DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA

Ciclo 32

Coordinatore Prof. Trillo Stefano

Development of new system for prediction of hematic parameters in extra-corporeal treatments

Settore Scientifico Disciplinare: ING-INF/02

Dottorando

Dott. Decaro Cristoforo

Tutore

Prof. Bellanca Gaetano

Anni 2017/2019

Contents

CHAPTER 1

Introduction

This dissertation describes the process behind the development of a real-time monitoring system of hematic parameters for extra-corporeal treatments.

The system is optimized for hemodialysis and it provides real-time measure of:

- Hematocrit (Hct)
- Oxygen saturation (sO_2)

There are other optical sensors on the market, with different measure ranges and accuracies, suitable to monitor parameters of blood during extra-corporeal treatment. The one proposed in this Thesis, due to the fact that it is based on a spectroscopic approach, which provides information on a wide spectral range instead of working at a single wavelength, presents some advantages that will be described in more details in next chapters.

Clinical needs require a better way to monitor hematic parameters of blood during hemodialysis without removing blood samples from patients and without interfering with dialysis treatment in order to reduce the risk of over-treatment that can cause severe injury for patients.

The setup is intended to assist clinicians during dialysis, it can prevent side effects on patients and optimize treatment duration and fluid removal rate in order to achieve a more effective dialysis.

However, the device is not intended to replace medical staff, but it could help clinicians, according with their experience, in making good decisions for patients safe.

The setup includes a white high-power LED, a mini-spectrometer, a fiber optic and a microcontroller.

The system exploits light to trans-illuminate the blood within a chamber, which represents the optical window for absorbance spectroscopic measurements.

The visible spectrum contains information about chemical composition of matter: in this case it includes the levels of hematocrit and oxygen saturation.

Machine learning algorithms have been developed to extract directly the parameter levels from visible spectra, in order to obtain prediction of hematocrits and oxygen saturation with high precision.

1.1 Acknowledgement

This research has been funded by Regione Emilia Romagna in the framework of the *PO Fse 2014/2020 Alte competenze per la ricerca, il trasferimento tecnologico e l' imprenditorialitá.*

The project has been carried out by Universitá degli studi di Ferrara leading by its team of integrated optics and in partnership with MISTER Smart Innovation [1] and MEDICA S.p.A. [2], two important industrial players in Emilia Romagna with experience in industrial research and blood purification treatments, respectively.

All players have contributed during the project with their experiences and allowing the use of their laboratories and instruments, they have also added a commercial effort to this work.

1.2 Challenges

During dialysis, waste is continuously filtered from the blood in a process called ultrafiltration (UF). This process must be monitored in order to optimize the results of the treatment and to stop it at the optimal time, but many times medical staff have not these information and their decisions are made by experience.

Moreover, nurses or other professionals commonly extract some blood during dialysis, then the sample is put into a capillary device and hematocrit is evaluated through cen-

trifugation. This technique is not efficient, because it is long, it could be repeated many times until the treatment is finished and it need a centrifugation device and a nurse or a physician to perform it.

Real-time monitoring through spectroscopy changes this procedure: information are provided continuously and the system is non invasive, because the device is not in contact with human blood. As previously stated, the system proposed here has some advantages with respect to other devices available in the market, all related to the fact that the hematic parameters are evaluated through machine learning models, which are very flexible and reliable. The setup has been optimized in all of its parts:

- mechanical components are designed with CAD software and they are realized with a 3D printer; they have been intended to mechanically optimized the path of the light across blood sample and reduce external noise;
- optical path has been optimized;
- electronic hardware has been designed to drive light and sensor;
- software has been implemented for post processing the data and for applying machine learning algorithms.

Finally, two different prototypes have been realized:

1. The first setup (Test bench prototype) is intended for creating a database. It includes a microcontroller for measuring spectrum and a laptop to store and post-process data. The data are collected during tests on bovine blood. This collection of all spectra forms a database, which is used for training machine learning models. Results are evaluated with trained models applied on new spectral measurements of blood.
2. The second setup (Stand-alone prototype) is intended for real time application on hemodialysis patients. In this final prototype, post-processing and machine learning models are provided by a Raspberry Pi connected to a microcontroller. The models are already trained before uploading into microcontroller, so the prototype will applied them on new samples.

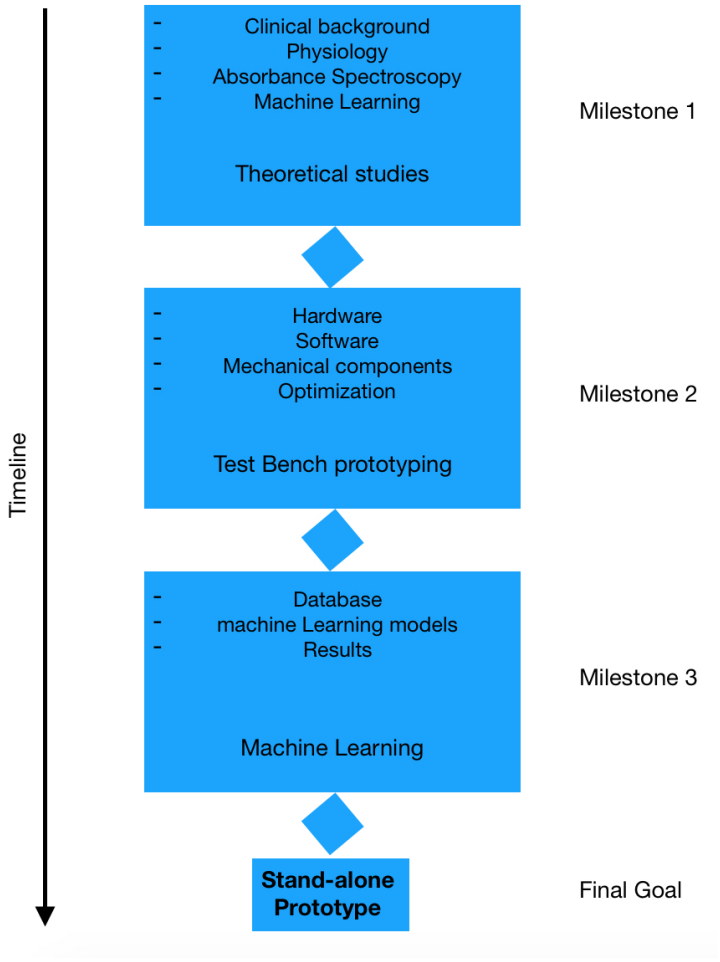


Figure 1.1: Project plan showing milestones and final goal

1.3 Workflow plan

The workflow plan has been designed to quickly deliver a working prototype for testing.

The activities have been summarized into four work-packages, the block diagram shown in figure 1.1 plots the milestones over the project timeline and the final project goal.

The project has been started with some theoretical studies, which cover the following fields:

- Clinical background
- Physiology
- Absorbance Spectroscopy
- Machine Learning

Milestone 2 includes the development of hardware and software for the first test bench prototype.

The aim was to test the prototype with bovine blood reproducing hemodialysis sessions, in order to generate a database of samples with different hematic characteristics.

The database has been used to train machine learning models; in this step different smart algorithms have been tested in order to find the best model which guarantees the highest accuracy.

The final goal of this research has been the development of a stand alone prototype, easily integrable with hemodialysis device, which allows the monitoring of hematocrits and oxygen saturation of blood during the treatment.

The tests have been carried out with bovine blood, because it is cheaper and widely more available comparing to the human blood, moreover MEDICA's laboratory is certified for these specific tests. The choice to use bovine blood does not affect the validity of the results, but same tests will be performed with human blood before using this setup with patients.

1.4 How to read this dissertation

The rest of the dissertation is organized as follows:

Chapter 2 provides the basic clinical knowledges and theories that are relevant for the understanding of this work. It includes clinical background and importance of kidneys diseases, the current state of art of optical sensors for monitoring hematic parameters with their advantages and limitations.

Chapter 3 starts with absorption spectroscopy theory and its principles. This is essential to understand why spectroscopy is one of the most widely used techniques to determine concentration of species in solutions. The chapter continues with basic elements of blood physiology and its optical behaviour in visible wavelengths.

Chapter 4 provides an overview of machine learning, its uses in medicine field and a description of the algorithms used in this project. In the last part of this chapter the balancing problem of a dataset is introduced with principal techniques to solve it.

Chapter 5 describes the development of two prototypes realized during the project with a fully description of the setup and all the software and hardware components.

Chapter 6 illustrates the methods used in this project, including the tests carried out in specialized laboratories in order to realize the database of bloods spectra. This database is used for training machine learning algorithms in order to find models for the prediction of target parameters of blood.

The results are shown in chapter 7, where the accuracies of the different models for the prediction of hematocrit and oxygen saturation are compared.

Finally, chapter 8 concludes the dissertation, describing the limitation of the system and suggesting forward paths for future improvements.

CHAPTER 2

Clinical Background

Kidneys are two important organs located at the back of abdominal cavity, below the rib cage and between the spine (figure 2.1); they are protected by fat and surrounded by fibrous renal capsule.

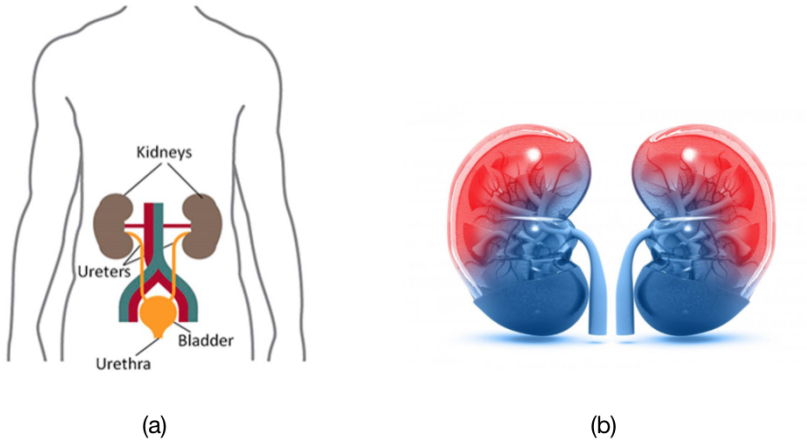


Figure 2.1: Reppresentations of human kidney (a) and its shape (b). Figures from web

Kidneys are characterized by bean shape with a series of lobes, they have an outer renal cortex and an intern medulla.

In both kidneys, there are about a million of nephrons, which are responsible of filtering waste [3].

Each nephron includes a glomerulus, which is the active part that performs filtration, and a tubule, which return the healthy substances back to the blood while removing waste (figure 2.2 (a)).

The glomerulus allows only smaller molecules to pass into the tubule like wastes and water, while the bigger molecules, like proteins and blood cells, remain into the vessel.

The tubule performs the second filtration: an inner blood vessel reabsorbs all the water with nutrients, while the remaining part become urine to be excreted.

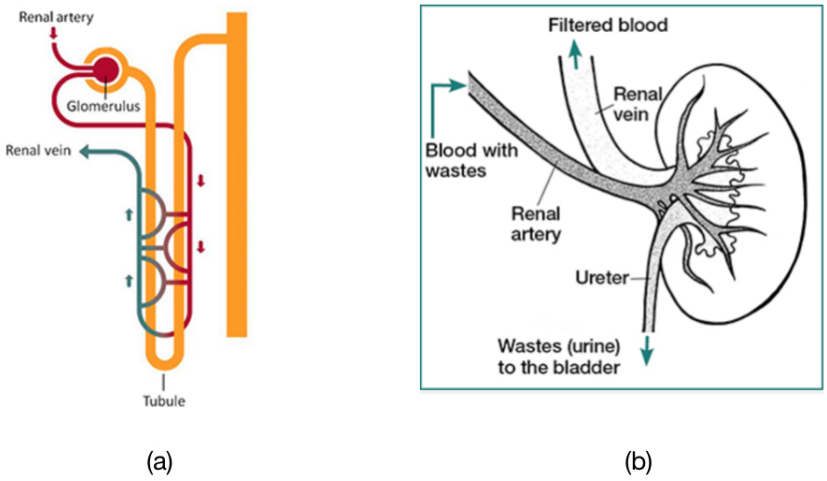


Figure 2.2: Nephron shape and structure (a). Blood flows into human kidney

The blood arrives to the kidneys through a network of branches starting from the renal artery, it is full of waste and it reaches the nephrons, where it can be filtered and then it flows out through the renal vein (figure 2.2 (b)).

These organs are able to filter over 140 litre of blood every day, extracting 1 or 2 litres of urine.

2.1 Kidney functions

Kidneys have different functions in human body: they remove the waste from blood and transform these substances into urine to be expelled.

Substances, removed by kidneys, include:

- urea, it is the result from the breakdown of proteins;
- uric acid resulting from the breakdown of nucleic acid.

Kidneys reabsorb nutrients, like glucose, amino acids, bicarbonate, in order maintain homeostasis [4].

Other functions include maintaining pH level (along with the lung) within the human body: kidneys rigenerate bicarbonate from urine and, based on pH level, they can release or retain it.

Moreover in human body, kidneys are responsible on the regulation of osmolality, which is a measure of electrolyte-water balance. If the level of osmolality rises, the kidneys can increase the concentration of urine, it can intensify reabsorption of water in order to regulate osmolality.

Finally, kidneys are responsible of blood pressure, because they produce erythropoietins, which are linked with the production of red blood cells; renin, which controls the size of arteries and the volume of blood plasma; and calcitriol, which increases the calcium.

Stage	Description	GFR* Level
Stage 1	Kidney damage with normal or high GFR	90ml/min or more
Stage 2	Kidney damage and mild decrease in GFR	60 to 89mL/min
Stage 3	Moderate decrease in GFR	30 to 59mL/min
Stage 4	Severe decrease in GFR	15 to 29 mL/min
Stage 5	Established kidney failure (ESKD)	Less than 15mL/min

Table 2.1: 5 Stages of Chronic Kidney Disease (* GFR means Glomerular Filtration Rate)

2.2 Chronic kidney disease (CKD)

As explained, kidneys play an important role in human body and their good functionalities are essential to maintain a good status of life.

Unfortunately, there are different possible diseases that can compromise their status.

Although, the risk connected with kidneys diseases are well-know, it is very difficult to reduce the number of death every year due to kidneys failure, because it is an underdiagnosed public health disease [5].

Many people do not know to be affected by some form of kidney disease, because the symptoms are difficult to evaluate, especially in early stages [6].

This continuous increase in the number of end-stage patients and death due to CKD has demanded new early detection diagnostic procedure.

Chronic diseases of kidneys start when some nephrons stop working properly, this damage can expand to other nephrons, in this case kidneys lose their ability to filter blood and waste starts to accumulate into the body.

There are different causes of kidney disease; the two most commons are: high blood pressure and diabetes [7].

Other minor causes are recurrent kidney infection or due to prolunged obstruction of the urinary system; in other case, CKD is inherited; finally smoking and use of drugs can increase the risk of chronic kidney disease.

Kidney disease is grouped in 5 stages (see table 2.1) [8], not everyone knows about their failure until it reaches higher stages, but progressively and silently the functionality starts to reduce. It is called kidney failure when kidney functions start to decrease their efficacy. This early stage affects the people life and it can turn into CKD.

Chronic kidney disease is a condition in which kidneys are losing their normal functions. The final stage is called End-Stage Renal Disease (ESRD).

Symptoms of CKD [7] may include:

- Fatigue, or tiredness
- Increasingly frequent need to urinate, especially at night
- Itching
- Nausea
- Shortness of breath
- Swollen feet, hands, and ankles

2.3 Clinical target and Cost

According to the U.S. Centers for Disease Control and Prevention [9], in 2018 about 30 million of American people are affected by kidney disease at different stages.

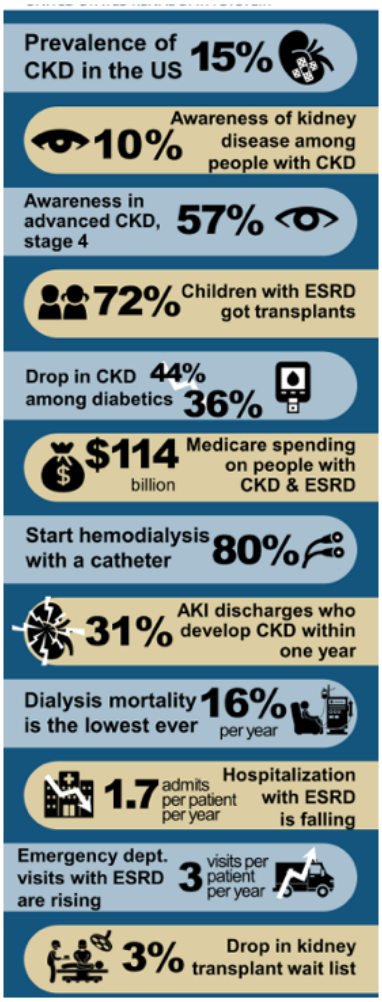


Figure 2.3: USRDS 2018 Annual Data Report

It represents over 10% of total population in USA. Moreover, 96% of people with CKD at early stages do not know their disease, while 48% with severe failure are not conscious about their problems.

In 1990, CKD were at position 27^{th} in the list of principal causes of deaths worldwide; the rank rose in 2010 to the number 18th, with an alarming increasing that was second only to HIV [10].

The national kidney foundation has reported that over 2 million of people in the world receive treatment for they kidney disease, but they represent 10% of all people who need treatments.

The spread of CKD is dramatical and the situation is more critical due to the cost of treatments. The long-term treatments are not affordable by low and middle economic

countries.

Cares like dialysis or transplantation are very expensive and they represent one of the highest percentage of health budget for many countries.

For example in USA, the treatments for CKD and ESRD have a cost of $114 billion$ in 2018 (data from: United States Renal Data System, figure 2.3). In Australia the cost of treatments for each patient is estimated between 50000 and 80000 AUD per year for a total of 12 billion AUD expected in 2020. In China,$558 billion$ are the money they economic loss in a decade due to death and disability for chronic cardiovascular or renal disease. Finally, in England, the renal disease are more expensive than breast, lung, colon and skin cancer combined [11].

People who receive treatment are concentrated in only five states: United States, Japan, Germany, Brazil, and Italy. In many other countries the treatment remains unfordable and it becomes very difficult to receive a renal replacement therapy, causing the death of all these final stage patients [12].

2.4 Current Treatments

The aim of treatments for CKD is only to reduce the progress failure of kidneys and prevent critical failures like ESKD. There is no cure for renal chronic disease.

In the early stages, it is recommended a regular life and a proper diet along with some medicines, in order to help kidneys in their regulation and filtration actions.

When condition becomes more critical and the waste starts to accumulate within the body, renal replacement therapies (RRT) are necessaries.

The waste must be removed from the human body, this is possible through long term dialysis. There are two types of dialysis:

- Hemodialysis
- Peritoneal dialysis

Finally, when the disease reaches the final stage and it turns to ESKD, a transplantation is required: it is a surgical treatment that can restore the normal functioning of the kidneys and it means a normal life for the patients.

2.5 Dialysis

Dialysis is a long term treatment for severe stages of kidneys disease. It is required when the kidneys have lost 85 to 90% of their functions [13].

Healthy kidneys filter from 113 up to over 144 litre of blood every day [3], but the waste remains in the blood when they do not work properly and it can be cause of coma and death.

Dialysis prevents the accumulation of waste within human body before reaching critical levels [14]; it replaces kidneys functions, such as:

- removing waste, water and unnecessary salt;
- maintaining the correct levels of sodium, bicarbonate, potassium and other chemicals;
- preventing regular blood pressure.

Dialysis helps patients to have a good level of life and it increase, up to 20 years, life expectation of people affected by renal disease.

Dialysis has also other side effects; for example, people who depend on dialysis may suffer of muscle cramps, low blood pressure, fluid overload and sleep problems.

There are two types of dialysis:

1. Peritoneal dialysis
2. Hemodialysis

2.6 Peritoneal dialysis

During a peritoneal dialysis [15], a dialysate solution is filled inside the peritoneal cavity through a tube. The solution is rich in minerals and glucose, it remains inside the peritoneal cavity for some time in order to absorb waste through an osmosis process. When the absorption process is completed, the solution is drained out from the abdomen and it is discarded (figure 2.4).

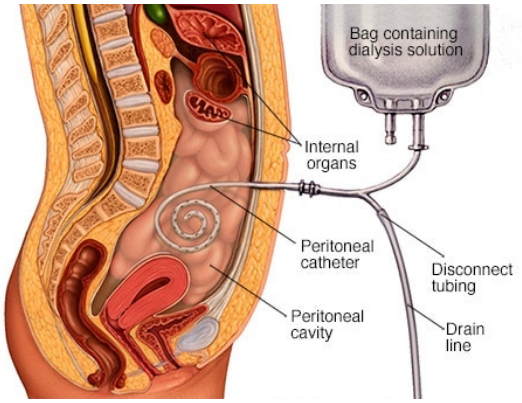


Figure 2.4: Principle of peritoneal dialysis. Photo from Mayo Clinic

A complete peritoneal dialysis is from 10 to 12 hours long, but it can be performed at home, at work or during travelling, so sometime it is preferred by patients, because it guarantees more freedom than hemodialysis.

Conversely, peritoneal dialysis costs of several cycles and it must be repeated many times per day, it is less efficient than hemodialysis, so it is not an option in case of severe CKD. Moreover, peritoneal dialysis can cause an infection of the abdominal lining.

2.7 Hemodialysis

Hemodialysis is the most common treatment for advanced kidney disease. Since 1960, hemodialysis has became a standard treatment for CKD; in recent years, new more effective and smaller machine were introduced in the market with the aim to reduce side effects of its long treatment [16].

People with chronic renal disease need hemodialysis 3 times a week and the treatment is 3 or 4 hour long; it can be performed in a special center or at home [17]. Hemodialysis performs extra corporal filtering of the blood.

The patient is connected, through a needle in the arm, to a circuit of tube; the blood flows from the vein to the external circuit and it goes to a filter (know as dyalizer), which removes the waste and the other unwanted fluids. The cleaned blood exits from the filter and it returns into the body.

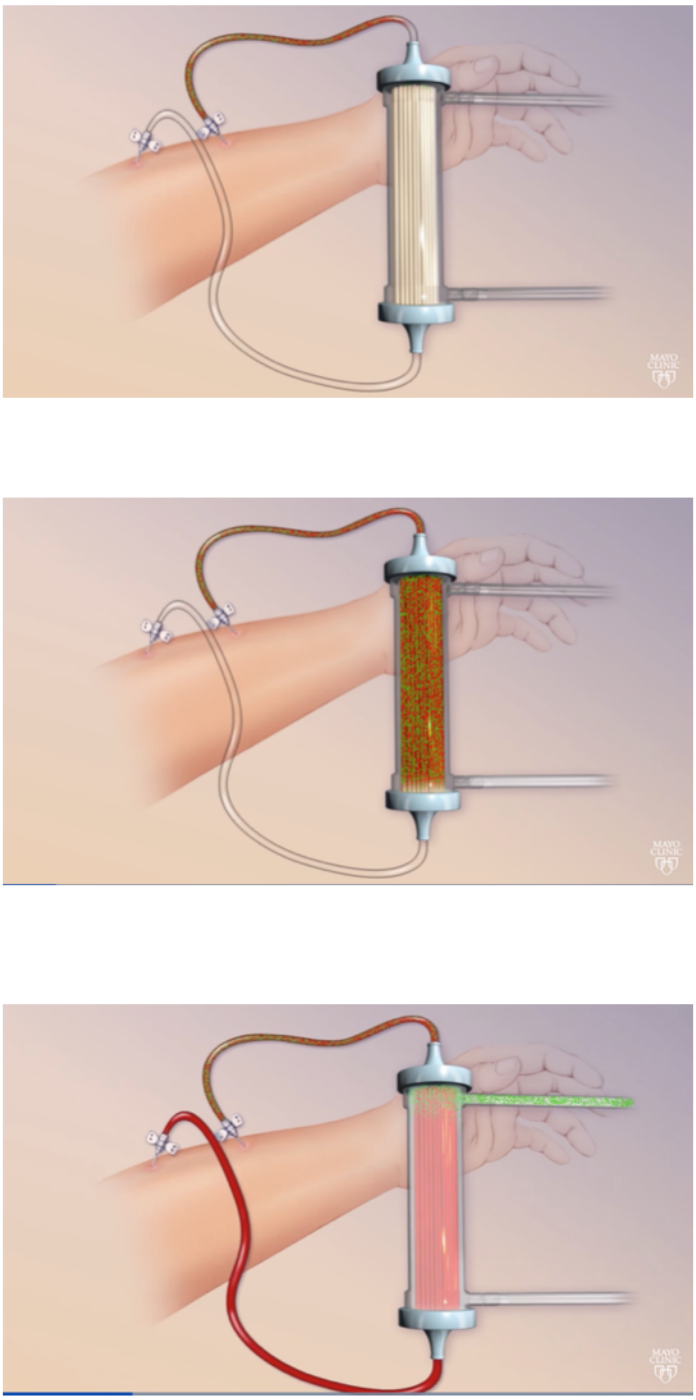


Figure 2.5: Hemodialysis steps. Pictures from Mayo Clinic

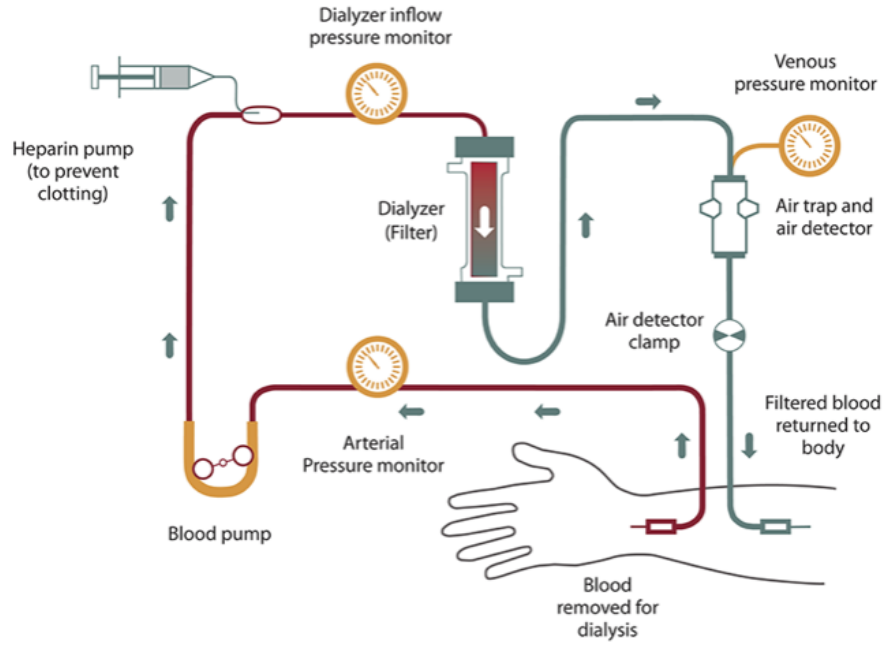


Figure 2.6: Components of a hemodialysis machine

2.8 Hemodialysis machine

Hemodialysis is performed with a complex setup that is continuously evolving. Nowadays, the system is optimised in order to facilitate the use for doctors and to reduce the pain for patient [18]. The machine has three aims:

1. Pump blood;
2. clean the blood from waste;
3. monitor blood pressure and the rate of filtration.

The system pumps the blood outside the human body; the blood flows thanks to a peristaltic pump, which drives the flowing rate of fluids in the machine.

The blood circulates along a series of hematic tubes and it enters into the dialyzer. The dialyzer [18] is an artificial filter which contains fine fibres with semi-permeable membranes.

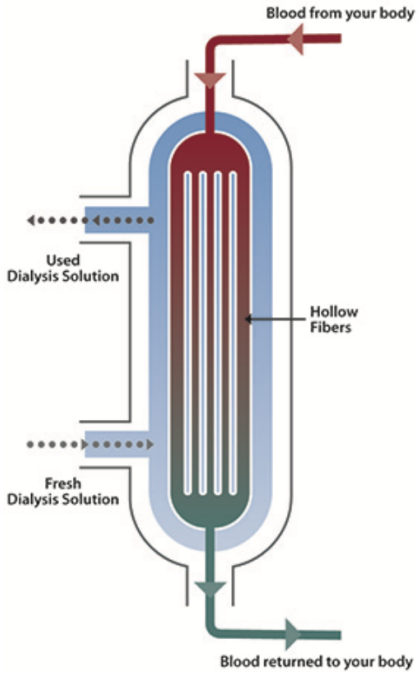


Figure 2.7: Dyalizer filter

The toxins are removed through diffusion process [17]: a fluid called dialysate is introduced into the filter and it bathes the fibres from outside; due to the semi-permeable membranes the fluid naturally crosses the membranes and it mixes with the dialysis fluid. The process depends on the different concentration of molecules; the blood cells and the proteins are too big to cross the fibres and they continue the flow outside the filter.

Along the hematic circuit, different sensors monitor the status of the patient (figure 2.6). For example, blood pressure sensors are introduced in different points in the extracorporeal circuit, other sensors include: dialyse pressure, temperature, O_2 saturation, dialyzer membrane pressure.

These sensors give all the information in order to adjust the machines settings during the treatment and to control the status of patient.

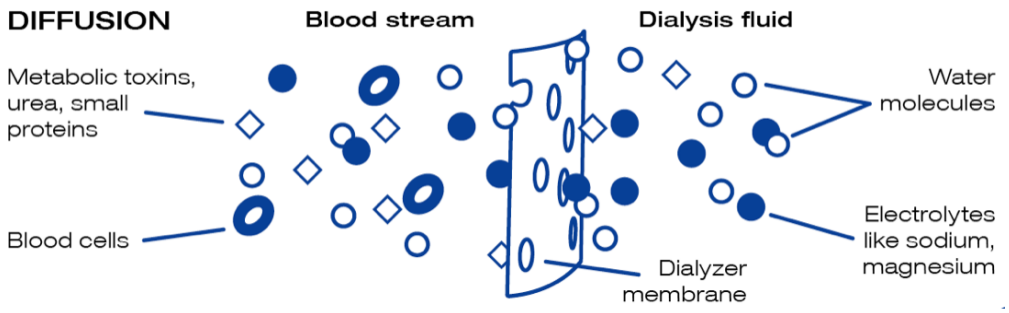


Figure 2.8: Diffusion process through dialyzer membrane

2.9 Standard procedure for hemodialysis

During the treatment, the patient remains sit in a chair while all the filtering operation is ended, meanwhile he is free to make simple actions (watching tv, speaking, ecc..) or even sleep [19]. The procedure consists in 5 steps:

1. Preparation: in this phase, some checking operations are performed before starting; weight, blood pressure, pulse and temperature are recorded and the vascular access is cleansed.

2. Starting: the two needles are inserted in the arms of the patient. The needles are connected to a series of tubes that connect to the dialyzer. One tube finishes at the inlet point of the dialyzer, where the blood arrives full of waste; while the second tube starts from the outlet point of the dialyzer, the cleaned blood finishes into the second vascular access.

3. Symptoms: many times, patients feel uncomfortable during hemodialysis, they can suffer of nausea and abdominal cramps. In this case, some operators can change some parameters of the machine to change the flowing of the blood or the medication.

4. Monitoring: this process is really important to reduce the side effects of hemodialysis. During the treatment only some parameters are monitored, like blood pressure and heart rate. The operators do not have enough information about the real-time status of the patient in terms of haemoglobin or oxygen saturation, it has to extract a sample of the blood from the patient and it analyses the hematic values.

5. Finishing: when the blood is cleaned, the treatment stops and the needles are removed. After some other checks on blood pressure and weight, the treatment is finally finished.

In order to reduce the pain for the patient and to stop the treatment just when the blood is fully cleaned, the operators need real time information about the blood itself. The sensor, here proposed, will be very useful to reduce the side effects on the patient and to make easy operators actions.

2.10 Risks connected to hemodialysis

Hemodialysis helps people affected by kidney disease to prolong their lives, but the expected life is at least 20 years, because hemodialysis is a very effective care.

Althought recent technologies in hemodialysis machines have increased the efficiency and reduced the time for each session, the treatment is very expensive, long time consuming for patient and it can be painful [20]. These risks include:

- Low blood pressure
- High blood pressure
- Muscle cramps
- Itching
- Sleep problems
- Anaemia
- Bone diseases
- Fluid overload

Other risk for patients safe is due to over-treatment.

During each treatment, a medical operator takes a sample of blood from the extracorporeal circulation of hemodialysis and it proceed with a blood centrifugation to measure hematocrit level of the patient.

This operation can be repeated until the blood has reached the standard values and after that the hemodialysis session is terminated.

It is evident the importance of an opto-electronic sensor to monitor the hematocrits level of blood during the treatment. Moreover, the monitoring setup here presented, allows the real-time analysis of hematocrits and oxygen saturation without stopping the circulation and avoiding any contact with the blood to prevent contamination.

The setup is integrable with hemodialysis machine: it is introduced along the tubes of hemodialysis close to the other sensors.

The information provided is useful for the operator, because it has a real indication of the purification level of blood and it can stop the treatment at the optimal time, without any further time loss for patient and reducing the risk due to an unnecessary long hemodialysis.

2.11 Other sensors on market

There are other optical sensors that can provides some information about hematic properties of the patient during hemodialysis. In this section, the working principles of the devices based on approaches different from the one proposed in this Thesis, i.e. spectroscopy, are discussed.

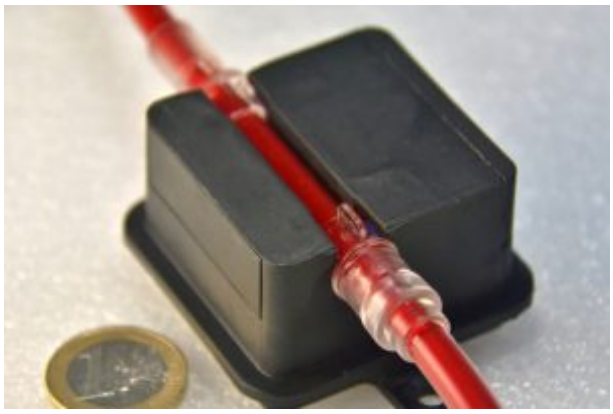


Figure 2.9: Picture of Obba optical analyzer made by Datamed

2.11.1 Obba made by DataMed

Obba [21] is an optical sensor to measure Hct, oxygen saturation, haemoglobin and temperature and it can be utilised in any different medical treatments where the extra corporeal circulation of blood is involved. This device is made by Datamed S.r.l and it is protected by IP. The technical specifications of the device are:

	Ranges	**Accuracy**	**Resolution**
Hematocrit	15 - 50% Hct	$\pm$ 5 Hct units (3σ)	0.1%
Hemoglobin	5 - 15 g/dl	$\pm$ 1.65 Hb (g/dl) (3σ)	0.1 g/dl
Oxygen Saturation	40% - 99% O_2 Sat	$\pm$ 5 O_2 Sat units (3σ)	0.1%
Blood Temperature	10 - 40^{o}C	$\pm$ 0.5^{o}C (3σ)	0.1^{o}C

Table 2.2: Specifications

This system involves optical measurements of blood in the range of visible and infrared in order to measure and monitor the hematic parameters.

Obba [22] exploits four different LEDs that emits light at different wavelength:

- LED A suitable for emitting light to 805 nm;
- LED B suitable for emitting light to 660 nm;
- LED C suitable for emitting light too 1450 nm;
- LED D suitable for emitting light too 1550 nm.

These narrow spectrum LEDs are switched on at different times to provide the parameters calculation.

Each spectrum is detected by InGaAs sensor which is placed opposite to the LEDs. This sensor is able to measure the amount of light absorbed by blood sample within a range from 600 nm up to 2600nm.

In between LEDs and InGaAs, there is a disposable (called CB002) made in PETG (Polyethylene terephthalate with Glycol) which is the measurement window along the tubes. The sensor and the LEDs are placed close to the disposable in order to reduce the noise from environmental light.

The measurements of hematocrits and oxygen saturation involve electromagnetic radiation of at least two LEDs at different wavelengths and then the radiation diffused in the blood is detected. The radiation diffuse is then normalized with a reference value; finally the parameter will be the ratio between the intensity of the electromagnetic radiation detected.

For example, Hct is evaluated through the ratio between three LEDs:

$$R_{Hct} = \frac{I\chi(A)}{I\chi(C) + I\chi(D)} \tag{2.1}$$

Where $I\chi(A)$, $I\chi(C)$ and $I\chi(D)$ are the intensity of detected electromagnetic radiation approximated with a mathematical function of the second order.

Finally, an empirical correlation function is applied to measure the level of Hct.

Same mathematical approach is performed for the measurement of oxygen saturation:

$$R_{sO_2} = \frac{I\chi(A)}{I\chi(B)} \tag{2.2}$$

In (2.2), $I\chi(A)$ and $I\chi(A)$ are the intensity of two detected electromagnetic radiations approximated with a mathematical function of the second order.

Figure 2.10: Picture of Cirt-Line 3 monitor system

2.11.2 CRIT-LINE made by Fresenius

Crit-Line [23] is a non invasive sensor which provides levels of hematocrit, oxygen saturation and percent change in blood volume.

The system has been developed by Fresenius Medical Care, a leader in renal therapies.

The device includes a photo detectors, a light system and a blood chamber, which is a disposable attached to the line of blood, it provides the optical viewing point for the measurements [24].

Both light and sensor are included in a custom clip, which covers the blood chamber, reducing the noise from external light and assures repetitive measurements.

The light system is composed by 3 electrodes:

- LED 1 emits light with a narrow spectrum with a peak around 660 nm;
- LED 2 emits light around 810 nm, this is the asbestic for re blood cells;
- LED 3 has peak around 1300 nm, which is asbestic for water.

LED 2 and LED 3 are involved in the measurement of hematocrit, while oxygen saturation is evaluated with LED 1 and LED 2. Crit-Line has two sensors:

1. A silicon photodetector to detect absorbed light emitted from LED 1 and LED 2;
2. an InGaAs sensor to detect the intensity of the light provided by LED 3.

The device evaluates the level of hematic parameters as ratio of intensity light that is not absorbed by blood sample:

$$Hct = f\left[\frac{ln(\frac{i_{810}}{i_{0-810}})}{ln(\frac{i_{1300}}{i_{0-1300}})}\right] \tag{2.3}$$

$$sO_2 = g\left[\frac{ln(\frac{i_{660}}{i_{0-660}})}{ln(\frac{i_{810}}{i_{0-810}})}\right] \tag{2.4}$$

In (2.3) and (2.4), f and g are empirical formulae; i_{660}, i_{810} and i_{1300} are the intensity light respectively emitted by LED 1, LED 2 and LED 3, while i_{0-660}, i_{0-810} and i_{0-1300} are absorbed lights at the respectively wavelengths by blood sample.

Device specifications are:

Oxygen Saturation	**@Hct**	**Acc within ±3%**	**Acc within ±5%**
Instrument Range	45 - 60	60 to 100	50 to 100
and Accuracy	20 - 45	50 to 100	30 to 100
	10 - 20	Not specified	40 to 100
	Range	**Accuracy**	
Hematocrit	10 to 60	±1	

Table 2.3: Crit-line range parameters and accuracy

CHAPTER 3

Absorbance spectrum of blood

Absorbance spectroscopy is a qualitative and quantitative measure of absorption radiation as a function of wavelength.

Thanks to its sensitivity, which is the ability to detect small quantities of compound, and selectivity, which is the ability to distinguish different compounds in a solution; absorbance spectroscopy becomes very popular as analytical chemistry tool, especially in biomedical applications.

For example, microbiology medicine exploits the advantages of spectroscopy [25] along with applications for cancer diagnosis [26].

Moreover, spectroscopy is commonly used in pharmaceutical analysis [27], because it is non-contact and non-destructive for the samples.

Some principles are introduced in order to give basic knowledge about spectroscopy, focusing on visible and near-IR spectroscopy, which is the range of wavelengths covered by the sensor used in this work.

In the second part of the chapter, the optical and physical characteristics of blood are described.

3.1 Theory of absorbance spectroscopy

Electromagnetic radiation is composed by electric (E) and magnetic (M) fields, these are oscillating waves which are oriented perpendicular to each other.

The energy of the electromagnetic radiation and the wavelength are related by the

expression:

$$E = \frac{hc}{\lambda} \tag{3.1}$$

where h is Plancks constant (6.62×10^{-34} Js) and c is the speed of light (3×10^{8} m/s).

During spectroscopy analysis, the electromagnetic radiation of multiple wavelengths is directed to a sample, it interacts with the medium and some energy is absorbed in a quantized manner while some other passed unaltered.

The energy absorbed in the range of visible and near-IR light results in changing of energy levels of electrons in the molecules sample.

This interaction is very selective, because the light energy has to be equal to the energy required for a specific electronic transition, otherwise it is not absorbed [28].

Thanks to this energy, electrons in the molecules jump from a ground state to an excited one.

The orbitals in the ground state are σ, π and n these are bonding orbitals, while the transitions involve the anti bonding orbitals, which are σ^* and π^* [28].

The allowed transitions are showed in figure 3.1.

For example, the transitions $n \rightarrow \sigma^*$ and $\sigma \rightarrow \sigma^*$ require energy of short wavelength (150 and 200 nm), while the transitions $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ are characterised by higher absorbrvity [29].

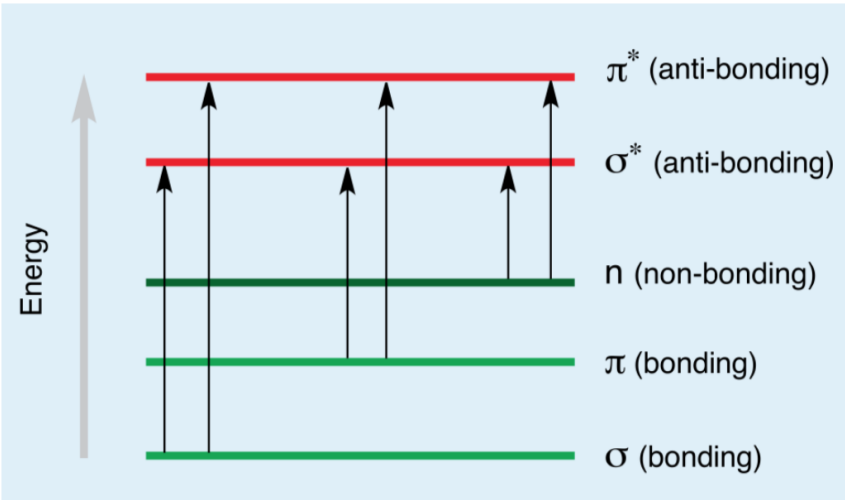


Figure 3.1: Electronic transition of π, σ and n electrons

3.2 Interpretation of an absorbance spectrum

Spectroscopy is a non destructive measurement technique for quantitative and qualitative analysis of matter through its interaction with electromagnetic radiation.

Successive radiation frequencies are used to scan a sample and measure which frequencies are absorbed and which are transmitted.

Detectors compare the attenuation of transmitted light with the incident one.

The outcome of this analysis is the resulting spectrum, which represents the amount of radiation absorbed as a function of the wavelength (expressed in μm) or, alternatively wavenumber (expressed in cm^{-1}).

The spectrum is full of information: it is considered the fingerprint of a substance because it is unique for each material. It is directly connected to the electronic and molecular composition of sample, because different substances absorb different wavelengths of light.

In an absorbance spectrum, there are some wavelengths of lights which are absorbed (peaks) and other wavelengths that are transmitted (troughs).

The presence of peaks and troughs allows the identification of group of atoms which are present or absent in the samples, while the quantitative analysis depends on the intensity of peaks, because these information are linked to the quantitative of light absorbed thanks to Lambert-Beers law.

Each absorption band is linked to a specific functional group and it is proportional with the concentration of that group within the sample.

3.3 Beer Lambert Law

In absorption spectroscopy, the Beer Lambert law [30] relates the absorption of light with the properties of the material where the light passes through.

According to Beer Lambert law, the absorbance is directly proportional with the concentration of a substance in a sample.

Beer Lambert law can be expressed as:

$$A(\lambda) = \epsilon(\lambda) l c \tag{3.2}$$

where:

- $A(\lambda)$ is the absorbance value at λ wavelength,
- l is the optical path or the dimension of the cuvette,
- c is the concentration of solution,
- $\epsilon(\lambda)$ is the extinction coefficient at λ wavelength

Transmittance is defined as the ratio between incident and transmitting radiation:

$$T = \frac{I}{I_0} \tag{3.3}$$

While the absorbance is equal to the logarithmic ration of the two radiations:

$$A = log_{10}\frac{I_0}{I} \tag{3.4}$$

where:

- I_0 is the intensity of incident radiation at a given wavelength passing through the sample
- I is the remaining transmitting radiation exiting from the sample

According to (3.2), (3.3) and (3.4), Beer Lambert law can be expressed as follows:

$$A = \epsilon lc = -log_{10}\frac{I}{I_0} = log_{10}\frac{1}{T} \tag{3.5}$$

Equation (3.5) expresses the linear relation between absorbance and concentration, but it is valid only under certain condition. This relation assumes that each photon is either absorbed or transmitted when it encounters an absorbing particle of the matter, but the scattering occurs in a real situation.

In Lambert Beer law, the phenomenon of scattering is ignored, this is predominant at high concentration solutions, where the linear relation is not valid and it becomes:

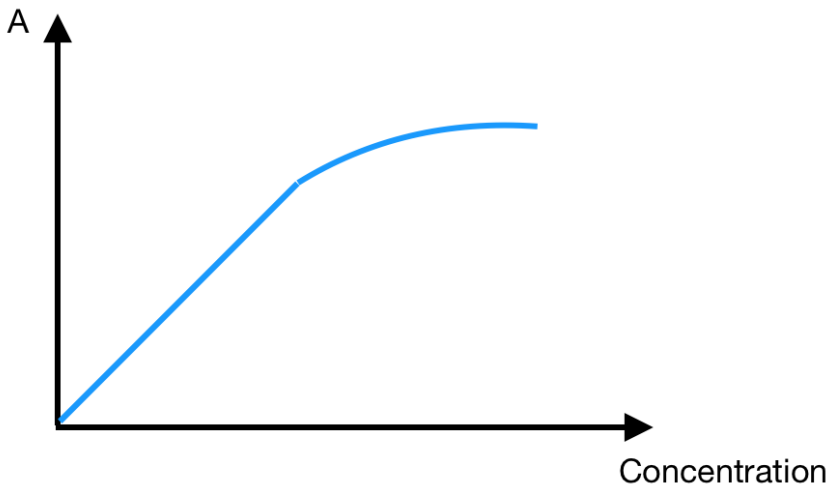


Figure 3.2: Effect of concentration on Lambert-Beer law

Scattering [31] is a physical phenomenon where radiations are deviated due to non-uniformities in the medium in which it passes through.

This phenomenon can be expressed as a series of elastic collisions between photons and particles, it results in a longer path along the material.

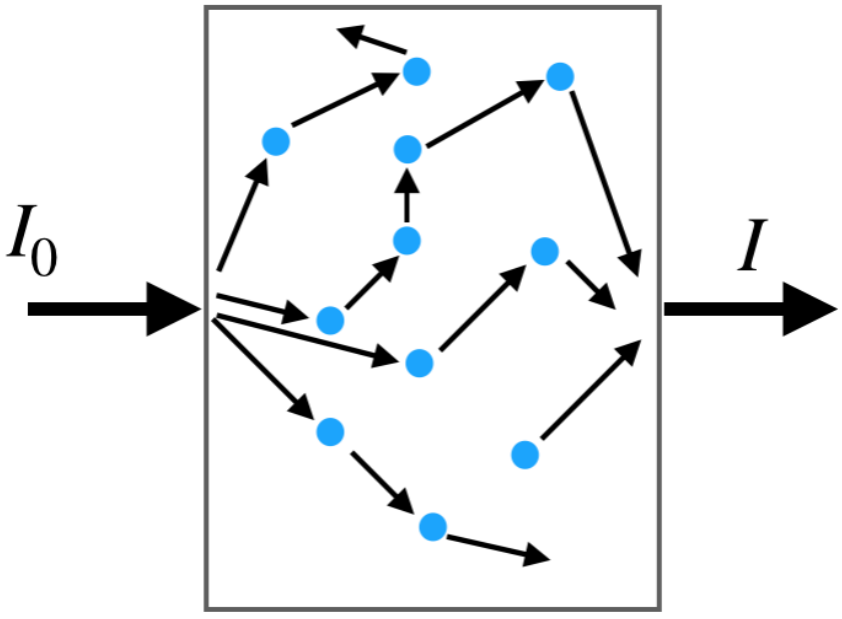


Figure 3.3: Propagation of photons in a scattering material

Scattering of light, in a non absorbed medium with a thickness length equals to L, modifies the absorbance of the material, which can be expressed as:

$$A(\lambda) = log_{10}\frac{I_0}{I} = \mu_s(\lambda)L \tag{3.6}$$

where:

- I_0 is the initial radiation of light,
- I is the transmitted light exiting the medium,
- μ_s is the probability of a photon to be deviated in its trajectory.

When absorbance spectroscopy is performed on a sample which does not satisfy Beer law's conditions, both phenomena of absorption and scattering have to be considered.

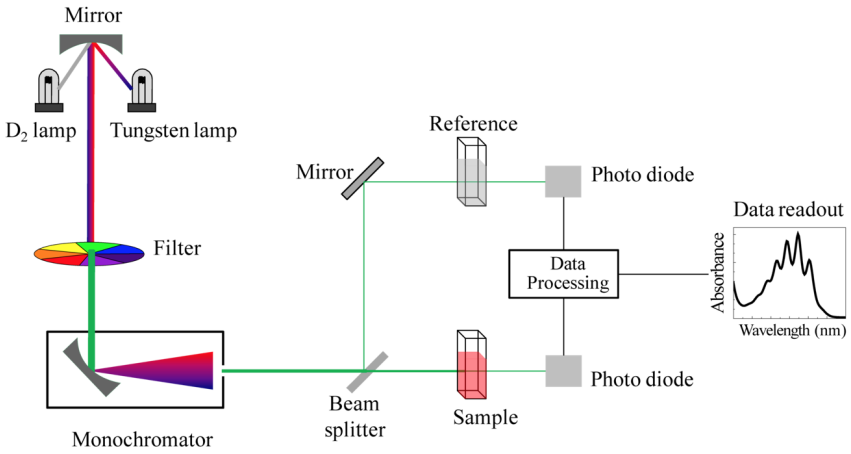


Figure 3.4: Block diagram of a double beam spectrophotometer

3.4 Spectroscopy measurement

Absorbance spectroscopy measure is performed through a spectrometer. A spectrophotometer consists of some basic components, which are:

- light source, generally it is a deuterium or halogen lamp;
- a monochromator allows only certain wavelength to go through the aperture, otherwise all wavelengths go to the sample;
- the light passes through the sample which is contained into a cuvette, which is invisible to light, so the cuvette does not absorb radiation;

- a photodiode, that is placed after the cuvette and it records the light that is not absorbed by sample;
- many spectrophotometer have a beam splitter, it splits the light into two radiation energies and two measurements are contemporary performed: one on a reference and the second on a sample.

The sample does not require any preparation; this analysis does not involve any risk and it is not destructive for the sample itself.

In this work, a micro spectrometer detector is used, this has the same principle of photodiode present in the schematic before, but the lamp is not integrated into it.

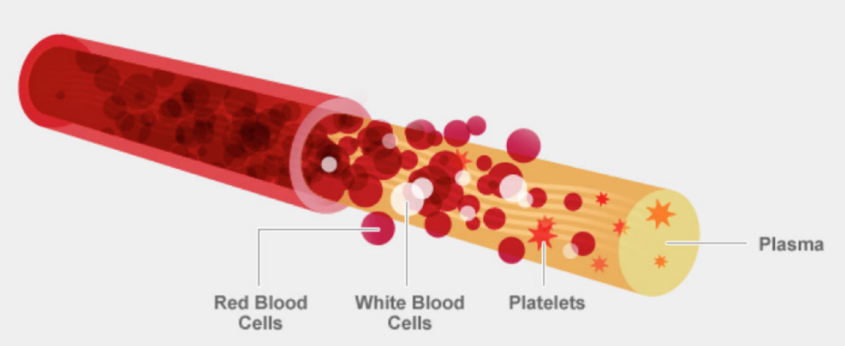


Figure 3.5: Model of blood

3.5 Physiology and Optical Response of Human Blood

Blood is an essential fluid for human life; it carries oxygen and nutrients to living cells. An average adult has more than 5 litres of blood in his body [32]. Blood has many different functions, most important are:

- it transports oxygen and nutrients to living cells;
- it is responsible for the formation of clots to reduce blood loss;
- it carries cells and antibodies to contrast infection;
- it is responsible of temperature regulation;
- it also brings waste to the kidneys and liver.

Blood is composed by four main components:

- plasma

- red blood cells
- white blood cells
- platelets

White blood cells, or leukocytes, represents only 1% of total volume of blood, they protect from infection and contrast viruses and bacteria.

Platelets, also called thrombocytes, are the smallest part of the blood, they have the function of clotting when an injury occurs. Coagulation prevents excessive blood loss, moreover fibrin, which is produced during coagulation, promotes the creation of new tissue.

Plasma is the liquid part of blood, it is a mixture of water with sugar, fat, protein and salts. It represents $55 - 60\%$ of the total volume of blood. The aim of plasma is to carry blood cells, nutrients, proteins and waste throughout the body.

Red Blood Cells, also known as erythrocytes, represent $40\% - 45\%$ of total blood volume, they represent the most common element of blood.

One drop of blood normally contains millions of erythrocytes and thousands of leukocytes. The hormone called erythropoietin is responsabile for the production of red blood cells, erythropoietin is produced by the kidneys; normal life cycle for erythrocytes is 120 days long [33].

The principal function of red blood cells is to transport oxygen from lungs to the rest of the body, meanwhile they pick up carbon dioxide from tissue and they transport back to the lungs.

Red blood cells are small cells with a diameter of about $7 - 8\mu$m with a characteristic biconcave shape with flattened center.

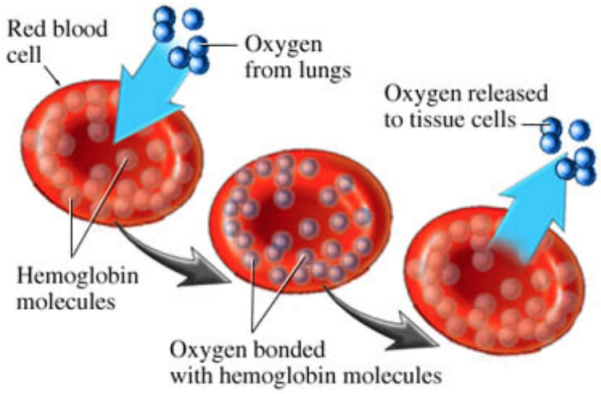


Figure 3.6: Process of oxygen carried by red blood cells

This shape enhances gas exchange thanks to a bigger surface area. The oxygen is carried by the red blood cells, then it is released into the plasma and finally, it reaches the cells. At the same time, carbon dioxide is picked up by the erythrocytes.

Red blood cells do not have nucleus, but they contain a protein called haemoglobin, which is responsible to transport oxygen within the body.

3.6 Hemoglobin

Hemoglobin [34] is a big molecules formed by proteins and iron; it is the responsible of red colour of blood, its aim is to carry oxygen throughout the body.

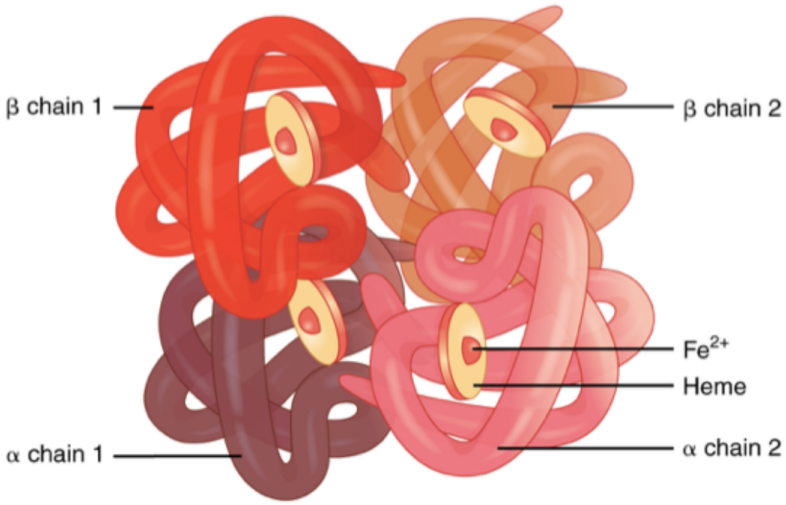


Figure 3.7: Structure of hemoglobin

The structure consists of four chains of protein, called globin, connected with red molecule called heme, which contains iron (Fe^{2+}).

The iron ion contained in the heme group can create a connection with one molecule of oxygen; therefore, each hemoglobin can transport four oxygen molecules.

About 300 million of hemoglobin molecules are contained in a single erythrocyte, thus 1.2 billion oxygen molecules can be transported [35].

When the iron ion binds with oxygen, it forms the oxyhemoglobin (or HbO_2), this chemical process happens in the lungs. The oxyhemoglobin is characterised by a bright red colour. The colour changes when the molecule releases the oxygen and it becomes deoxyhemoglobin (or HHb) and it turns to a characteristic darker red.

Meanwhile, CO_2 enters in the blood stream: this is waste that must be expelled; about $76 - 77\%$ dissolves in plasma, while $23 - 24\%$ binds with hemoglobin, it forms carbaminohemoglobin molecule, which is release in the lungs [35].

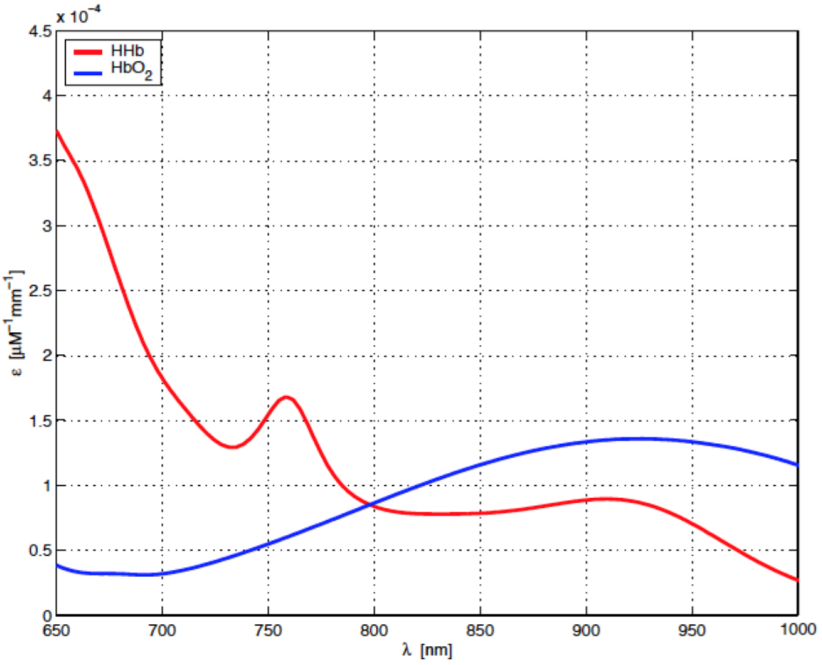


Figure 3.8: Molar absorption coefficient of oxyhemoglobin and deoxyhemoglobin

3.7 Oxygen Saturation

Patients with insufficient hemoglobin in blood cells are affected by anaemia and the tissues does not receive sufficient oxygen.

The determination of percentage of oxygen of tissues is knows as oxygen saturation. This values is commonly monitored in healthcare through an instrument called oximeter. This device exploits lights at two different wavelengths (normally one red and one infrared lights) and it evaluates absorbance with a photodetector.

Standard values of oxygen saturation are between 95% up to 100%, while lower percentages can be symptom of hypoxemia ($< 93\%$ of oxygen saturation) [36].

Normally, kidneys filter about 180 litres of blood, each day, in response to hypoxemia, less oxygen arrives to the kidney; in this case, kidneys secretes Erythropoietin (EPO) to produce more erythrocyte in order to restore the right level of oxygen.

Critical condition of hypoxemia can be connected to disease such as asma, lung cancer or chronic obstructive pulmonary disease.

Oxyhemoglobin and deoxyhemoglobin absorb light in a different way, so concentrations of Oxyhemoglobin or deoxyhemoglobin result in different optical absorbance spectra.

This different optical behaviour is showed in figure 3.8, where the coefficient of molar extinction is showed.

It is therefore possible to measure the ratio between oxyhemoglobin and deoxyhemoglobin in blood exploiting their different optical behaviour.

3.8 Hematocrit

Hematocrit (or Hct) is the percentage of red blood cells over whole blood volume, so it is possible to define it as:

$$Hct = \frac{V_{red\ blood\ cells}}{V_{tot}} = \frac{V_{red\ blood\ cells}}{V_{red\ blood\ cells} + V_{plasma}} \tag{3.7}$$

where $V_{red\ blood\ cells}$ is the volume of the red blood cells, while V_{tot} is the total volume of blood, that can be approximate as the sum of V_{plasma} and $V_{red\ blood\ cells}$, that are the main components of blood.

Red blood cells are mainly hemoglobin, that can be present in two different configurations: oxyhemoglobin and deoxyhemoglobin; while plasma is principally composed by H_2O.

The formula (3.7) can be rewritten as:

$$Hct = K * \frac{[HHb] + [HbO_2]}{[HHb] + [HbO_2] + [H_2O]} \tag{3.8}$$

where K is a constant which approximates the volumes with the corresponding concentrations. H_2O represents 90% of the three concentrations, so Hct is, approximately, equal to:

$$Hct = K * \frac{[HHb] + [HbO_2]}{[H_2O]} \tag{3.9}$$

The optical behaviour of oxyhemoglobin and deoxyhemoglobin is showed in figure (3.8).

H_2O is not completely transparent to light: it absorbs at near-infrared wavelengths, while the absorbance contribution in visible light is lower. The optical behaviour of water is showed in figure (3.9).

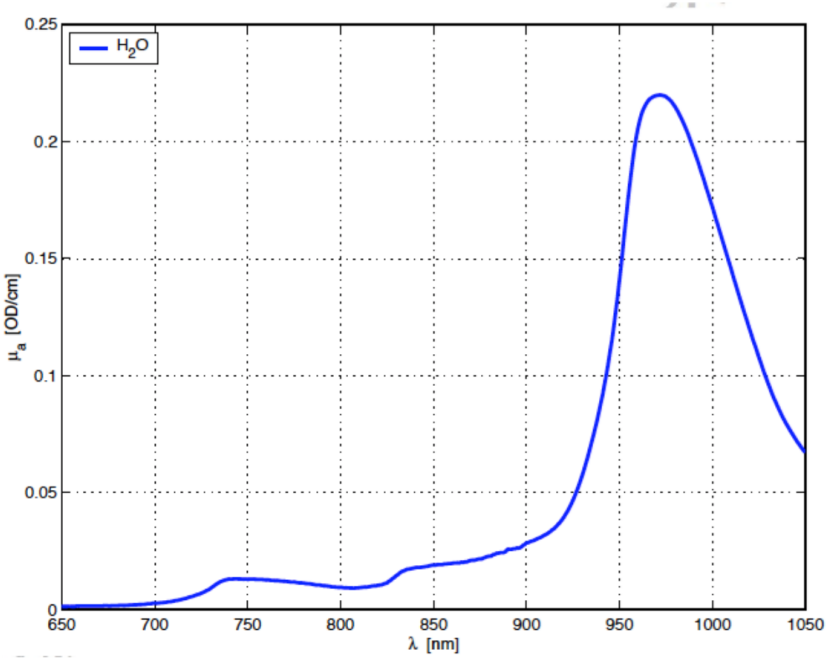


Figure 3.9: Molar absorption coefficient of H_2O

Generally, values of Hct are from 38 up to 50 for men and 35 up to 45 for women [37].

Measuring hematocrit is very important in healthcare, because its value can help doctor to make diagnosis or to monitor the response of a treatment.

Low values of hematocrit [37] can be connected to:

- anaemia
- infections
- leukemia or lymphoma
- vitamin or mineral deficiencies

Conversely, higher values of hematocrit than normal can be symptoms of:

- disorder in red blood cells production
- lung disease
- hurt disease

Usually, Hct is measured during a complete blood count (CBC) test: a blood sample is taken from the vein and it is analysed to evaluate hematocrit, hemoglobin, white blood cells, and the platelets throughout an automated haematology analyzer.

Another standard test to evaluate hematocrit is through centrifugation. The blood sample is placed in a thin capillary tube and then it is inserted in a centrifuge for 5 minutes at 10000 rpm in order to separate blood into layers: the red blood cells place at the bottom and they are well separated from the rest of the blood, so hematocrit is equal to the red blood cells height divided by the total fluid in the capillary tube.

CHAPTER 4

Machine Learning

Machine learning (or ML) is commonly used in academic and industrial fields; big companies, such as Google, IBM, Amazon, Facebook heavily invest in machine learning [38], these attention opens the ways to new possibilities and applications in this field.

People become familiar with machine learning applications and some of them are considered indispensable in everyday life.

These applications include:

- web page ranking used to find a page with search engines, like Google [39];
- security applications, such as face recognition [40];
- vocal assistant [41];
- fraud detection used by banks and credit companies [42].

Other applications will be achievable in few years, for example:

- self driving cars [43];
- voice recognition for people with degenerative diseases [44];
- improving healthcare with minimally invasive brain-machine interfaces [45].

4.0.1 Machine Learning in Healthcare

Machine learning has been successfully used in a wide range of healthcare or biomedical uses. Current ability to record massive amount of data has deeply changed healthcare and this has helped machine learning to find widespread applications in this field.

Every year, the US healthcare systems generates approximately one trillion of gigabytes of data; these data come from different sources, such as laboratory results, clinical records, medical imaging.

Machine learning can integrate data and then use in order to make every diagnosis and decision, or it can also provide personalized therapies.

Even though, ML require less human guidance, clinical experts must work together in order to include in the databases the relevant variables, data, examples and to find out the relationship between dependent and independent variables.

Many different applications of machine learning in medicine have demonstrated accurate results; for example: machine learning and big data are able to detect diabetic retinopathy and diabetic macular edema in retinal funds photography with as accuracy as human physicians [46]; or it is possible to identify skin cancer from images using intelligent models [47]; or predict cardiovascular attack based on patients characteristics, such as biometric data, clinical history and lab test results [48].

These are only some examples of machine learning results that have improved the life of patients, but more other results will be possible.

Despite of its success in many areas from speech recognition to autonomous driving vehicle, machine learning has encountered different impediments when applied to medicine.

These algorithms could save the life of millions of people, because they exploit personalized data along with samples from collective experience, because it presents unique challenges and scenarios. These obstacles are mainly due to the impossibility to have large and high quality data in order to correctly train the algorithms.

For example, machine learning for image recognition require massive amount of a data due to the complexity and variety of the tasks.

In other cases, data are sufficient, but they do not represent the entire possible scenario or they are not uniformly distributed.

4.1 Definition of Machine Learning

Machine learning is the science of programming computers in order to learn from data [49]. According to Arthur Samuel, a pioneer of self-learning computers, machine learning is:

[ML is] the field of study that gives computers the ability to learn without being explicitly programmed

Arthur Samuel, 1959

Machine learning changes the way of programming: instead of writing, line by line, a program for a specific task (figure 4.1), machine learning algorithms learn from a collection of examples where the correct outputs are already known (figure 4.2).

The program need a learning process and then it produces a model that works for new samples.

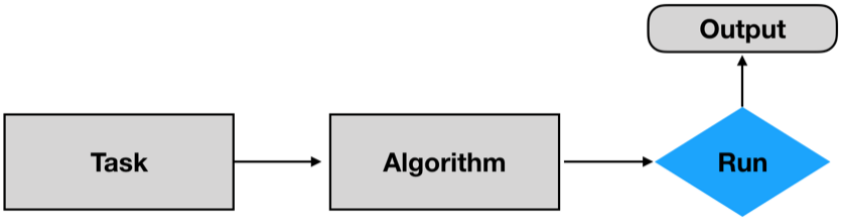


Figure 4.1: Standard way of programming

In last years, machine learning has become very popular due to the large amount of structured and unstructured data that are available.

Therefore, machine learning offers a way to build intelligent algorithms in order to transform data into knowledge and, iteratively, improve the performances of predictive models and make data-driven decisions.

The development of an accurate ML models is divided into stages; figure 4.3 shows a typical roadmap for building machine learning systems.

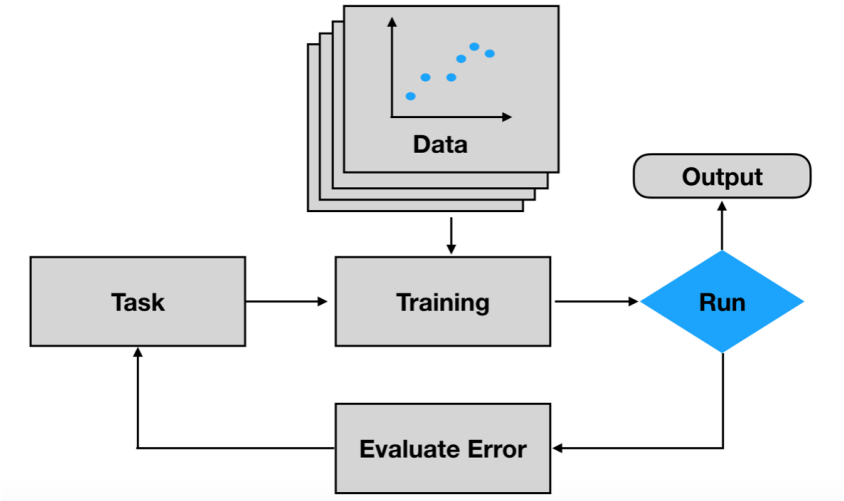


Figure 4.2: Machine learning new way of programming

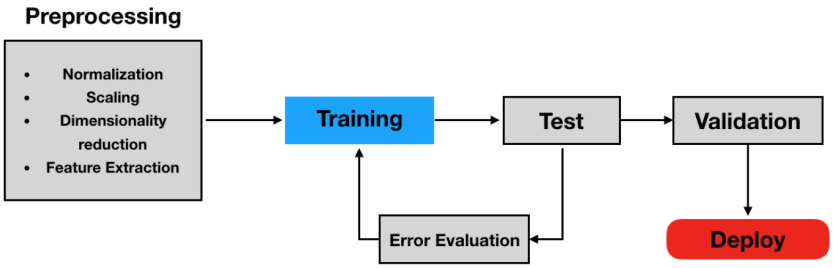


Figure 4.3: Roadmap for a standard development of machine learning systems

Preprocessing is a crucial step, because models are trained through data of different form and shape, so they must be manipulated.

It is important to use data on the same scale in order to improve predictivity of the model, so many machine learning algorithms require normalization or scaling of data before training.

In other cases, some data are redundant and they produce overfitting models: machine learning developers usually analyze the statistical correlation between inputs and eventually they reduce the dimensionality of the problem avoiding noisy data from the system.

Finally during preprocessing, the entire dataset is randomly divided into three subdatasets:

- Train dataset

- Test dataset
- Validation dataset

The first dataset is used during the second step of the roadmap in figure 4.3: training.

Training is an iterative process where each complete iteration is called epoch.

During epochs, a model is implemented and adapted through examples from training dataset. Test dataset is used to estimate how much accurate the model is with unseen examples, in order to evaluate the generalization error.

The iterative procedure of fitting and estimating error continues until accuracy performance is satisfied.

There are different machine learning algorithms, although there are some common rules for selecting the best algorithm, it is therefore essential to compare different predictive models in order to select the best performing one.

Moreover, each algorithm has different parameters to tune in order to increase predictive performance. These parameters are called hyperparameters and some optimization techniques help to fine-tune the performance of the model.

When the model satisfied the performance and the general error is acceptable, the validation step begins. The model is used to predict from new data, this is the final stage before deployment.

After testing and validation, the model is ready for deployment: it can be released and integrated into a real production environment.

4.2 Evaluation Metrics

In machine learning field, there are different parameters for the evaluation of model's accuracy [50]. They all compute the error between desired and predicted values.

The most used performance parameters are:

- Mean Squared Error (MSE) estimated over samples. It is defined as:

$$MSE(y, \widehat{y}) = \frac{\sum_{i=1}^{n-1}(y_i - \widehat{y_i})^2}{n} \tag{4.1}$$

- Mean Absolute Error (MAE) estimated over samples. It is defined as:

$$MAE(y, \hat{y}) = \frac{\sum_{i=1}^{n-1} |y_i - \widehat{y_i}|}{n} \tag{4.2}$$

- Coefficient of determination (r^2). It is mathematically defined as:

$$r^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n-1}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n-1}(y_i - \tilde{y})^2} \tag{4.3}$$

  where $\tilde{y}$ is equal to:

$$\tilde{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i \tag{4.4}$$

(4.3) provides a measure of how well future samples are likely to be predicted by the model: r^2 equals to 1 means the model can predict exactly every solution.

In (4.1), (4.2) and (4.3), y_1, y_2, $\cdots$ y_n, are n observed targets and $\widehat{y_1}$, $\widehat{y_2}$, $\cdots$ $\widehat{y_n}$ are the corresponding predicted values.

Calculation of MSE , MAE and r^2 allows a statistic evaluation of model's performance, giving a comparison between them and helping during selection of the algorithm.

4.3 Python for Machine Learning

Python has been used for this project during the implementation of ML models. Python is a powerful and versatile programming language; it has become the most popular for machine learning and data science [49], because it combines many advantages, such as being open-source with a big and dynamic community of developers, with clear syntax.

Moreover, there are plenty of libraries, which makes faster each step of the development of machine learning models.

For the realization of this work, some useful libraries have been used in order to reduce time cost. These libraries are:

- Numpy [51] is an open source library, which implements high-level mathematical functions for multi-dimensional arrays and matrices.

- Pandas [52] is a library developed for data manipulation and analysis. It allows the implementation of structured databases.

- Scikit-Learn [53] is one of the most popular open-source libraries for machine learning. It includes standard preprocessing functions and various classification and regression algorithms.

- Keras [54] is a high-level API, which allows the integration of python with popular machine learnings framework, such as Tensorflow, Theano, CNTK. Keras is intended for fast prototyping, academic research and production, it allows a rapid implementation of neural network, which is one of the algorithms used in this project.

- Talos [55] is a hyperparameter tuning library for Keras. It automatically configures, performs and evaluates hyperparameter evaluating in parallel different keras models.

4.4 Machine Learning algorithms

The aim of this project is to apply machine learning algorithms in order to predict the hematic parameters from visible spectrum of blood.

The target values are hematocrit and oxygen saturation, these variables have different complex relationships with the feature values of spectrum, so two machine learning models predict the two outputs separately.

Different algorithms have been investigated in order to find the best possible approach compatible with the dataset already realized.

The prediction of hematocrit has required a deep analysis, because it is more complex for machine learning the prediction of hematocrit than oxygen saturation due to the more complex connection between Hct with the spectra and due to the imbalanced composition of the dataset.

Therefore, oxygen saturation has been predicted with only two machine learning algorithms: support vector machine and artificial neural networks, because they both have reached high level of accuracy.

Hematocrit was first predicted with SVM and ANN, but imbalanced techniques have been applied to manipulate the dataset and to improve the prediction accuracy for a

standard human range of hematocrit. After balancing dataset, four different machine learning algorithms have been compared: Ridge regression, Elastic Net, Random Forest and ANN.

Name of the algorithm	**Predicted parameter**
Ridge Regression	Hct
Elastic Net	Hct
Random Forest	Hct
SVM	sO_2
ANN	Hct and sO_2

Table 4.1: Machine Learning algorithms

The table 4.1 summarizes the algorithms investigated during this work and the target parameter for which they have been used.

All the different algorithm used in this work are explained in the following paragraphs along with some basic theory about balancing techniques.

The hyperparameters of all the algorithms are reported in the methods section.

4.4.1 Linear regression and regularization

Linear regression [50] is a supervised learning techniques, which assumes that the relationship between the features and the target vector is approximately linear, so it considers constant effect of the features on the target.

If we consider only two features, the linear model is equal to:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \epsilon \tag{4.5}$$

where:

- $\hat{y}$ is the target
- x_i is the data for i_{th} feature
- β_i is the coefficient (or weight) which value is identified by fitting the model

- β_0 is the bias term
- ϵ is the error

Training a linear regression model means setting the parameters until the model best fits the data in training set. Mean squared error (MSE) is the common performance value used to evaluate the generalization error.

In other words, training a linear regression model is an iterative process of minimization of MSE, which is computed by the equation (4.1).

However this minimization process could lead to overfit the data, the model will not fit as well with new data because it is specifically implemented for data already analysed during training.

In order to prevent overfitting, some regularization techniques are introduced.

During regularization, a constrain (or penalty factor) is introduced in the model equation with the aim to reduce the influence of weights.

The result is a more general model which fits less the training data, but it is more general and more accurate with new examples.

There are three types of regularization learners:

1. Ridge Regression
2. Lasso Regression
3. Elastic Net

They all differ for the penalty factor added to the model.

In ridge regression the penalty term is equal to:

$$\alpha \sum_{i=1}^{n} \beta_i^2 \tag{4.6}$$

α is a tuning parameter that controls how much we want to regularize the model. If $\alpha = 0$ there is no regularization, so Ridge regression becomes a Linear regression, while if α is large, the weights are close to zero and higher is the smoothness constraint.

Ridge Regression cost function equation becomes:

$$MSE + \alpha \sum_{i=1}^{n} \hat{\beta}_i^2 \tag{4.7}$$

Lasso (Least Absolute Shrinkage and Selection Operator Regression) regression is the second regularization method for linear regression.

Like Ridge regression, this method adds a penalty factor to the cost function. The penalty term is the norm of the weight vector:

$$\alpha \sum_{i=1}^{n} |\hat{\beta}_i| \tag{4.8}$$

and the cost function becomes:

$$MSE + \alpha \sum_{i=1}^{n} |\hat{\beta}_i| \tag{4.9}$$

α is the tuning parameter to control the level of model regularization.

Finally in Elastic Net, the regularization term is a mix of Ridge and Lasso and it can be controlled by the ratio of two contributions. When r = 0 the Elastic Net becomes a Ridge regression, while for r = 1 the Elastic Net is just a Lasso regression.

The penalty factor is equal to:

$$r\alpha \sum_{i=1}^{n} |\hat{\beta}_i| + \frac{1-r}{2} \alpha \sum_{i=1}^{n} \hat{\beta}_i{}^2 \tag{4.10}$$

The fine tuning of hyperparameter α (or r in Elastic Net) is critical to find an accurate and general model.

4.4.2 Decision Tree and Random Forest

Decision trees [49] become a popular machine learning algorithm thanks to their easy interpretability, moreover they are the basic blocks for random forest.

Decision trees are used in decision analysis to visualize and represent a decision making root.

Decision trees is a recursive partition of space in order to minimize the generalization error, it includes a series of nodes, which form the rooted tree. The starting node has

no incoming edges, while all the others have only one incoming edge. All the nodes are called internal nodes, except for the terminal ones that are called leaves (see figure 4.4).

In a decision tree, all the internal nodes split data into two or more sub-space and the split is based on a discrete function applied on input values.

Each leaf includes samples corresponding to the same target value or sample with higher probabilities to belong to the same output.

The simplest example of tree is the binary tree, where each node has at most two children.

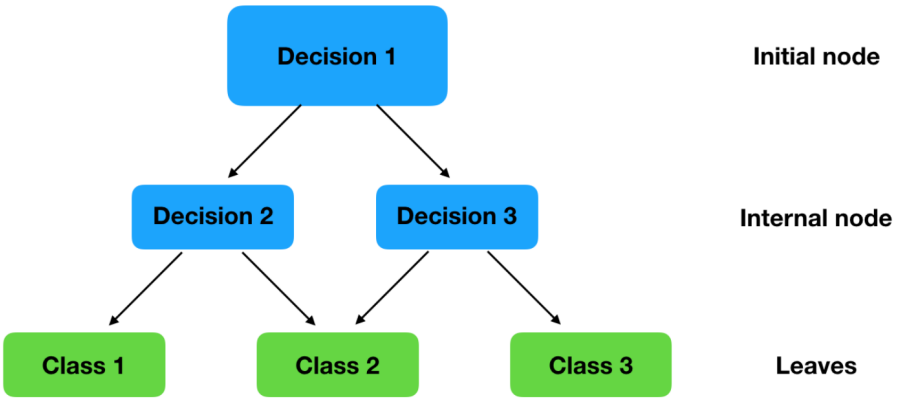


Figure 4.4: Basic scheme of a binary decision tree

The decision algorithm split, iteratively, the data in order to obtain the largest Information Gain (IG); this iterative process can create deep decision trees with many nodes leading to overfitting.

IG is the impurity-based criterion that is maximized at each splitting node and it is defined as:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{4.11}$$

where:

- f is the feature
- D_p and are the dataset of the parent and j^{th} child node
- I is the impurity [47]

- N_p is the total number of samples
- N_j is the number of samples in the j^{th} child

Lower is the impurity at each node and larger is the information gain. For a simple binary tree, IG becomes:

$$IG(D_p, f) = I(D_p) - \frac{N_{LEFT}}{N_p} I(D_{LEFT}) - \frac{N_{RIGHT}}{N_p} I(D_{RIGHT}) \tag{4.12}$$

Where D_{LEFT} and D_{RIGHT} are the two child nodes.

There are two common impurity measures :

1. ENTROPY:

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2 p(i|t) \tag{4.13}$$

   Where $p(i|t)$ is the proportion of samples to class c for the node t. The entropy is zero if all samples belong to the same class.

2. GINI IMPURITY:

$$I_G(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{4.14}$$

(4.14) maximizes the probability of misclassification.

Random Forest is an ensemble of decision trees (figure 4.5). The algorithm implements multiple decision trees in parallel in order to find a more robust model, improving generalization and avoiding overfitting.

In a random forest, features are randomly sampled and passed to different trees in a process called bootstrap [49], which increases their splitting randomness of the random forest and it prevents from overfitting.

Each decision tree makes its decision taking into account the bootstrap sample and maximizing the information gain. This step is repeated iteratively many times and finally, the results from each tree are aggregated and the target is assigned by evaluating the confidence of every prediction.

Random Forest is used for both regression and classification tasks, it has different hyperparameters, that can be tuned to determine the highest performance of accuracy.

The most influent hyper parameters are:

- the number of maximum features to consider at each node;
- the number of decision trees;
- the max depth of the tree;
- the minimum required samples to split.

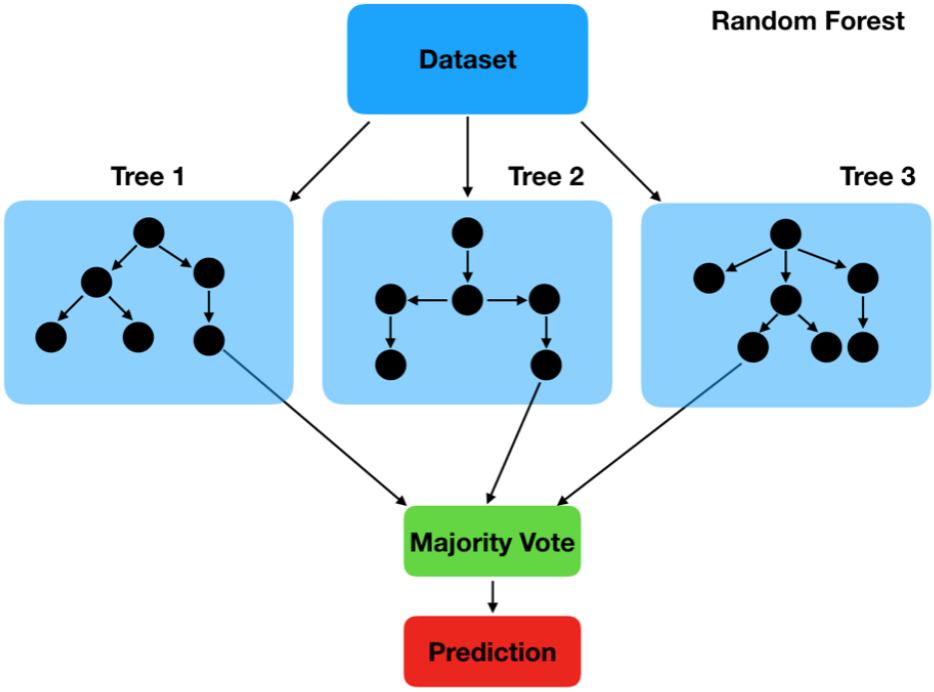


Figure 4.5: Example of Random forest with three decision trees

4.4.3 Support Vector Machine

Support Vector Machine (SVM) [56] was introduced in 1992, by Boser, Guyon, and Vapnik. It is a common machine learning supervised algorithm thanks to its versatility, because it can be used for classification and regression tasks.

The optimization objective of this algorithm is to maximize the margin, which is the distance between separating hyperplanes, the closest samples to the hyperplane are called support vectors.

SVM finds the most confident decision boundary for the correct prediction of training samples.

For example, in the figure 4.6, point A is distant from the decision boundary, so it is confident to predict that A belongs to class +1.

Point C is very close to the decision boundary, so it belongs to class +1 but a small change in the decision boundary can change the prediction of this point.

Intuitively, the prediction of A is more confident than the prediction of C.

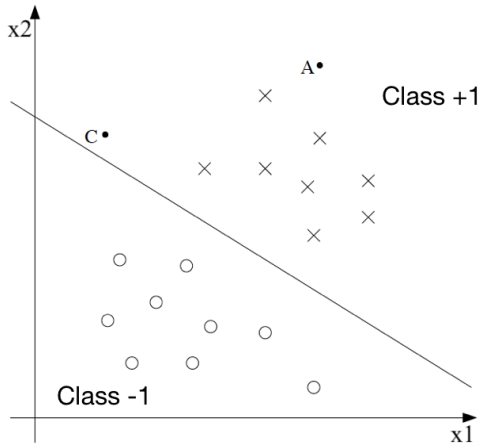


Figure 4.6: Example of separating hyperplane

Mathematically it is possible to formally express this concept through the definition of functional and geometrical margin [56].

Consider a linear binary classification problem with a training $(x^{(i)}, y^{(i)})$, the functional margin of (ω, b) is defined as:

$$\hat{\gamma}^{(i)} = y^{(i)}(\omega^T x + b) \tag{4.15}$$

Where $y^{(i)}$ belongs to -1,+1 and it denotes the class labels.

If $y^{(i)} = 1$, then $(\omega^T x + b)$ need to be large and positive in order to have large functional margin.

If $y^{(i)}$ = -1, then $(\omega^T x + b)$ need to be large and negative in order to have large functional margin.

If $y^{(i)}(\omega^T x + b)$ ¿ 0, it represents a correct prediction, so a large and positive values of functional margin is a confident and correct evaluation of label.

In general, for a training set $S = \{(x^{(i)}, y^{(i)}) \quad \text{with}\, i = 1 \cdots m\}$, the functional margin (ω, b) is the smallest functional margin among all the training examples and it can be written as:

$$\hat{\gamma} = min_{i=1\cdots m}\hat{\gamma}^{(i)} \tag{4.16}$$

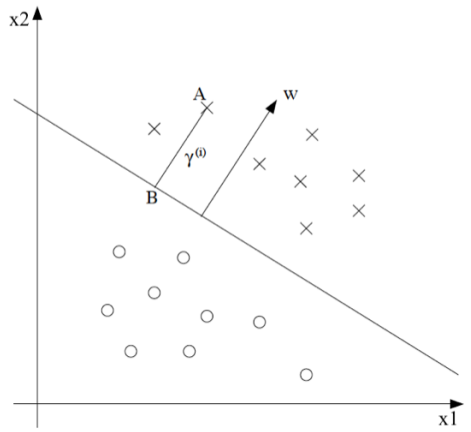


Figure 4.7: Example of geometric margin

In the figure 4.7, it is shown the decision boundary between the two classes.

A is the input of the training example $x^{(i)}$ and its label is $y^{(i)} = 1$, while the distance from the boundary is the segment $\overline{AB}$.

B is equal to:

$$x^{(i)} - (\gamma^{(i)} \cdot \frac{\omega}{\|\omega\|}) \tag{4.17}$$

where $\frac{\omega}{\|\omega\|}$ is a unit-length vector point with ω direction.

Moreover, B is on the boundary line so, it satisfies the condition $\omega^T x + b = 0$ Therefore, it is possible to write:

$$\omega^T \left(x^{(i)} - \gamma^{(i)} \cdot \frac{\omega}{\|\omega\|} \right) + b = 0 \tag{4.18}$$

(4.18) can be rewritten to find $\gamma^{(i)}$ as:

$$\gamma^{(i)} = \frac{\omega^T x^{(i)} + b}{\|\omega\|} \tag{4.19}$$

(4.19) is only valid for positive training examples, in general (4.19) becomes:

$$\gamma^{(i)} = y^{(i)} \left(\frac{\omega^T x^{(i)} + b}{\|\omega\|} \right) \tag{4.20}$$

(4.20) represents the geometrical margin of (ω, b) for the training sample $(x^{(i)}, y^{(i)})$.

Generally, the geometric margin in case of training set $S = \{(x^{(i)}, y^{(i)}) \quad \text{with}\, i = 1 \cdots m\}$ is the smallest geometric margin between all the individual samples:

$$\hat{\gamma} = min_{i=1 \cdots m} \hat{\gamma}^{(i)} \tag{4.21}$$

That is equal to the functional margins if $\|\omega\| = 1$.

SVM finds the optimal margin, which maximizes the confidence of the prediction of samples.

In order to find the hyperplane, the following optimization problem must be solved:

$$\begin{gathered} max_{\gamma,\omega,b}\, \gamma \\ Subject\, to\, y^{(i)}(\omega^T x^{(i)} + b) \geq \gamma\,, i = 1 \cdots m;\ \|\omega\| = 1 \end{gathered} \tag{4.22}$$

However, it is computationally easier to solve this problem minimizing the reciprocal one:

$$\begin{gathered} min_{\omega,b} \frac{1}{2} \|\omega\|^2 \\ Subject\ to\ y^{(i)}(\omega^T x^{(i)} + b) \geq 1\,, i = 1 \cdots m \end{gathered} \tag{4.23}$$

In some case it is not possible to find an hyperplane, because data are not separable. So the variable ζ was introduced by Vladimir Vapnik [57] in order to make the algorithm available for non-separable data and less sensitive to outliers.

The objective to minimize becomes:

$$\begin{gathered} min_{\omega,b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} \zeta_i \\ S.t.\ y^{(i)}(\omega^T x^{(i)} + b) \geq 1\,, i = 1 \cdots m; \zeta_i \geq 0,\ i = 1, \cdots, m \end{gathered} \tag{4.24}$$

The parameter C in (4.24) controls the penalty for misclassification. Large values of C mean large error penalties; otherwise small values of C correspond to less strict misclassification errors (figure 4.8).

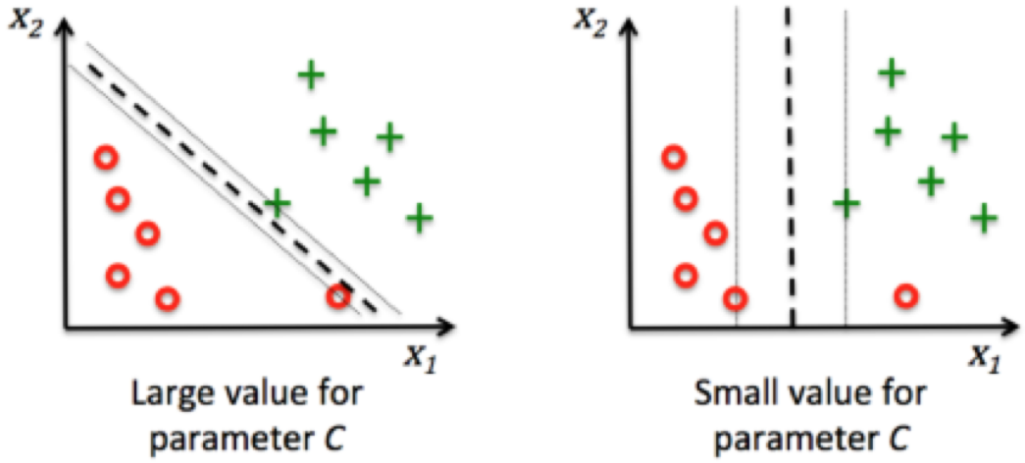


Figure 4.8: Influence of C on the width of the margin

SVM becomes very popular because it can solve non linear problems thanks to kernel methods.

The kernel method creates non-linear combinations of linearly non-separable variables and move the problem to higher dimensional space where the data becomes linearly separable.

However higher dimensional problem corresponds to more expensive computational cost.

One of the most used kernel is the Radial Basis Function (RBF):

$$K(x^{(i)}, x^{(j)}) = exp\left(-\frac{\|x^{(i)} - x^{(j)}\|}{2\sigma^2}\right) \tag{4.25}$$

(4.25) is often simplified as:

$$K(x^{(i)}, x^{(j)}) = exp\left(-\gamma\|x^{(i)} - x^{(j)}\|\right) \tag{4.26}$$

Where $\gamma = \frac{1}{2\sigma^2}$.

SVM can also be applied to regression problems: in this case, the penalty function is modified in order to include a distance measure ϵ and the penalty is not assigned if the predicted value is less than a distance equals to ϵ from the actual value.

4.4.4 Artificial Neural Networks

An artificial neural networks (ANN) [58] are data driven algorithms which learn from a dataset of examples and try to find out hidden functional relations, even if physics is not explicitly provided. The name is inspired by biological neural networks because ANN tries to replicate their structure and functionalities.

There are many different topologies of ANN, but all of them are based on artificial neurons, which are simple precessing elements. Many different neurons are arranged together to create a network.

Each neuron (figure 4.9) is able to compute three basic mathematical operations: multiplication, summation and activation.

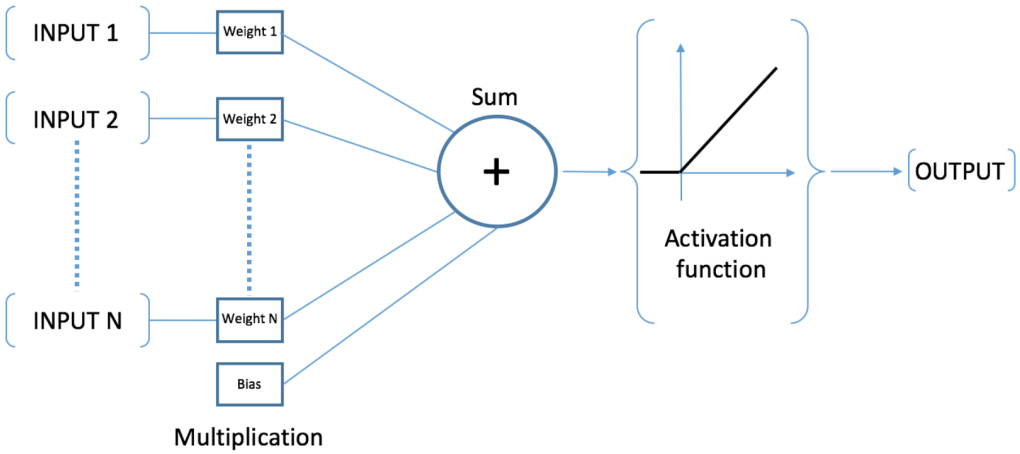


Figure 4.9: Working principle of neural network

Firstly, neurons multiply inputs with different individual eights. Weighted inputs are then summed together with a bias. Finally, an activation function is applied on inputs and bias. Mathematically, a neuron can be described as:

$$y(k) = F\left(\sum_{i=0}^{m} w_i(k)x_i(k) + b\right) \tag{4.27}$$

where:

- $x_i(k)$ is the feature value
- $w_i(k)$ is the weight value

- b is the bias term
- F is the transfer (or activation) function
- $y_i(k)$ is the output

The transfer function characterized the neuron and its efficacy, so it is important to choose the right one. Transfer function can be any mathematical functions, but the most used to solve machine learning tasks are: step functions, linear functions and sigmoid (figure 4.10)

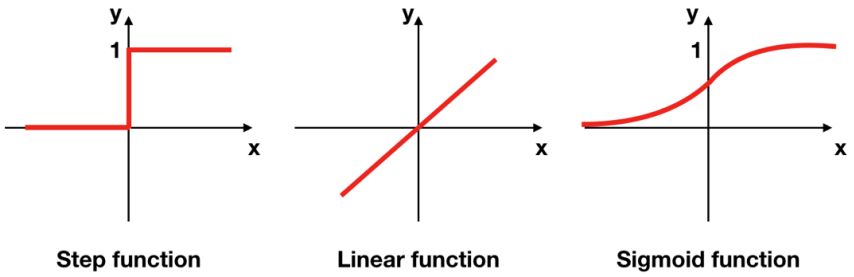


Figure 4.10: Examples of transfer functions

There are different architecture of neural networks, these are characterised by the way neurons are connected.

Two fundamental topologies are:

- Feed-forward Neural Network (FNN)
- Recurrent Neural Network (RNN)

In a FNN, the information flows in one direction: from input to output; while in RNN the information can flow in both directions.

In both configurations, neurons are arranged into layers:

- the first one is called input layer;
- the final one is the output layer;
- all the other layers are called hidden layers.

The information flows from neuron on one layer to the next one through the activation function (figure 4.11).

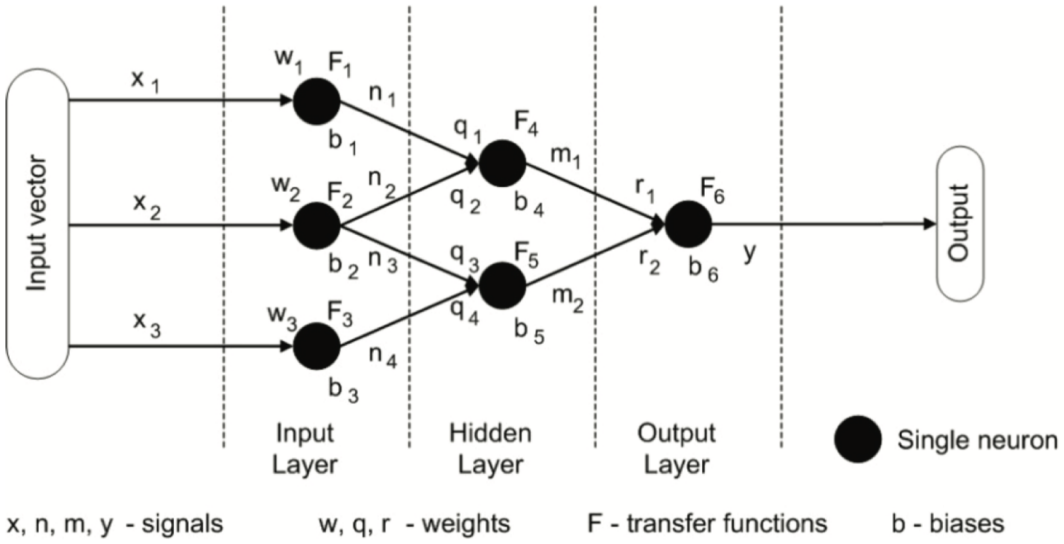


Figure 4.11: Example of simple FNN

A simple FNN can mathematically be written as follow:

$$
\begin{aligned}
n_1 &= F_1(w_1 x_1 + b_1) \\
n_2 &= F_2(w_2 x_2 + b_2) \\
n_3 &= F_3(w_3 x_3 + b_3) \\
n_4 &= F_4(w_4 x_4 + b_4) \\
m_1 &= F_4(q_1 n_1 + q_2 n_2 + b_4) \\
m_2 &= F_5(q_3 n_3 + q_4 n_4 + b_5) \\
y &= F_6(r_1 m_1 + r_2 m_2 + b_6)
\end{aligned}
\tag{4.28}
$$

$$
y = r_1(F_4[q_1 F_1[w_1 x_1 + b_1] + q_2 F_2[w_2 x_2 + b_2]] + b_4) + r_2(F_5[q_3 F_2[w_2 x_2 + b_2] +
$$

$$
\dots
$$

$$
+ q_4 F_3[w_3 x_3 + b_3] + b_5]) + b_6
$$

The topology of ANN used in this project is a FNN with back-propagation error. It is a typical FNN (the information flows in only one direction), where outputs are compared

with targets at the end of each epoch; the error is then back-propagated and the weights are adjusted in order to reduce the error or until some stopping-criterion is satisfied.

The back-propagation computes the gradient descent optimization of networks error and how much each output connections contribute to the error and change the weights during next epoch.

ANN has many advantages, three of them are:

- the ability to solve non linear and complex tasks;
- learning from samples, implementing general models, that can be applied with unseen examples;
- working in parallel with many neurons at the same time, because each one computes only simple operations;

ANN training step could be computational expensive, but most of the computational cost is spent during this step. Once ANN is trained for a particular task, then it can be quickly employed to solve same problems.

Conversely, there are no standard rules to determine the best topology of a neural network, neither its best parameters.

Basic approach is trial and error starting from a simple and easy structure and increasing complexity when results are not satisfactory.

When the topology is fixed, then the parameters are fine-tuned through optimization techniques.

4.5 Imbalanced dataset

Spectral dataset of blood realized during this project is imbalanced, because data are not uniformly distributed.

An imbalanced dataset occurs when the distribution of target is not uniform among the different classes; most of data in real world are often imbalanced, for example fraud detection datasets or spam mail detection datasets are commonly imbalanced.

This problem is recurrent in different healthcare applications where machine learning could be easily applied. For example, many available datasets of skin cancer images are

imbalanced: they have a different number of photo of non-disease examples over disease samples.

The different distribution of negative and positive examples makes harder the implementation of accurate models which are not able to generalize predictions.

Moreover, the error related to the minority class is often critical, because the model misunderstands people who are really affected by disease. In terms of machine learning, the aim is to implement an automated model with highest accuracy and highest sensitivity, due to the importance of possible related consequences on human health.

The sensitivity is the hardest challenge for automated machine algorithms when train data are imbalanced, because data belonging to the most frequent class have a negative effect on the predictions.

Simple predictive accuracy is clearly not appropriate in such situations, while higher sensitivity and highly rate of correct detection in the minority class are more desirable.

Different proposals have been provided to reduce the effects of imbalanced dataset on machine learning models [59] [60].

There are two main different approaches: under sampling and oversampling [61].

Undersampling involves a random removal of samples belonging to most frequent class. There are two types of under sampling methods: random-under-sampling, where the deleted data are chosen randomly and focused-under-sampling, where data are removed when they are located on the border between two classes.

The result is a more balanced dataset, but the data size becomes smaller. Therefore, undersampling is the best approach for big dataset, where removing some data cannot lead to loss of information.

The second technique is oversampling; it involves the duplication of some data belonging to minor classes. These examples can be chosen randomly (Random-Over-Sampling) or among data located on the borders between classes (Focused Over Sampling).

Oversampling is the best choice with limited size datasets, but it produces overfitting; for this reason, the algorithm will not be able to implement a general model.

Basic oversampling techniques generate overfitting, but there are more advanced approaches for balancing a dataset avoiding overfitting, for example SMOTE and SMOTE

+ ENN techniques.

Synthetic Minority Oversampling Technique (SMOTE) is an advanced oversampling method, which creates synthetic samples in the minority class of imbalanced datasets. It avoids overfitting because data are not already present in the dataset.

SMOTE was developed by Chawla in 2002 [62] who proposed, for the first time, an oversample of minority class by creating synthetic examples. SMOTE algorithm takes data from the minority class and introduce synthetic examples along the segments joining any of the minority class nearest neighbours.

The steps for generating synthetic samples are:

- the algorithm considers the difference between the feature vectors and their nearest neighbors;
- it multiplies the difference by a random number between 0 and 1 and add it to the feature vector;
- a random point along the line segment between two specific features is selected.

SMOTE forces the decision region of minority class to become more general and, consequently, more robust. In literature, there are other advanced techniques to balance datasets.

These are a combination of SMOTE followed by cleaning data techniques, such as Edited Nearest Neighbour (ENN) [63]. ENN deletes all the misclassified data from training set using KNN optimization technique.

K-NN (K Nearest Neighbour) [64] is a supervised learning algorithm commonly used for classification; this algorithm finds k-samples in the training dataset that are closest to the point that we want to classify. The class label of the new data point is assigned by majority vote among the k nearest neighbours. The algorithm is summarized in figure 4.12.

ENN removes all the misclassified samples, it optimally eliminates outliers and possible overlap samples among the different classes.

The combined approach of SMOTE + ENN is promising for imbalanced dataset as it improves the final accuracy of the model.

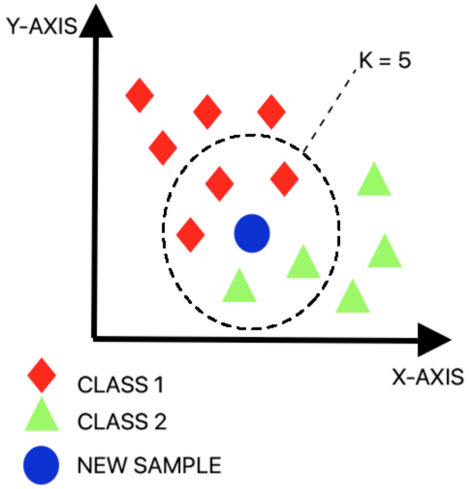


Figure 4.12: K Nearest Neighbor. The algorithm classifies data points by searching within the training set for the K most similar cases (neighbors) and assigning class labels based on the most common among them

CHAPTER 5

Prototyping

A prototype helps for a better understanding of the design, it gives early feedback about every concept stage and it provides validation before the development of the final product, but it requires time and cost [65].

In this project two prototypes are created, but both are based on the same electronic and hardware components. The first prototype (Test-bench prototype) testes the electronic, mechanical and optical components of the setup. This mock-up is an intermediate level of prototyping, because it has limited functionalities, moreover electronic circuit is realized on a breadboard with discrete components.

The second prototype (Stand-alone prototype) is a forward step to the final product: a PCB is realized with integrated components, all software functionalities are provided and all the optimizations are made. This prototype gives a near realistic experience of the final product.

In this chapter, both setups are presented; all components are listed in figure 5.1 and they are discussed in next sections including the spectrometer, which is the core of this work, the hardware, the optical components and finally the mechanical ones.

In the second part of the chapter, the microcontroller firmware is described, both prototypes exploit this firmware, which has all the final functions of this work.

The two prototypes differ for the software: in the first prototype, the software is able to post-process, visualize and store data into a laptop. This data will be used to create a database for machine learning analysis.

The second prototype is intended for the final usage of the setup: the data are post-

processed with a raspberry Pi and they are analyzed through machine learning models in order to extract the hematic information.

The difference between the two prototypes and their softwares will be discussed in the final part of this chapter.

Component	Name
Micro-spectrometer	C12880MA by Hamamatsu
Microcontroller	16 Bit Arduino-based microcontroller
Buffer Amplifier	MCP6001R
Digital Buffer	MC74VHCT125A
LED	High Power white LED
Optic Fiber	Multimode fiber by Thorlabs
Cuvette	Custom PLA cuvette
Cuvette and Spectrometer Holder	PLA parts made custom by Ultimaker 3D printer

Figure 5.1: List of electronic (red), optical (yellow) and mechanical (green) setup components

5.1 Prototype components

5.1.1 Micro-Spectrometer

The spectrometer, selected for this project, is the C12880MA micro-spectrometer developed by Hamamatsu [66].

This module is the smallest spectrometer on the market, it is a high sensitive sensor and its spectral range covers wavelengths from 340 nm up to 850 nm with a spectral resolution equals to 15 nm.

The micro spectrometer has a fixed and integrated slit, a diffraction grating and a CMOS linear sensor.

The light enters the spectrometer, it is reflected by the diffraction grating and it is directed towards the CMOS sensor, which converts the diffracted light into analog signal (see figure 5.2).

CMOS sensor is composed by 288 cells, that are sensitive to light at different wavelengths; each cell generates a voltage signal which is converted with an external ADC.

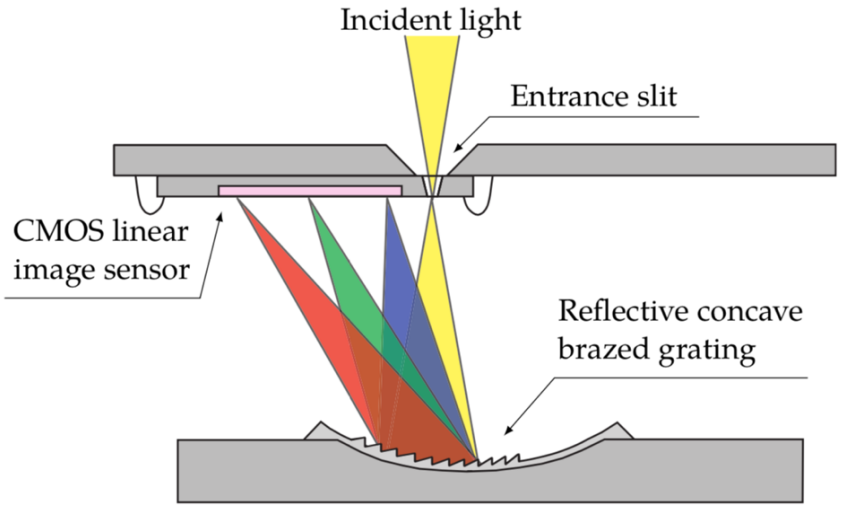


Figure 5.2: Optical component layout of Hamamatsu C12880MA [64]

The spectrometer works with a stable and constant clock in the range between 0.2 up to 5 MHz.

The slits aperture is not tunable, but the integration time can be set. Thanks to this setting, it is possible to perform spectral measurements with different light source levels: for a low light source, the integration time must be set to high values to increase signal; instead for direct and high light source, the integration time can be set to lower values to avoid saturation of photosensor.

The integration time is equal to high level of start signal (ST) plus 48 clock cycles (see the timing chart in figure 5.3).

ST has to be high at least 6 clock cycles, so the shortest integration time is about $10.8\mu s$, that is possible to reach at 5 MHz clock frequency.

A longer integration time can be set by increasing the high-level time of ST.

Video signal is referred to the output analog voltage, there are 288 pixels, the first one is ready at the 87^{th} clock pulses when ST level is low.

The measurement routine is fully described by timing chart reported in the datasheet of Hamamatsus C12880MA (see figure 5.3), it consists of two different sequences:

1. measure, when the image sensor captures light;

2. output, when the sensor sends out data pixel by pixel for each clock cycle.

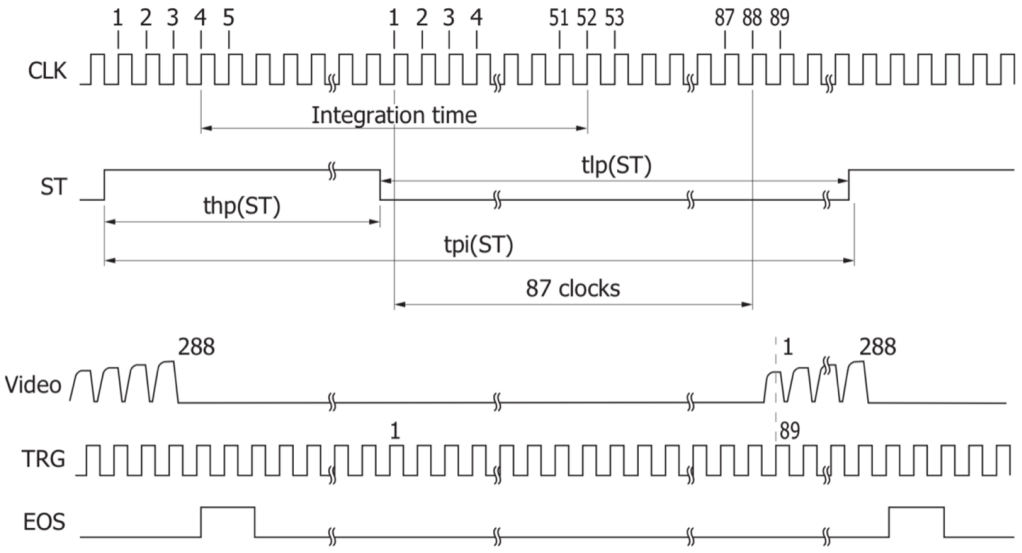


Figure 5.3: Timing chart of Hamamatsu C12880MA

5.1.2 ADC Selection

After the selection of spectrometer, the driver circuit has been designed and realized in order to optimize the performance of the setup.

The prototypes were first realized on a breadboard, then on a soldering board and, finally, a PCB has been realized for the stand-alone prototype.

A guideline for the driver circuit is provided by Hamamatsu [66], it requires: a micro-controller as timing generator, a digital buffer which provides CLK, ST, EOS and TRG signals to the spectrometer, a buffer amplifier and an ADC to convert video output signal.

In order to accurately convert the data, ADC must have high resolution: it depends on the ratio between voltage saturation and readout noise, which is generated within CMOS sensor during the readout process [67].

Readout noise is provided in the datasheet, so the ADC must have at least effective resolution equals to:

$$log_2 \frac{V_{saturation}}{V_{Noise,RMS}} = log_2 \frac{5V}{1.8mV_{RMS}} = 11.4bits \tag{5.1}$$

Therefore, ADC must have a nominal resolution of 14-16 bit with at least 12 usable bits, in order to have higher dynamic range than spectrometer.

5.1.3 Microcontroller

Microcontroller and ADC have to satisfy some critical requirements in order to maximise the performance of the spectrometer.

First of all, ADC must have at least 12 usable bits in order to have as dynamic range as spectrometer, it can be either external from microcontroller or integrated into it.

An Arduino-based microcontroller is chosen for its technical and electrical characteristics: this board is equipped with 32 bit ARM cortex M4 microprocessor overclockable up to 96 MHz. RAM is 64 Kbytes, while flash memory is equal to 256 Kbytes.

The board allows different communication protocols:

- UARTs;
- SPI;
- I^2C;
- CAN BUS.

The microcontroller has a micro-USB port to connect the micro controller with a laptop, the firmware can be upload in the board using Arduino IDE.

The microcontroller is Arduino-compatible, so most of Arduino's libraries and functions work on this board; it works at 3.3 V, but all digital pins are 5 V tolerable. Microspectrometer C12880MA works at 5 V, so a direct connection with the microcontroller is possible using a logic level shifter.

Moreover, the microcontroller has a 16 bits integrated ADC with 13 usable bits, so it is possible to use the internal ADC for converting the analog video from spectrometer to digital signal without reducing dynamic range of the signal.

Thanks to these characteristics, this microcontroller is a good compromise between electrical characteristics and costs, without decreasing the performance.

In this schematic circuit, the microcontroller is connected to a digital buffer where:

- ST and CLK are output signals for microcontroller;
- EOS and TRG are input signals from spectrometer and they arrive to the microcontroller.

The microcontroller is also connected to the output of a buffer amplifier and it receives the video signal pixel by pixel when data are ready.

Finally, the microcontroller controls the light source: it turns ON the LED when data is acquiring while it turns OFF the LED when measurement is finished. A precise synchronization between spectrometer and LED is required in order to save power and use the LED only when necessary, because the light provided by LED can increase the temperature of the setup if it is ON for a long time.

The wiring connections are showed in appendix 1.

5.1.4 Buffer amplifier

MCP6001R [68] is selected as buffer amplifier, it is a general purpose operational amplifier made by Microchip with a gain bandwidth equals to 5 MHz.

The operational amplifier requires a supply voltage from 2.2 V to 6.0 V.

MCP6001R is connected in buffer mode: the input is the analog video provided by spectrometer and it exits from the buffer to be sampled by ADC.

Buffer amplifier is introduced in order to avoid excess current consumption that could increase the noise due to increase temperature.

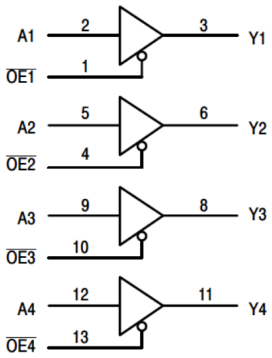
5.1.5 Digital Buffer

MC74VHCT125A [69] is a CMOS technology high speed buffer. It has 3-state control input to set output signals.

Logic diagram and function table in figure 5.4 show its logic functionalities, the buffer is an active-low logic digital port; it achieves high speed with low power dissipation and it can be used to interface signals with different levels from 2.0 V up to 5.5 V.

This digital buffer allows to interface 3.3 V microcontroller with spectrometer, which works at higher level.

Four signals pass through the digital buffer: CLK and ST come from the microcontroller and they arrive to the spectrometer to start the measure; EOS and TRG are input signals for micro-controller.



(a)

VHCT125A		
Inputs		Output
A	$\overline{OE}$	Y
H	L	H
L	L	L
X	H	Z

(b)

Figure 5.4: Logic diagram (a) and functional table (b) of digital buffer

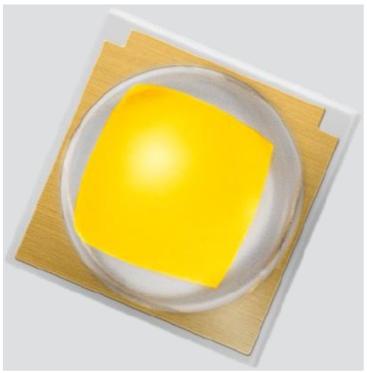


Figure 5.5: Photo of Samsung LED with ceramic package

5.2 Optical components

5.2.1 Light Source

Halogen lamps are usually used as light source for different spectroscopic applications thanks to their continuous emission spectrum; this property makes halogen lamp suitable for visible absorbance spectroscopy measurements, due to the high power efficiency light.

In incandescent lamp, the light is produced due to the chemical reaction between halogen gas and the tungsten filament, it happens at high temperature (about 250^oC [70]), so the outer bulb glass reaches high temperature and the heat remains concentrated on its small surface.

In the proposed setups, the light source is placed close to the cuvette where the blood

is flowing during hemodialysis session.

An incandescent light could obstruct the blood flow within the cuvette due to rapid heat produced compromising the measurement and the treatment, so LED is used instead of halogen lamp (see Appendix 2).

LEDs are semiconductor devices where light is emitted when electrons in the semiconductor recombine with electron holes. There are LED of different colours, for example white LED is obtained using together phosphors with a short wavelength LED. This LED technology results in a broad spectral power distribution with a peak in blue wavelength (450-470 nm) and a regular luminous light distribution in the range of visible wavelengths; this power distribution allows LED to be used as source light for spectral applications.

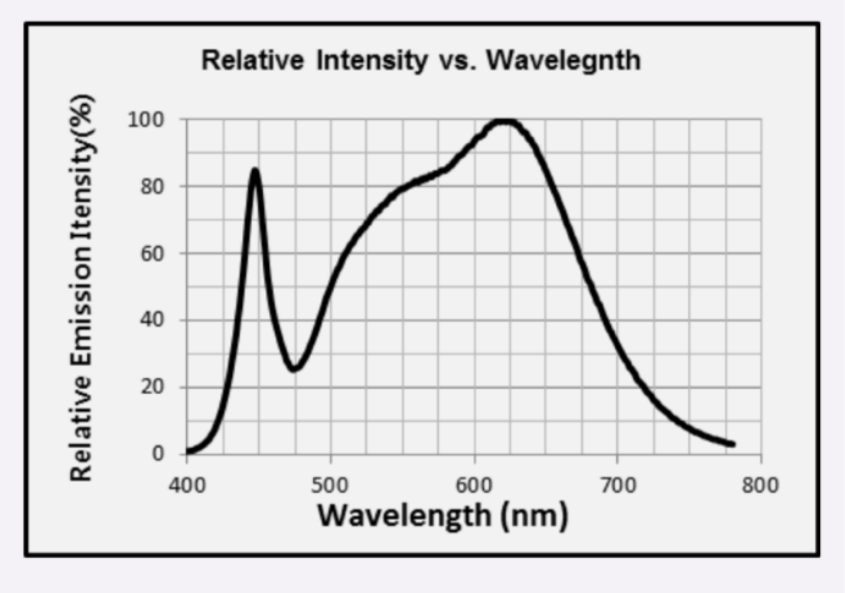


Figure 5.6: Emission profile of Samsungs LH351B LED

Samsungs LH351B LED [71] is used in this work as light source; it is high efficacy and high quality white LED, which provides uniform light distribution for spectral analysis of blood. Figure 5.6 shows the relative emission intensity along the spectral range of visible wavelength.

The relative luminous flux is linearly dependent from the forward current (figure 5.7), the LED operates up to 1.5 A and it also includes a ceramic packaging for heating dissipation (see figure 5.5).

LH351B is a family name of LED made by Samsung, there are different models with

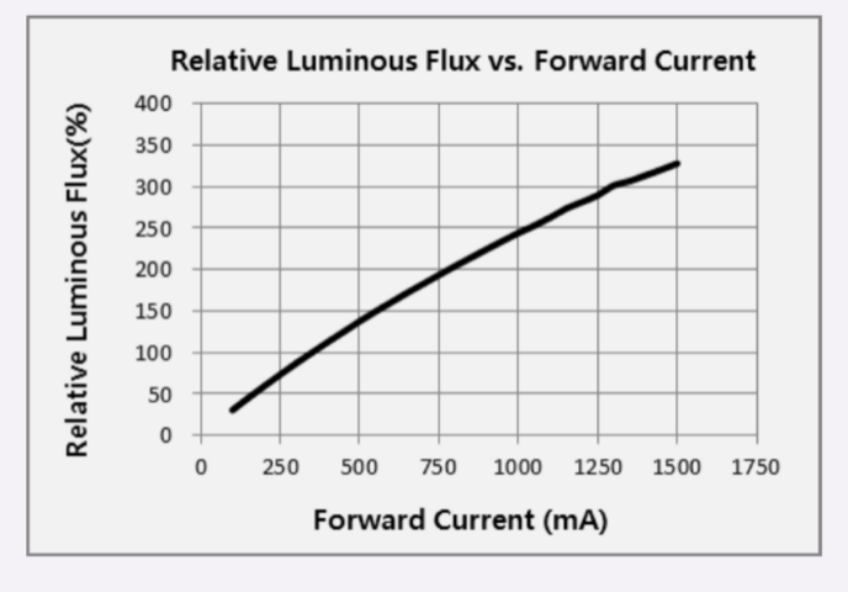


Figure 5.7: Linear behaviour of relative luminous flux with forward current

different spectral distribution; the prototype is equipped with CRI90 warm white light model, because it has the best spectrum distribution for this application: the blue peak has lowest intensity than the other models, so it does not saturate the spectrometer.

LED is driven by the microcontroller through a BJT driver circuit (see Appendix 1 - Wiring).

5.2.2 Optic Fiber

A multimode fiber is used to collect the light transmitted by the sample. The light from LED is focused on the optical window of the cuvette, it is absorbed by sample and the remaining part enters in the fiber optic which is connected to the slit aperture of the spectrometer.

The fiber optic is a multimode fiber mode by Thorlabs, the core is pure silica and the numerical aperture is 0.39, while the core diameter is $600 \pm 10\mu m$ [72].

This model is specified for visible to near-IR transmission of light. Two SMA connectors provide mechanical fixing in order to increase reproducibility of the measurements, the connectors are integrated within the cuvette and the spectrometer holders.

5.2.3 Cuvette

A cuvette is a sterile blood chamber designed for spectroscopic analysis. There are different type of cuvette, they can differ for: optical path length, material, dimension, etc..

The light crosses the walls of the cuvette and the sample filled into it, the path light is the inner length of the cuvette. The optical path represents the signal level of a spectrometer measurement: if the optical path is too long the light will be absorbed and very low signal reaches the detector, while if the optical path is too short, so low energy will be absorbed and it saturates the detector.

The material depends on the type of spectral analysis we want to perform: for example Quartz cuvettes are used for UV spectroscopy, while plastic and glass are transparent to visible and IR lights [73].

In this project, a custom cuvette is designed in order to optimize the optical path and the blood flowing (see figure 5.8).

The cuvette is designed with CAD software and realized with Form 2, this is a stereolitography printer made by Formlabs [74]. Stereolitography is more expensive than filament 3D printers, but it allows higher resolution (layer thickness from $25\mu m$ to $300\mu m$ [74]).

Moreover, the cuvette integrates two terminals that help the integration of the cuvette along the hematic circuit. The cuvette has been suitably designed to allow both an efficient flow of the blood and an effective illumination through the light source. For this, accurate CAD design and different concurrent processes are needed. The shape of the cuvette is important for a continuous blood flow in the blood path. In particular, the choice of the materials is fundamental to allow low absorption of the light. The problems of having different materials in different parts of the cuvette, thus involving a series of technological issues, has also been addressed and conveniently solved. The structure has been printed in 3D through stereo-litography, allowing to obtain suitably a low-cost efficient solution for the development of the final device.

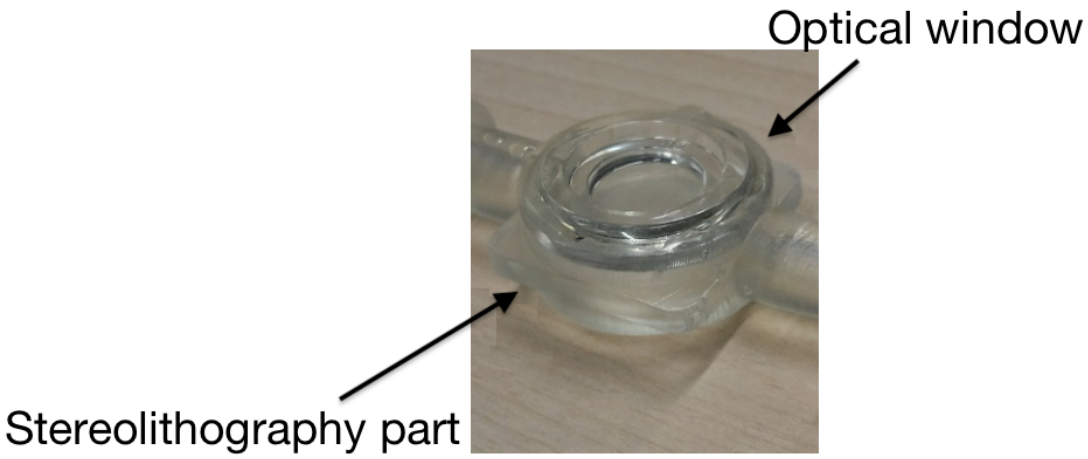


Figure 5.8: A photo of the final structure of cuvette

5.3 Mechanical components

5.3.1 Holders

Custom holders are designed and realized with PLA through 3D printer. Two holders are realized:

1. Cuvette holder
2. Spectrometer holder

The first one (figure 5.9) includes a site for LED and it integrates a slit for SMA connector in order to mechanically fix the fiber optic. Fiber optic is placed perpendicular to the cuvette in order to capture all the transmitted light.

The design is suitable for the custom cuvette in figure 5.8: the holder covers the optical window of the cuvette, it avoids the ambient light to alter measurements.

LED is placed next to the cuvette but the holder excludes contact between LED and cuvette to avoid distortion of light incident flux and to protect LED from damage.

The second holder is designed as case for the spectrometer (see figure 5.10). The holder avoids ambient light to enter and alter the spectral response of the sensor, it also

fixes the distance between slit aperture and SMA connector of the fiber optic.

The holders are designed with Solidworks CAD and they are printed with Ultimaker 3D, which allows a layer resolution of 0.25 mm [75], this printer exploits FFF (fused filament fabrication) technology to realize high quality manufacturing.

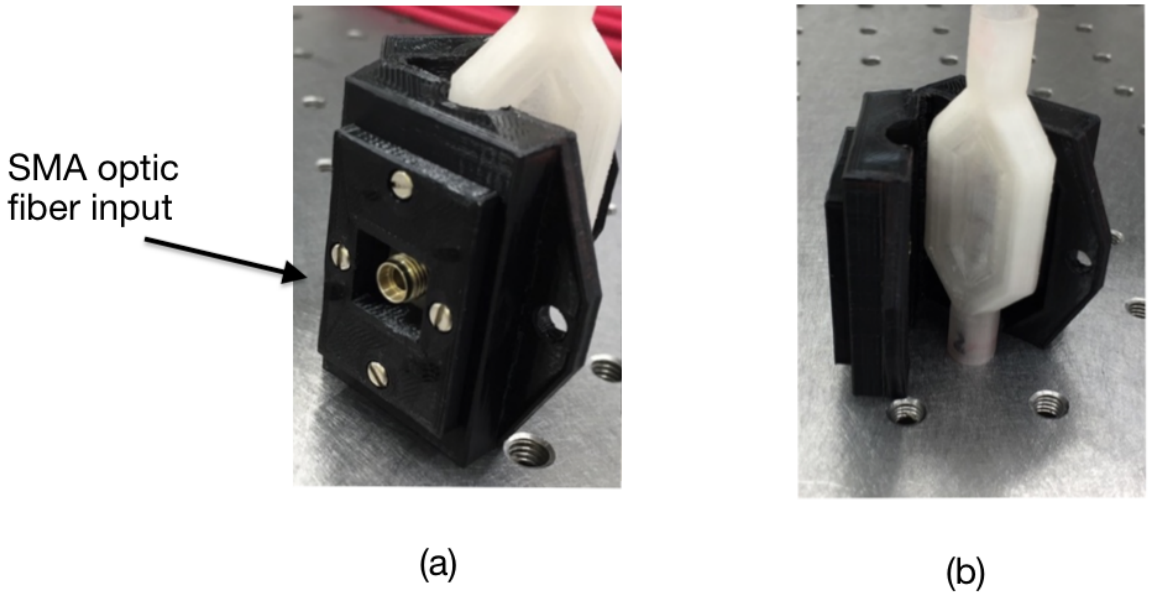


Figure 5.9: Cuvette holder with SMA connector for optic fiber and LED source light

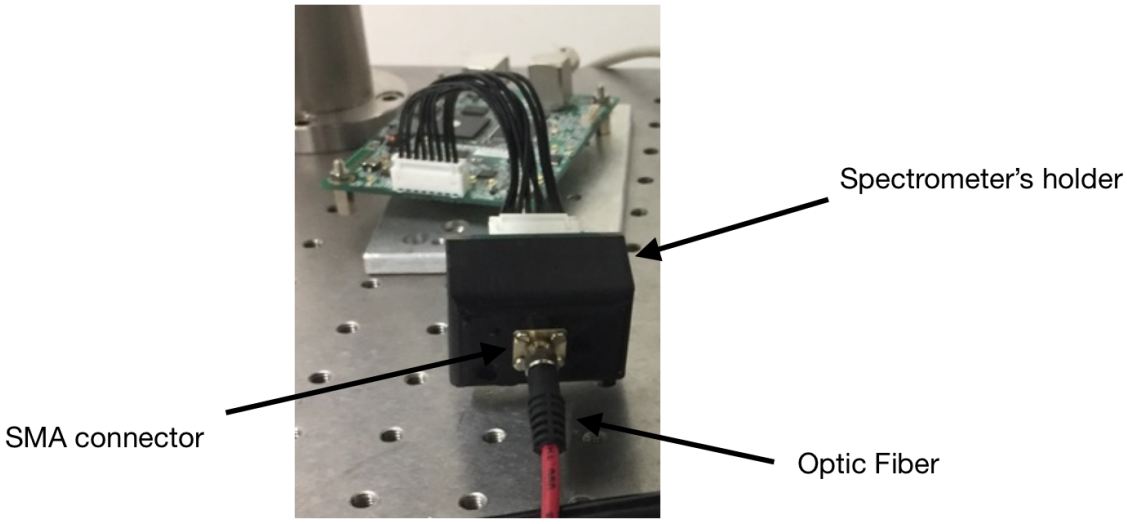


Figure 5.10: Holder for the spectrometer with optic fiber connected through SMA

5.4 Microcontroller firmware

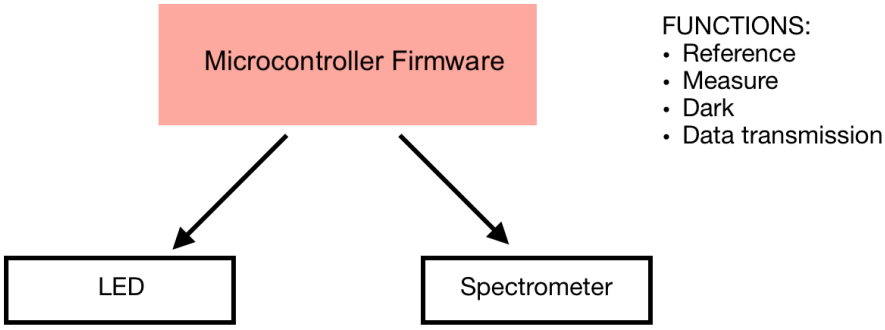


Figure 5.11: Block diagram of microcontroller firmware

As described in the previous chapter, the spectrometer is controlled by the microcontroller.

Arduino code-like has been written and uploaded within the microcontroller. The code includes different functions, which is possible to select through a serial command coming from the Python software.

The functions implemented in the microcontroller are:

- Reference
- Measure
- Dark
- Data transmission

These are included in both prototypes.

5.4.1 Dark

A measurement of dark noise is performed before the real spectral analysis of sample or reference.

LED light is off during this measurement routine and the spectrometer records only the electronic noise and the possible ambient light, which is excluded thanks to the mechanical optimization of setup.

Dark noise is subtracted from reference and sample spectra, in order to avoid noise from final absorbance spectrum.

5.4.2 Reference

Two different measurements must be performed, in order to obtain the final transmission or absorbance spectrum of a sample.

A spectral analysis is first performed on a reference sample, which is usually the cuvette itself or the disposable filled with physiological solution. This measure represents the reference spectrum to whom the spectrum of sample will be compared.

In this work, physiological solution is used as reference, because the blood is mostly composed by water, so all the samples in the dataset are referred to their references. The reference must be performed before the treatment.

Physiological solution transmits unaltered the light of LED, because it is transparent in the range of visible wavelengths, so a low integration time is needed in order to avoid saturation of the spectrometer.

When the reference mode is selected, the microcontroller provides the signals to start this measure: ST and CLK are sent to the spectrometer, ST also includes the information about integration time, which is fixed for reference. When data is ready, spectrometer sends EOS to the microcontroller along with reference spectrum.

The microcontroller drives the light source: it turns on LED before measurement and it turns OFF when data is available.

5.4.3 Measurement routine

This routine performs the spectral analysis on a sample in the cuvette.

The data is composed by 288 pixels of spectrometer, which records the light transmitted by blood. The integration time for blood measurements is longer than reference, because the sample absorbs more light than the transparent physiological solution.

In this routine, the signals provided by microcontroller are the same as reference; the difference is ST at high level that is not fixed, but it is provided through serial command, this information is indicated in μs.

5.4.4 Data Transmission

Dark, sample and reference data are composed by 288 values which represent the transmission of light at different wavelengths.

When data are ready, they are serially sending out from microcontroller.

5.5 Python software for test-bench prototype

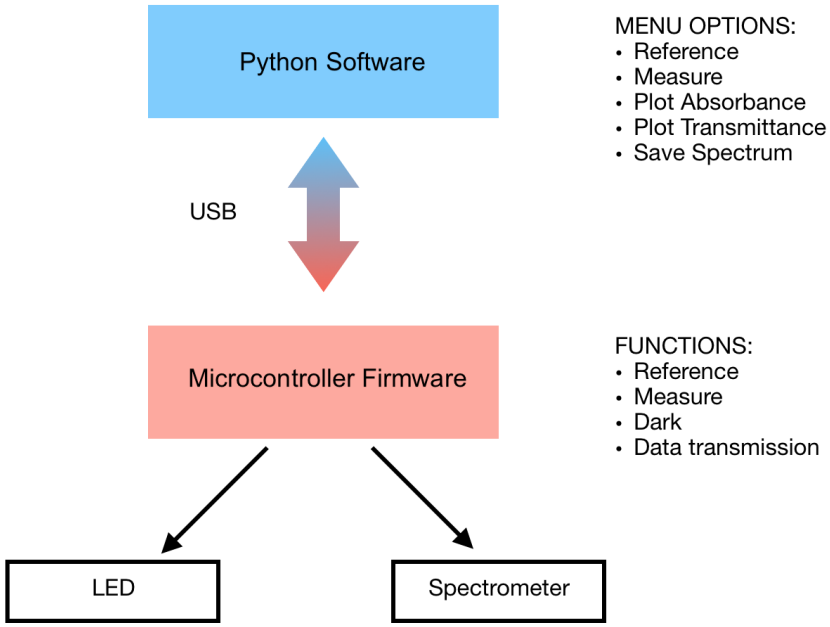


Figure 5.12: Block diagram of the two levels of software programming

The first version of the software is implemented to interface microcontroller with laptop; it is intended to develop a database of blood spectra.

The software exploits a simple interface to help the user, who can choose the available functions through a menu option and the command is sent to microcontroller for starting the related sequence.

The software stores the spectral result from the microcontroller and it operates all the post processing steps in order to plot the final absorbance spectrum of the sample and it finally allows the creation of a database of blood spectra with different hematocrits and oxygen saturation; the database will be analysed through machine learning algorithms in order to obtain a smart model for the prediction of hematic parameters for new blood samples.

The software is written in Python 3. This is a very versatile programming language with a lot of libraries that reduce development and time costs.

Before starting dialysis session some physiological solution is introduced to the circulation tubes, the spectrum of water is stored during this process and it represents the reference. Reference routine also includes the acquisition of dark spectrum of water, that is performed with LED light turned off.

To increase repeatability, the routine performs 100 different measures of water absorption spectrum, these measures are all saved in the laptop and the mean value is evaluated for each wavelength. The final spectrum of reference is the mean result of all measures without dark noise, which is subtracted:

$$T_{REF\lambda} = \sum_{i=1}^{100} T^*_{\lambda_i} - T_{Dark_{\lambda_i}} \tag{5.2}$$

Where:

- T_{REF} is the final transmission of the water spectrum for each λ;
- $T^*_{\lambda_i}$ is the i^{th} measure of transmission of reference sample for each λ;
- $T_{Dark_{\lambda_i}}$ is the i^{th} measure of dark transmission for each λ;

When T_{REF} is evaluated, the hemodialysis session and the spectral acquisition of samples can start.

The measure begins with dark transmission for the evaluation of electronic noise.

When dark spectrum is stored in the laptop, 100 measures of blood spectrum are performed storing the light source passing through the cuvette. The two different spectra of reference and blood can be compared after a linearization post-processing operation due to the different integration times used for the two measures.

The software allows the visualization of both plots (absorbance and transmission) and the data are finally stored.

5.6 Spectrometer Linearity

Linearity is the property of the spectrometer to respond linearly to an increase of integration time. In an ideal linear spectrometer, a change in the measure corresponds to a change of the measured light and it is independently from integration time.

Hamamatsu provides technical information about linearity: the company shows the typical error of the sensor which is from 0 to -8% from the typical value, the test was conducted for a variable integration from 1 to 770 ms.

The linearity has been tested tuning the integration time with a constant light, A/D output is the output after dark reduction, the difference between ideal and typical value contains the measurement error.

This test is performed by Hamamatsu using a different spectrometer (C13016), while similar test result is not provided for C12880MA [66].

Reference and measurement of blood have different integration times. The reference is the physiology solution which need a low integration time because it does not absorb light in visible wavelengths, while blood samples are not transparent so they absorb more light and they require higher integration times. Therefore, it is not possible to directly compare reference and measurements of blood.

When different integration times are used, the software introduces a corrective factor τ to compare them:

$$\tau = \frac{IntTime_{Blood}}{IntTime_{Ref}} \tag{5.3}$$

where:

- $IntTime_{Blood}$ is the integration time used during blood measurement
- $IntTime_{Ref}$ is the integration time for spectral acquisition of reference.

This term (τ) is introduced in the Beer-Lambert law to evaluate the transmission and the absorbance spectrum of blood samples:

$$T\% = \frac{I_1}{I_0} \times \tau \times 100 \tag{5.4}$$

$$A = -Log_{10}(\frac{1}{T} \times \tau) \tag{5.5}$$

To verify spectrometer linearity, a test was performed using C12880MA and the evaluation board developed by Hamamatsu. The test consists in the acquisition of absorbance spectra of same mixture recorded using different integration times, while the light is keeping constant.

The absorbance spectra in figure 5.13 are overlapped because the linearization was applied.

The test shows high linearity of the spectrometer and it validates the linearization techniques described in (5.3), (5.4) and (5.5).

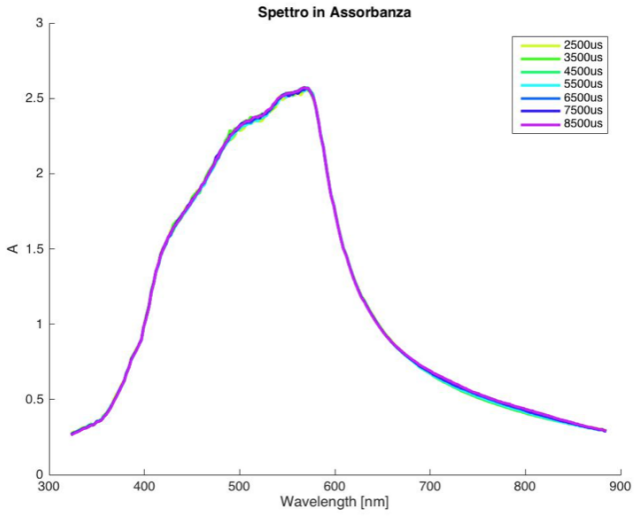


Figure 5.13: Absorbance spectra of same sample with different integration time

5.7 Stand-alone prototype

A second prototype is built to integrate machine learning models within the system without using a laptop, increasing portability and usability of whole setup. The models are already trained, so they do not need a powerful computing performance, so the laptop is replaced by a smaller Raspberry Pi as showed in the figures 5.14 and 5.15.

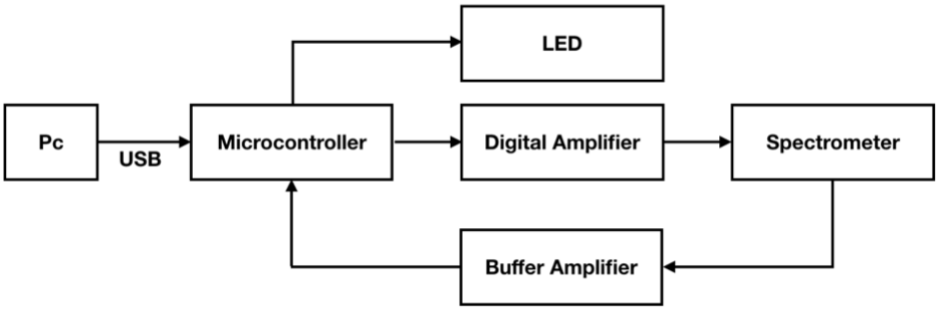


Figure 5.14: Hardware block diagram of the first prototype. A laptop communicates with microcontroller through USB

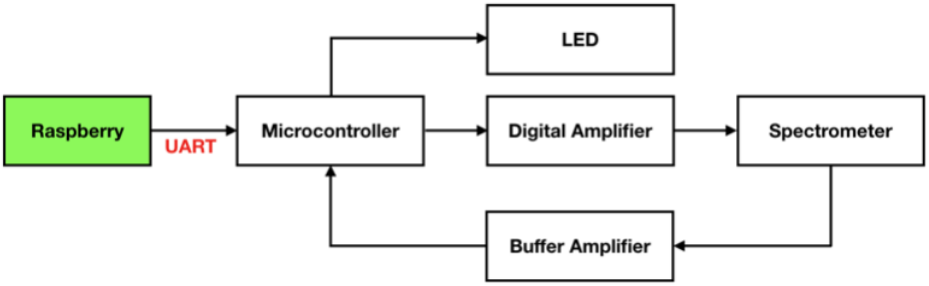


Figure 5.15: Hardware block diagram of the stand-alone prototype. The laptop is replaced by a Raspberry

Microcontroller and Raspberry Pi [76] communicate through UART protocol, while the electronic and mechanical components in the final prototype are the same described in the previous sections of this chapter.

The aim of this second prototype is to develop a setup that is stand-alone, without using pc, focusing on the needs of the patients and doctors: the new prototype integrates all the components in a single box that is compact and easy to use during hemodialysis treatment.

The prototype is user-friendly, because a menu shows the modalities to use that can be selecting by physical buttons, the results of hematocrit and saturation are shown on a LCD display.

The user must be only aware to follow the procedure of reference and measurement,

but a step-by-step procedure guides it (see appendix 3).

This version of the setup includes some new characteristics that will be explained in the following sections.

5.7.1 Raspberry

Raspberry Pi 3 is a low-cost, single board computer. Unless is small size, Raspberry has high computing power with a 64-bit ARM Cortex processor and it is able to connect to internet through ethernet and wireless.

In the second prototype, Raspberry replaced laptop, increasing the portability and reducing the cost of the setup. Moreover, Raspberry board has two built-in UART ports, which allow the communication between Raspberry and the microcontroller.

5.7.2 UART

Universal Asynchronous Receiver-Transmitter (or UART) [77] is an inbuilt circuital block which is implemented within the board.

UART enables a serial data communication between two devices:

- a transmitter (Tx)
- a receiver (Rx)

The data flows from transmission device to the receiver one. Tx and Rx work at the same speed and each bit is transmitted for a fixed duration.

This transmission time is expressed in terms of baud rate, which is the rate at which information is transmitted.

For example:

$$9600 baudrate => T = \frac{1}{baudrate} = 104\mu s \tag{5.6}$$

UART communication is commonly used for the communication between two devices, thanks its easy interface that is showed in figure 5.16.

UART can be implemented using only three signals:

1. Rx
2. Tx
3. GND

The data transmission is then guaranteed due to fixed baud rate of both devices.

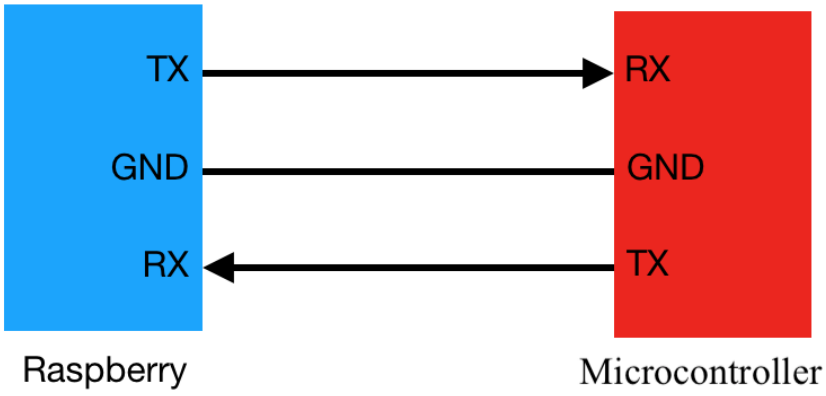


Figure 5.16: UART interface

UART shows multiple advantages:

1. CLK is not required, because communication is asynchronous;
2. it is easy to interface, because it requires only two signals;
3. synchronization is guaranteed by start and stop sequence bits;
4. baud rate is the only parameter to be seated in order to obtain a proper communication;
5. the circuital interface is already integrated in the microcontroller.

5.8 Software for stand-alone prototype

Raspbian has been installed on Raspberry Pi, this is an operating system optimized for this board. Python software runs on this operating system, so a newer version of the software is implemented and installed on Raspberry.

The software has the same routine for measurement and pre-processing of data as first prototype. The differences are:

- implementation of a routine for automatic integration time selection;
- application of trained machine learning models.

The routine for the automatic integration time selection includes three steps:

1. it starts with a one-shot spectrum measurement of sample;
2. max value is found from the raw signal measurement of first step;
3. evaluation of max value

The max value must be under the saturation level of sensor, but higher than values of dark measurement in order to optimize the signal/noise ratio. If the value satisfies these conditions, the normal measurement will start, otherwise the routine will start again with a different integration time.

The new value of integration time depends on which condition is not satisfied:

- if the signal is too high, the integration time will decrease;
- if the signal is too low, the integration time will increase.

This routine makes the system self-executing: users do not need to try different integration times, while the machine provides the best value for the measurement of the sample. Moreover, the routine consists only on one shoot measure, so this procedure is very fast, although it can be repeated many times for each measure, it does not compromise any functionality of the system.

The second main different in the final prototype is in the use of machine learning. Unless the computational capabilities of Raspberry, this prototype is not intended for training new machine learning models, but it only applies models on new samples.

The models are pre-trained with the database created thanks to the first prototype, these models are deployed into the stand-alone prototype.

The deployment of a machine learning model is the integration process of machine learning models into existing production environment in order to start using them.

The prototype does not have the database into it, because it does not need to train new models, it only applies pre-trained models.

Keras library helps to save the architecture and weights of models, these are then load into raspberry before using it.

CHAPTER 6

Methods

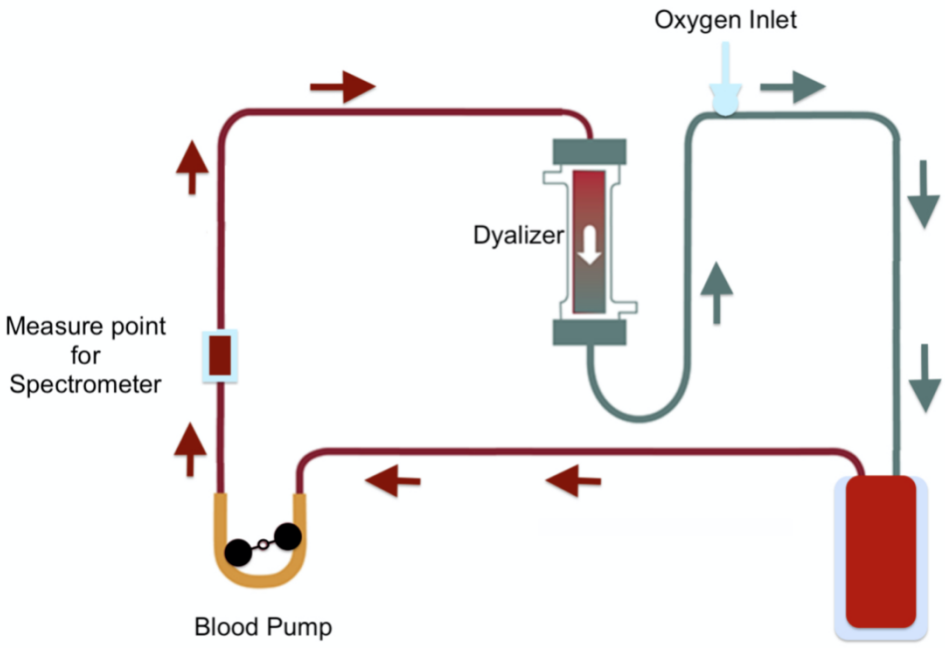


Figure 6.1: Operational environment for tests

The next step into the development of a smart system for the prediction of hematic parameters of blood is the testing part [78]. The prototypes, described in chapter 5, need to be validate in an operational environment in order to test and demonstrate their functionalities.

The challenge is to integrate the prototype within a hemodialysis machine in order to record visible spectra of blood, performing the post-processing operations, creating a

database with all the spectra and finally deploying the models using with new samples.

The development of the dataset was carried out with spectral acquisition of blood.

The hemoglobin absorption and the scattering properties of red blood cells determine a visible spectrum used as input for machine learning algorithms, to determine hematic properties.

Considering the hemodialysis machine as in figure 2.6, the optical setup is integrated along hematic tubes as described in figure 6.1.

The session starts with the acquisition of reference spectrum, so physiological solution flows along the tubes.

The blood is from bovine and it circulates through the hematic tubes reproducing a real-hemodialysis treatment; spectra are recorded after the peristaltic pump and before the dyalizer.

A dyalizer is used to change the hematocrit concentration, while an oxygen inlet is introduced to increase oxygen saturation in the blood.

Machine learning requires a big amount of data to define an accurate predictive model, so a database with spectra of different blood samples has been developed.

These data must have different values of hematocrits and oxygen saturation in order to create be as general as possible.

The models for the prediction of hematic targets are supervised learning algorithms, because they are based on input examples where output labels are already provided. In this case, the models learn from this input-output pairs, while they try to find the functional relation between them.

Therefore, database must include absorbance spectra at different wavelengths as input informations and output values, that are the values of hematocrit and oxygen saturation evaluated with standard techniques: hematocrit is measured through blood fractionation, while oxygen saturation through hemogas analyzer.

Blood fractionation [79] is a process of separating whole blood into its different components: the separation occurs after a centrifuging treatment of samples.

A fractionated blood shows an upper plasma layer, a thin interface of white blood cells and red blood cells in the lower part as described in figure 6.2.

Blood components are well separated and hematocrit is evaluated as the ratio between

red blood cells over the total volume.

Saturation has been measured through GEM Premiere 3000, which is an electrochemical sensor that measures pH, electrolytes and other parameters of blood such as oxygen saturation. GEM Premiere 3000 has a resolution of 1% for sO_2 in the range between 0 and 100% [80].

We used a rigid protocol in order to increase measurement repeatability and to avoid human error.

The protocols steps are the following:

1. Spectral acquisition of sample
2. Show the absorbance spectrum on the laptop and verify if the signal/noise ratio is acceptable, otherwise repeat the measure
3. Collect about 5 ml of blood, which is extracted from hematic tubes
4. A bit of sample is processed through GEM Premiere 3000
5. A bit of blood is filled in a capillary tube and it is centrifuged at 10000 rpm for 5 minutes at room temperature
6. Record oxygen saturation and hematocrit

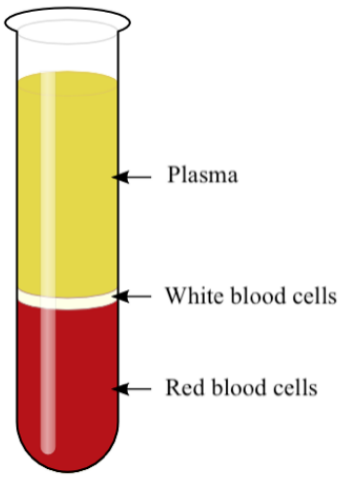


Figure 6.2: Fractionate blood with upper plasma and lower red blood cells layers

Two datasets have been developed: the first is composed by 160 samples and it is used for the prediction of oxygen saturation through SVM and artificial neural networks.

The second dataset is composed by 270 different samples and it was manipulated in order to increase the accuracy and to avoid the development of biased models for hematocrit.

The preprocessing steps and training of machine learning models for prediction of the sO_2 and Hct will be discussed separately.

6.1 Composition of first dataset

Five simulated dialysis sessions were performed, resulting in a first dataset composed by 160 different spectra of bovine blood. Every spectrum consists of 288 values, which represent transmittance levels at specific wavelengths. Each spectrum is the average of 100 scans at the highest sensor resolution.

A reference spectrum was acquired and subtracted from measurements.

Spectrometer provides the transmittance value of light and post-process operations are performed to evaluate absorbance spectra.Different combinations of sO_2 and hematocrit were tested to provide several possible scenarios. All samples are plotted in figure 6.3.

Samples ranging from 5 up to 100% for the sO_2, and from 9 up to 70 for hematocrit, are considered. This is a full exhaustive range, because it covers all the possible common situations.

However database is not uniform, as most of spectra have sO_2 over 90%, because this is the most frequent range in hemodialysis patients, while most of hematocrit samples are under human standard level, because bovines normally have hematocrit level lower than humans [81].

6.2 Preprocessing for oxygen saturation

Two different models, SVM and neural network, are implemented for the prediction of oxygen saturation. The dataset is pre-processed in order to enhance the predictive power of machine learning.

Figure 6.3: Hct and sO_2 combinations of tested samples

Savitzky-Golays filter [82] is applied on dataset. This is a digital filter, commonly used in spectroscopy, which removes noise while preserving the characteristics of a signal spectrum [83].

Many machine learning estimators require normalized data. Scikit-learn provides different standardization techniques: robust scaler was chosen for this task. This scaler removes the median values and scales the data according to the quantile range, these operations are performed independently on each feature using statistics that are robust to outliers.

Normalized dataset is then randomly splitted into two parts: the training set and the test set.

The training set, a fraction representing 85% of whole data, is used to fit the models, while the remaining 15% of the data, the test set, is used to evaluate the models performances.

The split is performed pseudo-randomly because a seed is used to obtain always the same sequence of training and test sets. This is important to compare different models with the same training and test samples.

Parameter	Value
C	10^3
γ	10^{-3}

Table 6.1: SVM Hyperparameters

6.3 Training Machine Learning models for sO_2

Support vector machine, artificial neural network algorithms and all other preprocessing operations are developed in Python 3.7.SVM was fitted with the following hyperparameters:

k-fold cross validation is used to avoid overfitting in SVM.

For each setting of parameter, the k-fold algorithm follows these steps:

- inputs are splitted in k parts (in this case k=3);
- fitting the algorithm for k-1 parts of inputs (training set);
- evaluation of score for the remaining part (validation set);
- iteration of algorithm for the others k-1 parts;
- evaluation of mean score error for training and validation

Different values of k are tested, but the best result is achieved with k=3.

Artificial neural networks have a set of hyperparameters, as a consequence the optimization process can be long time consuming.

Talos library is used in order to fine tune hyperparameters of neural network. Talos is compatible with Keras and it trains neural networks with different hyperparameters finding the best model solution implementing a Grid Search algorithm.

The hyperparameters list includes: number of hidden layers, learning rate, epochs, activation function and number of neurons. The best solution is finally re-trained by Keras.

In artificial neural networks, the problem of overfitting is overtaken with early stopping criterion. This method stops the training when the error increases, this is a form of regularization used to prevent overfitting.

Keras also provides reduce learning rate on plateau technique: it adjusts the learning rate while monitoring the loss each epoch.

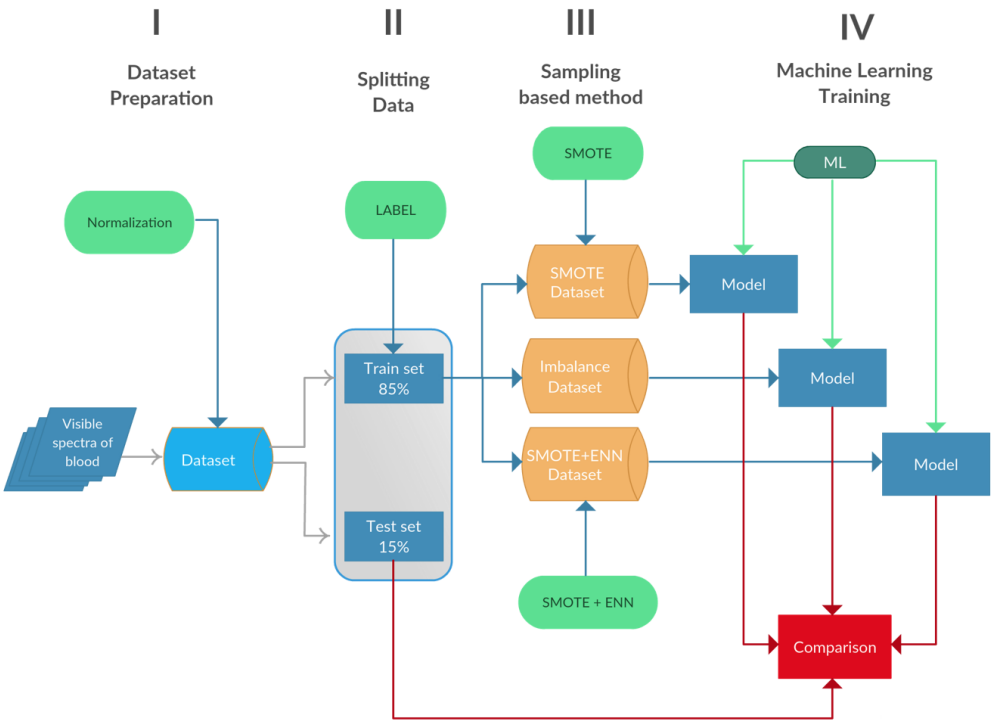


Figure 6.4: Block diagram describing the methodology the workflow for the development of machine learning models for the prediction of hematocrit

6.4 Preprocessing for hematocrit

The size of first dataset of blood spectra have made possible the implementation of algorithms for prediction of oxygen saturation; these machine learning models are very accurate and the results are shown in chapter 7.

Despite high accuracy results for the prediction of sO_2, a more accurate analysis is performed for the prediction of Hct, because the dataset requires more examples for the prediction of hematocrit: the models results inaccurate and the imbalanced data generates a bias in the system which alters the predictions for samples with high hematocrit level.

To overcome this issue and increase prediction accuracy of hematocrit in the whole range from 9 up to 70, a different approach is followed: the dataset is increased and new spectra are added into it, moreover the dataset is manipulated with balancing techniques.

The block diagram in figure 6.4 summarizes the methodology proposed for the prediction of hematocrit.

A simple preprocessing normalization is used before splitting the data into train and test set.

Normalization changes the values of the features to a common scale, without distorting the ranges of values, in order to enhance the accuracy of machine learning.

After normalization, labels are added to the train set to divide the data into two classes:

- spectra with hematocrit level lower or equal to 35 belong to Class 0;
- Spectra who belong to hematocrit greater than 35 belong to Class 1.

Class 0 represents the most frequent class, while class 1 has lower number of samples in it.

The SMOTE and SMOTE+ENN are applied to the same training set. Three different training sets are prepared:

- the original imbalance training set,
- the balanced training set where SMOTE is applied;
- the balanced training set obtained through SMOTE and ENN as data cleaning technique.

The class labels are then removed and the training sets are all separately fitted to implement models. The comparison of models accuracy involves evaluation of regression score function (r^2) and mean squared error (MSE) for all the models. Both parameters evaluate the error between desired and predicted values.

However, MSE and r^2 give only a statistical evaluation of the overall error. In classification tasks, there are different evaluation parameters, such as sensitivity, specificity or ROC curve. These evaluation metrics reflect more accurately the performance on imbalanced dataset than the standard ones, because they take into account of minority class which are harder to detect due to the smaller number of samples in training set.

In regression, there are not advanced evaluation metrics that are focused on less frequent samples in training set.

Therefore, it can be convenient to measure the accuracy of the models using the standard evaluation parameters for regression (MSE and r^2), but they only evaluate error on less frequent data.

A subset is generated: it includes all the data from test set with hematocrit over the threshold value (35), which is the standard hematocrit human level. The histogram in figure 6.5 shows the distribution of data in the subset. It is crucial to focus on these data, because during real dialysis treatment on ex-vivo human blood most of data will be within this range.

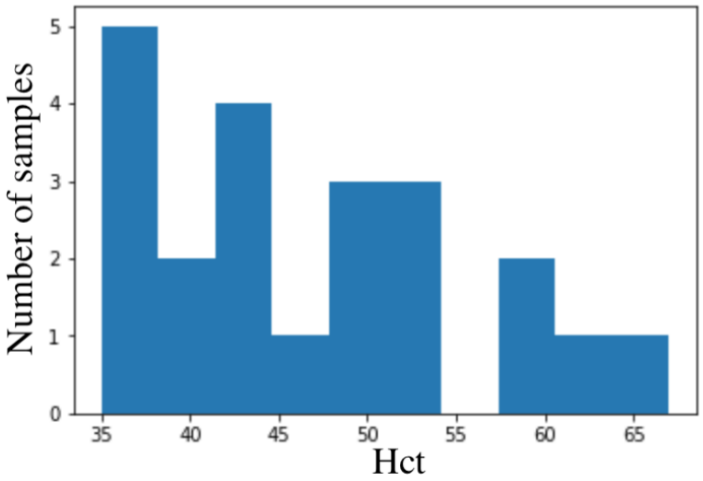


Figure 6.5: Test set with values within standard human range

6.5 Composition of second dataset

In the second dataset, there are 293 different spectra of animal blood at different hematocrit levels.

Each sample is composed by 288 values of absorbance at different wavelengths in visible range. The dataset is scaled using Robust Scaler normalization [84] provided by Scikit-learn library.

After normalization, the dataset is randomly split in train and test datasets. The train dataset is composed by 249 samples: it represents the 85% of whole data, the remaining

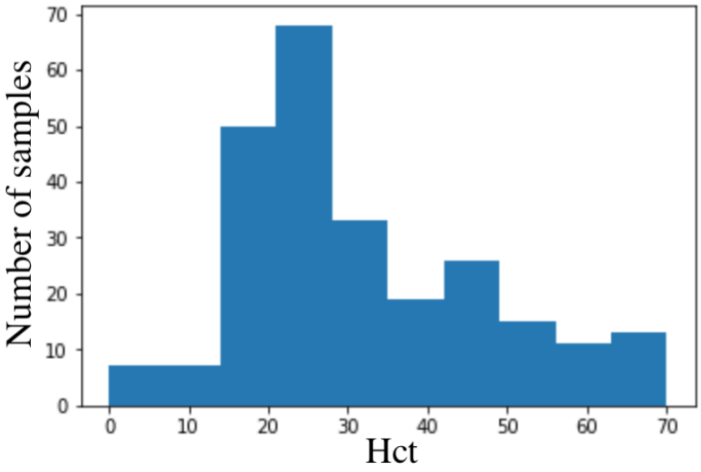


Figure 6.6: Distribution of samples in training set

15% representing the test set.

The histogram in figure 6.6 shows the distribution of samples in the training dataset. The dataset is imbalanced: there are a lot of samples with hematocrit level close to 25.

Models are influenced by these data and it will be difficult to predict accurately samples with higher Hct.

After splitting the data, it is possible to add label class to train dataset choosing 35 as threshold value.

The training set is consequently composed by:

Imbalanced Dataset	
# Total training samples	249
# Data in class 0	150
# Data in class 1	99

Table 6.2: Class distribution samples of Imbalanced dataset

SMOTE and SMOTE+ENN methods are applied using Imbalanced-learn library [85]. The SMOTE dataset is composed by:

SMOTE Dataset	
# Total training samples	300
# Data in class 0	150
# Data in class 1	150

Table 6.3: Class distribution samples of SMOTE dataset

while SMOTE+ENN dataset is composed by:

SMOTE + ENN Dataset	
# Total training samples	290
# Data in class 0	145
# Data in class 1	145

Table 6.4: Class distribution samples of SMOTE + ENN dataset

The SMOTE+ENN dataset size is lower than SMOTE dataset one because ENN operates a data cleaning from both classes, if these data belong to border class decision.

The class labels are then removed and the datasets are trained.

The oversampling technique balances the data increasing the number of data in class 1.

Ridge Regression	
α	0.1
Tolerance	0.00001
Max Iteration	None
Solver	Auto

Table 6.5: Ridge Regression Hyperparameters

Elastic Net	
α	1
R	0.8
Max Iteration	10000
Tolerance	0.0001
Selection	Cyclic

Table 6.6: Elastic Net Hyperparameters

Random Forest	
Criterion	MSE
Number of estimators	100
Max depth	15
Min samples split	2
Max features	log2

Table 6.7: Random Forest Hyperparameters

Artificial Neural Network	
Number of hidden layer	1
Number of neurons in hidden layer	16
Activation function	Elu
Kernel initializer	Normal
Optimizer	Adam
Epochs	2000

Table 6.8: Artificial Neural Network Hyperparameters

6.6 Training Machine Learning models for Hct

In order to find the best machine learning model for the prediction of hematocrit, four different machine learning algorithms are implemented with the three different training

datasets in tables and, then, compared with same test samples.

The investigated machine learning techniques are:

1. Ridge Regression
2. Elastic Net
3. Random Forest
4. Artificial Neural Network

These four different algorithms are all optimized through hyper parameter optimization techniques. In linear models, different values of penalty factor are manually tested, to finally obtain the most accurate result.

Moreover, a grid search is used to fine tune hyper parameters in random forest. The hyper parameter optimization of ANN is trickier, because there are a lot of parameters to take into account.

Talos library is used to find the best combination of parameters by performing a grid search in ANN.

All the details about the final hyperparameters of the machine learning models with the applied parameters are reported in tables 6.5, 6.6, 6.7 and 6.8

CHAPTER 7

Results

In this chapter, the results of machine learning models for the prediction of oxygen saturation and hematocrit will be discussed.

The aim is to find the best machine learning technique for the prediction of both parameters; the best solution is the one which provides the highest accuracy with test samples.

The comparison between different models is possible using evaluation metrics explained in chapter 4 section Evaluation Metrics.

Prediction of oxygen saturation has been performed with SVM and ANN, while four different techniques have been investigated for hematocrit.

7.1 Results for sO_2

Support vector machine and artificial neural network have been compared through evaluation parameters, to verify the accuracy of both models for the prediction of oxygen saturation.

Table 7.1 shows the overall performance of support vector machine and neural networks for prediction of oxygen saturation on test set, both models show similar accuracy performances.

Figure 7.1 shows the regression plot on the test set. Regression plot analysis function compares actual outputs of two algorithms with the corresponding desired ones (targets). In figure 7.1, x-axis represents the target values, y-axis represents the predicted values,

ML algorithm	MSE	MAE	r^2
SVM	0.51	0.55	0.99
ANN	2.4	1.2	0.99

Table 7.1: Performance results for oxygen saturation

line represents the perfect fitting between target and predicted values, while scatter points represent test samples. Results are excellent for both the machine learning-based algorithms, they provide very accurate predictions. For both algorithms, the coefficient of determination is equal to 99%, so models report very high performance, but SVM is the best machine learning algorithm, because MSE and MAE are lower than the ones of ANN. These results show that both SVM and ANN techniques are able to predict accurately oxygen saturation.

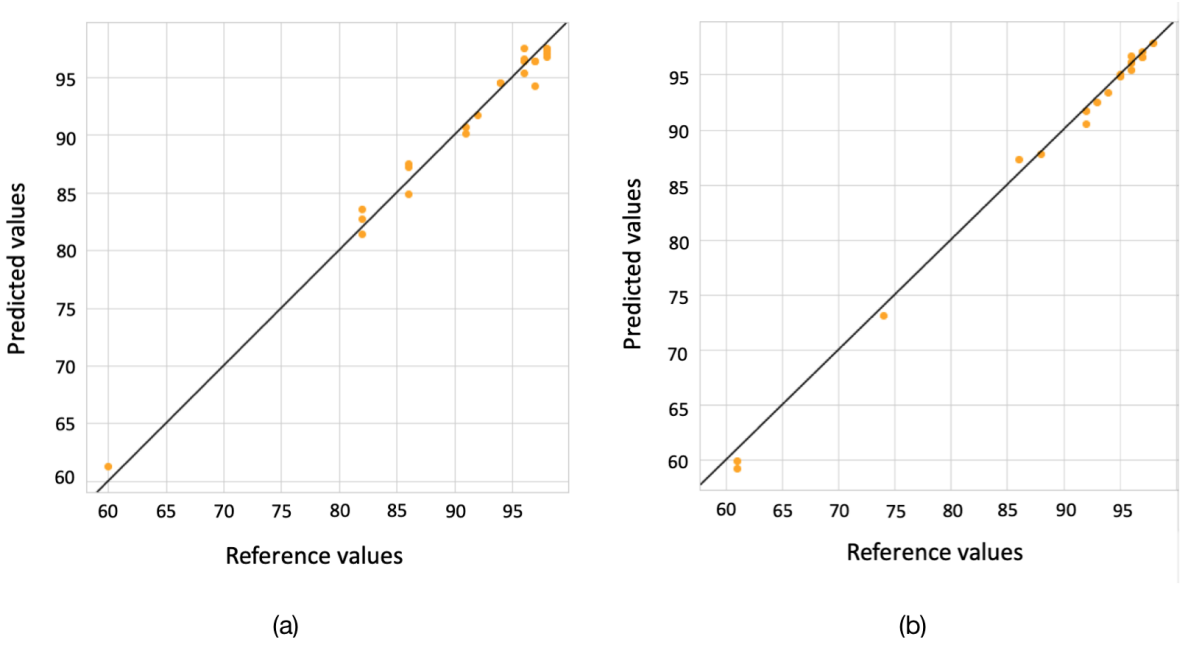


Figure 7.1: Regression plots of SVM (a) and ANN (b) on test set

7.2 Results for Hct

Four models are trained with all the three datasets and the performance of each model is evaluated on the same test set. Results are reported in tables 7.2, 7.3, 7.4, 7.5:

Dataset	MSE	r^2
Imbalanced Dataset	21.57	0.90
SMOTE	27.20	0.87
SMOTE+ENN	27.96	0.87

Table 7.2: Performances of Ridge Regression on test sets

Dataset	MSE	r^2
Imbalanced Dataset	23.06	0.89
SMOTE	26.70	0.88
SMOTE+ENN	27.39	0.87

Table 7.3: Performances of Elastic Net on test sets

Tables 7.2, 7.3, 7.4, 7.5 show comparative performance results of different machine learning techniques fitted with different training dataset. Ridge regression and Elastic Net show close results; both linear models are very accurate, showing a small error and high r^2. Moreover, the balancing techniques do not increase the performance, because the model trained with imbalanced datasets shows lower MSE in both Ridge and Elastic Net. Despite of hyperparameter optimization, Random Forest shows lower performance than linear models. In this case, the best model in terms of r^2 and MSE is the one fitted with the imbalanced dataset. Moreover the performance on model trained with balanced datasets is very low. ANN is the most promising machine learning technique for prediction of hematocrit. The models are very precise with the highest r^2 and the lowest MSE among all the models. Figure 7.2 shows the linear regression plots of the models implemented by ANN techniques. They are trained with SMOTE dataset (a) and SMOTE + ENN

Dataset	MSE	r^2
Imbalanced Dataset	37.30	0.82
SMOTE	52.13	0.76
SMOTE+ENN	49.56	0.77

Table 7.4: Performances of Random Forest on test sets

Dataset	MSE	r^2
Imbalanced Dataset	16.81	0.92
SMOTE	15.01	0.93
SMOTE+ENN	11.11	0.95

Table 7.5: Performances of Neural Networks on test set

(b) and represent the most accurate models, with a r^2 equals to 0.93 (a) and 0.95 (b). The test set has the same distribution of the training set: therefore, the models are fitted with training data similar to the test set. The result shows higher accuracy for models trained with imbalanced dataset. The results are different if the same statistical analysis is

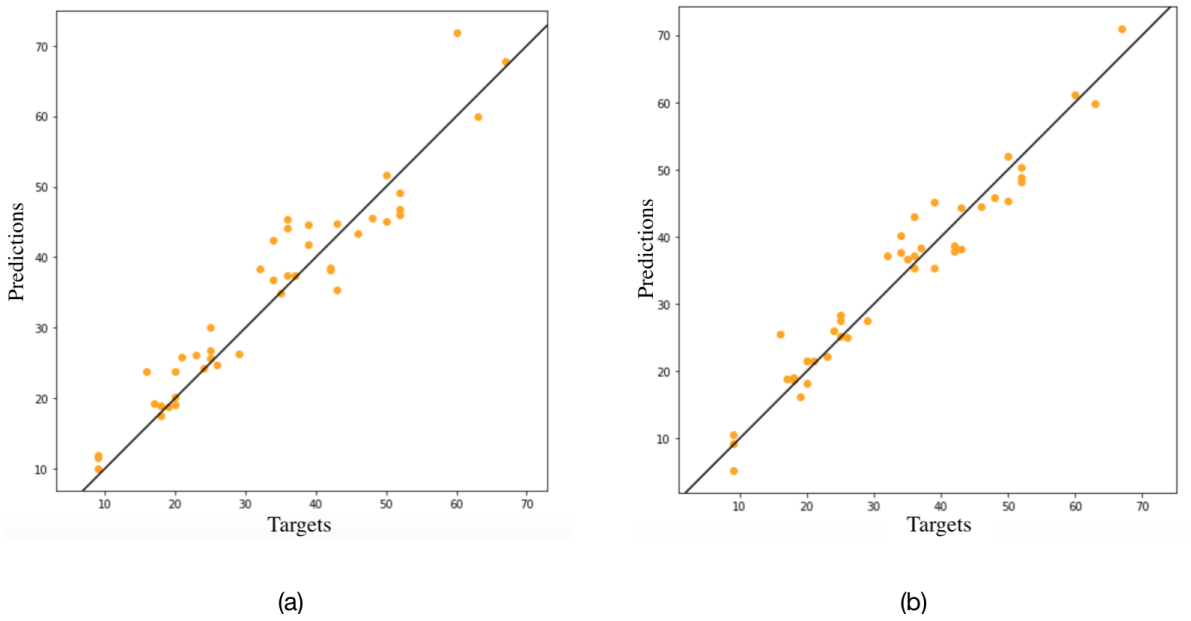


Figure 7.2: Regression plot of ANN fitted with SMOTE (a) and SMOTE+ENN (b) dataset

conducted on only tests that are within human standard range for hematocrit. Therefore, same statistical performance analysis is carried out on a smaller test set, where samples with Hct in the range between 35 and 67 are considered. The results are reported in the tables 7.6, 7.7, 7.8 and 7.9: The values of r^2 are generally lower than the ones evaluated with entire test set. This shows the difficulty of machine learning models to predict data that are less frequent during training. Ridge and Elastic Net show better prediction if

Dataset	MSE	r^2
Imbalanced Dataset	22.46	0.72
SMOTE	28.08	0.65
SMOTE+ENN	29.26	0.64

Table 7.6: Performances of Ridge Regression on human standard range samples

Dataset	MSE	r^2
Imbalanced Dataset	25.54	0.69
SMOTE	27.45	0.66
SMOTE+ENN	28.85	0.65

Table 7.7: Performances of Elastic Net on human standard range samples

Dataset	MSE	r^2
Imbalanced Dataset	65.02	0.18
SMOTE	79.02	0.01
SMOTE+ENN	71.36	0.10

Table 7.8: Performances of Random Forest on human standard range samples

Dataset	MSE	r^2
Imbalanced Dataset	23.25	0.72
SMOTE	12.87	0.86
SMOTE+ENN	13.49	0.85

Table 7.9: Performances of Neural Networks on human standard range samples

they are fitted with imbalance dataset, while random forest is not able to predict samples in this target range. Once again, ANN is the best technique in terms of error and r^2. It is therefore evident the positive effect of balancing the dataset: SMOTE and SMOTE with ENN allow a significant improvement in model accuracy. Figure 7.3 shows the linear regression fitting of the model implemented by ANN technique and trained with balanced datasets. The plot shows high accuracy of the two models in the prediction of hematocrit with human standard range samples. The network fitted with both SMOTE and SMOTE with ENN training dataset are the best solutions for the prediction of samples belonging to human range of hematocrit.

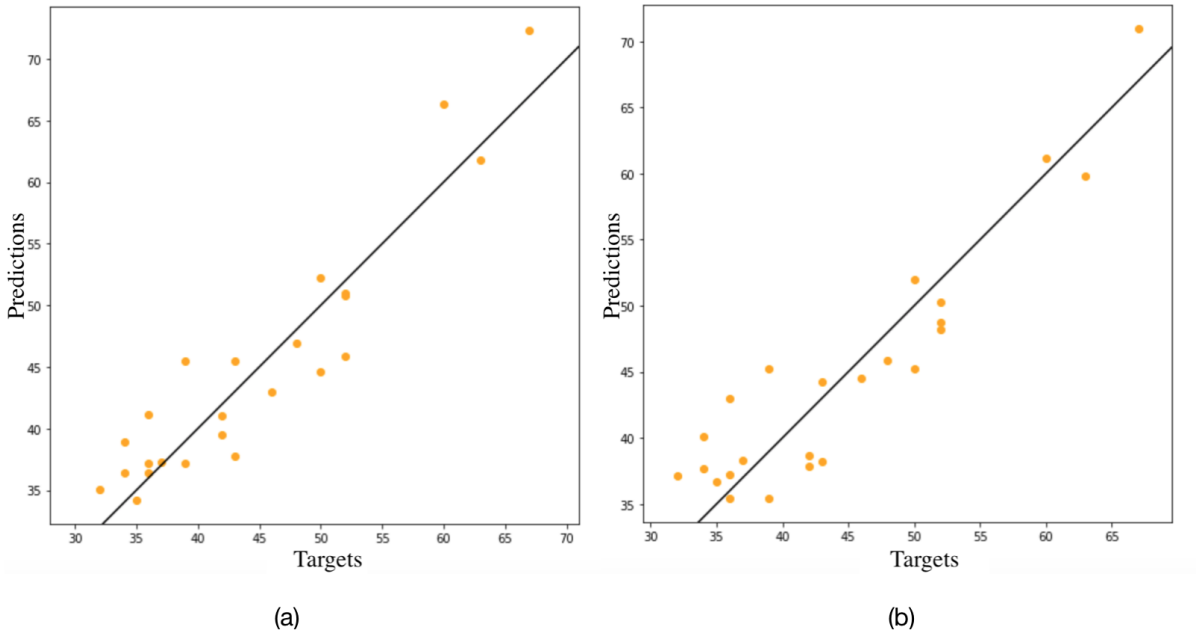


Figure 7.3: Regression plot of SMOTE (a) and SMOTE+ENN (b) on human standard range

CHAPTER 8

Conclusion

This thesis is the result of a three year design, prototype and testing efforts on the development of new real-time monitoring system of hematic parameters for extra-corporeal treatments. The system satisfies all the clinical needs, because it shows a better way to monitor hematocrit and oxygen saturation during hemodialysis without any contact with blood and it easily integrates itself within dialysis machines without interfering with the treatment in order to reduce the risk of over-treatment. The proposed setup shows some advantages determined by the fact that, being based on a spectroscopic approach, leads to an increase amount of information that can be profitably managed by machine learning techniques to obtain the parameters of interest. Two different prototypes have been realized and their hardware and software have been shown in chapter 5, they have been tested in a significant operational environment (chapter 6) in order to validate the system and create a database that is used to train different machine learning models. The results in chapter 7 shows the application of a machine learning combined with a low cost spectroscopic based setup for monitoring hematic parameters of blood. A support vector machine and an artificial neural network have been implemented and applied to data obtained through spectrometry in the visible and near infrared of different blood samples. Results demonstrate that SVM and ANN models achieves good learning performances and both show the ability to learn relationship between input and sO_2. In term of accuracy, the most promising algorithm is SVM, but both machine learning methods are able to elaborate accurate predictive models. The dataset is imbalanced and this has a negative impact on the prediction of Hct, but two oversampling methods have been analyzed

and successfully applied to increase accuracy. SMOTE and SMOTE+ENN have been applied on same dataset, resulting in two different datasets. These datasets have been fitted, along with imbalance one, with different machine learning techniques to compare the performance. The results show an increase in performance for ANN models fitted with balanced dataset for human values of hematocrit. SMOTE or SMOTE with ENN allows the implementation of more accurate neural network models, improving the performance of machine learning models and reducing the error for the prediction of hematocrit. The workflow depicted in chapter 1 (figure 1) was followed and all the milestones have been accomplished including the realization of the stand-alone prototype, because it includes all the electronics, hardware components and the trained models, so it allows the real-time monitoring of hematocrit and oxygen saturation.

8.1 Future works

The combination of spectrometer and machine learning algorithms shows accurate measurements for Hct and sO_2, but further studies could be conducting to use the same setup along with machine learning in order to measure other different blood analytes. This represents a decisive improvement, as allows the development of more effective devices able to process the large amount of data contained in the spectrum providing more information at comparable costs.

More measures and tests can be conducted to increase the size of database: new samples will increase the accuracy of models, tests on human patients will allow the creation of new database that will be more balanced on human standard range of hematic parameters. This device, allowing a real-time monitoring of the blood during the treatment, can also be used to verify the regular process development and alert health personnel in case of anomalies. Moreover, the sensitivity of the system can be increased and the machine learning algorithms can be made more efficient to allow a more precise detection.

This approach can be useful to measure the hematocrit before and after the dialyzer and verify its performance. Moreover, measurements can be performed to show the presence of blood and inform about possible breakage.

Finally, the cuvette can be realized through injection molding in order to have a more

reliable disposable for measurements. This work demonstrates that machine learning applied to the analysis of spectroscopic data can be of great help in diagnosis of human health diseases. This is only one of the possible applications of this approach, but many other applications can be developed in the next future.

Bibliography

[1] MISTER Smart Innovation. Mister smart innovation website. `https://www.laboratoriomister.it`. Accessed on 31.01.2020.

[2] Medica spa. Medica spa website. `https://www.medica.it`. Accessed on 31.01.2020.

[3] The National Institute of Diabetes, Digestive, and Kidney Diseases Health Information Center. Your kidneys & how they work. `http://www.niddk.nih.gov/`, June 2018.

[4] National Kidney Foundation. How your kidneys work. `https://www.kidney.org/`. Accessed on 31.01.2020.

[5] Remuzzi G. Schieppati A. Chronic renal diseases as a public health problem: epidemiology, social, and economic implications. *Official Journal of the international society of nephrology*. Doi:10.1111/j.1523-1755.2005.09801.x.

[6] P. Czarniak E. Kraszewska S. Lizakowski R. Szubert S. Czekalski W. Sulowicz E. Krl, B. Rutkowski and A. Wilcek. Early detection of chronic disease: results of the polnef study. *American Journal of Nephrology*, 29:264–273, 2009.

[7] Mayo Clinic Staff. Chronic kidney disease. `https://www.mayoclinic.org/`. Accessed on 31.01.2020.

[8] American Kidney Fund. Stages of chronic kidney disease (ckd). `https://www.kidneyfund.org/`. Accessed on 31.01.2020.

[9] Jina Sawani. New report captures the high burden, high cost and low awareness of kidney disease in the united states. `http://www.uofmhealth.org/`. based on report from American Journal of Kidney Diseases.

[10] Iseki K et al. Jha V., Garcia G. Chronic kidney disease: global dimension and perspectives, May 31, 2013. Doi:10.1016/S0140-6736(13)60687-X.

[11] World Kidney Day: Chronic Kidney Disease. Chronic kidney disease. `https://www.worldkidneyday.org`. Accessed on 31.01.2020.

[12] Mendis S Tonelli M Couser WG, Remuzzi G. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney International*, 80, 12 October 2011. Doi: https://doi.org/10.1038/ki.2011.368.

[13] National Kidney Foundation. Dialysis. `https://www.kidney.org/`. Accessed on 31.01.2020.

[14] NHS. Overview dialysis. `https://www.nhs.uk/`, 14 June 2018. Accessed on 31.01.2020.

[15] Mayo Clinic Staff. Peritoneal dialysis. `https://www.mayoclinic.org/`, 24 April 2019. Accessed on 31.01.2020.

[16] NDDK. Treatment methods for kidney failure hemodialysis, December 2006.

[17] Fresenius Medical Care. What is dialysis and how does dialysis work? `https://www.freseniusmedicalcare.com`. Accessed on 31.01.2020.

[18] NIH National Institute of Diabetes, Digestive, and Kidney Diseases. Hemodialysis, 2018.

[19] Mayo Clinic Staff. Hemodialysis. `https://www.mayoclinic.org/`, 2019. Accessed on 31.01.2020.

[20] R. Jasmer L. Maarks. What is hemodialysis. `https://www.everydayhealth.com/`, 10.23.2015. Accessed on 31.01.2020.

[21] Datamed. Obba. `https://datamedsrl.com/`. Accessed on 31.01.2020.

[22] Roberto Pozzi Giampiero Porro and Alessandro Torinesi, 15.02.2018. WO Patent 2010136962A1.

[23] Fresenius. Crit-line iv monitor: Specification sheet. https://fmcna.com/. Accessed on 31.01.2020.

[24] Sammann K. L. Barrett, D. Peterson. Measuring hematocrit and estimating hemoglobin values with a non-invasive, optical blood monitoring system, 10/16/2012. Application number: 13/366119.

[25] M. Knapp E. Gaubitzer M. Puchinger A. Shahzad, G. Kohler and M. Edetsberger. Emerging applications of fluorescence spectroscopy in medical microbiology field. *Journal of Translational Medicine*, 26 November 2009. 10.1186/1479-5876-7-99.

[26] C Boone C. Crum R. Dasari I. Georgakoudi K. Keefe K. Munger S. Shapshay E. Sheetse M. Feld K. Badizadegan, V. Backman. Spectroscopic diagnosis and imaging of invisible pre-cancer. *Faraday Discussion*, 126, 2004. Doi: 10.1039/b305410a.

[27] Nancy Lewen. The use of atomic spectroscopy in the pharmaceutical industry for the determination of trace elements in pharmaceuticals. *Journal of Pharmaceutical and Biomedical Analysis*, 55(4):653 – 661, 2011. Doi: https://doi.org/10.1016/j.jpba.2010.11.030.

[28] Sanjay M Nilapwar, Maria Nardelli, Hans V Westerhoff, and Malkhey Verma. Absorption spectroscopy. *Methods in enzymology*, 500:59?75, 2011. Doi: 10.1016/b978-0-12-385118-5.00004-9.

[29] RSC Advancing Chemical Sciences. Ultraviolet-visible scpectroscopy. *The royal society of chemistry.*

[30] Pierce Biotechnology Inc. Extinction coefficients-technical resource. October 2002.

[31] A. Timothy Lovell, Jeremy C. Hebden, John C. Goldstone, and Mark Cope. Determination of the transport scattering coefficient of red blood cells. 3597:175 – 182, 1999. Doi: 10.1117/12.356795.

[32] G. H. Orians W. K. Purves and H. C. Heller. Corso di biologia. *Zanichelli Editore, Bologna*, 1995.

[33] Andreina e Alma Tagliabue A.P. Baracchi. Elementi di chimica. 2012. Editore Lattes.

[34] E. Johnson J. Johnson O. Korol D. Kruse B. Poe J. Wise M. Womble K. Young J. Betts, P. Desaix. Anatomy and physiology. `http://cnx.org/contents/14fb4ad7-39a1-4eee-ab6e-3ef2482e3e22@8.24.`, Feb 2016.

[35] Alana Biggers. Is my blood oxygen level normal? `http://www.healthline.com/health/normal-blood-oxygen-level`.

[36] William C. Shiel. Hematocrit. `http://www.medicinenet.com/hematocrit/article.htm/what_is_the_hematocrit`. Accessed on 31.01.2020.

[37] Laura del Senno Gian Luigi Castoldi. Erythrocytes. *Encyclopedia of Immunology.*, 1998.

[38] Alex Hern. Partnership on ai' formed by google, facebook, amazon, ibm and microsoft, 2016.

[39] Junaid Khan Vijay Chauhan, Arunima Jaiswal. Web page ranking using machine learning approach. *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. Doi: 10.1109/ACCT.2015.56.

[40] Mohammed Mehedi Hassan, Md. Zia Uddin, Amr Mohamed, and Ahmad Almogren. A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, 81:307 – 313, 2018. Doi: https://doi.org/10.1016/j.future.2017.11.029.

[41] E. V. Polyakov, M. S. Mazhanov, A. Y. Rolich, L. S. Voskov, M. V. Kachalova, and S. V. Polyakov. Investigation and development of the intelligent voice assistant for the internet of things using machine learning. In *2018 Moscow Workshop on Electronic and Networking Technologies (MWENT)*, pages 1–5, 2018. Doi: 10.1109/MWENT.2018.8337236.

[42] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International*

Conference on Computing Networking and Informatics (ICCNI), pages 1–9, 2017. Doi: 10.1109/ICCNI.2017.8123782.

[43] Q. Rao and J. Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, pages 35–38, May 2018.

[44] Julie Cattiau. How ai can improve products for people with impaired speech. `https://blog.google`, 7 May 2019. Published on bloog.google.

[45] Musk Elon. An integrated brain-machine interface platform with thousands of channels. *bioRxiv*, 2019. Doi: 10.1101/703801.

[46] Marc Coram M. et al. Varun Gulshan, Lily Peng. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, December 13, 2016. Doi: 10.1001/jama.2016.17216.

[47] Mohammad Norouzi George E. Dahl Timo Kohlberger Aleksey Boyko Subhashini Venugopalan Aleksei Timofeev Philip Q. Nelson Greg S. Corrado Jason D. Hipp Lily Peng Martin C. Stumpe Yun Liu, Krishna Gadepalli. Detecting cancer metastases on gigapixel pathology images. *Computer Vision and Pattern Recognition*, March 2017. `http://arxiv.org/abs/1703.02442`.

[48] Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12(4):1–14, 04 2017.

[49] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt Publishing, 2017.

[50] Aurelien Geron. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2017.

[51] Travis E. Oliphant. *Guide to NumPy*. CreateSpace Independent Publishing Platform, 2nd edition, 2015.

[52] Wes Mckinney. *Python for Data Analysis*. O'Reilly, 2nd edition, September 2017.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning research*, 2011.

[54] François Chollet et al. Keras. `https://keras.io`, 2015.

[55] Autonomio. Talos [computer software]. `http://github.com/autonomio/talos`. Accessed on 31.01.2020.

[56] Andrew Ng. Support vector machine. `http://cs229.stanford.edu/notes/cs229-notes3.pdf`. CS229 Lecture Notes Part V.

[57] Vladimir Vapnik Corinna Cortes. Support vector networks. *Machine Learning*, 20:273–297, September 1995. Doi: 10.1007/BF00994018.

[58] Andrej Krenker, Janez Bes?ter, and Andrej Kos. Introduction to the artificial neural networks. In Kenji Suzuki, editor, *Artificial Neural Networks*, chapter 1. IntechOpen, Rijeka, 2011. Doi: 10.5772/15751.

[59] Nov 2016 Bartosz Krawczyk, 5. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5:221–232, 2016. Doi: 10.1007/s13748-016-0094-0.

[60] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6:429–449, 11 2002. Doi: 10.3233/IDA-2002-6504.

[61] Salvador Garca Julin Luengo, Alberto Fernndez and Francisco Herrera. Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling. *Soft Computing*, 15:1909–1936, 2011. Doi: 10.1007/s00500-010-0625-8.

[62] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 01 2002. Doi: 10.1613/jair.953.

[63] Gustavo Batista, Ronaldo Prati, and Maria-Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6:20–29, 06 2004. Doi: 10.1145/1007730.1007735.

[64] Ivan Tomek. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(6):448–452, June 1976. Doi: 10.1109/TSMC.1976.4309523.

[65] Martin Grimheden Anders Berglund. The importance of prototyping for education in product innovation engineering. 2011. Research Publishing ISBN:978-981-08-7721-7.

[66] Hamamatsu Datasheet. Finger-tip sized, ultra-compact spectrometer head supporting high sensitivity and long wavelength region. `http://www.hamamatsu.com/resources/pdf/ssd/c12880ma_kacc1226e.pdf`. Accessed on 31.01.2020.

[67] A. Boukhayma, A. Peizerat, and C. Enz. Temporal readout noise analysis and reduction techniques for low-light cmos image sensors. *IEEE Transactions on Electron Devices*, 63(1):72–78, Jan 2016. Doi: 10.1109/TED.2015.2434799.

[68] Microchip. Mcp6001r datasheet 1 mhz, low-power op amp. `https://ww1.microchip.com/downloads/en/DeviceDoc/21733j.pdf`.

[69] ON Semiconductor. Quad bus buffer with 3-state control inputs. `https://www.onsemi.com/pub/Collateral/MC74VHCT125A-D.PDF`.

[70] J. Andrew. Sexton. Vibration and thermal vacuum qualification test results for a low-voltage tungsten-halogen light. Publication date: February 01, 1991. Accessed on 31.01.2020.

[71] Samsung. Lh351b datasheet. high power led series 3535 ceramic hot binning. `https://docs-emea.rs-online.com/webdocs/14f6/0900766b814f6951.pdf`.

[72] Thorlabs. Optic fiber datasheet 0.39 na tecs hard-clad step-index, multimode fiber. `https://www.thorlabs.com/drawings/660bad8eb02230ee-8530D70F-D363-D1E0-8C6C873DFA526BD8/FT600EMT-SpecSheet.pdf`. Accessed on 31.01.2020.

[73] Hellma Analytics. Optical components for uv/vis/nir spectroscopy. https://www.hellma.com/en/laboratory-supplies/. Accessed on 31.01.2020.

[74] Formlabs. Formlabs website. https://formlabs.com. Accessed on 31.01.2020.

[75] Ultimaker. Ultimaker website. https://ultimaker.com. Accessed on 31.01.2020.

[76] Raspberry Pi Foundation. Raspberry pi 3 model b+. https://static.raspberrypi.org/files/product-briefs/Raspberry-Pi-Model-Bplus-Product-Brief.pdf. Accessed on 31.01.2020.

[77] U. Nanda and S. K. Pattnaik. Universal asynchronous receiver and transmitter (uart). In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 01, pages 1–5, Jan 2016. Doi: 10.1109/ICACCS.2016.7586376.

[78] C. Decaro, G. B. Montanari, R. Molinari, A. Gilberti, D. Bagnoli, M. Bianconi, and G. Bellanca. Machine learning approach for prediction of hematic parameters in hemodialysis patients. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–8, 2019. Doi: 10.1109/JTEHM.2019.2938951.

[79] Blood fractionation. *Journal of the American Medical Association*, 156(8):772–772, 10 1954. Doi: 10.1001/jama.1954.02950080020009.

[80] Sohrab Mansouri. Kevin Fallon. The gem premier 3000 with intelligent quality management: Features. technical description & statistical validation. Accessed on 31.01.2020.

[81] Leonie Roland, Marc Drillich, and Michael Iwersen. Hematology as a diagnostic tool in bovine medicine. *Journal of Veterinary Diagnostic Investigation*, 26, 08 2014. Doi: 10.1177/1040638714546490.

[82] Jianwen Luo, Kui Ying, and Lijing Bai. Savitzky?golay smoothing and differentiation filter for even number data. *Signal Processing*, 85:1429–1434, 07 2005. Doi: 10.1016/j.sigpro.2005.02.002.

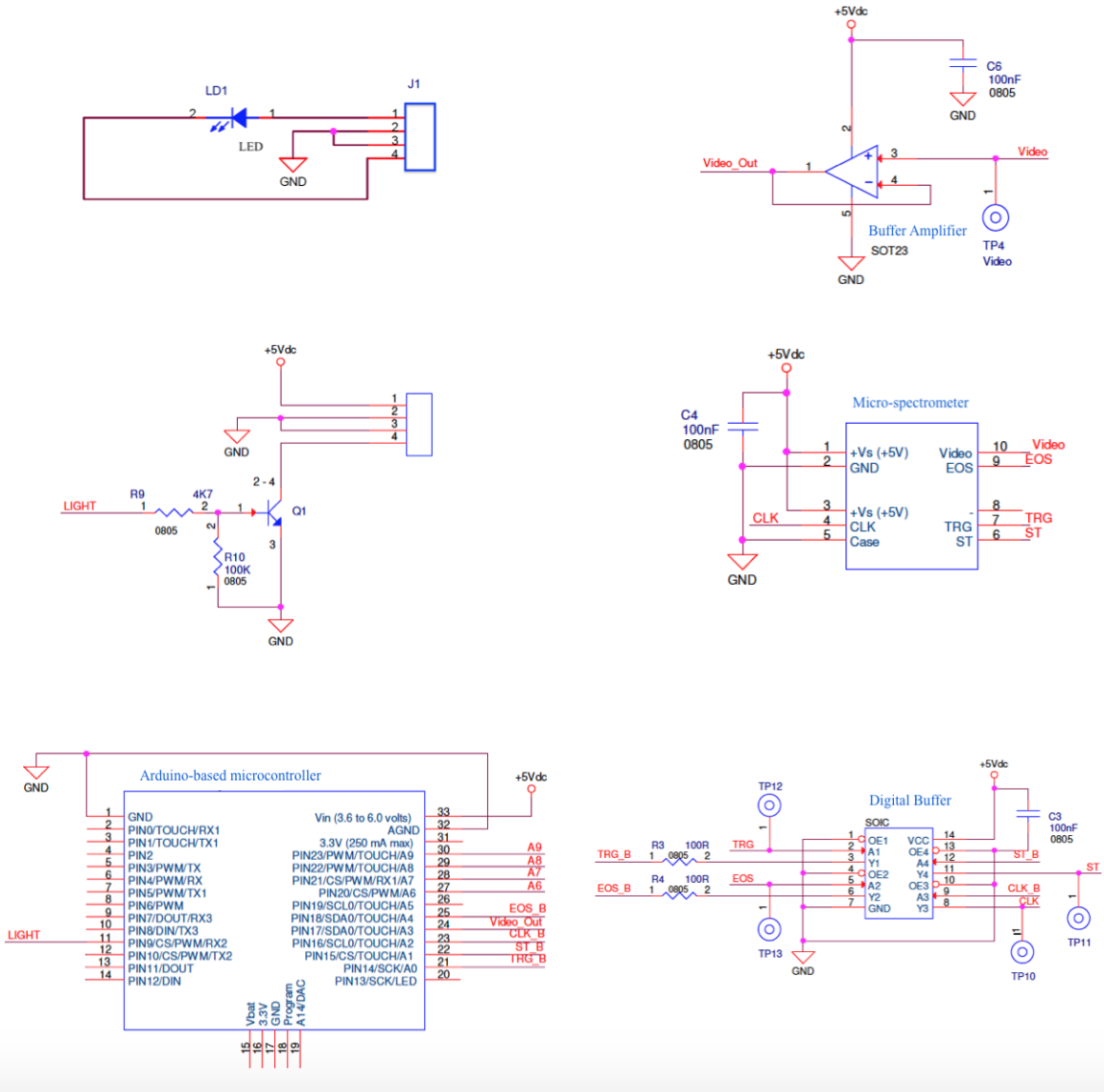
[83] Chris Ruffin and Roger King. The analysis of hyperspectral data using savitzky-golay filtering-theoretical basis. 1. volume 2, pages 756 – 758 vol.2, 02 1999. Doi: 10.1109/IGARSS.1999.774430.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit preprocessing. `https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.robustscaler.html`. Accessed on 31.01.2020.

[85] Guillaume Lematre, Fernando Nogueira, and Christos Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. 18, 09 2016.

Appendix 1

Wiring



Appendix 2

Coagulation

High temperature can cause coagulation of blood in the cuvette, that can alter the measurements and can be dangerous for patients safety, because coagulation can obstruct the flowing within the hematic circuit of hemodialysis machine.

The coagulation occurs due to the increasing temperature provided by the light source, which is placed close to the cuvette. The blood changes its fluidic behaviour, it becomes viscous and it starts to deposit into the cuvette. In this case the optical window is compromised, because a layer of coagulated blood covers the window and the absorption is altered. This phenomenon appeared for the first time during a research test when an halogen lamp was used, instead of LED.

During first attempts, the halogen lamp was turned on before starting each session test and it remained switched on during the whole test, causing a critical increase of temperature. The photo of the cuvette after that session is showed in figure 1, the coagulation obstructs the measurement point and all the measures were altered.

To avoid coagulation, new setup exploits LED, which is driven by microcontroller and it turns on/off the light source only when a measure is performing.

The cuvettes have been verified after each dialysis test sessions and the presence of coagulation never appeared with final prototypes.

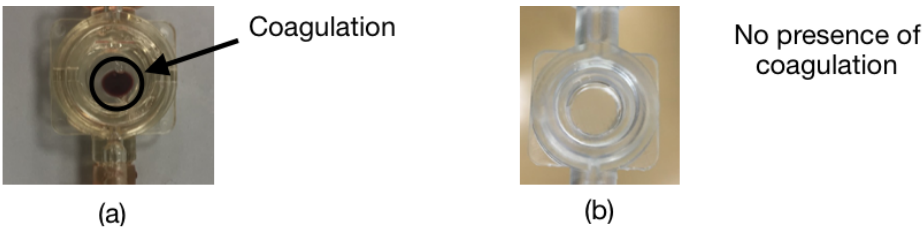


Figure 1: Two cuvettes after a session of dialysis. (a) shows the presence of a coagulation, while (b) does not show the presence of coagulation

Appendix 3

Instructions for the use of stand-alone Prototype

Figure 2: Front view of stand alone prototype

In figure 2 there is the front view of final prototype. It is a stand-alone prototype: raspberry, microcontroller and spectrometer are within the box. When all the connections

are completed, it is possible to switch on the power supply and start with monitoring. The display will assist the operator with all the steps. Before starting with the real measurement of blood, the system asks for a reference value, pushing the reference button, the operator will start the procedure:

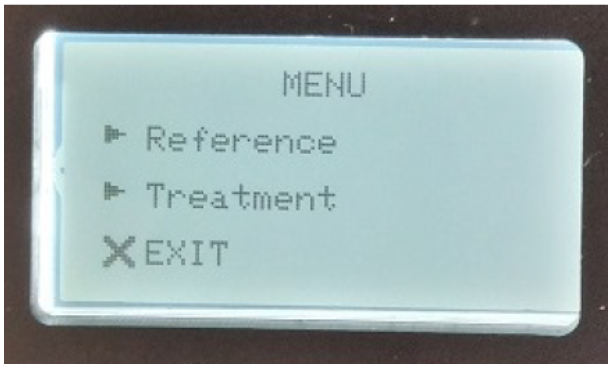


Figure 3: Menu screen

During the measurement the display shows a waiting screen

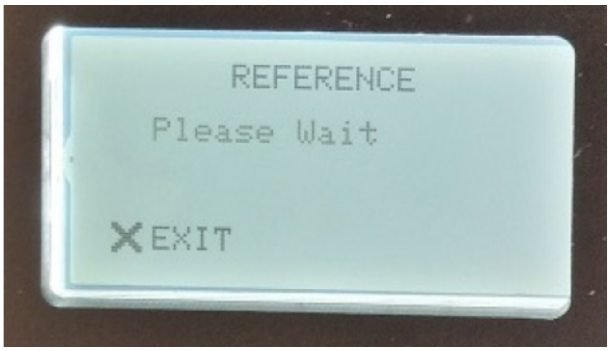


Figure 4: Reference screen

And when the reference is done, the display shows a note to inform the operator:

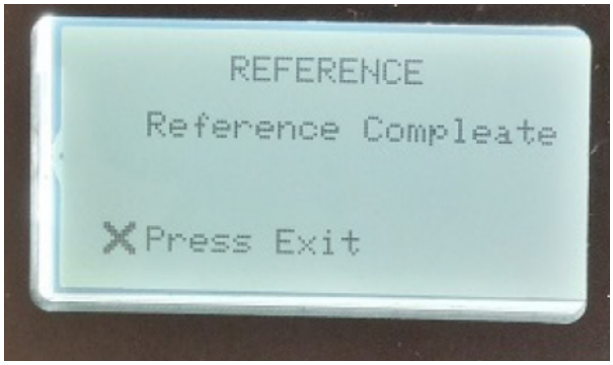


Figure 5: Reference complete screen

The treatment can now start pushing the treatment button:

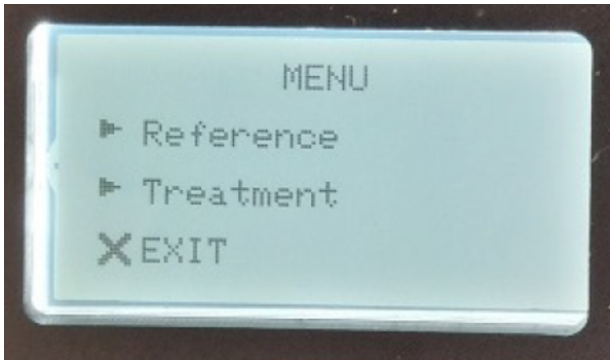


Figure 6: Menu screen

When the measurement is performed, a screen shows Hct and sO_2 values:

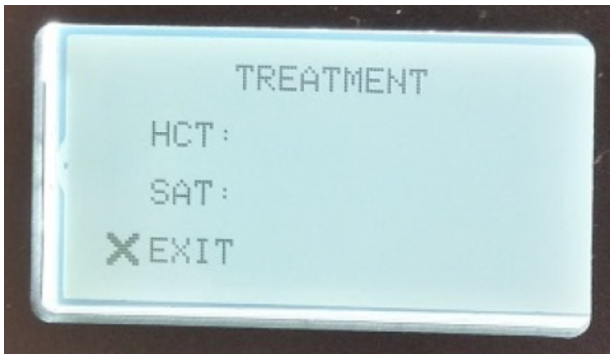


Figure 7: Treatment screen

The exit button stops the treatment.