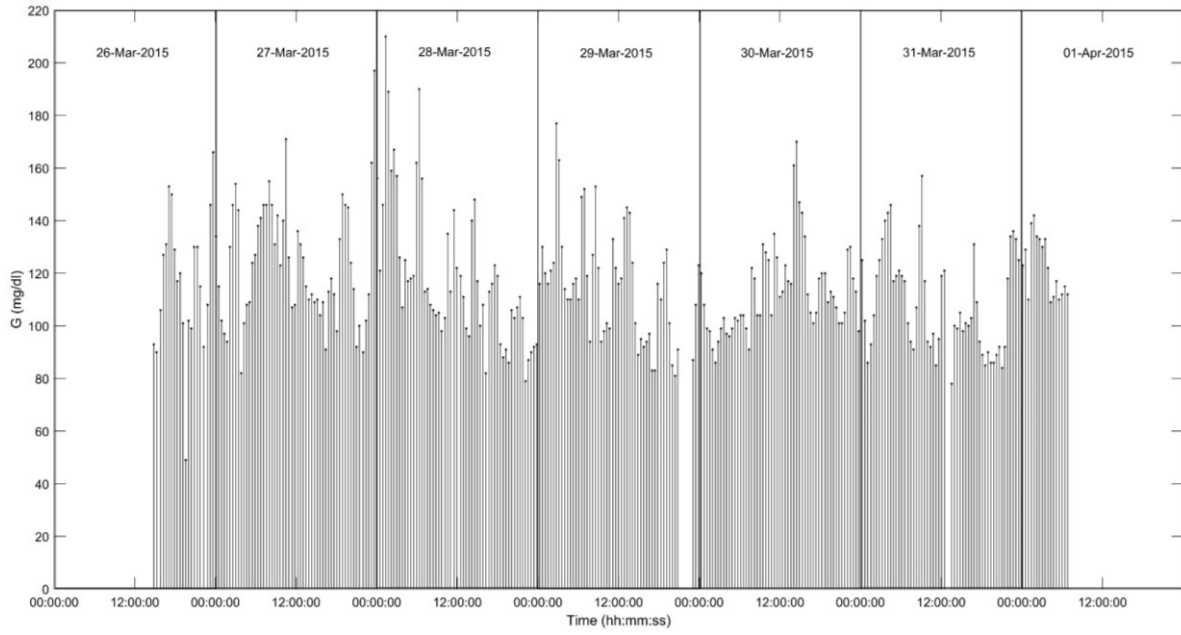


FIGURE 21 - CGM RECORDING ACROSS 7 DAYS OF A NORMOGLYCEMIC SUBJECT

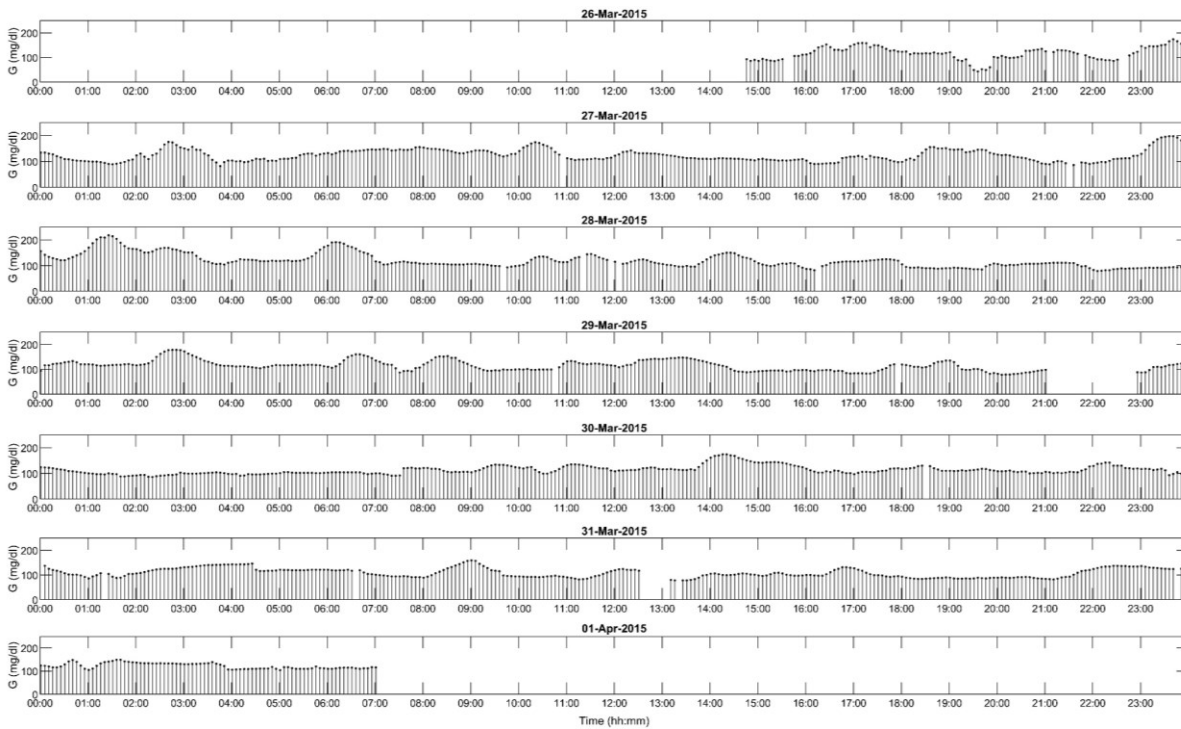


FIGURE 22 – CGM RECORDING OF A NORMOGLYCEMIC SUBJECT STRUCTURED BY DAYS (DAILY CGMS)

### 5.1.2 Classification problem

The classification in “control” or “highRisk” group of the instances of the dataset has been performed through logistic regression (see paragraph 4.2.6), which was implemented by fitting data to a generalized linear model (GLM) with binomial distribution of the error around the response variable and logit function as link function. A GLM is a generalization of the ordinary linear regression model, that allows some degrees of choice on the error distribution of the response variable and on the function that links the linear model to the response variable (link function) [34]. In our case, since the response variable is bounded between 0 and 1 (probability), the related error will have a binomial distribution. The link function for logistic regression is the sigmoid function, indeed the logit function is exactly its inverse.

The function used to fit the GLM is *fitglm*, part of MATLAB® Statistics and Machine Learning Toolbox. The decision threshold for classification was set to 0.5.

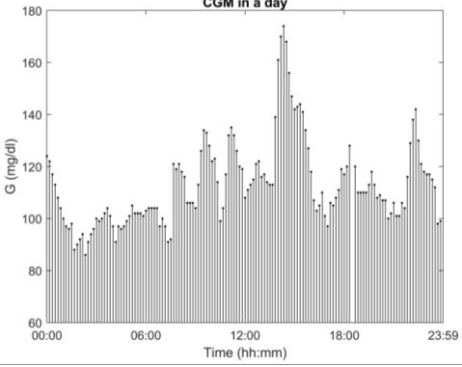
### 5.1.3 Feature extraction and selection

Even though the dataset contains multiple daily CGM recordings for each participant, they haven't been considered as single instances of the dataset, but they are all related to the subjects to which they belong. Indeed, the instances of the dataset are exactly the subjects, each with his/her set of daily CGM recordings. Given this premise, each subject of the dataset will be represented by a unique set of features, extracted from his/her overall CGM timeseries. In particular, the features for a given subject will be extracted from each of the single daily CGM recordings and then averaged across days.

For instance, with reference to Figure 22, these daily CGM recordings belong to a single person, and we want to extract two features  $F_1$  and  $F_2$  from his/her overall CGM signal; first, we extract the features  $f_1^{(i)}$  and  $f_2^{(i)}$  from each of the seven daily recordings, i.e.  $i = [1, 2, \dots, 7]$ , and then we average them among the seven days, i.e.  $F_1 = \frac{1}{7} \sum_{i=1}^7 f_1^{(i)}$  and  $F_2 = \frac{1}{7} \sum_{i=1}^7 f_2^{(i)}$ . These two features  $F_1$  and  $F_2$  will represent in the model the subject to which they belong.

The features extracted from the CGM recordings were meant to either characterize the overall signal, quantify the glucose variability or measure the complexity of the dynamic. They are summarized in Table 4, and their detailed description is reported below.

TABLE 4 - FEATURES EXTRACTED FROM DAILY CGM TIMESERIES

	Feature	Description
	mean <sub>G</sub>	<i>Mean of glucose signal</i>
	AUC <sub>G</sub>	<i>Area Under the glucose Curve</i>
	sd <sub>G</sub>	<i>Standard deviation of glucose</i>
	TBR	<i>Time Below Range (BG &lt; 70 mg/dL)</i>
	TIR	<i>Time In Range (70 ≤ BG ≤ 180 mg/dL)</i>
	TAR	<i>Time Above Range (BG &gt; 180 mg/dL)</i>
	HBGI	<i>High Blood Glucose Index</i>
	LBGI	<i>Low Blood Glucose Index</i>
	mean <sub>GD</sub>	<i>Mean of glucose difference signal</i>
	max <sub>GD</sub>	<i>Maximum of glucose difference signal</i>
	sd <sub>GD</sub>	<i>sd of glucose difference signal</i>
	MSE <sub>s1</sub>	<i>Sample entropy</i>
	MSE <sub>C15</sub>	<i>Multiscale Entropy Complexity Index</i>

- **Indices of the overall glucose dynamic**

These indices are the most simple and intuitive ones, and they try to quantify characteristics of the glucose dynamic that are usually detectable by visual observation.

- **mean<sub>G</sub>, AUC**

The most common and simple indicator of the overall glucose dynamic is the mean value (Eq. (7)). It is highly correlated with HbA1c and it is especially influenced by hyperglycemia [35]. AUC is the area under the glucose curve (Eq. (8)), it is indicative of whole glucose excursion and it is often used as a clinical parameter to diagnose impaired glucose tolerance.

$$mean_G = \frac{1}{N} \sum_{i=1}^N BG(i) \quad (7)$$

$$AUC_G = \int_1^N BG(t) dt \quad (8)$$



- **Indices of glucose variability**

In the detection and management of diabetes, high blood glucose level is not the only risk factor to be considered, but also the occurrence of abnormal glucose variability (GV), which is indeed thought to be related to the development of diabetes complications [35]. Several indices have been proposed to quantify GV, but here we report only those used as features of the model.

- **sd<sub>G</sub>**

The most common statistical measure that quantifies the variability of the signal is the standard deviation (Eq. (9)), which is a measure of the dispersion of data around the mean value. It is influenced by non-Gaussian distributions and outliers [35].

$$sd_G = \sqrt{\frac{\sum_{i=1}^N (BG(i) - mean_G)^2}{N}} \quad (9)$$

- **TBR, TIR, TAR**

This international consensus report [36], distinguishes 5 levels of glucose ranges (Figure 23): Very low ( $BG < 54$  mg/dL), Low ( $54 \leq BG \leq 69$  mg/dL), Target ( $70 \leq BG \leq 180$  mg/dL), High ( $181 \leq BG \leq 250$  mg/dL) and Very high ( $BG > 250$  mg/dL).

Given a CGM recording of arbitrary length, the percentage of time spent in each of these levels is an important indicator of the metabolic condition of an individual, and new strategies for blood glucose regulation in diabetic patients are based on the maximization of the time spent in the target range alongside with the minimization of the time spent in the low and high ranges [36].

These percentages are referred to as: TBR\_L2 (Time Below Range, level 2; time (%) spent in the very low range), TBR\_L1 (Time Below Range, level 1; time (%) spent in the low range), TIR (Time in Range; time (%) spent in the target range), TAR\_L1 (Time Above Range, level 1; time (%) spent in the high range), TAR\_L2 (Time Above Range, level 2; time (%) spent in the very high range).

However, other works [35] define only three time in ranges: Time Below Range (TBR, that is the time (%)  $BG < 70$  mg/dL), Time in Range (TIR, that is the time  $70 \leq BG \leq$

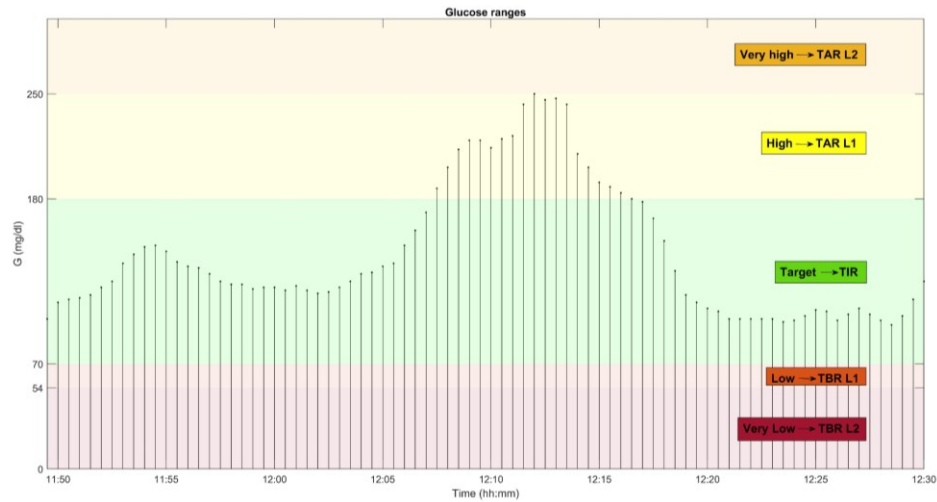


FIGURE 23 - GLUCOSE RANGES ACCORDING TO INTERNATIONAL CONSENSUS REPORT [36]

180 mg/dL) and Time Above Range (TAR, that is the time (%) BG > 180 mg/dL). In other words, TBR joins together TBR\_L1 and TBR\_L2 and TAR joins together TAR\_L1 and TAR\_L2.

#### ► HBGI, LBG1

High Blood Glucose Index (HBGI) and Low Blood Glucose index (LBGI) quantify respectively the risk of hyperglycemia and hypoglycemia from glucose timeseries [35]. They are derived from a logarithmic transformation of the BG range, described in details by Kovatchev et al. in [37], which aims to symmetrize the range of all possible BG values around zero [35]. Indeed, assuming the range of all possible BG values to be 20-600 mg/dL, the hypoglycemic (BG < 70 mg/dL) and hyperglycemic (BG > 180 mg/dL) ranges have significantly different sizes, and this causes the target range (70-180 mg/dL) to be not centered [35]. As a direct consequence, the center of the target range (about 112.5 mg/dL) doesn't correspond to the centre of the range of all possible BG values (about 300 mg/dL) and this makes the BG distribution skewed and asymmetric [35]. For BG values expressed in mg/dL, the nonlinear transformation is given by (Eq. (10)), where log is the natural logarithm, and its graphic representation is showed in Figure 24, left. It is straightforward noticing that the target range has been mapped between -1 and 1 on a new symmetric centered scale (y axis). Based on this transformation, a risk function was designed to assigns a certain degree of risk to each BG reading (Eq. (11)), i.e. maximum risk assigned to hypoglycemic and hyperglycemic

BG levels to 20 and 600 mg/dL and zero risk assigned to 112.5 mg/dL (Figure 24, right).

$$f(BG) = 1.509 \cdot ([\log(BG)]^{1.084} - 5.381) \quad (10)$$

$$r(BG) = 10 \cdot f(BG)^2 \quad (11)$$

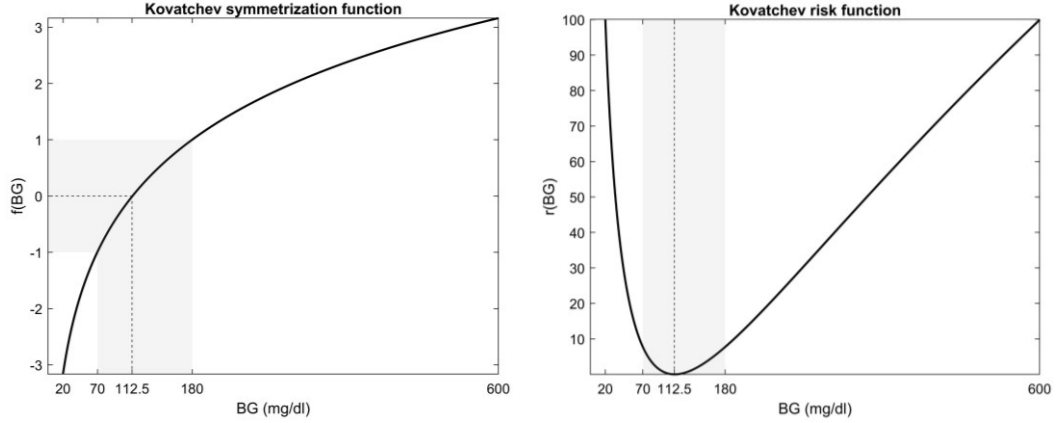


FIGURE 24 - KOVATCHEV SYMMETRIZATION AND RISK FUNCTIONS

Finally, HBGI is obtained by averaging the risk associated with each BG reading greater than 112.5 mg/dL while LBGI is obtained by averaging the risk associated with each BG reading lower than 112.5 mg/dL.

$$\begin{cases} HBGI = \frac{1}{N} \sum_{i=1}^N r(BG), & BG > 112.5 \frac{mg}{dL} \\ LBGI = \frac{1}{N} \sum_{i=1}^N r(BG), & BG < 112.5 \frac{mg}{dL} \end{cases} \quad (12)$$

$$\begin{cases} HBGI = \frac{1}{N} \sum_{i=1}^N r(BG), & BG > 112.5 \frac{mg}{dL} \\ LBGI = \frac{1}{N} \sum_{i=1}^N r(BG), & BG < 112.5 \frac{mg}{dL} \end{cases} \quad (13)$$

► **mean<sub>GD</sub>, max<sub>GD</sub>, sd<sub>GD</sub>**

Given a glucose time series  $G = \{g_1, g_2, g_3, \dots, g_N\}$  of  $N$  samples, we define the *glucose difference* signal (Figure 25) as the signal of the differences between consecutive samples  $GD = \{(g_2 - g_1), (g_3 - g_2), (g_4 - g_3), \dots, (g_N - g_{N-1})\}$ .

The idea is to use this signal to quantify how rapidly the dynamic of the signal changes, e.g. large values of glucose difference indicate a fast changing dynamic of the signal. From the glucose difference signal we extract 3 features: the mean (mean<sub>GD</sub>), the standard deviation (sd<sub>GD</sub>) and the maximum of its absolute value (max<sub>GD</sub>).

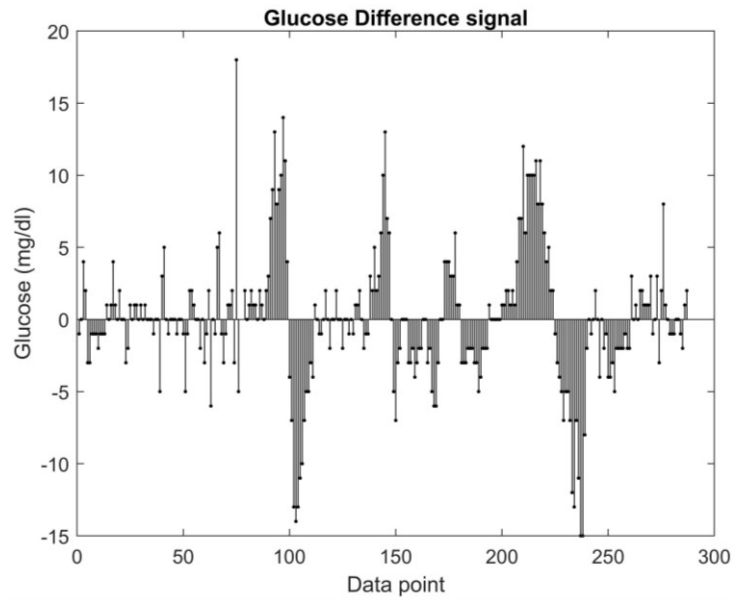


FIGURE 25 - GLUCOSE DIFFERENCE SIGNAL

- **Indices of glucose dynamics complexity**

Due to the complexity of biological systems, physiological signals contain information that might not be detectable with traditional signal processing algorithms. In the last decades, algorithms based on the entropy of the signal are emerging as an appropriate technique to quantify the complexity of physiological time series dynamics [38, 39].

- **MSE<sub>s1</sub>, MSE<sub>CI5</sub>**

The concept of entropy is deployed by a wide range of domains, such as information theory, statistics and neural networks, but regardless the field of application, it is used to determine the degree of disorder or uncertainty within a system [38]. For physiological timeseries, this translates in the possibility of quantifying the randomness of the signal, or inversely, of determining the regularity of data based on the existence of patterns [39], which is finally related to the complexity of the dynamics.

A number of studies [40, 41, 42] have found that diabetic patients show a decreased complexity in their glucose dynamics with respect to control subjects due to an increased repetition of patterns in the timeseries, probably attributable to impairments in underlying glucose control mechanisms [40]. Over the years, several definitions of entropy have been proposed, and those used to quantify the complexity of a time series are the approximate entropy (ApEn) and the sample entropy (SampEn). SampEn is a

variation of ApEn, which was proposed to replace this one as it overcomes its disadvantages [38, 39, 43]. Here we report the mathematical formulation of SampEn [39, 43].

Given a time series of  $N$  uniformly sampled points  $u(j) = \{u(1), u(2), \dots, u(N)\}$ , a positive integer  $m$  (with  $m < N$ ), and a positive real number  $r$ , we define  $N-m+1$  vectors  $X_m(i) = \{x_m(1), x_m(2), \dots, x_m(N-m+1)\}$  where  $x_m(i)$  is a vector of  $m$  data points from  $u(i)$  to  $u(i+m-1)$ . Then, we calculate the distance between any pair of such vectors as the maximum distance of their corresponding scalar component  $d[x_m(i), x_m(j)] = \max(|u(i+k) - u(j+k)| : 0 \leq k \leq m-1)$ .

Afterwards, we compute, for each template vector  $x_m(i)$ , how many vectors  $x_m(j)$  are within distance  $r$ , i.e.  $B_i = d[x_m(i), x_m(j)] \leq r$ , with  $j \leq N-m+1$  and  $j \neq i$ .

At this point we repeat the same steps using  $m+1$  instead of  $m$ , hence obtaining  $A_i = d[x_{m+1}(i), x_{m+1}(j)] \leq r$ , with  $j \leq N-m+1$  and  $j \neq i$ .

The sample entropy is finally defined as:

$$\begin{aligned} \text{SampEn}(m, r, N) &= -\log \frac{A}{B} = \\ &= -\log \frac{\sum_{i=1}^{N-m} \sum_{j=1, j \neq i}^{N-m} [\text{number of times that } d[x_{m+1}(i) - x_{m+1}(j)] < r]}{\sum_{i=1}^{N-m} \sum_{j=1, j \neq i}^{N-m} [\text{number of times that } d[x_m(i) - x_m(j)] < r]} \end{aligned} \quad (14)$$

Since  $A$  is always smaller or equal to  $B$ , SampEn will be always greater or equal to zero. The parameters on which SampEn depend are: the number or samples of the signal,  $N$ , the length of the sequences to be compared,  $m$  and the tolerance value for accepting sequence matchings,  $r$ . In order to allow the comparison of entropy measures between datasets with different amplitudes, the tolerance is usually set equal to  $p \cdot sd(u)$  [43], where  $sd(u)$  is the standard deviation of the signal and  $0 < p < 1$ . Using as tolerance a fraction of the standard deviation of the signal is equivalent to normalizing the signal to unit variance and zero mean before calculating SampEn [40]. Typical values of  $p$  are 0.15 or 0.20, while  $m$  is usually set to 2 [40]. A schematic representation of the steps of SampEn calculation is shown in Figure 26.

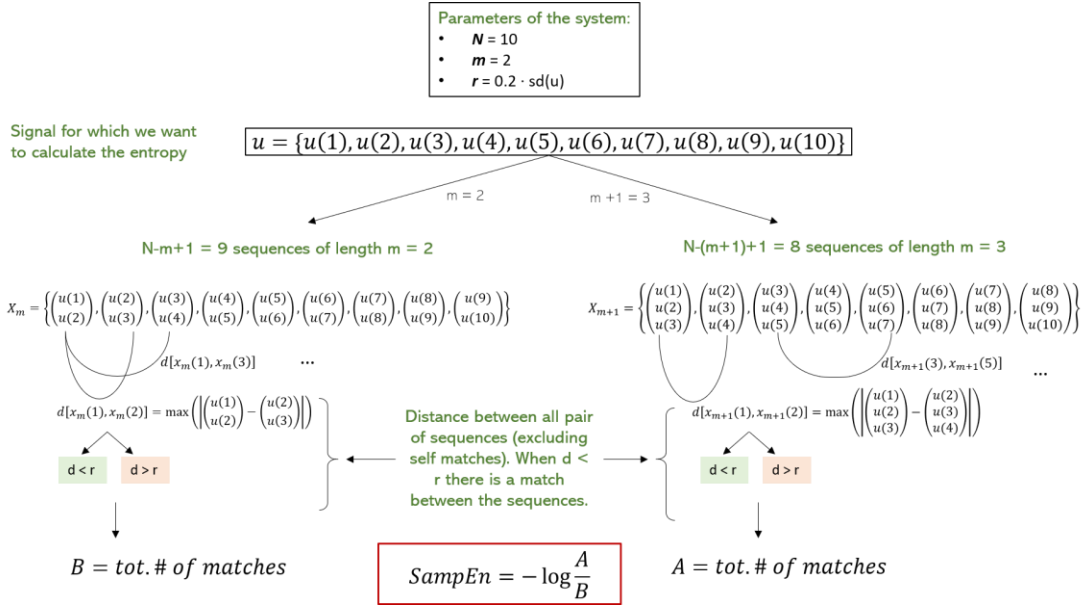


FIGURE 26 - SAMPEN ALGORITHM

It is clear from the definition (Eq. (14)) that the lower is SampEn, the more regular and repetitive is the signal.

The complexity of biological systems lies in the fact that non-random fluctuations are present at different time scales [41] and this is also certainly true for glucose dynamics.

Multiscale entropy (MSE) analysis has been extensively used to measure the complexity of physiological signals on different time scales, and it is based on SampEn algorithm. Besides its application in diabetes, MSE analysis has been used to quantify the complexity of EEG signal, gait, posture etc. [41].

Given the signal  $u(j)$  and a positive integer value , the method to compute its MSE involves the following steps [40]:

- 1 – Derivation of a set of time series from  $u(j)$  representing the signal on different time scales through a coarse-graining procedure (Figure 27).
- 2 – Calculation of SampEn for each time series extracted in the previous step.
- 3 – Computation of complexity indices as the sum of SampEn values over a given number of time scales (Eq. (15)).

$$\begin{aligned}
\text{Scale 1} \quad u_{s1} &= \{u(1), u(2), u(3), u(4), u(5), u(6), u(7), u(8), u(9), u(10), u(11), u(12)\} \\
\text{Scale 2} \quad u_{s2} &= \left\{ \frac{u(1) + u(2)}{2}, \frac{u(3) + u(4)}{2}, \frac{u(5) + u(6)}{2}, \frac{u(7) + u(8)}{2}, \frac{u(9) + u(10)}{2}, \frac{u(11) + u(12)}{2} \right\} \\
\text{Scale 3} \quad u_{s3} &= \left\{ \frac{u(1) + u(2) + u(3)}{3}, \frac{u(4) + u(5) + u(6)}{3}, \frac{u(7) + u(8) + u(9)}{3}, \frac{u(10) + u(11) + u(12)}{3} \right\} \\
\text{Scale 4} \quad u_{s4} &= \left\{ \frac{u(1) + u(2) + u(3) + u(4)}{4}, \frac{u(5) + u(6) + u(7) + u(8)}{4}, \frac{u(9) + u(10) + u(11) + u(12)}{4} \right\}
\end{aligned}$$

FIGURE 27 - COARSE-GRAINING PROCEDURE FOR SCALES FROM 1 (ORIGINAL SIGNAL) TO 4.

$MSE_{CI5}$  (Eq. (15)) and  $MSE_{s1}$  (Eq. (16)) have been used as features of the model, with  $m = 2$  and  $r = 0.1 \cdot \text{sd}_G$ .

$$MSE_{CI5} = \sum_{s=1}^5 SampEn(s) \quad (15)$$

$$MSE_{s1} = \sum_{s=1}^1 SampEn(s) = SampEn(s1) \quad (16)$$

Note that  $MSE_{CI5}$  is the complexity index calculated as the sum of  $SampEn$  over 5 scales of time, while  $MSE_{s1}$  is simply the sample entropy of the original glucose signal (time scale = 1).

After selecting a total of 13 features from the CGM time series (Table 4), it was carried out a wrapper type feature selection based on the maximization of the sensitivity of the model (see paragraph 4.2.4). Figure 28 shows the pseudocode of the implemented algorithm: given the initial set of 13 features, a minimum and a maximum number of selectable features has been fixed. For any number of selectable features in the range between minimum and maximum, the algorithm finds the combination of features which allows the highest classification sensitivity.

The minimum and the maximum number of predictors have been set to 6 and 10 respectively; the resulting models have been named according to their number of predictors:  $M_6$  (with 6 features),  $M_7$  (with 7 features),  $M_8$  (with 8 features),  $M_9$  (with 9 features) and  $M_{10}$  (with 10 features).

```

1 Features = extracted features;
2 min_feat = minimum number of features to select;
3 max_feat = maximum number of features to select;
4 for n_feat = min_feat:max_feat
5     find the subset of n_feat which allows
6     the highest sensitivity
7 end

```

FIGURE 28 - PSEUDOCODE OF SENSITIVITY-BASED WRAPPER TYPE FEATURE SELECTION

After feature selection,  $M_6$  and  $M_7$  have been selected as the optimal ones, and subsequently, through the Kaiser-Meyer-Olkin (KMO) criterion, which is based on correlation, it was assessed whether or not it was appropriate to apply PCA on these two models [29]. The threshold value for KMO was set to 0.6, i.e. values greater than 0.6 indicate that PCA is feasible.

Given that KMO criterion showed results greater than the threshold value, PCA was finally applied to  $M_6$  and  $M_7$ . PCA was applied to normalized data during model training, and the obtained coefficients have been used to apply the same transformation to the validation set. Data normalization was obtained by removing the mean and dividing by the standard deviation, feature-wise,  $\frac{X - \bar{x}}{sd(x)}$ . The number of components retained after PCA ensured the 95% explained variance.

#### 5.1.4 Model validation and performance measures

Results have been validated using a 5-fold cross validation protocol, which takes 46 samples for training and 11 for validation for each fold. It was ensured the presence of instances of both classes in training and validation sets for every fold.

The model has been evaluated in terms of ACC, SE, SP, PPV, NPV and bACC. These measures were extracted for every fold, and finally averaged to have an estimation of the overall model performances. Besides these metrics, also the deviance of the fit, the ROC curve and its AUC will be used to evaluate the performances of the models.

## 5.2 Results and discussion

In this chapter we propose a logistic regression model to classify a subject as prediabetic/diabetic (highRisk) or healthy (control), based on CGM data recorded in a timespan



of at least 7 days. The choice of merging the prediabetic and diabetic classes has been driven by the following reasons:

- 1- The original dataset contains 38 instances of healthy subjects, 14 of prediabetic and 5 of diabetic subjects. Performing a multiclass classification on this dataset would be not feasible because of the highly imbalanced class distribution and the inadequate representation of the diabetic class.
- 2- A preliminary analysis on the dataset revealed that the difference between the prediabetic and diabetic classes was not statistically significant. For all 13 features extracted from CGM signals, the Kruskal-Wallis test failed to reject the null hypothesis that the data from the two groups (prediabetic and diabetic) belonged to the same population, hence it confirmed that the two groups were not statistically significant different. The Kruskal-Wallis test is a nonparametric form of ANOVA (ANalysis Of VAriance) test, used to compare the medians of two or more groups of data and determine whether they belong to the same population or not.
- 3- The subjects labeled as “prediabetic” or “diabetic” actually were not aware of their condition before being involved in the study. This suggests that probably the dynamic of glucose fluctuations of these subjects might be very similar to each other. Given this premise, it makes more sense to define a highRisk group, comprising both prediabetic and diabetic subjects.

The choice of logistic regression for this classification task meets the requirement of having a simple model, which at the same time decreases the risk of overfitting and is very intuitive and versatile.

The initial set of features has been carefully selected trying to avoid redundancy or irrelevant information, however, using all of them to train our classifier turned out to be not the best option in terms of performances (mean ACC = 70%). For this reason, it was necessary to perform feature selection. The selection strategy, implemented through a simple algorithm shown in Figure 28, was based on the maximization of the model sensitivity; indeed, when it comes to applications in the healthcare domain, we definitely prefer to misclassify a negative (healthy) sample rather than a positive (disease) sample, which would imply missing the diagnosis (see paragraph 4.2.4). In other words, we want to penalize false negatives and one way to achieve this result is by enhancing the sensitivity of the model.

Based on the results of feature selection shown in Table 5, we can notice that some features were never or nearly never selected, i.e.  $\text{mean}_{\text{GD}}$  and  $\text{sd}_{\text{G}}$ ,  $\text{AUC}_{\text{G}}$ ,  $\text{sd}_{\text{GD}}$  while other ones were often or always selected. This suggests that we could rank the features based on how often they were selected.

TABLE 5 - FEATURE SELECTION BASED ON SENSITIVITY MAXIMIZATION

		Number of selected features				
		6	7	8	9	10
<b>Features</b>	$\text{mean}_{\text{G}}$					
	$\text{AUC}_{\text{G}}$					
	$\text{sd}_{\text{G}}$					
	TBR					
	TIR					
	TAR					
	HBGI					
	LBGI					
	$\text{mean}_{\text{GD}}$					
	$\text{max}_{\text{GD}}$					
	$\text{sd}_{\text{GD}}$					
	$\text{MSE}_{\text{s1}}$					
	$\text{MSE}_{\text{CI5}}$					
<b>SE</b>	<b>0.85±0.18</b>	<b>0.85±0.18</b>	<b>0.85±0.18</b>	<b>0.80±0.28</b>	<b>0.77±0.28</b>	
<b>ACC</b>	0.79±0.10	0.80±0.10	0.80±0.10	0.77±0.12	0.74±0.15	
<b>SP</b>	0.67±0.25	0.70±0.28	0.70±0.28	0.67±0.25	0.62±0.30	
<b>PPV</b>	0.73±0.05	0.80±0.10	0.80±0.10	0.70±0.10	0.68±0.10	
<b>NPV</b>	0.78±0.19	0.79±0.19	0.79±0.19	0.77±0.20	0.70±0.27	
<b>bACC</b>	0.76±0.15	0.77±0.15	0.77±0.15	0.73±0.18	0.70±0.20	
<b>DEV</b>	35.63±4.41	35.89±5.31	35.89±5.31	34.75±4.13	34.28±5.04	

SE = sensitivity; ACC = accuracy; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; bACC = balanced accuracy; DEV = deviance of the fit. Values are expressed as mean ± standard deviation.

The lower section of Table 5 shows the performances of the models  $M_n$  (with  $n$  = number of features, from 6 to 10) cross-validated on the corresponding selected predictors: the best

sensitivity is ensured by  $M_6$ ,  $M_7$  and  $M_8$  (85%), so we could put aside  $M_9$  and  $M_{10}$ ; the performances of  $M_7$  and  $M_8$  are actually identical, hence we can infer that the predictor “TAR”, that is what makes these two model different, doesn’t add any information to the system, hence we can undoubtedly discard  $M_8$ .  $M_6$  and  $M_7$  show very comparable results, hence they can be interchangeably used as our definitive model.

DEV is the deviance of the fit and it quantifies the “goodness” of the fit. A more detailed explanation of this parameter is given in the next chapter, paragraph 6.1.3.

Regardless the relevance of the features for the classification model, it is worth making few remarks on some features:

- Using TBR\_L2, TBR\_L1, TIR, TAR\_L1, TAR\_L2 (see paragraph 5.1.3) was not feasible because the outer levels were almost empty vectors, i.e. there were only few samples in the whole dataset that were included in the upper and lower levels. At the same time, TBR\_L1 and TAR\_L1 were poorly populated. In order to overcome this issue without excluding these important indices of glycemic variability, TBR\_L1 has been joined with TBR\_L2 and TAR\_L1 has been joined with TAR\_L2, hence obtaining TBR, TIR and TAR.
- For what concerns the choice of the indices related to the entropy of the signal, it was based on empirical observation of the MSEs for control vs highRisk groups (Figure 29). The x axis represent the timescale expressed in minutes, and it is equivalent to the scale from 1 to 7, where 1 corresponds to CGM readings every 5 minutes (original signal) and 7 corresponds to CGM readings every 35 (=5·7) minutes. As we could have expected from the premises given in paragraph 5.1.3, the MSE for the control group shows higher values of entropy, indicating a greater complexity in the glucose dynamics compared with highRisk group. Moreover, we can notice that the two curves cross each other at timescale equal to 30 minutes, (scale = 6), hence we compute the complexity index  $MSE_{CI5}$  as the sum of the SampEn from scale 1 to scale 5.  $MSE_{s1}$  has been selected as a feature because the control group and the highRisk group show the maximum difference at scale = 1, hence it seemed to be a good candidate to separate the two classes.

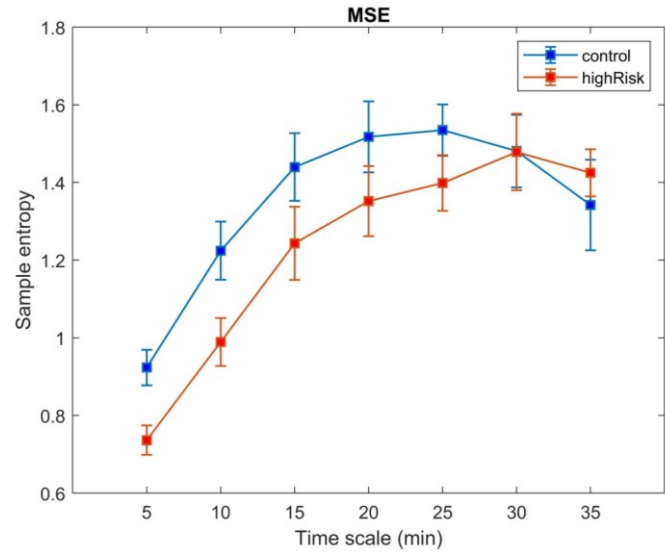


FIGURE 29 - MULTISCALE ENTROPY FOR CGM SIGNAL. TIME SCALE FROM 5 TO 35 MINUTES. SQUARES ARE MEAN VALUES AND VERTICAL BARS ARE VARIANCES

Due to the limited dimensions of the dataset and the problem of class imbalance, the holdout of a portion of the dataset for testing purposes was considered not feasible; for this reason, model validation has been carried out with a 5-fold cross validation, which ensures the presence of “enough” examples in each fold to train the model as well as “enough” examples in each fold to validate the model (a 10-fold cross validation probably wouldn’t ensure this condition). Although we can’t rigorously test the performances of the model without a test set, cross-validation can give a qualitative measure of the classifier performances without subtracting data for training.

Figure 30 shows ROC curves of  $M_6$  and  $M_7$  for the 5-folds of cross validation. In accordance with the results above, the behaviour of the two models are pretty much the same, having the mean AUC equal to 0.89 and 0.88 respectively. The curves have this discontinuous appearance due to the reduced size of the training sets (46 samples), and their values suggest that the classes are not well separable, probably due to class imbalance or anomalous dynamics among the healthy population.

Even though the dimensionality of the feature space would be adequate for the dataset size, we may want to find out whether the model can benefit from dimensionality reduction. However, it is not guaranteed that the transformation applied by PCA leads to better performances of the model, simply because PCA is “blind” to the output variable (unsupervised), hence the new axes might not be consistent with the discriminatory features of the classification problem [29].

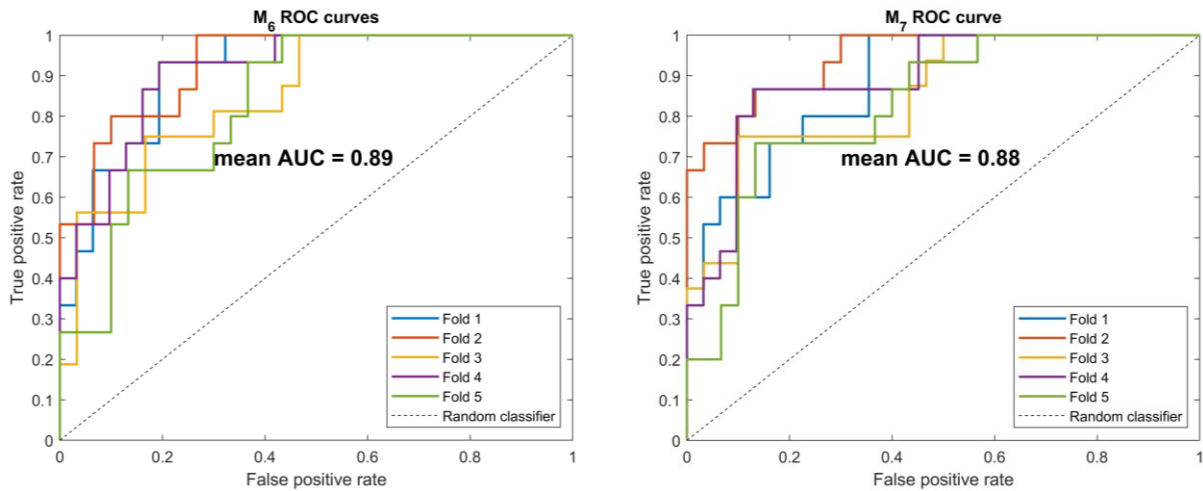


FIGURE 30 - ROC CURVES FOR  $M_6$  AND  $M_7$

The values of KMO for  $M_6$  and  $M_7$  were equal to 0.64 and 0.63 respectively, hence PCA was feasible and it was applied to both models.

Table 6 compares the performances of  $M_6$  and  $M_7$  before and after applying PCA; we can notice a slight degradation of the performances of  $M_{7+PCA}$ , and a general improvement of the performances of  $M_{6+PCA}$ , in particular for the specificity (from 67% to 70%) and positive predictive value (from 73% to 80%). The ROC curves across the folds for  $M_{6+PCA}$  and  $M_{7+PCA}$  are very similar to the corresponding models before PCA (shown in Figure 30), namely they have mean AUCs equal to 0.88 and 0.87 respectively.

In general, the performances of the output models are influenced by the reduces size of the dataset and this reflects in a quite high variability of the results across the folds (standard deviation) and suboptimal performances (for each fold, the model is validated on 11 samples only, therefore even misclassifying few of them will result in poor classification scores) .

The class imbalance also plays a role in the performances of the classifier, especially for the correct classification of positive samples (highRisk): since the positive class has an occurrence of about 30% in the dataset, in both training and validation sets there are far fewer positive samples than negatives, so when the model is training it doesn't have "enough" instances to learn how to identify the positive class, and when it's validating, its performances tend to be poor.

Now that we have gathered an idea of the classifier performances through cross validation, we want to learn our final classifier on the whole dataset.

TABLE 6 - CLASSIFICATION PERFORMANCES BEFORE ( $M_6$  AND  $M_7$ ) AND AFTER PCA ( $M_{6+PCA}$  AND  $M_{7+PCA}$ )

	SE	ACC	SP	PPV	NPV	bACC
$M_6$	0.85±0.18	0.79±0.10	0.67±0.25	0.73±0.05	0.78±0.19	0.76±0.15
$M_{6+PCA}$	0.85±0.18	0.81±0.10	0.70±0.28	0.80±0.10	0.79±0.19	0.77±0.15
$M_7$	0.85±0.18	0.80±0.10	0.70±0.28	0.80±0.10	0.79±0.19	0.77±0.15
$M_{7+PCA}$	0.80±0.20	0.77±0.15	0.70±0.28	0.79±0.12	0.74±0.28	0.75±0.19

SE = sensitivity; ACC = accuracy; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; bACC = balanced accuracy. Values are expressed as mean ± standard deviation.

The set of features of this final model is the same used in  $M_{6+PCA}$  (see Table 6). Performances are shown in terms of ROC curve and its AUC, confusion matrix and its related metrics (Figure 31 and Figure 32). The issues due to small dataset size and class imbalance hold also for the final model.

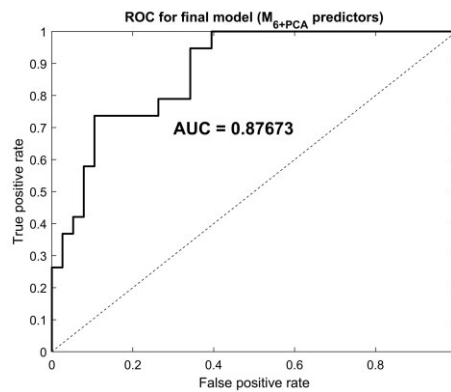


FIGURE 31 - ROC CURVE FOR THE FINAL MODEL (TRAINED ON THE WHOLE DATASET, USING  $M_{6+PCA}$  PREDICTORS)

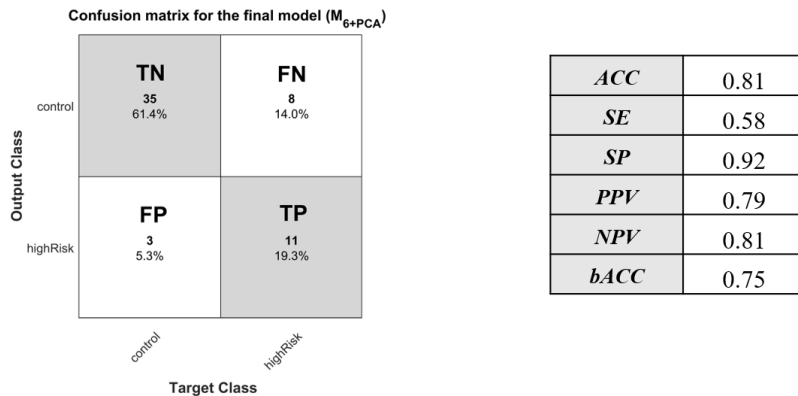


FIGURE 32 - CONFUSION MATRIX AND RELATED PERFORMANCE MEASURES FOR THE FINAL MODEL (TRAINED ON THE WHOLE DATASET, USING  $M_{6+PCA}$  PREDICTORS)

## 6. POSTPRANDIAL GLYCEMIC RESPONSE CLASSIFICATION

As already mentioned in the introduction of chapter 5, in this chapter we want to build a classification system based on postprandial glyceemic response of normoglycemic vs non-normoglycemic individuals. Note that these data were still recorded through a CGM device, but only for the duration of a standardized meal test, and not for one or more days, as it was supposed to be for a CGM recording.

### 6.1 Methods

Data visualization, preprocessing and manipulation have been entirely handled in MATLAB<sup>®</sup> R2020b environment.

#### 6.1.1 Dataset

The dataset used in this chapter was found from a collection of links to publicly available CGM datasets [33], and it goes under the name © 2018 Hall et al. [7]. The article related to this dataset is open access and distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>).

The dataset contains CGM recordings of 30 healthy adult participants without prior diagnosis of diabetes, aged between 25 and 65 years old (median 37), of which 20 are females and 10 males. CGM data were recorded during a standardized meal test, using Dexcom G4 CGM devices, which record interstitial glucose every 5 minutes. The postprandial glyceemic response was recorded from 30 minutes prior to the start of the meal until 2.5 hours after the start of the meal [7] (Figure 33).

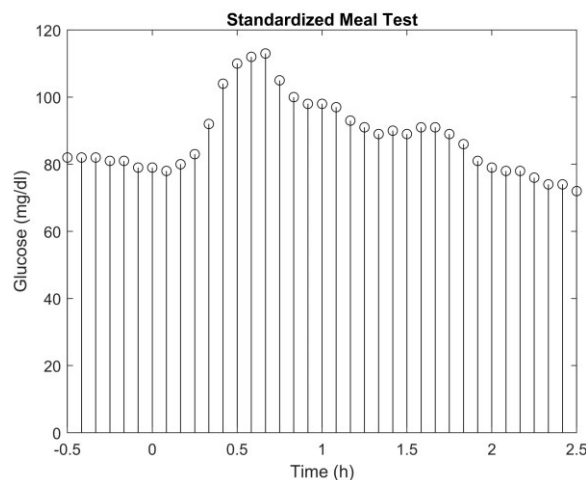


FIGURE 33 - GLYCEMIC RESPONSE TO STANDARDIZED MEAL TEST. THE RECORDING STARTS 30 MINUTES BEFORE THE MEAL INTAKE AND IT ENDS 2H30 AFTER IT

Although all participants were supposed to be healthy, during screening tests, 7 of them met criteria for having prediabetes and 3 T2D, while the remainder 20 subjects were normoglycemic [7]. The diagnosis has been made on the basis of ADA guideline reported in Table 1. Table 7 resumes the participant’s characteristics as mean  $\pm$  standard deviation.

There were 3 types of standardized meal test, with similar caloric load but different amounts of proteins, fat and fibers: cornflakes and milk, bread and peanut butter and a protein bar (Table 8). Theoretically, every participant received each type of meal twice, always at breakfast, for a total of six standardized meal tests per participant; however, some recordings were missing from the dataset, which contains instead a total of 176 standardized meal tests among the 30 participants.

Data inspection and pre-processing revealed that, some CGM recordings were shorter or longer than expected due to the presence of doubled and/or missing samples . Given that the sampling time is 5 minutes, and that the expected duration of the standardized meal test is 3h, we should have recordings of 37 samples ( $\frac{60}{5} \cdot 3 + 1$ , where 1 has been arbitrarily added to account for the time spent to actually eat the meal).

TABLE 7 - STANDARDIZED MEAL TEST PARTICIPANTS’ CHARACTERISTICS

Participants	All	Healthy	Prediabetes	T2D
n	30	20	7	3
Age [years]	42 $\pm$ 14	38 $\pm$ 13	48 $\pm$ 14	55 $\pm$ 9
BMI [kg/m <sup>2</sup> ]	25.93 $\pm$ 5.531'	23.08 $\pm$ 2.65	32.23 $\pm$ 5.85	30.2 $\pm$ 5.18
SSPG [mg/dL]	119 $\pm$ 81.89	80.69 $\pm$ 35.07	159.5 $\pm$ 105.05	231 $\pm$ 90.27
FBI [mIU/L]	6.79 $\pm$ 4.08	4.78 $\pm$ 2.05	10.29 $\pm$ 5	10.67 $\pm$ 3.51
FBG [mg/dL]	94.24 $\pm$ 14.34	86.84 $\pm$ 4.65	105 $\pm$ 11.80	116 $\pm$ 25.16
2h-OGTT [mg/dL]	122.67 $\pm$ 49.12	100.21 $\pm$ 21.63	'152 $\pm$ 40.05'	248 $\pm$ 11.31
HbA1C [%]	5.36 $\pm$ 0.40	5.16 $\pm$ 0.17	5.67 $\pm$ 0.38	5.83 $\pm$ 0.67
hsCRP [mg/L]	1.55 $\pm$ 2.10	0.9 $\pm$ 1.0	3.04 $\pm$ 3.24	2.37 $\pm$ 2.83
Tri/HDL [-]	1.39 $\pm$ 1.13	0.93 $\pm$ 0.62	2.26 $\pm$ 1.53	2.43 $\pm$ 1.17

BMI = body mass index; SSPG = steady state plasma glucose concentration; FBI = fasting blood insulin concentration; FBG = fasting blood glucose concentration; 2h-OGTT = 2-hour plasma glucose after oral glucose tolerance test; HbA1c = glycated hemoglobin; hsCRP = high-sensitivity c-reactive protein; Tri/HDL = triglyceride to high-density lipoprotein ratio, an approximation of insulin resistance. Values are expressed as mean  $\pm$  standard deviation.



TABLE 8 - NUTRITION FACTS FOR STANDARDIZED MEALS

Nutrients	Bread and	Protein bar	Cornflakes and milk
Calories (kcal)	430	370	280
Fat (g)	20	18	2.5
Carbohydrates (g)	51	48	54
- Sugar (g)	12	19	35.2
- Fiber (g)	12	6	3.3
Protein (g)	18	9	11

Figure 34 shows the length of the recordings in number of samples; there are some recordings significantly shorter than expected, e.g. a recording of 15 samples corresponds to a duration of 75 minutes, which is less than 50% of the expected duration. Since it would not be possible to identify the meal timing in such short recordings, we discard those which are at least 25 minutes shorter than expected, i.e. less than 33 samples. To standardize the length of the recordings, doubled samples have been replaced by their average, missing values have been filled by computing the mean value between the previous and the following sample or by repeating the last available sample in case of missing values at the end of the signal.

After length standardization, the dataset contains 151 postprandial glycemc response signals, of which 105 belong to healthy subjects, 32 to prediabetic and 14 to diabetic subjects (according to ADA guidelines). However, the instances of the dataset have been manually labelled with two classes: “control” (105 samples) for healthy subjects and “highRisk” (46 samples), for diabetic or prediabetic subjects.

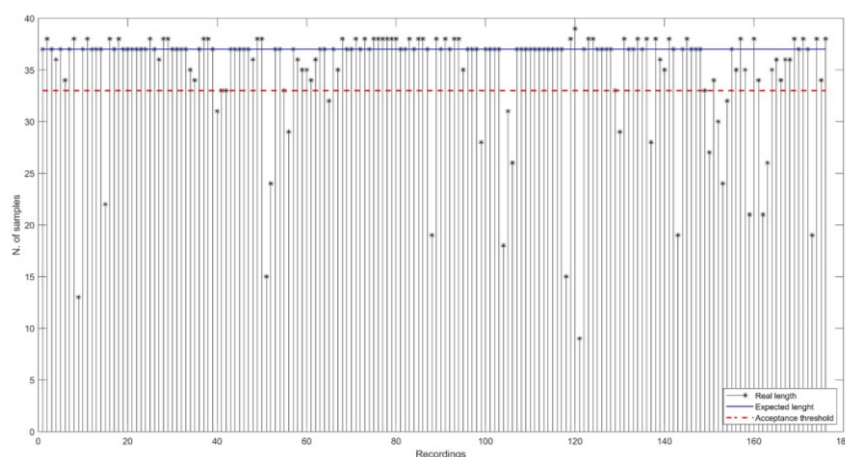


FIGURE 34 - NUMBER OF SAMPLES FOR EACH RECORDING. EXPECTED LENGTH: 37 SAMPLES; ACCEPTED LENGTH: GREATER OR EQUAL TO 34 SAMPLES

### 6.1.2 Classification problem

The classification model used to classify each CGM recording in the control or highRisk class was based on a logistic regression model (see paragraph 4.2.6), which was implemented by fitting data to a generalized linear model (GLM) with binomial distribution of the error around the response variable and logit function as link function. A GLM is a generalization of the ordinary linear regression model, that allows some degrees of choice on the error distribution of the response variable and on the function that links the linear model to the response variable (link function) [34]. In our case, since the response variable is bounded between 0 and 1 (probability), the related error will have a binomial distribution. The link function for logistic regression is the sigmoid function, indeed the logit function is exactly its inverse.

The function used to fit the GLM is *fitglm*, part of MATLAB® Statistics and Machine Learning Toolbox. The decision threshold for classification was set to 0.5.

### 6.1.3 Feature extraction and selection

A total of 19 features were extracted from each recording, and they can roughly be divided in two groups:

- **Statistical-signal features** →  $\text{mean}_G$  (*i.e.*, mean), [44],  $\text{median}_G$  (*i.e.*, median),  $\text{min}_G$  (*i.e.*, minimum),  $\text{max}_G$  (*i.e.*, maximum) [45],  $\text{maxGidx}$  (*i.e.*, timing of the peak after the meal) [45],  $\text{std}_G$  (*i.e.*, standard deviation) [44],  $\text{range}_G$  (*i.e.*, difference between maximum and minimum) [44],  $\text{prePmean}_G$  and  $\text{postPmean}_G$  (*i.e.*, preprandial and postprandial mean) [44],  $\text{prePtime}$  and  $\text{postPtime}$  (*i.e.*, fraction of time spent below the preprandial mean and above the postprandial mean, respectively) [46],  $\text{slope}_G$  (*i.e.*, slope between meal time and maximum peak) [46],  $\text{AUC}_G$  (*i.e.*, area under the curve) [44] and  $\text{skew}_G$  (*i.e.*, skewness) of the CGM recording.
- **Spectral features** →  $\text{mean}_{\text{FFT}}$  (*i.e.*, mean),  $\text{median}_{\text{FFT}}$  (*i.e.*, median),  $\text{min}_{\text{FFT}}$  (*i.e.*, minimum),  $\text{max}_{\text{FFT}}$  (*i.e.*, maximum) and  $\text{std}_{\text{FFT}}$  (*i.e.*, standard deviation) of the Fast Fourier Transform of the CGM recording [47].

Since we don't know a-priori if all these features are relevant to this classification task or if there are highly correlated features, we want to perform feature selection (see paragraph 4.2.2).

Feature selection was carried out in two steps: the first step was a filter type feature selection, and it was based on visual inspection of the scatterplot matrix (Figure 35), which shows pairwise relationships between any combination of features. For a given pair of highly correlated features, point biserial correlation (equivalent to the Pearson's correlation) to class was computed and the one showing highest correlation was retained and the other discarded (Table 9, second column). For instance,  $std_G$  and  $range_G$  seems high correlated, and the correlations to the target variable are 0.37 and 0.33 respectively, hence  $range_G$  has been discarded. Note that the correlation to the target alone would not be a reliable method to assess the relevance of a feature because a combination with other features might instead be highly correlated to the target.

The second step of feature selection was a wrapper type or sequential feature selection, based on the performances of the classification. It was carried out through a recursive algorithm based on the evaluation of the deviance of the fit (Eq. (17)), which is a generalization of the residual sum of squares [48] used to quantify the goodness of the fit of a given model. It is defined as twice the log-likelihood ratio of the saturated model compared to the reduced model, where the saturated model is a model having a parameter for every observation, hence perfectly fitting the data, while the reduced model is the model of which we want to estimate the goodness of the fit [48]. In other words, the deviance of the fit quantifies how much the model deviates from a perfect fit. The pseudo-code of this sequential feature selection is available in Figure 36.



FIGURE 35 - SCATTERPLOT MATRIX FOR FEATURE CORRELATION VISUALIZATION

```

1  features = features after filter
   selection;
2  fullModel = f(features);
3  dev0 = fullModel.Deviance;
4  tolerance = chi2inv(0.95,1);
5  sort features in ascending order of
   correlation with the target;
6  for f in features
7     features = features - f;
8     reducedModel = f(features);
9     if reducedModel.Deviance > dev0 +
       tolerance
10        features = features + f;
11    end
12 end

```

FIGURE 36 - PSEUDO-CODE OF DEVIANCE-BASED WRAPPER TYPE FEATURE SELECTION

Beginning with the full set of features selected in the previous step, the deviance of the fit was computed and used as reference (dev0); one feature at a time was removed from the model and the deviance of the reduced model (devR) was calculated and compared with the deviance of the full model: if devR was higher than dev0 augmented by a tolerance value, the feature just excluded was considered significant for the classification task, hence it was reintroduced in the model, otherwise it was judged irrelevant and it was discarded. The tolerance value was set to the 95th percentile for the chi-square distribution with 1 degree of freedom [48], i.e. tolerance equal to 3.84.

$$D(y, \hat{\mu}) = 2(\log(p(y|\hat{w}_s)) - \log(p(y|\hat{w}_r))) \quad (17)$$

Where  $\hat{w}_s$  and  $\hat{w}_r$  are the fitted parameters of the saturated and reduced models respectively and  $\hat{\mu}$  are the predictions to the observations  $y$ .

#### 6.1.4 Model validation and performance measures

We can distinguish two models for this classification task (Figure 37):

- **M<sub>FT</sub>** is the model with the features selected after the filter type feature selection
- **M<sub>FT+WT</sub>** is the model with the features selected after the two consequential feature selections, the filter type and the wrapper type.

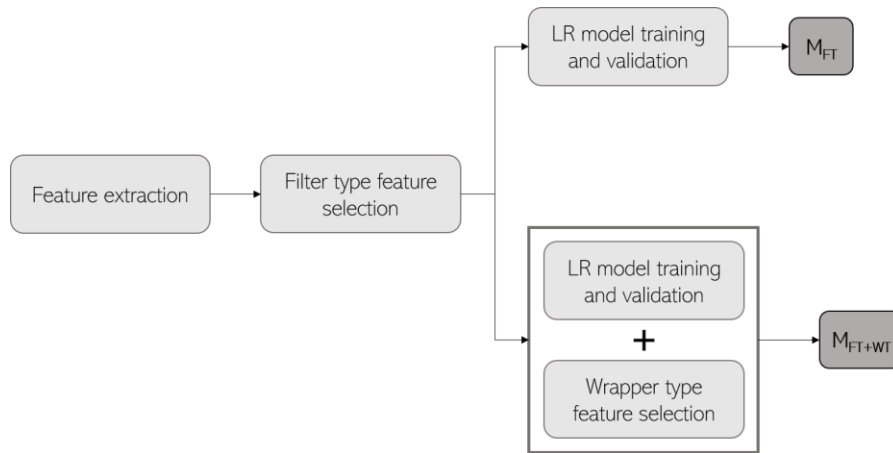


FIGURE 37 –  $M_{FT}$  AND  $M_{FT+WT}$  FEATURE SELECTION TIMELINES. LR STANDS FOR LOGISTIC REGRESSION.

The models have been validated using a 5-fold cross-validation protocol (see paragraph 4.2.3). Given that there were multiple recordings for each participant and that they were assumed as distinct observations in the dataset, whenever the dataset was split in training and validation sets during cross-validation, the observations of the same participant were included in either one of the two sets and not split between them.

The classification performances of the two models have been evaluated and compared by means of ROC curve, AUC, confusion matrix and the related metrics (see paragraph 4.2.4)

## 6.2 Results and discussion

In this chapter we propose a logistic regression model to classify a subject as prediabetic/diabetic or healthy, based on CGM data recorded during a standardized meal test. Despite a multinomial classification (healthy vs prediabetes vs diabetes) would be preferable from the point of view of diagnosis and prevention, the prediabetic and diabetic classes have been merged in the highRisk class for two main reasons:

1. After discarding short recordings (see paragraph 6.1.1), the dataset contains 105 recordings of healthy, 32 of prediabetic and 14 of diabetic subjects. Performing a multinomial classification with this dataset would arise in technical problems related to class imbalance and inadequate representation of the minority class (see paragraph 4.2.1).

2. All the subjects involved in the study were supposed to be healthy, but during screening tests some of them turned out to have one or more physiological parameters in the prediabetic or diabetic range. This means that these subjects were not aware of their condition, and an official and rigorous diagnosis would require further tests which are not available in our case. For this reason, rather than defining these people diabetic or prediabetic, it would be more accurate to consider them having a high risk of developing impairments in glucose regulation.

Point 1 of the above list states that the dataset contains 151 recordings; this is indeed true, but we should remind that the subjects involved are just 30, therefore each of them contributes to the dataset with 5 recordings on average. The assumption is to consider the 151 recordings as independent instances, e.g. an healthy subject with 4 recordings contributes to the dataset with 4 instances, each labelled as “control”.

The choice of logistic regression for data classification has been determined by the intention of having a simple model that is both intuitive and that generalizes well.

The initial set of features extracted from CGM data accounts for 19 predictors, which seems to be an adequate number with respect to the number of samples (see paragraph 4.2.3). However, we still need to verify the presence of highly correlated features, and eventually discard them (filter type feature selection); furthermore, we might want to minimize the number of features while keeping the quality of the model as high as possible (wrapper type feature selection). After the filter type feature selection, 14 features out of 19 were retained and the rest discarded, and the wrapper type feature selection furtherly dropped 5 features, hence arriving to 9 features. The deviance of the fit is equal to 89.44 for  $M_{FT}$  and 93.21 for  $M_{FT+WT}$ . This means that  $M_{FT+WT}$  “fits worst” than  $M_{FT}$ , but its deviance is still within the allowed value which is 93.28 (= 89.44 + tolerance, where tolerance = 3.84). The timeline of feature selection is schematized in Table 9.

Model validation has been carried out with a 5-fold cross validation, which ensures the presence of “enough” examples in each fold to train the model as well as “enough” examples in each fold to validate the model (a 10-fold cross validation probably wouldn’t ensure this condition). Due to the limited dimensions of the dataset and the problem of class imbalance, the holdout of a portion of the dataset for testing purposes was considered not feasible; however, cross-validation can give a qualitative measure of the classifier performances without subtracting data for training.

TABLE 9 - TIMELINE OF FEATURE SELECTION.  $M_{FT}$  INCLUDES THE 14 HIGHLIGHTED FEATURES, OBTAINED AFTER FILTER TYPE FEATURE SELECTION;  $M_{FT+WT}$  INCLUDES THE 9 HIGHLIGHTED FEATURES, OBTAINED AFTER WRAPPER TYPE FEATURE SELECTION.

Features	Point biserial correlation	Description	$M_{FT}$	$M_{FT+WT}$
mean <sub>G</sub>	0.59	Mean glucose		
median <sub>G</sub>	0.58	Median glucose		
min <sub>G</sub>	0.51	Minimum glucose		
max <sub>G</sub>	0.49	Maximum glucose		
maxGidx	0.21	Timing of peak after meal		
std <sub>G</sub>	0.37	SD of glucose		
range <sub>G</sub>	0.33	maximum - minimum		
prePmean <sub>G</sub>	0.52	Pre-prandial mean glucose		
postPmean <sub>G</sub>	0.58	Post-prandial mean glucose		
prePtime	-0.19	% time below prePmean		
postPtime	0.12	% time above postPmean		
slope <sub>G</sub>	0.03	Slope between meal and peak		
AUC <sub>G</sub>	0.59	Area Under glucose Curve		
skew <sub>G</sub>	-0.26	Skewness		
mean <sub>FFT</sub>	0.48	Mean of Fourier transform		
median <sub>FFT</sub>	0.20	Median of Fourier transform		
min <sub>FFT</sub>	0.24	Min of Fourier transform		
max <sub>FFT</sub>	0.59	Max of Fourier transform		
std <sub>FFT</sub>	0.59	SD of Fourier transform		

The performances of the two models  $M_{FT}$  and  $M_{FT+WT}$  can be evaluated and compared by means of ROC curves and metrics derived from the confusion matrix: the mean AUC is 0.94 for  $M_{FT}$  and 0.93 for  $M_{FT+WT}$  (Figure 38), and this suggests that, in both cases, the classes are well separable, and the models are robust with respect to the choice of the decision threshold for classification.

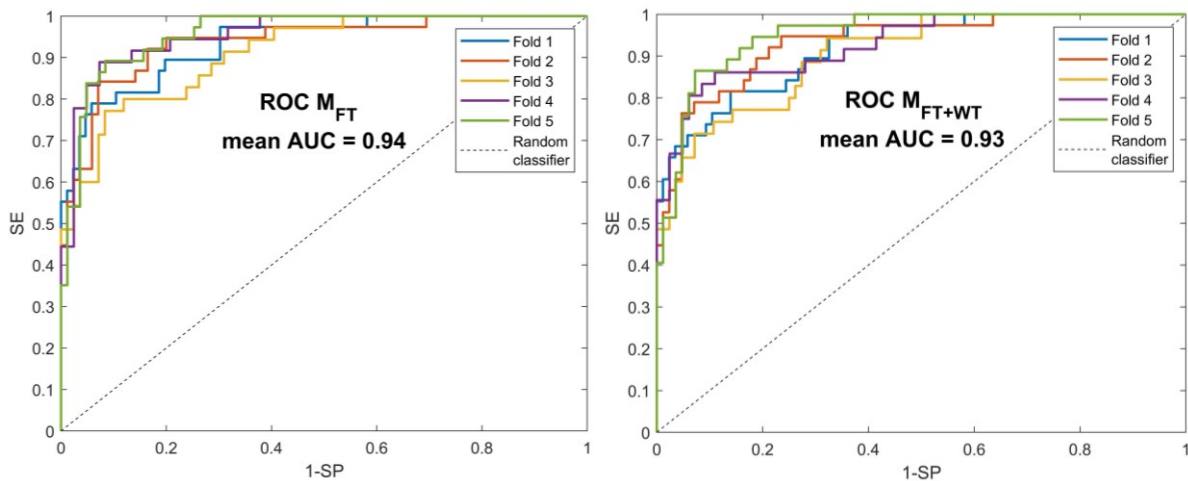


FIGURE 38 – ROC CURVES AND MEAN AUC FOR  $M_{FT}$  AND  $M_{FT+WT}$

Despite the higher deviance of the fit of  $M_{FT+WT}$  with respect to  $M_{FT}$ , the results reported in Table 10 reveal that  $M_{FT+WT}$  represents the optimal choice between the two models. Indeed, besides an improvement of  $M_{FT+WT}$  compared with  $M_{FT}$  in terms of mean ACC, SE and NPV, a lower standard deviation across the folds in almost all  $M_{FT+WT}$  measures (except SE) can be observed. This indicates that in  $M_{FT+WT}$  the performances of the classifier are quite independent on the training and validation partitions used.

TABLE 10 -  $M_{FT}$  AND  $M_{FT+WT}$  PERFORMANCE MEASURES

		<i>Fold 1</i>	<i>Fold 2</i>	<i>Fold 3</i>	<i>Fold 4</i>	<i>Fold 5</i>	<i>Mean ± sd</i>
<i>ACC</i>	<i>M<sub>FT</sub></i>	0.89	0.86	0.94	0.79	0.65	<b>0.82 ± 0.10</b>
	<i>M<sub>FT+WT</sub></i>	0.85	0.86	0.91	0.82	0.74	<b>0.84 ± 0.05</b>
<i>SE</i>	<i>M<sub>FT</sub></i>	0.84	0.63	0.95	0.96	0.73	<b>0.82 ± 0.13</b>
	<i>M<sub>FT+WT</sub></i>	0.84	0.63	0.95	1.00	0.82	<b>0.85 ± 0.13</b>
<i>SP</i>	<i>M<sub>FT</sub></i>	1.00	0.95	0.91	0.40	0.44	<b>0.74 ± 0.26</b>
	<i>M<sub>FT+WT</sub></i>	0.88	0.95	0.82	0.40	0.56	<b>0.72 ± 0.21</b>
<i>PPV</i>	<i>M<sub>FT</sub></i>	1.00	0.83	0.95	0.79	0.76	<b>0.87 ± 0.09</b>
	<i>M<sub>FT+WT</sub></i>	0.94	0.83	0.91	0.79	0.82	<b>0.86 ± 0.06</b>
<i>NPV</i>	<i>M<sub>FT</sub></i>	0.73	0.86	0.91	0.80	0.40	<b>0.74 ± 0.18</b>
	<i>M<sub>FT+WT</sub></i>	0.70	0.86	0.90	1.00	0.56	<b>0.80 ± 0.16</b>
<b>bACC</b>	<i>M<sub>FT</sub></i>	0.92	0.79	0.93	0.68	0.59	<b>0.78 ± 0.13</b>
	<i>M<sub>FT+WT</sub></i>	0.86	0.79	0.89	0.70	0.69	<b>0.78 ± 0.08</b>

ACC = accuracy; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value; bACC = balanced accuracy. Values are expressed as mean ± standard deviation

Since we have already gathered an idea of the classifier performances on unseen data with cross validation, we now want to learn our final classifier on the whole dataset (Figure 39 and Figure 40).

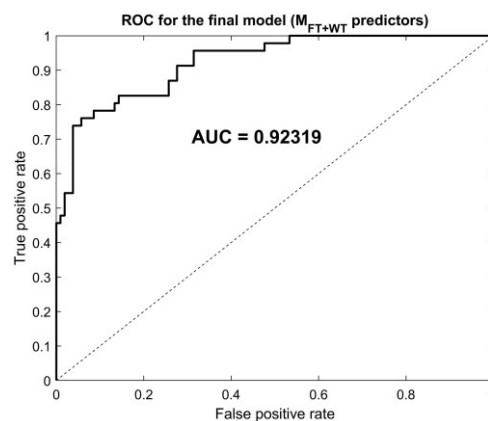


FIGURE 39 - ROC CURVE FOR THE FINAL MODEL (TRAINED ON THE WHOLE DATASET, USING  $M_{FT+WT}$  PREDICTORS)



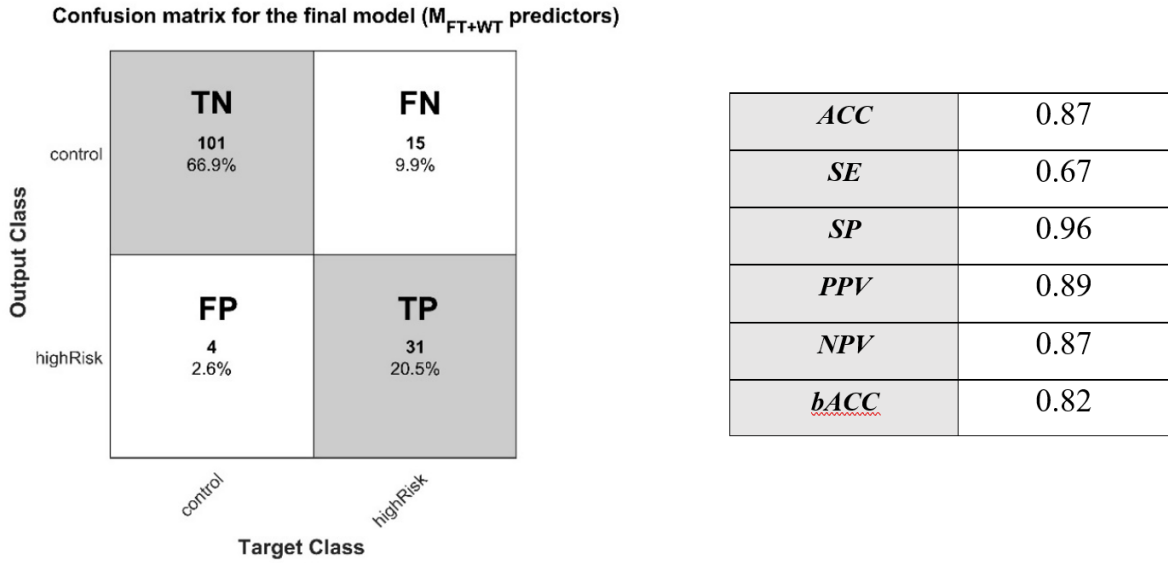


FIGURE 40 - CONFUSION MATRIX AND RELATED PERFORMANCE MEASURES FOR THE FINAL MODEL (TRAINED ON THE WHOLE DATASET, USING  $M_{FT+WT}$  PREDICTORS)

The set of features of this final model is the same used in  $M_{FT+WT}$  (9 predictors, see Table 9). The low sensitivity (67%) is attributable to two factors: firstly, the positive class is the minority class, accounting for just the 30% of the samples, hence the model probably doesn't have "enough" instances of this class to properly learn how to recognize it; secondly, the positive class exhibits a higher data variance since it includes CGM observations of both prediabetic and diabetic individuals (Figure 41).

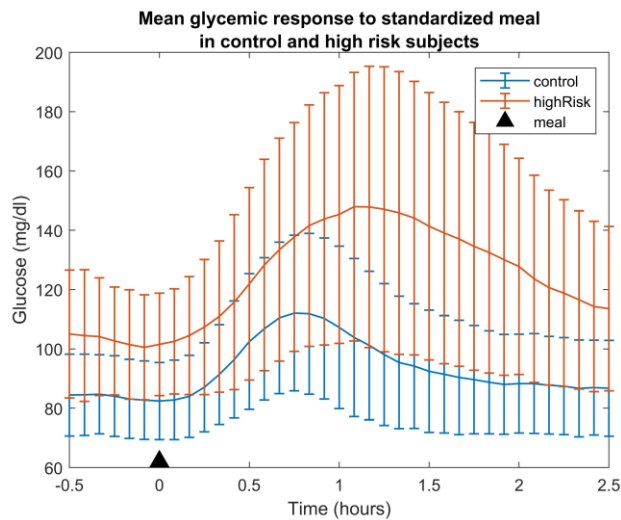


FIGURE 41 - MEAN GLYCEIC RESPONSES IN CONTROL AND HIGH RISK SUBJECTS. VERTICAL BARS ARE STANDARD DEVIATIONS OF THE MEASUREMENTS.

## CONCLUSION

The purpose of this thesis was the development of a classification model able to distinguish healthy from diabetic subjects based on their CGM recordings.

Previous works have dealt with similar problems with a variety of approaches: in Wang et al. [44], a double-class AdaBoost classifier was trained using 17 features extracted from (CGM) signals to classify patients in type 1 or type 2 diabetes classes, obtaining up to 90.3% accuracy; in Longato et al. [49, 50], support vector machine was used to classify CGM data as belonging to diabetic (T2) or impaired glucose tolerance subjects on the base of glycemic variability indices, obtaining up to 87.1% accuracy; Chen et al. [51], used physiological measurements and demographic data to build a hybrid prediction model that uses k-means clustering for data reduction and decision tree algorithm to classify patients as healthy or diabetic, obtaining an accuracy of 90.04 %.

This thesis tackled the problem of classifying subjects as diabetic or healthy with a dual approach: by using CGM data spanning over several days and CGM data recorded only for the duration of a standardized meal test (3h).

Both approaches used logistic regression, but the models were trained on different sets of features, extracted from the corresponding CGM signals and chosen through different feature selection strategies. Cross-validation of the two models, having respectively 6 and 9 predictors, led to 81% and 84% mean accuracy respectively, with 85% sensitivity in both cases. These results are quite encouraging considering the limitations of the present work, first among them the reduced dataset size and the class imbalance; some approximations on the dataset may also have had a negative impact, that are the combination of the prediabetic and diabetic class and, in the postprandial glycemic response classification model, having used the recordings of a single subject as independent instances of the dataset.

Despite the limitations, these models have the advantages of being very simple and intuitive and offering a large margin of improvement, for example by overcoming the present constraints, by using other types of classifier, etc.

Furthermore, with respect to other works based on physiological and demographical data [51], these models use CGM signals to discover hidden patterns that are otherwise not derivable from static blood glucose measurements. The use of such data to characterize the health status of an individual, has the potential to account for individual differences that standard medical

approaches might miss [25]. In this sense, this work can be framed in the domain of precision medicine, both because CGM allows for precision monitoring, that because the logistic regression model implies a probability-based decision which is compatible with a precision diagnosis [25].

Future developments of the present work should include the extraction of other meaningful features, already existing or new ones, or the use of deep learning, to extrapolate relevant insights about glucose dynamics. Indeed, CGM data are still a largely untapped source of information, with an enormous potential for diagnosis as well as for treatment applications, and the role of machine learning in the exploitation of this resource is simply crucial.

## BIBLIOGRAPHY

- [1] F. H. Martini, J. L. Nath and E. F. Bartholomew, *Fundamentals of Anatomy and Physiology*, 11th ed., Pearson Education, 2018.
- [2] P. Mergenthaler and U. Lindauer, "Sugar for the brain: the role of glucose in physiological and pathological brain function," *Trends in neurosciences*, vol. 36, no. 10, pp. 587-597, 2013.
- [3] R. M. Berne, M. N. Levy, B. M. Koeppen and B. A. Stanton, *Physiology*, 5th ed., Elsevier, 2004.
- [4] M. N. Nakrani, R. H. Wineland and F. Anjum, "Physiology, Glucose Metabolism," Aug 2020. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK560599/>. [Accessed Jan 2021].
- [5] C. L. Adams, *An Extensible Mathematical Model of Glucose Metabolism*, Dissertations, Ed., Old Dominion University: Mathematics & Statistics, 2011.
- [6] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karura, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9 th edition," *Diabetes Res Clin Pract.*, 2019.
- [7] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin and M. Snyder , "Glucotypes reveal new patterns of glucose dysregulation," *PLoS Biol* , vol. 16, no. 7, 2018.
- [8] American Diabetes Association , "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes - 2021," *Diabetes Care*, vol. 44, pp. S15-S33, 2021.
- [9] R. I. Holt, C. S. Cockram, A. Flyvbjerg and B. J. Goldstein, *Textbook of Diabetes*, 4 ed., Wiley, 2010.

- [10] A. Arsyad, I. Idris, A. A. Rasyid, R. A. Usman, K. R. Faradillah, W. O. U. Latif, Z. I. Lubis, A. Aminuddin, I. Yustisia and Y. Y. Djabir, "Long-Term Ketogenic Diet Induces Metabolic Acidosis, Anemia, and Oxidative Stress in Healthy Wistar Rats," *Journal of Nutrition and Metabolism*, vol. 2020, pp. 1-7, 2020.
- [11] Osmosis, "Diabetes mellitus (type 1, type 2) & diabetic ketoacidosis (DKA)," Sept 2019. [Online]. Available: <https://www.youtube.com/watch?v=-B-RVybvfU&t=484s>. [Accessed Feb 2021].
- [12] M. Y. Donath and S. E. Shoelson, "Type 2 diabetes as an inflammatory disease," *Nature Reviews Immunology*, vol. 11, pp. 98 - 107, 2011.
- [13] World Health Organization, "Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO Consultation," Geneva, 2011.
- [14] S. E. Oberfield, M. P. Gallagher and D. E. Hale, "How is hemoglobin A1C helpful for monitoring diabetic control?," in *Pediatric Secrets*, Fifth Edition ed., Mosby, 2011, pp. 197-228.
- [15] American Diabetes Association, "Understanding A1C: Diagnosis," Feb 2021. [Online]. Available: <https://www.diabetes.org/a1c/diagnosis>. [Accessed Feb 2021].
- [16] M. J. Fowler, "Microvascular and Macrovascular Complications of Diabetes," *Clinical Diabetes*, vol. 26, no. 2, pp. 77-82, 2008.
- [17] Mayo Clinic, "Diabetic neuropathy," Mar 2020. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/diabetic-neuropathy/symptoms-causes/syc-20371580>. [Accessed 02 2021].
- [18] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), "Continuous Glucose Monitoring," June 2017. [Online]. Available: <https://www.niddk.nih.gov/health-information/diabetes/overview/managing-diabetes/continuous-glucose-monitoring>. [Accessed Feb 2021].
- [19] World Health Organization, "Diabetes," May 2020. [Online]. Available: [https://www.who.int/health-topics/diabetes#tab=tab\\_3](https://www.who.int/health-topics/diabetes#tab=tab_3). [Accessed Feb 2021].

- [20] A. T. Soliman, V. DeSanctis, M. A. Yassin, R. Elalaily and N. E. Eldarsy, "Continuous glucose monitoring system and new era of early diagnosis of diabetes in high risk groups," *Indian Journal of Endocrinology and Metabolism*, vol. 18, no. 3, pp. 274 - 282, 2014.
- [21] T. Siegmund, L. Heinemann, R. Kolassa and A. Thomas, "Discrepancies Between Blood Glucose and Interstitial Glucose—Technological Artifacts or Physiology: Implications for Selection of the Appropriate Therapeutic Target," *Journal of Diabetes Science and Technology*, vol. 11, no. 4, pp. 766-772, 2017.
- [22] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Commun. ACM*, vol. 55, p. 78–87, 2012.
- [23] E. Kavlakoglu, "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?," IBM, May 2020. [Online]. Available: <https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks>. [Accessed Apr 2021].
- [24] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Gangu, S. Shekhar, N. Samatova and V. Kumar, "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318-2331, 2017.
- [25] W. K. Chung, K. Erion, J. C. Florez, A. T. Hattersley, M.-F. Hivert, C. G. Lee, M. I. McCarthy, J. J. Nolan, J. M. Norris, E. R. Pearson, L. Philipson, A. T. McElvaine, W. T. Cefalu, S. S. Rich and P. W. Franks, "Precision medicine in diabetes: a Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)," *Diabetologia*, vol. 63, p. 1671–1693, 2020.
- [26] S. J. Russell and P. . Norvig, *Artificial Intelligence: A Modern Approach*, ed., vol. , , : Pearson Education, Inc., 2010, p. .
- [27] MathWorks, "Introduction to feature selection," [Online]. Available: <https://www.mathworks.com/help/stats/feature-selection.html>.

- [28] V. Spruyt, "The Curse of Dimensionality in classification," Apr 2014 . [Online]. Available: <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>. [Accessed Apr 2021].
- [29] M. Pechenizkiy, A. Tsymbal and S. Puuronen, "PCA-based Feature Transformation for Classification: Issues in Medical Diagnostics," *Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems*, pp. 535-540, 2004.
- [30] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Phil. Trans. R. Soc. A*, vol. 374, p. 20150202, 2016.
- [31] A. Sap, "(Machine)Learning = Representation + Evaluation + Optimization," Jan 2018. [Online]. Available: <https://annisap.medium.com/learning-representation-evaluation-optimization-c73fce64281f>. [Accessed May 2021].
- [32] Y. S. Abu-Mostafa, M. Magdon-Ismail and H.-T. Lin, *Learning from data*, AMLbook.com, 2012.
- [33] M. Martin, E. Chun, D. Buchanan, E. Wang, S. Senthil and I. Gaynanova, *irinagain/Awesome-CGM: List of public CGM datasets (Version v1.0.0)*, Zenodo, 2020.
- [34] MathWorks, "fitglm," [Online]. Available: <https://www.mathworks.com/help/stats/fitglm.html>. [Accessed Jan 2021].
- [35] C. Fabris, *Glucose variability assessment in diabetes mellitus monitoring and control*, 2015.
- [36] T. Battelino and T. e. a. Danne, "Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range," *Diabetes Care*, vol. 42, p. 1593–1603, 2019.
- [37] B. P. Kovatchev, D. J. Cox, L. A. Gonder-Frederick, D. Young-Hyman, D. Schlundt and W. Clarke, "Assessment of Risk for Severe Hypoglycemia Among Adults With IDDM," *Diabetes Care*, vol. 21, no. 11, p. 1870–1875, 1998.
- [38] M. Borowska, "Entropy-Based Algorithms in the Analysis of Biomedical Signals," *Studies in logic, grammar and rhetoric*, vol. 43, no. 56, pp. 21-32, 2015.

- [39] A. Delgado-Bonal and A. Marshak, "Approximate Entropy and Sample Entropy: A Comprehensive Tutorial," *Entropy*, vol. 21, no. 6, p. 541, 2019.
- [40] M. D. Costa, T. Henriques, M. N. Munshi, A. R. Segal and A. L. Goldberger, "Dynamical Glucometry: Use of Multiscale Entropy Analysis in Diabetes," *Chaos*, vol. 24, no. 3, p. 033139, 2014.
- [41] J.-L. Chen, P.-F. Chen and H.-M. Wang, "Decreased complexity of Glucose Dynamics in Diabetes: Evidence from Multiscale Entropy Analysis of Continuous Glucose Monitoring System Data," *Am J Physiol Regul Integr Comp Physiol*, vol. 307, no. 2, pp. R179-83, 2014 .
- [42] K.-D. Kohnert, P. Heinke, L. Vogt, P. Augstein and E. Salzsieder, "Applications of Variability Analysis Techniques for Continuous Glucose Monitoring Derived Time Series in Diabetic Patients," *Frontiers in physiology*, vol. 9, p. 1257, 2018.
- [43] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am J Physiol Heart Circ Physiol.*, vol. 278, no. 6, pp. H2039-49, 2000.
- [44] Y. Wang, S. Liu, R. Chen, Z. Chen, J. Yuan and Q. Li , "A Novel Classification Indicator of Type 1 and Type 2 Diabetes in China," *Sci. Rep.*, vol. 7, no. 1, p. 17420, 2017.
- [45] K. G. Brodovicz, C. J. Girman, A. M. C. Simonis-Bik, J. M. Rijkkelijkhuiz, M. Zelis, M. C. Bunck, A. Mari, G. Nijpels, E. M. W. Eekhoff and J. M. Dekker, "Postprandial metabolic responses to mixed versus liquid meal tests in healthy men and men with type 2 diabetes," *Diabetes Res. Clin. Pract.*, vol. 94, no. 3, p. 449–455, 2011.
- [46] D. Dave, D. J. DeSalvo , B. Haridas, S. McKay , A. Shenoy, C. J. Koh, M. Lawley and M. Erraguntla , "Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction," *J Diabetes Sci Technol*, 2020.
- [47] G. Fico , L. Hernández, J. Cancela, M. M. Isabel, A. Facchinetti , C. Fabris, R. Gabriel, C. Cobelli and M. T. A. Waldmeyer, "Exploring the Frequency Domain of Continuous Glucose Monitoring Signals to Improve Characterization of Glucose Variability and of Diabetic Profiles," *J Diabetes Sci Technol*, vol. 11, no. 4, pp. 773-779, 2017.



- [48] MatWorks, "Sequential Feature Selection," [Online]. Available: <https://www.mathworks.com/help/stats/sequential-feature-selection.html>. [Accessed Dec 2020].
- [49] E. Longato, G. Acciaroli, A. Facchinetti, L. Hakaste, T. Tuomi, A. Maran and G. Sparacino, "Glycaemic variability-based classification of impaired glucose tolerance vs. type 2 diabetes using continuous glucose monitoring data," *Computers in Biology and Medicine*, vol. 96, pp. 141-146, 2018.
- [50] E. Longato, G. Acciaroli, A. Facchinetti, A. Maran and G. Sparacino, "Simple Linear Support Vector Machine Classifier Can Distinguish Impaired Glucose Tolerance Versus Type 2 Diabetes Using a Reduced Set of CGM-Based Glycemic Variability Indices," *J Diabetes Sci Technol.*, vol. 14, no. 2, pp. 297-302, 2020 .
- [51] J. Chaki, T. S. Ganesh, S. K. Cidham and A. S. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, pp. 2-22, 2020.