Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Università degli Studi di Padova

Dipartimento di Ingegneria Dell'Informazione

Corso di Laurea Magistrale in
Ict For Internet And Multimedia

Machine learning applied to Raman gas spectroscopy for biological survey

Relatore:
Prof. Maria pelizzo
Corelatore:
Dott. Lorenzo Cocola

Laureando:
Daniele Barbiero
Matricola: 2027879

Anno accademico 2021/2022
7 Settembre 2022

Contents

List of Figures

Abstract (IT)

I metodi di rilevamento microbiologico tradizionali utilizzati nelle industrie lattiero-casearie per rilevare le spore richiedono molto tempo e sono limitati in termini di efficienza e sensibilità. Secondo la produzione fermentativa di acido butirrico, durante la fermentazione vengono prodotti anche anidride carbonica e idrogeno come sottoprodotti. È stato proposto un sensore multigas basato sulla spettroscopia Raman per la misurazione della concentrazione di CO_2 e H_2 nello spazio di testa delle fiale, poiché la presenza di questi gas indica la contaminazione da batteri. Lo strumento proposto sarà dotato di un caricatore per la misura automatizzata di più provette e di un motore passo-passo che permette la rotazione della singola provetta. L'obiettivo di questa tesi è quello di sviluppare il software dello strumento che controlla la rotazione della provetta, che serve a trovare posizioni in cui lo spettro acquisito non sia degradato da fluorescenza e altri artefatti legati a difetti del vetro. Attraverso l'elaborazione di immagini e algoritmi di machine learning, viene proposta una possibile soluzione (sviluppata con fiale contenenti aria) per il controllo automatico della rotazione della provetta per massimizzare la precisione di misura.

Abastract (EN)

Traditional microbiological detection methods used in dairy industries to detect spores are time consuming and limited in efficiency and sensitivity. According to the fermentative butyric acid production, carbon dioxide and hydrogen are also produced as byproducts during fermentation. A multi-gas sensor based on Raman spectroscopy has been proposed for the measurement of CO_2 and H_2 concentration in the vials headspace, as the presence of these gases indicates contamination by bacteria. The proposed instrument will be equipped with a loader for the automated measurement of several test tubes and a stepper motor that allows the single tube to rotate. This thesis aims to develop the instrument's software that controls the rotation of the test tube, which is done to find positions in which the acquired spectrum is not degraded by fluorescence and other artefacts related to glass defects. Through image processing and machine learning algorithms, a possible solution (developed with vials containing air) is proposed for the automatic control of the test tube rotation to maximise measurement accuracy.

Chapter 1

Introduction

1.1 Project overview

Despite the advances in the dairy industry, the contamination of milk (and therefore its derivatives) by *clostridia* remains both a public health and an economic problem. Current microbiological investigation techniques are costly both in terms of money and time (usually a few days). Furthermore, given the growing demand for raw milk, the dairy industry has the problem of not being able to analyze all the milk samples for the reasons mentioned above, but this market will represent an ever greater opportunity. Also hard and semi hard cheeses, such as Grana Padano and Parmigiano Reggiano suffer late-blowing defect cause by *Clostridium tyrobutyricum* that is resistant to whole-cheese manufacturing [4].
During the fermentation of butyric acid, carbon dioxide and hydrogen are also produced. [5]. If the gases in question are kept in an airtight container, they can be analyzed and therefore indicate whether or not contamination has occurred. Absorption laser spectroscopy would represent an excellent solution but unfortunately, hydrogen is difficult to be detected with this technique.
Headspace Raman spectroscopy offers a non-invasive, low-cost and very fast measurement (in the order of seconds). Raman spectroscopy is based on Raman scattering, which is the phenomenon that occurs when the light scattered by a molecule has a

different frequency from the light incident on it (usually generated by a laser). The shift in frequency contains information on the vibrational states of the molecule and therefore by measuring the shift it is possible to determine the presence or not of a molecule in the sample under examination. The measuring process can be carried out automatically, enabling measurement on a large number of samples and frequent sampling during the incubation period [1]. Furthermore, the samples do not require special procedures before analysis and it could be in principle applied even with other samples, for example, blood. This thesis aims to develop the software to control the rotation of a vial of a Raman spectrometer capable of measuring (mainly, for this application) CO_2 and H_2 in the headspace of test tubes and be able to discriminate between tubes containing clostridia contaminated milk samples, those contaminated with other bacteria and those not contaminated. Since Raman spectroscopy provides information on multiple gases (H_2, O_2, N_2, CO_2), the instrument in its final stage could be applied in other fields, such as in the health sector.

1.2 Milk spoilage

Milk, as a complex natural food matrix, is one of the most important dietary products, which contains nearly all the nutrients necessary to sustain life [7], in addition to the main milk sugar lactose, it also contains proteins (caseins, whey proteins, and minor proteins), essential amino acids, fats, minerals, and vitamins. However, milk is not only highly nutritious for humans, but also an excellent growth substrate for microorganisms [6].
Most frequently microorganisms found in milk can be divided into two groups:

- Pathogenic microorganisms, (e.g., Escherichia coli, staphylococcus, streptococci) cause food poisoning and disease in man and should not be present in the milk.

- Spoilage microorganisms are the cause of the deterioration of food in a state where it is unsuitable for human consumption [13].

Contamination of food has a double negative effect, one economic due to the deterioration of food, one on public health. Microbial spoilage of food is an important issue in terms of economic loss and an estimated loss of almost 33% of the total food supply has been attributed to it [12]. Regarding public health, according to the WHO (World Health Organization) 2015 report, every year as many as 600 million people in the world fall ill after consuming contaminated food, of these, 420,000 people die [11].

In spite of quality control, pasteurization and Ultra-High Temperature (UHT) treatments, numerous outbreaks of foodborne illnesses due to the consumption of contaminated/spoiled dairy products were reported [14].

The dairy industry is in need of a fast, sensitive and cost effective technology for the detection of foodborne pathogens that could successfully address the tasks appointed by legislative bodies [10]. The industry's demand for new technologies is primarily motivated by financial considerations in relation to milk farming, quality control and production accuracy management, such as efficient use of resources and increased safety. Furthermore, the consumption of raw milk is constantly growing and would represent a market opportunity but the main problem is that in these conditions raw milk cannot be certified safe within its life cycle: raw milk is a highly perishable food whilst pathogen analyzes are time consuming. As consequence, the microbial analysis of raw milk is merely retrospective, it has statistical value but cannot prevent a disease outbreak. The milk processing industry also requires tools for the rapid detection of clostridia in milk, as these can bubble the cheese even when the cheese is already maturing, causing a large loss of product. In addition, the available methods are excessively expensive to apply to the certification of every raw milk consignment [13].

The main problem with the tests that look for bacterial contamination in milk is that they are time consuming, they also take several days to obtain confirmation and require trained personnel who can interpret the results.

1.3 Gas detection for biological survey

The correlation between gas production and bacterial infection is already used in some protocols for contamination detection. The sample is grown in a paraffin-capped vial and the swelling of the cap is used as an indication of the presence of bacteria. The main problem is that this procedure is not specific to the different types of bacteria, it takes about a week and is also difficult to repeat [1].
In this scenario, Raman spectroscopy could represent a very effective solution since it is a multigas analysis, as shown in Figure (1.1).
From the identification of the gases produced by the bacteria, the classes of possible infectants can be distinguished:

1. *clostridia*, produce CO_2 and H_2. They are anaerobes spore-forming. This class contains both pathogenic species such as C. botulinum, C. tetani and altering species, such as C. tyrobutyricum and C. sporogenes.

2. Others, produce only CO_2. They are aerobic spore-forming. Also in this case the class contains both pathogenic spores, for example, Bacillus cereus and B. thuringiensis, and alterants, for example B. licheniformis, B. pumilus.

In principle, the analysis of hydrogen would be enough as an indicator to distinguish the two cases, so a specific spectroscopy technique for that gas could be used, such as absorption spectroscopy, but hydrogen is particularly difficult to measure with this technique. Furthermore, the use of a spectroscopic procedure would make the analysis more repeatable and reliable than checking for swelling of the paraffin cap.

1.4 Scope of this thesis

The advantages of using Raman spectroscopy in this context are many: it is fast, inexpensive, it does not require sample preparation and it can provide information about many gases at the same time. This technique has not yet been used on the main samples contained in the vials, one of the main reasons is that the glass impurities cause fluorescence, which can sometimes overwhelm the Raman signal.
Actually, the situation is even more complicated, other sources can disrupt the measurement such as multiple reflections or dirt on the vial's surface.
Figure(1.1) shows a typical "good" image acquired by the instrument when the milk is infected by *clostridia* as there are both peaks related to carbon dioxide and to hydrogen. Each image row represents a spectrum: the intensity of the Raman scattering as a function of the pixels (related to a frequency).

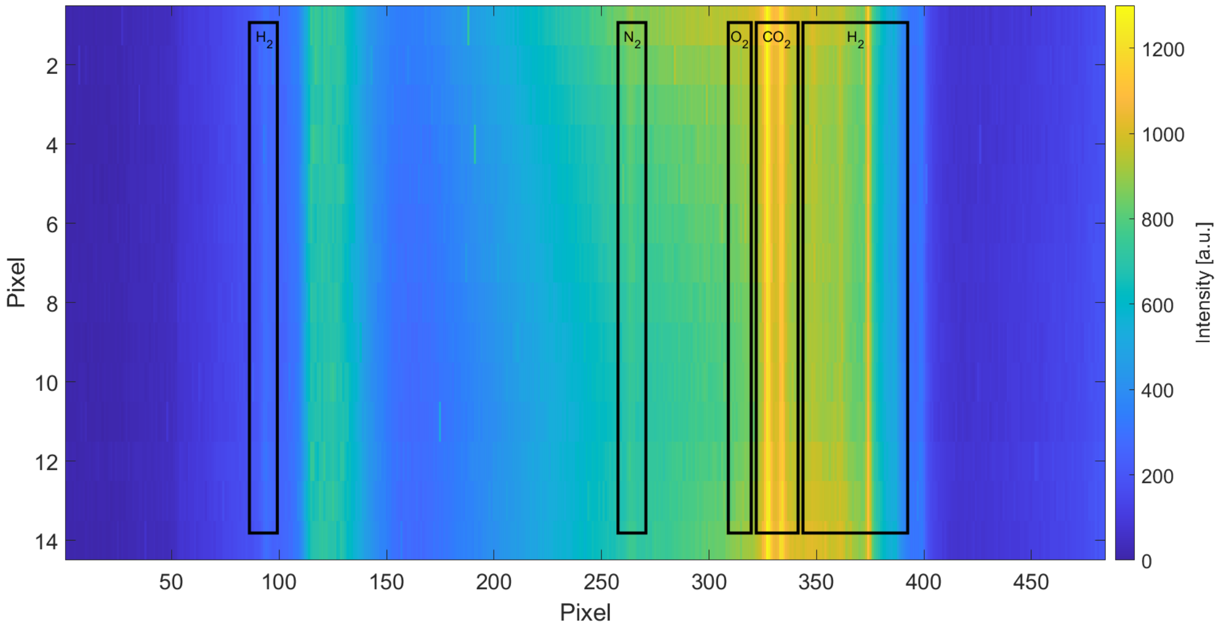


Figure 1.1: Typical good image of milk infected by *clostridia*

The so-called spectral lines are specific to each gas and represent a kind of fingerprint since the shift in frequency (or in pixels in this case) with respect to that of the laser is unique for each gas, so it is possible to associate a certain pixel position with a gas. The presence of gas can lead to the appearance of one or more spectral lines, the larger the intensity of a spectral line, the larger the concentration of that specific molecule in the vial; the intensity inequality of different spectral lines related to the same gas is caused by the different probability of that scattering process, more details are provided in the next chapter.

In particular, impurities in the test tube cause a lot of scattering, which is concentrated in the upper and lower part of the acquired image. The amount of scattered light is impossible to predict, as the light passes through the glass both in and out of the test tube and at each passage the direction of the scattering is random. Instead, what can be done is to try to rotate the tube until it is found in a position in which the scattered light does not corrupt the measurement too much.

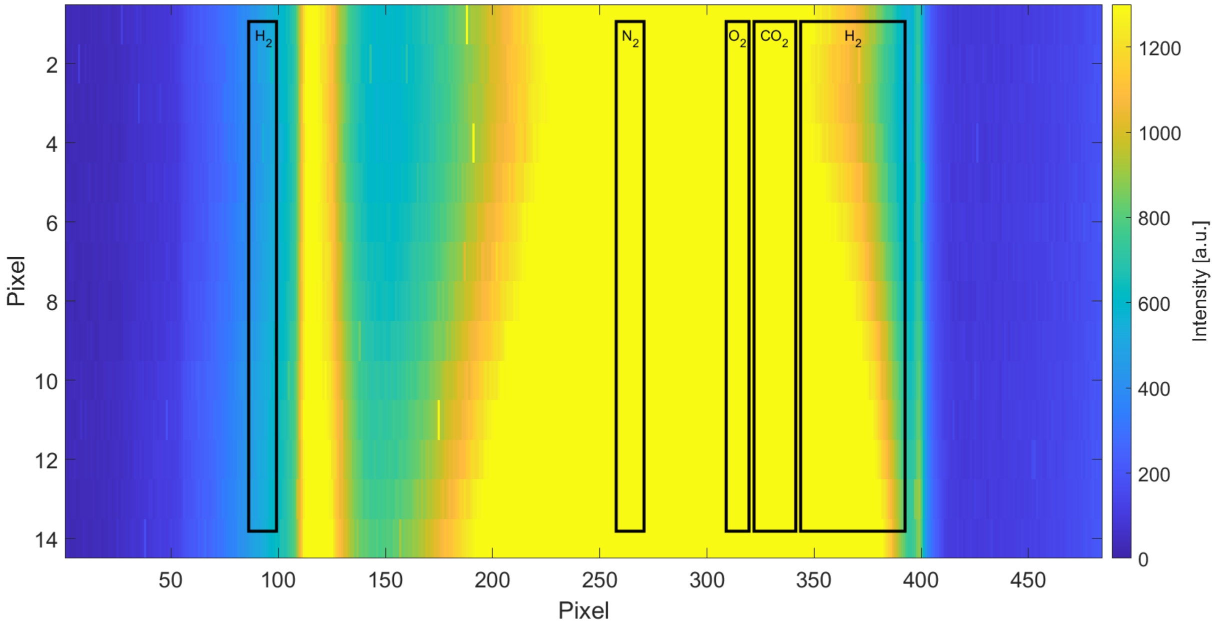


Figure 1.2: Typical bad image

Figure (1.2) shows a typical bad image of the same vial rotated to another position when a good part of the image is saturated with diffused light, in this case it is not possible to obtain useful information and it is necessary to rotate the vial. It must be said that in some cases it is possible to obtain an image in which even more than 90 % of the pixels are saturated, between the two conditions, a low saturated image and a completely saturated image, there are intermediate conditions that could be used.

The aim of the thesis is to develop an algorithm that controls the rotation of the test tube in order to acquire images only in the positions not saturated by diffused light. The approach is to find an internal standard (a parameter) that indicates how good an image is, or at least that can discriminate one from the other. The algorithm must satisfy the following conditions:

1. The quantifier of ”goodness” or the discriminator, must be created from a single measurement as it will be shown in the Sec(4), for some vials there could also be 90% of the positions to be discarded, so acquire multiple images to then decide whether performing the rotation would slow down the analysis of the samples too much.

2. The quantifier of ”goodness” or the discriminator cannot be created on the basis of the measured gas concentrations as these vary from sample to sample and over time.

3. The algorithm must be fast enough to be executed in real time, for this application, this means an execution time of the order of one second.

As will be presented in the course of the thesis, an approach inspired by image segmentation will prove useful, but machine learning will prove to be very efficient for this task.

To conclude the introduction, the structure of the thesis is presented below.

- Chapter two: Raman spectroscopy theory.

  The phenomenon of Raman scattering is presented: firstly according to the classical electromagnetic theory and then there is a brief recap of main quantum theory ideas at the basis of Raman scattering.

- Chapter three: Experimental setup

  This chapter describes the instrument with its components and how they are positioned to optimize the acquired image.

- Chapter four: Data acquisition and analysis.

  This chapter contains the description of the data taken and the related data analysis. Here it is addressed the main problem of this thesis: finding an algorithm that discriminates ”good” images from those saturated by diffused light. In particular, the problem of finding an internal standard to define a ”good” image is faced, and finally an algorithm capable of correctly classifying the images is proposed.

- Chapter five: Conclusions.

  This chapter contains the summary of this thesis work with the main results achieved and future developments.

Chapter 2

Raman spectroscopy theory

2.1 Basic concepts

When light interacts with matter, the first can be absorbed, diffused or pass through the second. If the energy of an incident photon equals the difference between two energy levels of a molecule, the photon is absorbed and the molecule is promoted into an excited state. This type of light-matter interaction is exploited in absorption spectroscopy, which quantifies the amount of absorption at each wavelength to characterize a molecule indeed the absorption as a function of the wavelength creates an absorption profile, called, spectra, that is unique for each molecule. Absorption spectroscopy is very widely used for its reliability, low cost and accuracy. Since the absorption is known at each wavelength, if the test substance is known, what is typically done, is to use a source at a specific wavelength at which there is a strong absorption. By measuring the transmittance through the sample it is possible to determine the amount of that molecule in that sample. This implies using a specific source for each molecule to avoid cross-talking between molecule's absorption therefore is not so recommended when the goal is to measure multiple gasses simultaneously.

Also light diffusion by matter is exploited in several spectroscopy techniques, one of these is the so called Raman spectroscopy, which has the great advantage of analyzing many types of molecules with the same source. In Raman spectroscopy, chemical characterization is performed by measuring the frequency shift of the scattered light with respect to that of the source, which provides information on the vibrational states of the molecules that make up the sample and allows their identification.

In a scattering process, the light interacts with the molecule and distorts the cloud of electrons around the nuclei, to form a short-lived state called a virtual state, that is not stable, so the photons are re-radiated; in other words, the molecule acts as an antenna. Most scattered photons have the same frequency as incident light since elastic scattering predominates, in this case, it is called Rayleigh scattering and their photons do not provide interesting information to characterize molecules in the sample. However, a small fraction of photons, one every $10^6 - 10^8$ scattering process, has a different frequency shift, this is because during the interaction a nuclear movement is induced, so a part of the energy is transferred from the incident photon to the molecule or from the molecule to the scattered photon. It is precisely the non-elastically scattered photons that carry the information on the vibrational states of the molecule. The very low probability of Raman scattering requires that the intensity of the incident light be very high, for this reason, a laser is used as a light source [16].

2.2 Raman scattering classical theory

When a molecule is subject to a varying electrical field $E(t)$, an electrical dipole moment $p(t)$ is induced:

$$p(t) = \alpha \cdot E_0 \, cos(2\pi\nu_0 t) \tag{2.1}$$

Where α is the polarizability of the molecule, that depends on the shape and the dimensions of the chemical bonds. The induced dipole moment oscillates with the same frequency of the incident electric field. More precisely, the polarizability is not always parallel to the incident electric field, since α is a tensor dependent on the normal coordinate Q of the molecule:

$$\alpha = \alpha_0 + \sum_k \left(\frac{\partial \alpha}{\partial Q_k} \right) \cdot Q_k + \frac{1}{2} \sum_{k,l} \left(\frac{\partial^2 \alpha}{\partial Q_k \, \partial Q_l} \right) \cdot Q_k \cdot Q_l + ... \tag{2.2}$$

The polarizability at equilibrium is α_0. Q_k and Q_l are the normal coordinates that correspond with the k^{th} and l^{th} normal vibration, corresponding to vibrational frequencies ν_k and ν_l . If we consider only the first term, so there are no cross-term, the equation becomes:

$$\alpha_\nu = \alpha_0 + \alpha'_\nu \cdot Q_\nu \tag{2.3}$$

Where α'_ν is the derivative of the polarisability tensor to the normal coordinate Q_ν. As a first approximation, the normal coordinate Q_ν oscillates as predicted by the harmonic oscillator model, hence it follows the law:

$$Q_\nu = Q_{\nu 0} \cdot cos(2\pi\nu_\nu t + \phi_\nu) \tag{2.4}$$

With $Q_{\nu 0}$ the amplitude of the normal vibration and ϕ_ν a phase term. The expression of the polarizability becomes:

$$\alpha_\nu = \alpha_0 + \alpha'_\nu \cdot Q_{\nu 0} \cdot cos(2\pi\nu_\nu t + \phi_\nu) \tag{2.5}$$

Now, is possible to write the induced dipole moment considering the series development of the polarizability term:

$$p(t) = \alpha_0 \cdot E_0 cos(2\pi\nu_0 t) + \alpha'_\nu \cdot Q_{\nu 0} \cdot cos(2\pi\nu_\nu t + \phi_\nu) \cdot E_0 cos(2\pi\nu_0 t) \tag{2.6}$$

Using:

$$cos(\alpha) cos(\beta) = \frac{1}{2}[\cos(\alpha + \beta) + \cos(\alpha - \beta)] \tag{2.7}$$

We get:

$$\begin{aligned} p(t) =&\alpha_0 \cdot E_0 cos(2\pi\nu_0 t) + \frac{1}{2}\alpha'_\nu \cdot Q_{\nu 0} \cdot E_0 \cdot cos[2\pi(\nu_0 + \nu_\nu)t + \phi_\nu]+ \\ &+ \frac{1}{2}\alpha'_\nu \cdot Q_{\nu 0} \cdot E_0 \cdot cos[2\pi(\nu_0 - \nu_\nu)t + \phi_\nu] \end{aligned} \tag{2.8}$$

The expression (2.8) is easier to interpret if it is divided in three term as:

$$p(t) = p(\nu_0) + p(\nu_0 + \nu_\nu) + p(\nu_0 - \nu_\nu) \tag{2.9}$$

The first term has the same frequency of the incident electric field and represents elastic scattering, which is the well know Rayleigh scattering term; the second and third terms have a different frequency from that of the incident light, thus representing inelastic scattering, they are called respectively Anti-Stokes and Stokes scattering contributions.

To summarize, the varying electric field induces a varying dipole moment, and the movement of charges in the molecule causes a periodic variation of the distances between its components and this leads to a vibration that can be sustained only at certain quantized frequencies. Since the polarizability depends on the shape and dimensions of chemical bonds, the vibration of the molecule arouses a small amount of the induced dipole moment to oscillate with frequencies different to the one of the electric field.

The idea of Raman spectroscopy is to exploit the Stokes and Anti-Stokes terms to obtain information about the chemical species in the sample because they are related to vibrational states of molecules. A strong beam is focused on the sample and the inelastic scattering is analysed as a function of the wavelength, this creates a Raman spectrum which is a unique fingerprint of chemical compounds.

2.3 Raman scattering quantum theory

The classical treatment used can be extended to the quantum one and this is necessary, for example, because the classical theory predicts the same intensity for the Stokes and Anti-Stokes scattering, but this is not true. Let's introduce the main steps to find the selection rules.

A molecule need to have, at least transiently, a dipole moment of frequency ν to emit or absorb a photon of the same frequency, in other words, a transition is allowed only if the accompanying charge redistribution is dipolar [9].

The expected value of a transient dipole moment is express as:

$$\mu_{if} = \int \psi_i^* \; \hat{\mu} \; \psi_f \; d\tau \tag{2.10}$$

With $\hat{\mu}$ the dipole moment operator, $d\tau$ is indicates that the integral is done over all space, ψ_i and ψ_f are respectively the initial and final state of the transition. It's clear that allowed transitions are the ones for which $\mu_{if} \neq 0$.

Expanding in Taylor series the dipole moment operator:

$$\hat{\mu} = \hat{\mu}_0 + \left.\frac{d\hat{\mu}}{dx}\right|_{x=x_{eq}} \cdot x \tag{2.11}$$

And substituing 2.11 in 2.10 we obtain:

$$\mu_{if} = \int \psi_i^* \; \hat{\mu} \; \psi_f + \left.\frac{d\hat{\mu}}{dx}\right|_{x=x_{eq}} \cdot \psi_i^* \; \hat{\mu} \; \psi_f \; d\tau \tag{2.12}$$

Since we are considering vibrational states, the first term is zero because ψ_i are eigenfunctions of the quantum harmonic oscillator so they are orthonormal to each other and $\hat{\mu}_0$ is a constant term, this is valid in the hypothesis that only the vibrational state of the molecule changes, sometimes formally expressed as $|\psi\rangle = |\psi_{vibr}\rangle$.

In analogy with what is valid for the classical treatment:

$$\left.\frac{d\hat{\mu}}{dx}\right|_{x=x_{eq}} = E \cdot \left.\frac{d\hat{\alpha}}{dx}\right|_{x=x_{eq}} \tag{2.13}$$

That provides the first selection rule, that is $\frac{d\hat{\alpha}}{dx} \neq 0$.
The derivation from the integral term to the related selection rule is quite tough and requires advanced quantum theory knowledge, so only the final result is reported. The selections rules of Raman spectroscopy are:

$$\begin{cases} \frac{d\hat{\alpha}}{dx} \neq 0 \\ \\ \Delta n = \pm 1 \end{cases} \tag{2.14}$$

With the quantum number n that is related to the vibrational energy state:

$$E_n = \left(n + \frac{1}{2}\right) h\nu \tag{2.15}$$

As expected the second selection rule predicts the two types of scattering:

- $\Delta n = -1 \implies$ anti-Stokes, the scattering wavelength is smaller than the input one.

- $\Delta n = 1 \implies$ Stokes, the scattering wavelength is larger than the input one.

Since the anti-Stokes scattering requires a molecule in an already excited state is less probable than the Stokes scattering. The ratio between the two is controlled by the temperature, in according to Boltzmann statistics:

$$\frac{I_{Stokes}}{I_{anti-Stokes}} = \left(\frac{\nu - \nu_1}{\nu + \nu_1}\right)^4 \cdot e^{\frac{h\nu_1}{kT}} \tag{2.16}$$

Eq(2.16) shows as the ratio between the two Raman scattering components can be used to determine the temperature of the sample.
Since the Stokes scattering is more intense and for the final application purpose the temperature is not required, the instrument measures only this scattering component.

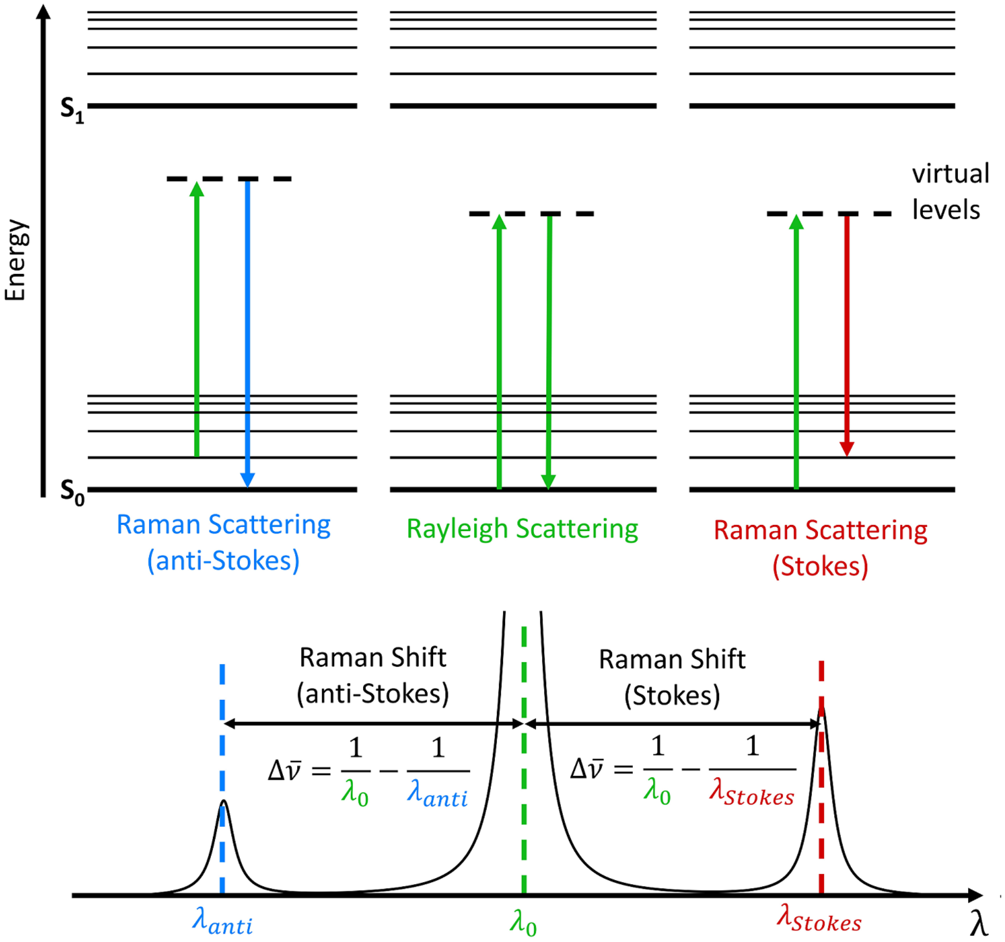

Energy
S_1
S_0
virtual levels
Raman Scattering (anti-Stokes)
Rayleigh Scattering
Raman Scattering (Stokes)
Raman Shift (anti-Stokes)
Raman Shift (Stokes)
$\Delta\bar{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda_{anti}}$
$\Delta\bar{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda_{Stokes}}$
λ_{anti}
λ_0
λ_{Stokes}
λ

Figure 2.1: Jablonski diagram of Raman scattering

2.4 Raman spectrum

A Raman spectrum is represented by the intensity of the Raman scattered light as a function of a wave number ($k = 2\pi/\lambda$) express in cm^{-1}, as per tradition in spectroscopy or in pixel. Figure 2.2 shows the Raman spectrum of air.

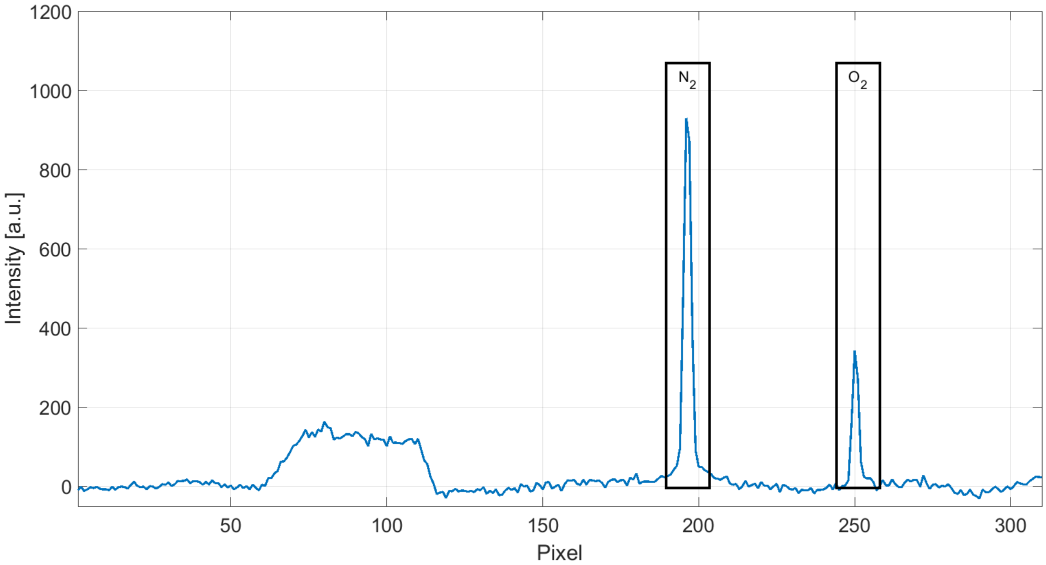


Figure 2.2: Example of a Raman spectrum

2.4.1 Raman intensity

The intensity of a Raman spectrum depends on the entire optical path from the laser to the camera, i.e. on both the sample and the instrument. The prediction of this value is almost impossible but considering a simple oscillating dipole (Sec 2.3) it's possible to understand its dependence by some parameters:

$$I_{raman} \propto \nu^4 \cdot p_0^2 \cdot \sin^2(\theta) \tag{2.17}$$

Where ν is the wavenumber of the incident light and p_0^2 is the amplitude of the oscillating induced dipole moment. Clearly the strong dependence on the wavenumber could suggest using a source with a small wavelength but this cause a lot of interference due to fluorescence, so in practice, this is not always a smart strategy. Clearly, the intensity of Raman scattering also depends on the density of the analyzed medium since it is a light-matter interaction phenomenon, if the matter is rarefied the probability of interaction is lower, for this reason, Raman spectroscopy is particularly difficult to use for gas analysis. By the same principle, if in the sample there is a concentration of one molecule more than others, its Raman signal will be greater, as can be seen from the Figure (2.2), where it is noted that nitrogen in the air is about four times greater than oxygen (the integral of the peak must be considered and not only the maximum value).

2.4.2 Raman bandwidth

The Raman bandwidth is due to several factors, even if most spectrometers are not able to resolve small contributions and the result is only the sum of many effects [19]. First of all, there is a natural broadening, that affects every spectroscopy spectrum that is the one predicted by Heisenberg's uncertainty principle, but its value is very small $\sim 10^{-8}\ cm^{-1}$. Gasses are affected by Doppler's effect that cause a broadening of the signal of $\sim 10^{-3}\ cm^{-1}$ at $300\,K$. Another source of broadening is the different chemical environments of molecules present in the sample, which modifies chemical strengths. Different isotopes produce different Raman shifts and this is another source of broadening. The truth is that spectral broadening is mainly due to experimental configuration. The most important parameters are the groove density of the diffractive element, the size of the aperture, the size of pixels of the detector and the focal length of the spectrometer.

2.4.3 Sources of interference

Raman spectra are affected a lot by fluorescence interference. Fluorescence is a phenomenon that occurs when a molecule is excited by a laser and reaches an excited electronic state, then the molecule decays into a lower state by a radiationless transition and then it may emit and in the end, reach the ground state. If the last transition is a radiation transition then the light emitted has a lower frequency than the one of the laser. To avoid fluorescence, in principle it is possible to use a laser with a lower frequency that therefore is not able to excite molecules, but the Raman scattering is proportional to the fourth power of the frequency (Eq 2.17) hence it is not a good choice. To avoid fluorescence is possible to use a pulsed laser since Raman scattering ($t_{rs} \sim 10^{-12}/10^{-13}\ s$) is faster than fluorescence ($t_{fl} \sim 10^{-7}/10^{-9}\ s$), but the signal is collected by a camera whose minimum exposure time is in the order of tens of microseconds [3], this possibility will be investigated in future developments. Raman spectra are not only influenced by fluorescence but also by multiple reflections, for example, as can be seen in the region between 80-120 pixels in Figure (2.2). That signal is not due to any gas but to a reflection inside the spectrometer.

Chapter 3

Experimental setup

3.1 Design and components

The experimental setup is typical of Raman spectroscopy with a Czerny-Turner spectrometer, if not for the presence of a stepper motor that allows the vial to rotate on itself.

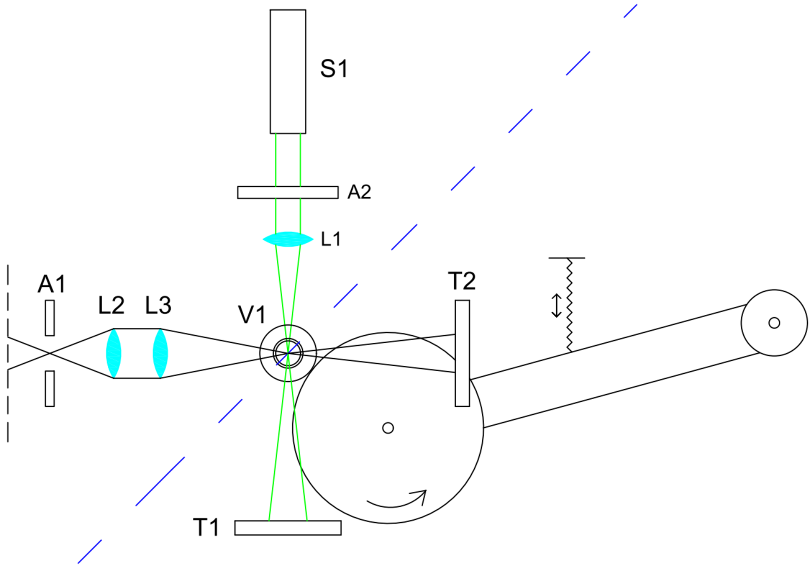


Figure 3.1: Instrument scheme

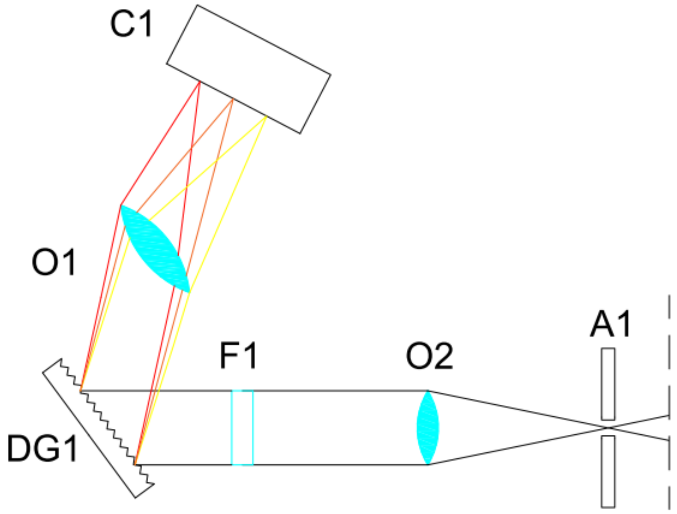


Figure 3.2: Spectrometer scheme

The light exits from the laser source (S1), passes through a focusing lens (L1) and then meets the vial, positioned at the focal distance. The 90-degree scattered light is firstly collected by two achromatic triplets (L2 and L3), then passes through a slit (A1), an objective lens (O2) and a filter that eliminates the laser component (F1), reaches a reflective diffraction grating (DG1) and finally is focused on the CMOS camera (C1). The rotation of the test tube is set in motion by a stepper motor, connected to a driver controlled through Arduino. As it can be seen, the is a shutter in front of the focusing lens (A2), because during future translations between different vials, the spot of the laser must be blocked to avoid the test tube holder catching fire or being damaged. The blue dashed line presented in Figure (3.1) is the direction over witch vials must be moved, this is very important because the vial must be very close to the aperture of the spectrometer to not lose too much Raman scattered light but it must leave enough space for the loader to move. It is also possible to notice two traps (T1 and T2) which are used to block the light and both avoid reflections and allow to operate on the optical bench in safety. In principle is possible to use a second spectrometer instead of the light trap T2, but this would

produce a larger and more expensive instrument and more difficulties in the design of the vial rotation system.
Figure(3.3) shows the instrument seen from above.

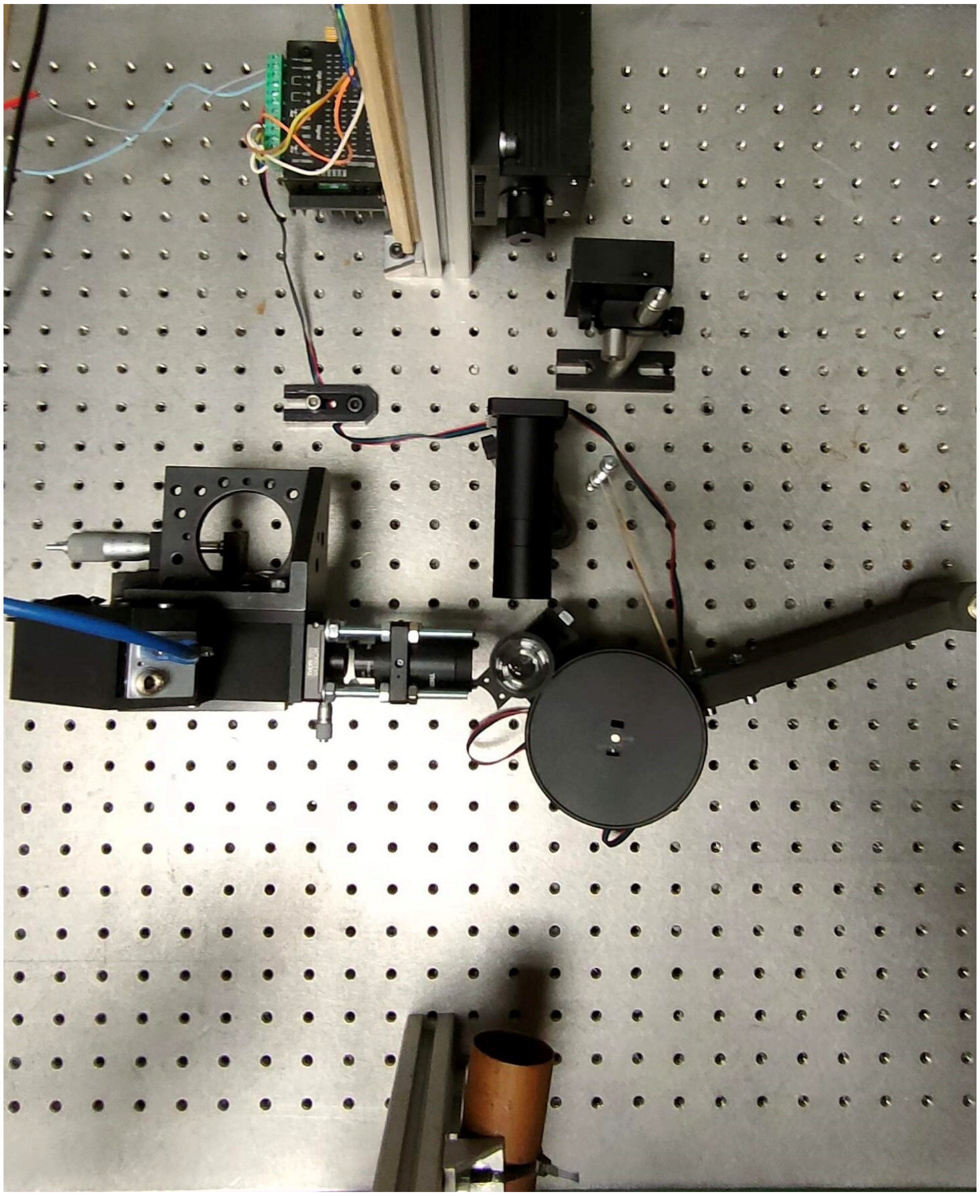


Figure 3.3: Experimental setup of the instrument

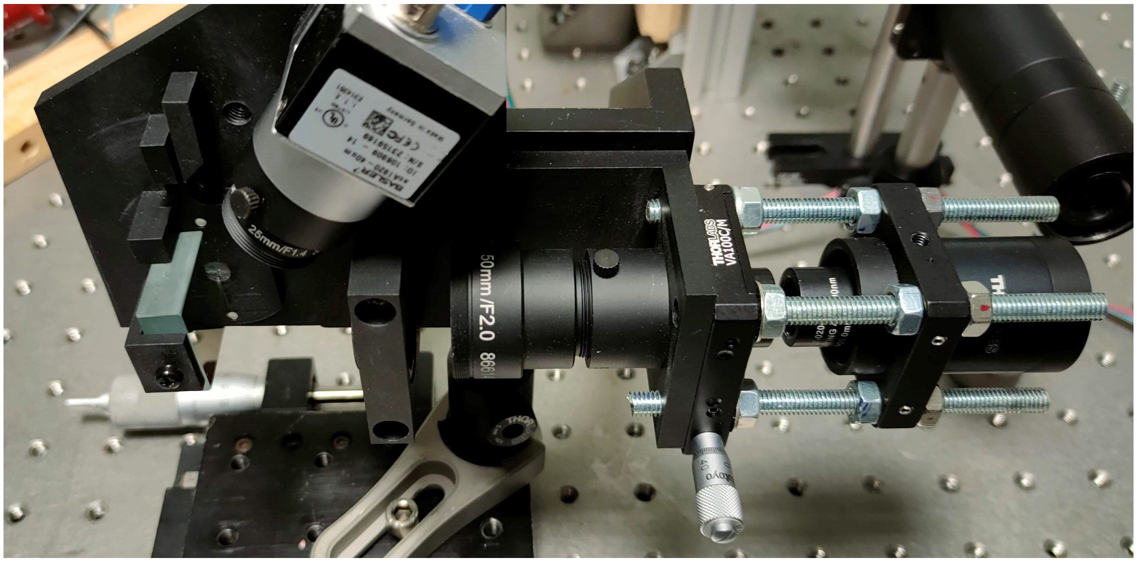


Figure 3.4: The spectrometer

Figure (3.4) shows the spectrometer with its coupling optic. It is a custom-made lens-based Czerny-Turner spectrometer (as it works in reflection), having a horizontal input slit. The spectrometer is the part of the instrument that separates and collects the different spectral components of the scattered Raman light creating an image, whose rows are the Raman spectra. This component is enclosed by a structure so as to let the light enter only through the slit, without disturbing the measurement and to protect it from dust. The inclination of the grating is such that the spectral lines of the first order of diffraction are focused on the camera. The micrometric screw that can be seen adjusts the opening of the slit at the entrance to the spectrometer, the adjustment of this parameter will be described later in this chapter. The camera is connected via USB to a PC.

Figure (3.5) shows the structure of the tube holder used and how the stepper motor is in contact with the tube to make it turn. The tube holder design is custom made to allow the laser beam to enter and exit but without letting in too much ambient light. In addition, there must be windows to allow diffused Raman light to exit from the test tube holder. There is also a window on the side without the spectrometer so that the diffused light does not undergo multiple reflections on the vial.

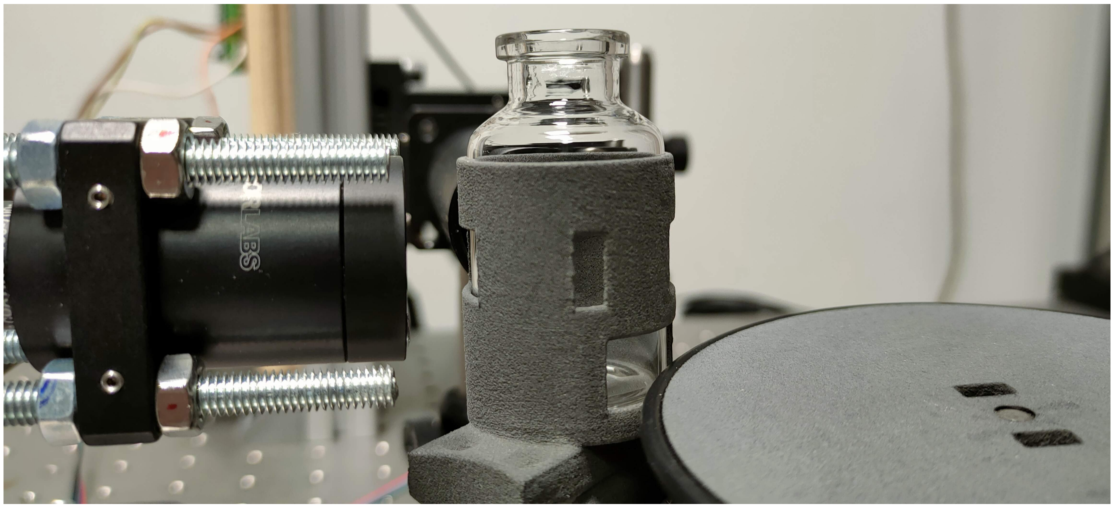


Figure 3.5: Tube holder

Given that the diameter of the test tube (d_1) is 30 mm, that of the pulley (d_2) is 96 mm and the minimum angle of the stepper motor is $\alpha_{min} = 0.9°$, it is possible to calculate the maximum number of positions that the test tube can assume:

$$N = \frac{360}{\alpha_{min} \cdot (\frac{d_2}{d_1})} - 1 = 124 \tag{3.1}$$

This value is valid only as a theoretical estimation since sometimes there is a slip between the pulley and the test tube, this leads to non-rotation and therefore the number of different positions obtainable cannot be estimated. It is also important to remember that the beam is focused before reaching the test tube so that a very small spot is obtained, it is unthinkable to reach the same position by making a rotation in one direction and one in the opposite direction. In addition, the rotation also induces small titling of the tube, which makes the positioning of the laser focus even more random. For these reasons it has been chosen to induce a rotation of 50 steps = 45 degrees at each time, sure that this does not make the acquisition pattern periodic. The angle of the induced rotation is not determinant for the instrument performance, but it has to be larger enough to ensure a rotation even if there is a small slipping effect.

The stepper motor is attached to a structure that has been designed ad hoc for this instrument with the software Fusion360 [2] and then printed with a 3D printer.

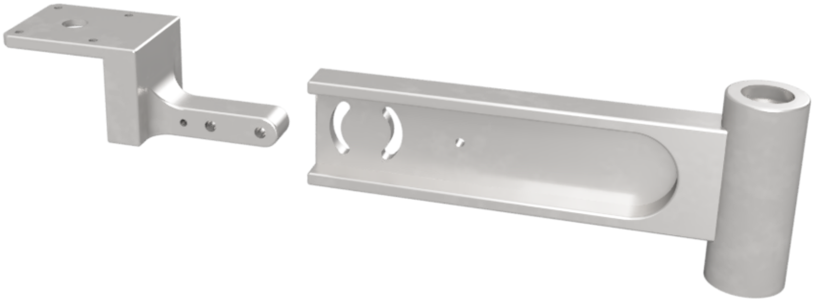


Figure 3.6: Rod

The two buttonholes are useful to allow the inclination of the plate to which the stepper motor is attached, in this way is avoided that the vial, during the rotation, escapes from the tube holder. The other holes are instead used to hook the spring to keep the structure pushed against the test tube. The system has been designed so that the thrust due to the advancement of the tubes causes the arm to rotate, which then pushes back against the tube. As it can be noticed, in the upper part of the cylinder there is a hollow, and another one is present also in the lower part of the same (not visible in Figure (3.6)), these two parts are designed to accommodate two ball bearings. Since the structure must rotate and withstand the weight of the stepper motor, the use of ball bearings is ideal for this application.

<table>
<tr><th colspan="4">Source</th></tr>
<tr><td>Code</td><td>Name</td><td>Wavelength [nm]</td><td>Power [mW]</td></tr>
<tr><td>S1</td><td>LASEVER LSR532H-2W+LSR-PS-FA</td><td>532</td><td>2</td></tr>
<tr><th colspan="4">Lenses</th></tr>
<tr><td>Code</td><td>Name</td><td>Focal length [mm]</td><td>Aperture [”]</td></tr>
<tr><td>L1</td><td>THORLABS LA1131-YAG</td><td>50</td><td>1</td></tr>
<tr><td>L2</td><td>THORLABS TRH127-020-A-ML</td><td>20</td><td>0.5</td></tr>
<tr><td>L3</td><td>THORLABS TRH254-040-A-ML</td><td>40</td><td>1</td></tr>
<tr><th colspan="4">Spectrometer objectives</th></tr>
<tr><td>Code</td><td>Name</td><td>Focal length [mm]</td><td>Aperture</td></tr>
<tr><td>O1</td><td>EDMUND 86-614</td><td>50</td><td>f/2.0</td></tr>
<tr><td>O1</td><td>EDMUND 85-355</td><td>25</td><td>f/1.4</td></tr>
<tr><th colspan="4">Diffraction grating</th></tr>
<tr><td>Code</td><td>Name</td><td>Blaze wavelength [mm]</td><td>Groove density [N/mm]</td></tr>
<tr><td>DG1</td><td>EDMUND 43-005</td><td>500</td><td>1200</td></tr>
<tr><th colspan="4">Filter</th></tr>
<tr><td>Code</td><td>Name</td><td>Cut-on wavelength [mm]</td><td>Aperture diameter[mm]</td></tr>
<tr><td>F1</td><td>THORLABS FELH0550</td><td>550</td><td>25</td></tr>
<tr><th colspan="4">Camera</th></tr>
<tr><td>Code</td><td>Name</td><td>Resolution [MP]</td><td>Fps</td></tr>
<tr><td>C1</td><td>BASLER acA1920-40um</td><td>2.3</td><td>41</td></tr>
<tr><th colspan="4">Test tube</th></tr>
<tr><td>Code</td><td>Name</td><td>Diameter [mm]</td><td>Material</td></tr>
<tr><td>V1</td><td></td><td>30</td><td>Pyrex Boro 3.3</td></tr>
</table>

Table 3.1: Table of components

3.2 Geometry optimization

3.2.1 Vial - spectrometer distance optimization

When positioning the spectrometer, it is necessary to pay attention to the compromise between the need to have an intense signal and being able to maintain a distance between the spectrometer (and its coupling optics) and the test tube such as to allow the analysis in a series of these. On the one hand, a spectrometer (and its coupling optics) very close to the vial captures a lot of signals, but will not leave enough space for the vials to move, on the other hand, a spectrometer too far from the vial will lose a lot of useful signals. The two degrees of freedom, on which is possible to act, are the distance between the spectrometer and the laser focus (where the centre of the test tube is located) and the distance between the aperture of the spectrometer and the coupling optics placed in front of the spectrometer. The first parameter controls the amount of light entering the spectrometer and the second controls the width of the spectral lines. It goes without saying that having a narrow spectral line results in a more intense signal peak and a lower error in determining its wavelength.
To optimize the components positioning, a Matlab code has been created that analyzes the acquired image in real time and, focusing on the oxygen spectral line, provides various information. Given an acquisition, the spectrum is created by averaging a number of lines compatible with that generated by the optimization of the window (Sec. 4.2.1), that algorithm cannot be applied since only one image is analyzed and real time spectral processing is needed. The interesting part of the spectrum is the region around the oxygen line. The baseline is subtracted by interpolating with a third degree polynomial only where there is no peak and an oxygen-related peak without baseline is obtained. The peak has been then interpolated with a spline function, increasing the number of points by a factor 5, to facilitate the subsequent Gaussian fit. The standard deviation and the value of the integral are extracted from the Gaussian fit (whose values are not real since a resize has been performed, but in any case, they are indicative of the signal size and width). An example of the program output is shown in Figure (3.7).

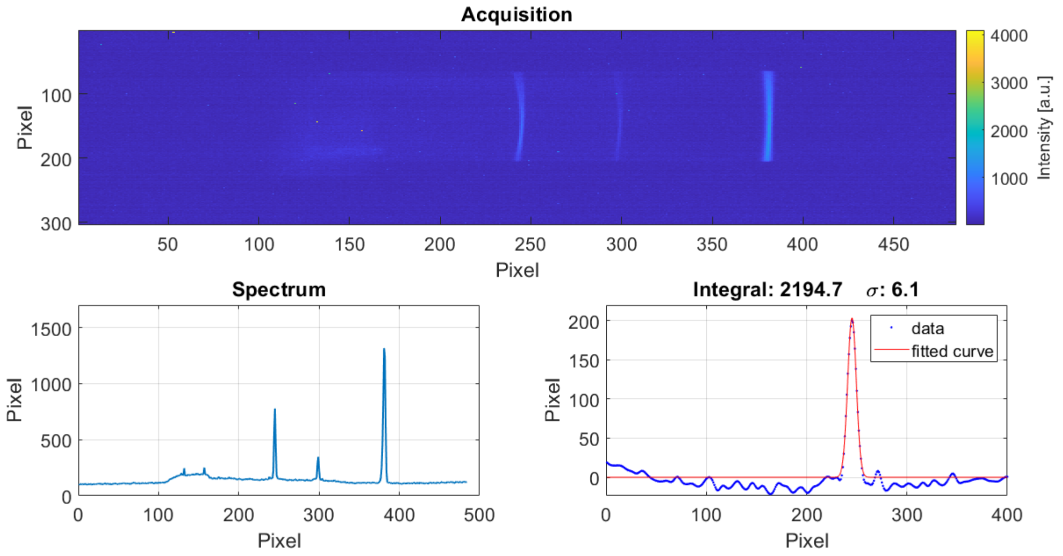


Figure 3.7: Geometry optimization

The laser focus distance - spectrometer and spectrometer distance - coupling optics were then varied in such a way as to maximize the integral and minimize the sigma of the fitting Gaussian. It is important to note that since the baseline is subtracted from the polynomial fit, the maximization of the integral is not affected by the diffuse light. The fact that the spectrum value is shifted in intensity and does not start from zero is due to the non-subtraction of the dark, this is not a problem since all images have this systematic shift. Negative intensities are due to baseline subtraction and spline interpolation.

3.2.2 Aperture's width optimization

Similar to what has been done previously, the opening of the slit in front of the spectrometer has been optimized. A small aperture lets in little light and implies having to use a longer exposure time, on the other hand, a large aperture lets in much more light but also diffused light. In this case, the important thing is not the total amount of light since it includes even diffuse light, but having the larger difference between signal and noise light. To do this, similarly to what has been done before, a code has been written whose output is the difference between the oxygen peak and the intensity of the lines distant three pixels from the peak and the difference between the oxygen peak and the intensity when there is no signal. As can be seen from Figure (3.8), the aperture has been opened to reach an intense signal but not too much to let in the diffused light, in this case the two parameters described above will both be high and almost equal. The peak of around 380 pixels corresponds to the laser so it should not be considered.

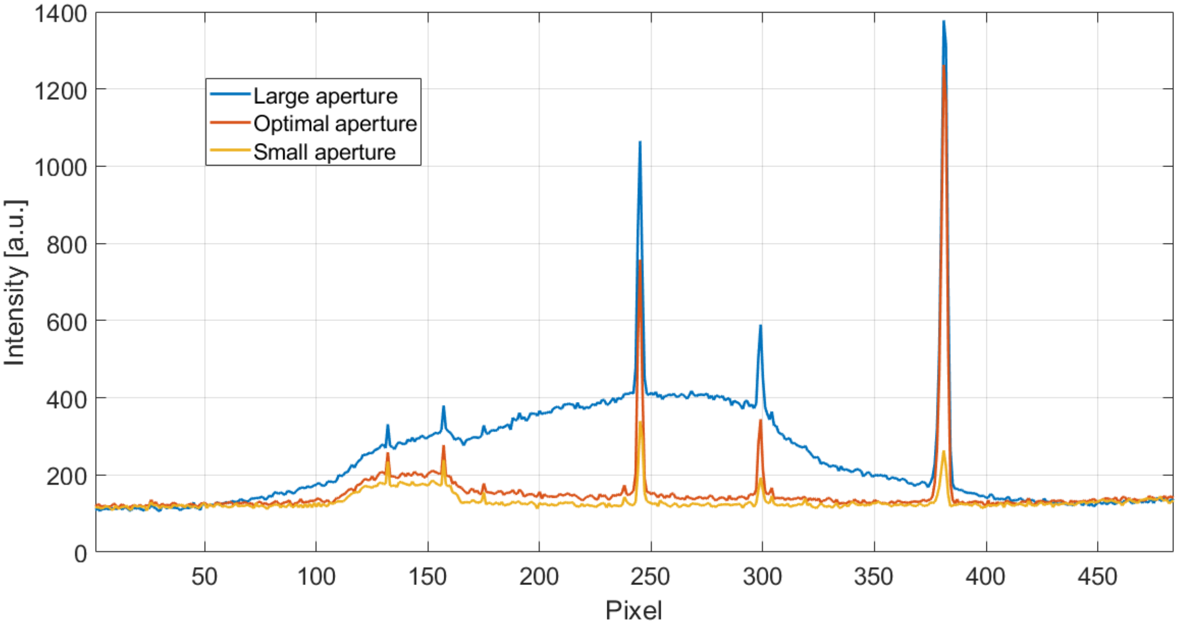


Figure 3.8: Aperture optimization

As in the previous case, the fact that the intensity where there is no signal is different from zero is due to the non-subtraction of the dark, but this is a constant value for each acquisition of the data socket so it does not affect the optimization in question.

Chapter 4

Data acquisition and analysis

4.1 Image acquisition

The acquisition process is performed by the CMOS camera and requires the laser turned on and stabilized. The acquisition time used is usually one second, with a gain of 24 dB, to make this process faster, in fact several rotations are required for each test tube before finding a good position.

Note: Spectral lines might be shifted across different images due to a different instrumental configuration, however, data analysis has been performed with groups of images originating from the same acquisition procedure with the same experimental setup.

4.1.1 Dark subtraction

The acquisition of an image with the laser turned off is called ”dark”. In principle, the resulting image should have all the pixels at zero intensity, but this does not happen for several reasons. Some detector's pixels might be saturated and might remain in this condition for all acquisitions, this is due for example to defects in the manufacture or degradation of the detector. The purpose of the dark measurement is to remove (ideally all) the light that is not part of the signal but that comes from the

environment or always saturated pixels. The analysis of the dark also allows you to check if there are intensity patterns that could be due, for example, to a non-optimal experimental setup.
The dark was acquired at the beginning and end of each data take, from which the average dark state is calculated, to compensate for the possible appearance of saturated pixels not due to the signal. Each image was then analyzed by first subtracting the average dark.

4.1.2 Noise

The total noise that affects the spectrum is given by three contributions:

$$\sigma = \sqrt{\sigma_s^2 + \sigma_i^2 + \sigma_p^2} \tag{4.1}$$

The first term σ_s^2, is the shot noise, it cannot be avoided and is given by:

$$\sigma_s = \sqrt{S} \tag{4.2}$$

With S that indicates the signal intensity. Since the relative error decreases as one over the square of the intensity, is recommended to achieve the greatest intensity possible. Given that the intensity is proportional to the integration time, it would suggest using a very long integration time, but this is not useful because most of the noise is due to the background noise[19], moreover the application of the instrument requires fast measurements, so it is not good strategy. It's important to remember that the total intensity is given by Raman scattering, fluorescence, ambient light and multiple reflections, all these components bring a certain amount of shot noise, therefore all of these which are unwanted cause an increase in the non useful signal (for example "fake" peaks in the spectrum) and also the increase of the total noise. The baseline subtraction can eliminate the contribution to the intensity, but cannot eliminate its noise, this leads to a decrease in the signal to noise ratio since the first becomes smaller and the second remains the same.
The term σ_i^2 refers to the instrumental noise, in particular, the dark signal is impor-

tant. The dark current is generated many electron-hole pairs are generated even if there is no photon captured by the camera, this effect could be mitigated using a cooled camera.

The term σ_p^2 refers to the signal processing, due to the error in the conversion of an analogue value to a digital one.

4.2 SNR as a quantifier of a position's goodness

As anticipated in Sec(1.4), the goal of this thesis is to find an algorithm that discriminates between images submerged by diffused light and useful ones. From the good images then, the Raman spectrum provides information regarding the gases present in the sample, to discriminate between non-infected samples, those infected with *clostridia* and those infected with *bacilli.*
The algorithm must satisfy the following conditions:

1. The quantifier of "goodness" or the discriminator, must be created from a single measurement, in fact, as it will be shown in the Sec(4), for some vials there could also be 90% of the positions to be discarded, so acquire multiple images to then decide whether performing the rotation would slow down the analysis of the samples too much.

2. The quantifier of "goodness" or the discriminator cannot be created on the basis of the measured gas concentrations as these varies from sample to sample and over time.

3. The algorithm must be fast enough to be executed in real time, therefore with an execution time of the order of one second.

4. The classification cannot be based only on the presence or absence of the spectral line as we are interested in how the concentration of the gas varies over time to determine if the milk is infected or not.

As previously shown, the image acquired by the camera strongly depends on the position in which the vial is located, even a small rotation involves the complete saturation of the detector.
It has been decided to investigate the use of the signal to noise ratio as a quantifier of the goodness of the measurement position, considering the integral of a peak linked to a certain gas as a signal and the standard deviation of the relative integrals as noise. Measurements made in a "good" position are therefore expected to have a high SNR (Signal to Noise Ratio) and vice versa. Since the Raman spectrum of

which the integral of a gas is calculated strongly depends on the number of lines used to create it, the question arises as to how many lines should be used to create a spectrum, this issue will be addressed in the next section.

The basic idea is to associate a signal-to-noise ratio to each position and then associate this value to a parameter (later described) that can be calculated from a single measurement. After correlating the parameter from the single image with the signal ratio of a position, it is possible, for example by using a threshold, to obtain a way to discriminate between saturated and unsaturated images.

4.2.1 Window optimization

Every time an image is acquired, the Raman spectrum is created from the average of several rows of the image, this allows to reduce the noise, but it remains therefore of fundamental importance to know how many rows must be used for this operation. Using all the rows of the image is not a good choice because, as shown above, the scattered light is mainly concentrated in the upper and lower part of the image (and usually its intensity is greater than that of the interesting signal). On the other hand, using only a small amount of rows to create a spectrum involves a lot of noise. To calculate the optimal window for the creation of the spectrum, a code has been developed that optimizes the window in order to maximize the signal-to-noise ratio of the integral associated with a peak of the desired gas. After acquiring repeated measurements, an optimal window can be associated with each gas and each position of the vial. If the interest is aimed at several gases at the same time, the window to be used is smaller of those of the individual gases. It has been decided to optimize the SNR for the oxygen line because the spectral lines related to carbon dioxide and to hydrogen are closer to it than to nitrogen (Figure 1.1).

The algorithm is based on maximizing the signal-to-noise ratio relative to the intensity of a peak associated with the oxygen line and requires repeated measurements at a fixed position, it is therefore not designed to be used in the final application during the real time analysis, but it is useful to tune the algorithm that chooses, from the parameter of a single image, whether a position is to be considered good or not. Indeed, the algorithm that discriminates the good positions, must lead to the position whose signal-to-noise ratio is greater than a certain threshold or at least greater than that of all the positions discarded.
To make the underlying idea of the algorithm clearer, a simplified flowchart of the algorithm is presented in the following Figure (4.1).
To develop this algorithm, twenty-five repeat images of the vial were acquired in one hundred different positions. The tested vial contains air.

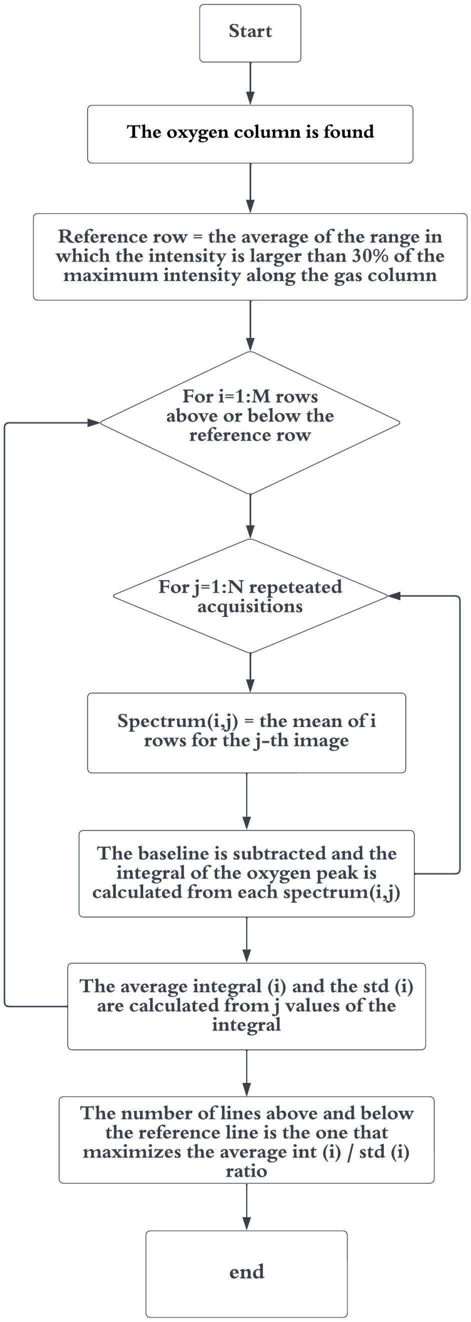

Start
The oxygen column is found
Reference row = the average of the range in which the intensity is larger than 30% of the maximum intensity along the gas column
For i=1:M rows above or below the reference row
For j=1:N repeteated acquisitions
Spectrum(i,j) = the mean of i rows for the j-th image
The baseline is subtracted and the integral of the oxygen peak is calculated from each spectrum(i,j)
The average integral (i) and the std (i) are calculated from j values of the integral
The number of lines above and below the reference line is the one that maximizes the average int (i) / std (i) ratio
end

Figure 4.1: Window optimization flowchart

The window optimization algorithm takes as input a set of repeated measurements at a fixed position and returns the best window and the spectrum generated by it. The dark must already be subtracted from the images.

Once the set of twenty-five images is loaded, its average image is created, this reduces the noise and it is used to search the oxygen spectral line. The user defines a range of columns within which the oxygen line is present, this range must be specified only the first time as it remains fixed until the experimental setup is changed. Within the range, the sum of the intensities of the pixels along the columns is calculated, the maximum will be found in the column corresponding to the oxygen line. In principle, the line could occupy several columns bu,t generally, it occupies two or three pixels, however, this operation does not require particular precision for how the algorithm was conceived. Figure (4.2) shows the acquired image and how the nitrogen spectral line is found at the 299 pixel position.

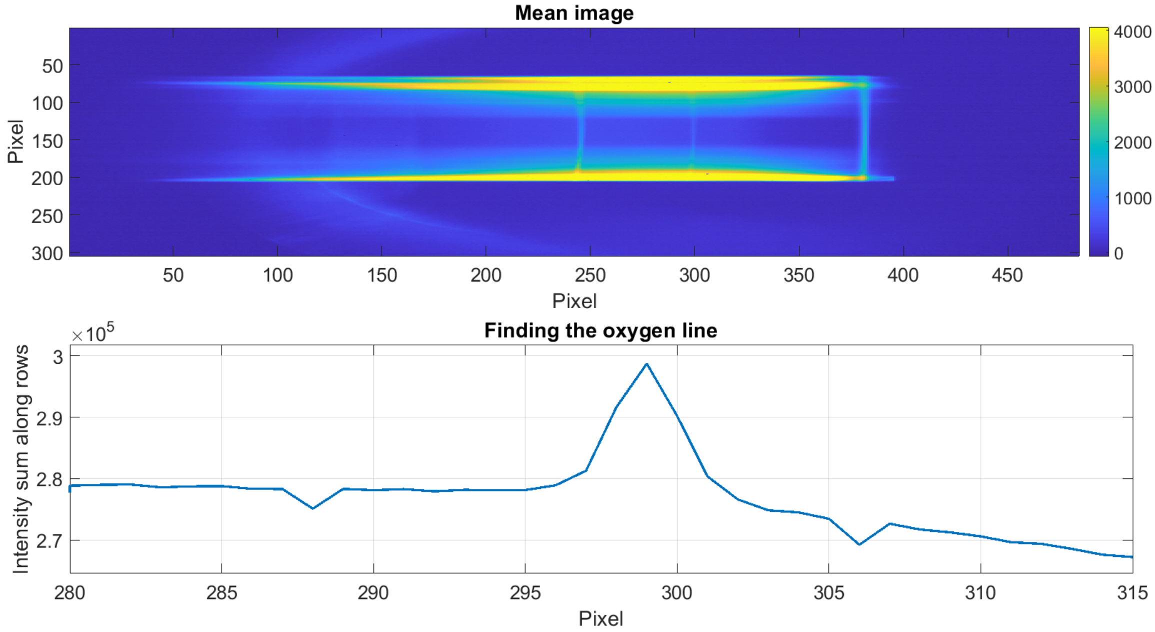


Figure 4.2: Search for the spectral line

Once the reference column has been found, the intensity profile along that column is analyzed, as shown in Figure (4.3).

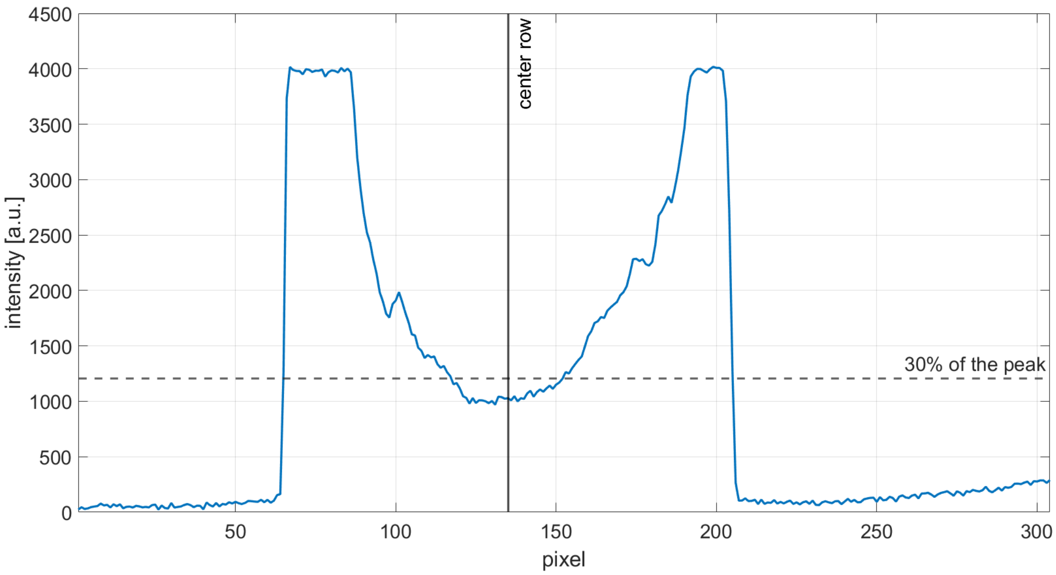


Figure 4.3: Oxygen line intensity profile

Largest peaks are relative to the scattered light, but as can be seen, in the central region there is a relatively high intensity, which is due to the gas signal. To choose which line to use as a reference, the zone in which the intensity is greater than the 30% of the maximum intensity along the line has been searched, and the average line (rounded to the nearest integer) has been calculated from that range. Once this operation has been done, a line is available from which to start searching for the window to be optimized.

For how the algorithm has been designed in its subsequent phases, performances are almost independent of the choice of the initial line as long as it is in the central band of the image, for this reason, more sophisticated systems have not been developed for its identification. In the beginning, it could be also have chosen to use the mid-spectrum row as a reference, but in case of big setup changes this would no longer be valid, but in this way, the algorithm provides some flexibility.

The fact that there is no very large peak in the central region of Figure (4.3) is due to the curvature of the spectral line, so it is visible on the column profile shifted one pixel above or below the one used here. This is not a problem as the reference line is placed in the center of this area.

Now, known the reference row, the real optimization can begin.
Given the vial in a fixed position, for each repeated image j of the set, many spectra are created using i rows below or above the reference row, in this way it obtained a spectrum for each i, with i that goes from 1 to the end of the region with the signal. For example, with $i = 1$ only two rows are used to create a spectrum, the reference one and the one above or below it; with $i = 2$, 3 rows are used and so on. In practice, for each i from 1 to M a spectrum is created for each of the twenty-five repeated images acquired with the vial in the fixed position with a number $i + 1$ of rows.
From each of the twenty five spectra (function of i), the goal is the evaluation of the integral of the oxygen's peak. First of all, it is needed to subtract the baseline, to do this a fit with a third-order polynomial is performed, but this fit should exclude the oxygen peak, to do so the standard deviation was calculated on a region to the left and right of the oxygen peak and the average of these two values is used as a threshold for the noise that should be exceeded to consider the intensity as a useful signal and this part of the spectrum has been excluded from the fit. Then, the baseline has been subtracted. It is important to notice that the number of noise-reference to be overcome to consider the useful signal is a tricky parameter, if it is too low then a lot of noise is included in the integral's calculus, if it is too high then a lot of interesting signals is excluded from the integral's calculus. The area of the useful signal to be integrated has been made started from the maximum of the peak to exclude possible regions whose noise is particularly high, as shown in Figure(4.4).

Since the number of pixels is quantized, in order to be more sensible with the thresholding, the signal, after the baseline subtraction, has been interpolated. As it can be seen from Figure(4.4), the spline-interpolation enables achieving a good result without losing too much signal between the pixel above and below the threshold and therefore a better evaluation of the integral.
Note: since the number of points has been increased with the interpolation, the

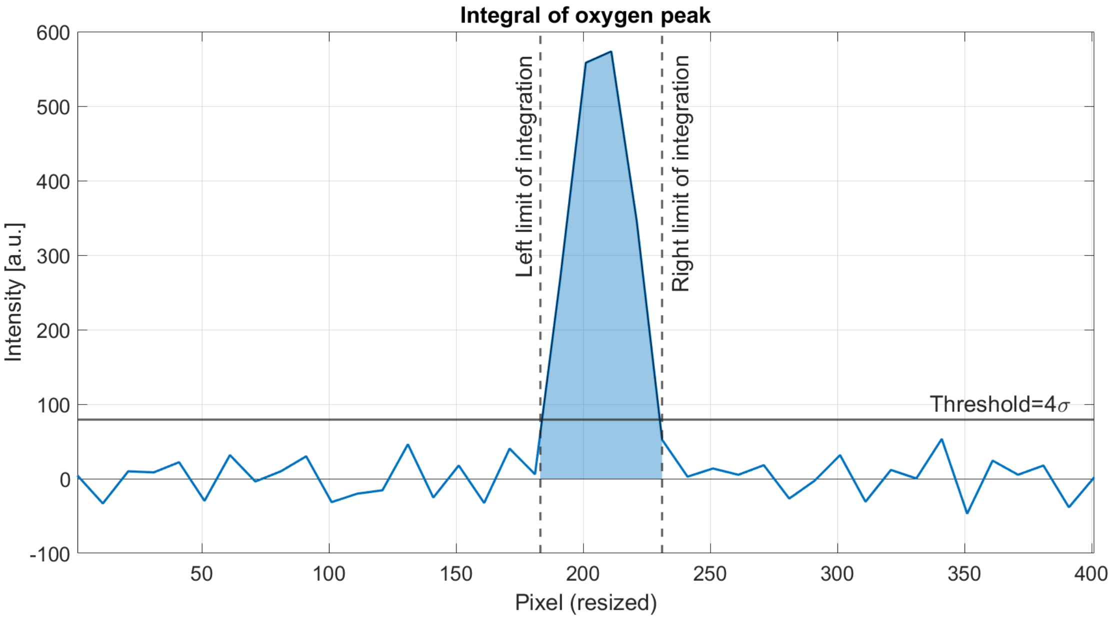


Figure 4.4: Integral of oxygen peak

value of the integral is not ”real”, but it will be compared only with other values obtained with the same procedure therefore it is consistent. The presence of negative intensities in Figure(4.4) is due to the baseline subtraction. The difference in signal strength between the images (4.3) and (4.4) is due to the subtraction of the baseline.

Once the integral is calculated for each i number of rows below and above the reference row, it can be compared with the same obtained from the other twenty five images. The result is the mean value of the integral over twenty five images, as a function of i. It is evaluated then the SNR as a function of i and the aim is the maximisation of this quantity, both for i above and below the reference row, as shown in Figure(4.5).

The final result is an asymmetrical window whose size leads to a maximisation of the SNR of the oxygen’s peak that is evaluated at the end over the optimized window.

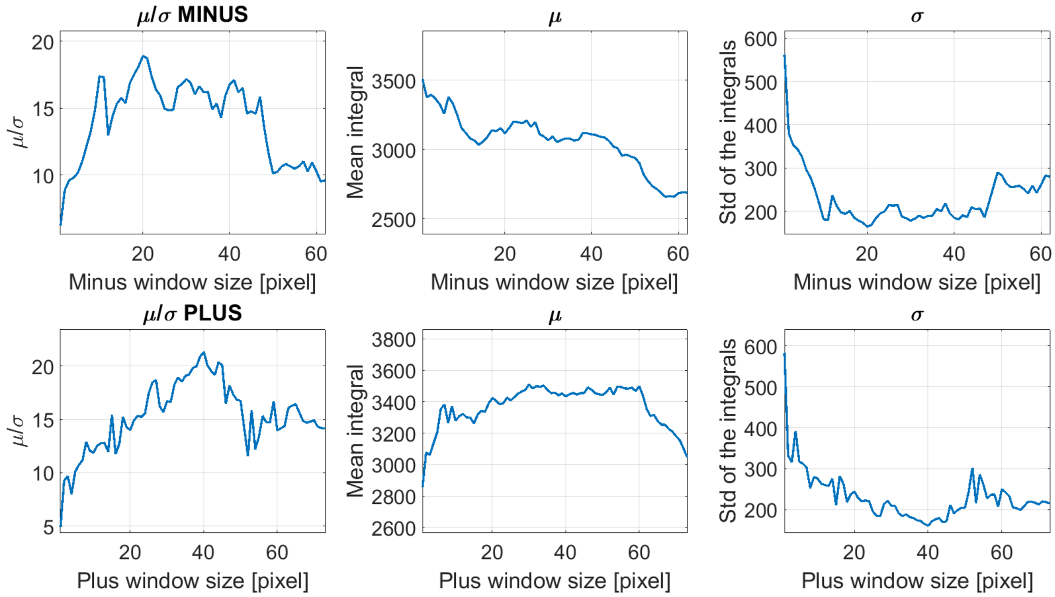


Figure 4.5: Window optimization

As it can be seen from the figure above the mean value of the integral is almost constant along the entire window, this is interesting because confirms that the signal is uniform along that direction (for small aperture angles). The average value drops only to the last values, this is because the light is cut off by the slit. The standard deviation of the integrals clearly decreases as a function of the number of rows as expected.

As can be clearly seen from the μ/σ ratio graphs, this procedure leads to a maximum that can be used to select the range of the optimal window, however, the algorithm is not so stable in the sense that there are maxima whose ratio is similar to that of the maximum peak but with a very different number of rows used, this problem is addressed in the next part.

Multiple maxima window optimization

The μ/σ ratio sometimes has multiple maxima, this is particularly dangerous when the two (or more) maxima vary little in the value of μ/σ but a lot in the number of lines. In some cases, there is a maximum precisely in correspondence of the appearance of diffused light since this contribution increases the value of the integral, while the std is monotonous decreasing. To make the algorithm more robust, the values of μ/σ have been evaluated within more than 95% of the maximum value and the average value in the range of smaller pixels has been taken as the definitive value. In practise, in practice, the window does not widen if there is no gain in the / ratio of at least 5% with respect to the maximum value, so if you can reduce the window by losing little relative SNR, this is done. Figure (4.6) shows a typical example of a situation solved thanks to this strategy.

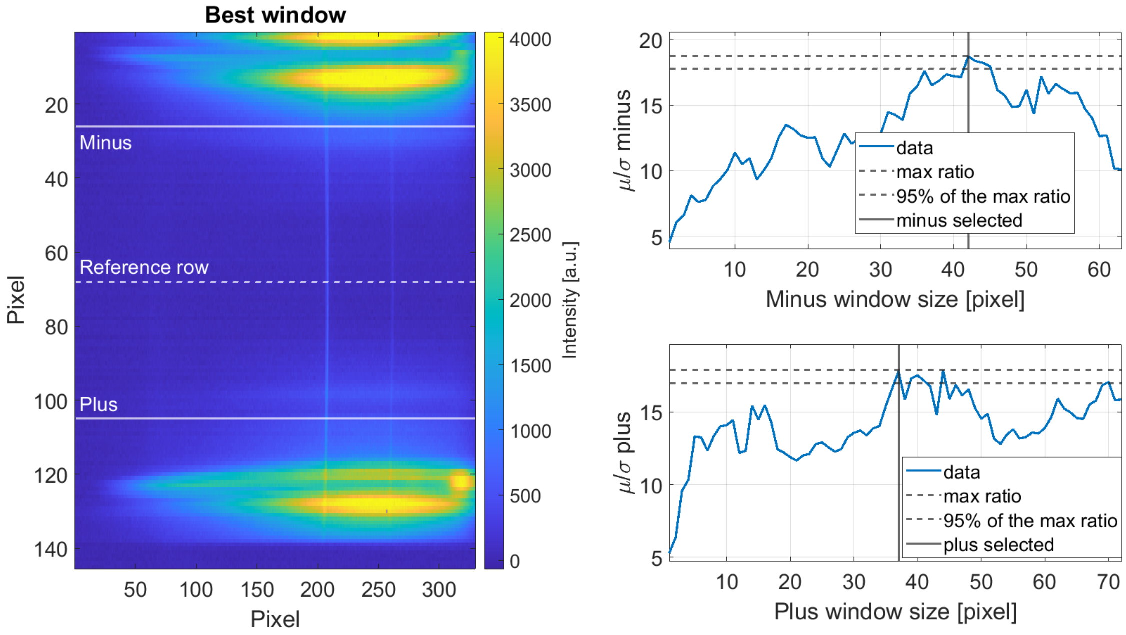


Figure 4.6: Window optimization with many local maxima

To have a further verification, in the best condition, that is the one without the test tube, it can be noted that the optimal window is particularly extended, given the lack (or in any case small quantity) of diffused light, shown in Figure (4.7), it is evident that the optimal window is much longer in this ideal case and also the SNR obtained is larger than in the previous case.

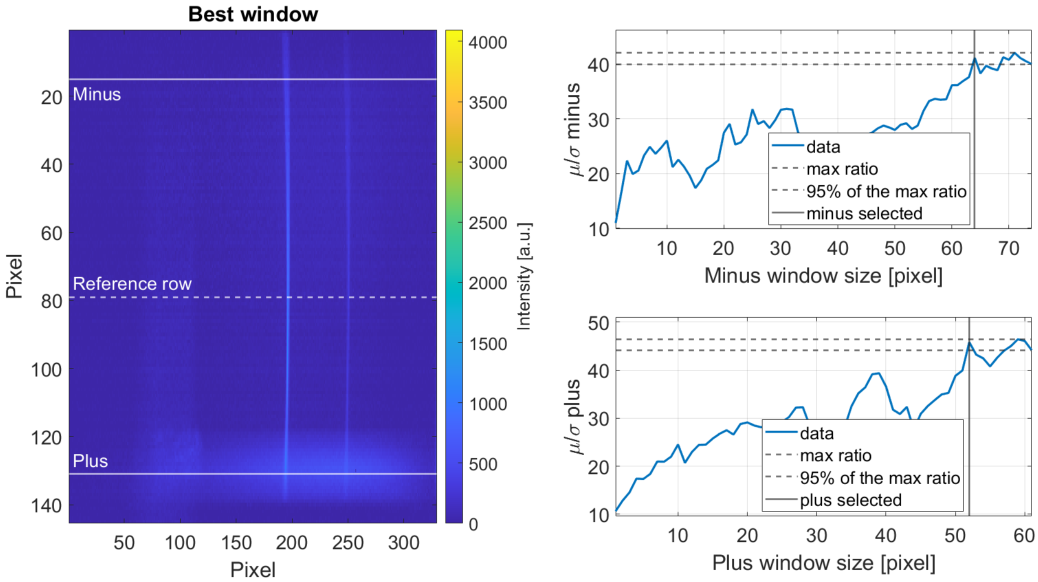


Figure 4.7: Window optimization without the vial

The idea of discriminating between good or bad positions based on the optimal window width is not the best as:

1. The final goal is not to have a large window, but a good spectrum, if an image with only a few lines is obtained but which produces a spectrum with a high signal-to-noise ratio, it makes no sense that this is eliminated.

2. The window optimization algorithm requires repeated measurements, which slows down the analysis process.

The window optimization algorithm is not trivial, in summary, a spectrum is created by averaging every possible number of lines above and below a reference line, then the integral of the oxygen peak is calculated for all images acquired in a fixed position and then the window that optimizes the μ/σ ratio is chosen.

To conclude, from a set of repeated acquisitions with the test tube at a fixed position it is possible to evaluate the best window from which to generate the spectrum, this procedure then leads to the estimation of an SNR for this specific position, which will be used to test the position selection algorithms. For each of the hundred positions, a signal to noise ratio has been associated as a goodness qualifier.

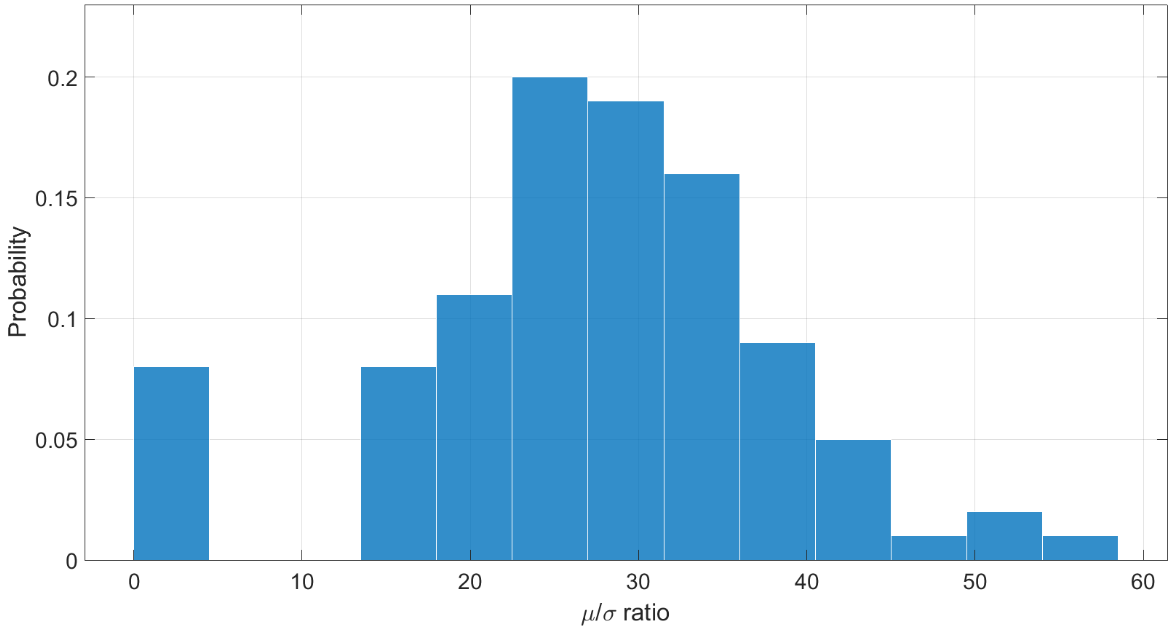


Figure 4.8: Distribution of the μ/σ ratio

Figure(4.8) shows the probability density function of the ratio μ/σ of the oxygen's peak integral, from the spectrum obtained using the best window for each position. As can be noted, there is no clear distinction between two regions, which can be classified as good and bad, on the contrary, the distribution appears to be random.

4.3 Parameter from a single image

Now, the SNR for each position is known, but since it is calculated from repeated measurements, it should be correlated to some parameter computable from a single measurement.

The aim is to find an algorithm that given only one image can produce a quantity that is proportional to the SNR for that position, or, at least, discriminate images with high SNR from that with a lower one. When choosing the discriminant parameter (for example the number of pixels over a threshold), care must be taken to avoid overfitting, in the sense that the algorithm might produce a parameter that works well only for the set of measurements acquired, but when it is tested on new data fails.

When the diffuse light is too much the whole image is saturated, for example as in Figure(1.2), it is not possible to apply the window optimization algorithm and a value of μ/σ equal to 0.1 has been associated with that position. This value is not real but is useful to test if the algorithm can predict very bad images.

It is important to specify that the final objective of the instrument has to discriminate the tubes containing infected milk from the non-infected ones, to do this, a precise estimate of the quantity of gas is not necessary, as it is sufficient to verify the increase in the presence or absence of CO_2 and/or H_2. For this reason, it is not of fundamental importance to be able to obtain positions with a very high noise signal, but a compromise must be found between the number of rotations required (and therefore the measurement time) and obtaining positions that give repeatable results.

4.3.1 First parameter: % of pixels above a threshold

The first simple algorithm tested involves the calculation of the percentage of pixels above a certain threshold. To make the method not dependent on the integration time, the threshold has been set as the half of maximum value of the image intensity. Furthermore, to make the calculus independent of the spectral lines, so it considers only diffuse light, the region with spectral lines has been zoroed. Since a vial containing air was measured throughout the analysis of this parameter, only the lines in the nitrogen and oxygen region have been masked.

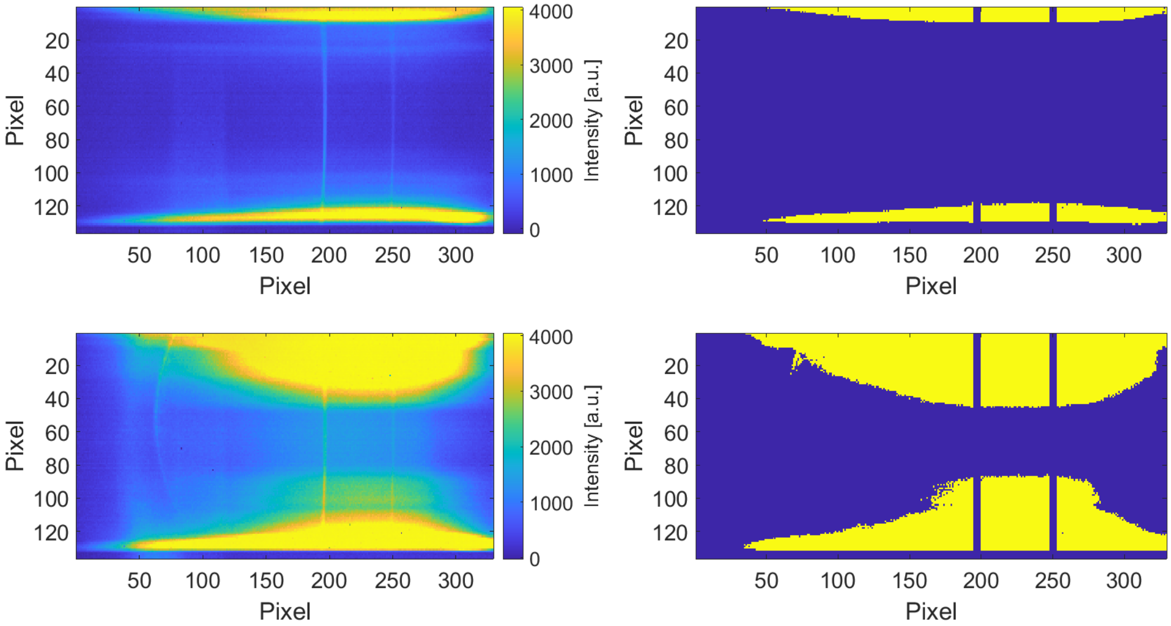


Figure 4.9: Percentage pixel above threshold algorithm

As can be seen from Figure (4.9), the algorithm can qualitatively identify the areas where diffused light is present. The algorithm is based on the assumption that the scattered light is in the greater half of the intensities and that an image with a low percentage of pixels corrupted by the scattered light implies a good spectrum (high μ/σ ratio).

To verify that a good spectrum corresponds to a low percentage of pixels above the threshold, the ratio μ/σ for the oxygen's peak, has been plotted as a function of the percentage of pixels above the threshold, as shown in Figure (4.10).

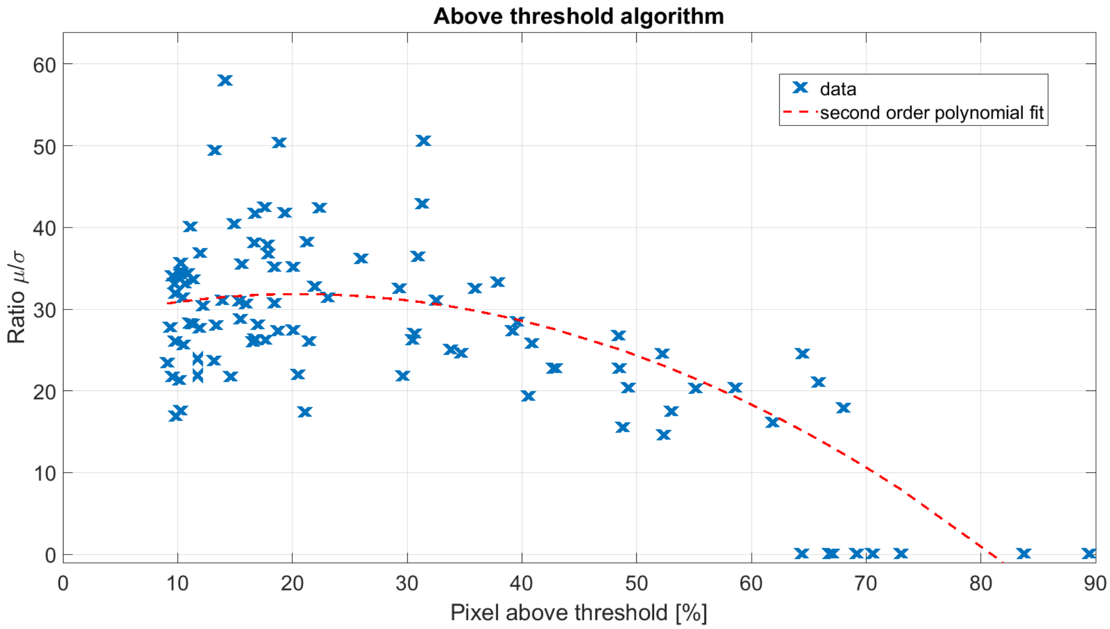


Figure 4.10: Above threshold test

The graph shows how although low SNR values can occur even for low percentages of pixels above the threshold, the algorithm shows the trend: at low percentages of pixels above the threshold it generally shows higher SNR values (the fit doesn't consider "fake" SNR equal to 0.1). Moreover, it can be seen how particularly high SNR values do not correspond to values of percentages greater than 50 (indicatively).

Advantages of this parameter:

- It is very fast and suitable for real-time applications.
- It captures the region with the largest noise.
- It is easy to interpret, easy to understand.

Disadvantages of this parameter:

- It does not capture the intensity's distribution of the noise, the output is the same if the intensity of the noise is just above the threshold or completely saturates the detector.
- It does not consider the spatial location of the scattered light, if it were all in a region where there are no spectral lines it would not be a problem, if instead it is located close to spectral lines, this makes the recognition of gasses more complicated.
- The threshold is simply a max operation, that is very sensitive to hot pixels and saturation.

The implementation in the case of real time analysis can be of two types:

1. A threshold is chosen from Figure (4.10) to ensure a high SNR and measurements are acquired by rotating the vial, until N acquisitions are obtained with a percentage of pixels within the threshold. Is is possible to set N equal to one
2. M images are acquired and the N images with the lowest percentage of pixels due to diffused light are kept.

To solve some cons of this algorithm, another approach has been developed, which is presented in the next section.

4.3.2 Second parameter: first PCA component

PCA applied to images

Principal component analysis (PCA) is a widely used technique for dimensionality reduction, it is also exploited to perform a preprocessing of data before it is given as input to artificial intelligence algorithms. As for images, PCA has also been used for facial detection and image classification in general [18].

The idea is that given a dataset of M points having N parameters, exists a reference system such that along one direction (principal component) there will be the greatest variance of the data, and along the second direction there will be the second largest variance and so on. This implies that if there is little variability along the latter directions, these directions are less useful than the former for distinguishing/separating data. Therefore it is not necessary to work with data in space $\mathbb{R}^N$ but it is enough to focus in the first p components, so M points in the space $\mathbb{R}^p$ will be obtained (with $p << N$).

PCA is the linear transformation that allows you to rotate the space from the original reference system to the one described above. The directions along which to project the data are called principal components and correspond to the eigenvectors of the covariance matrix.

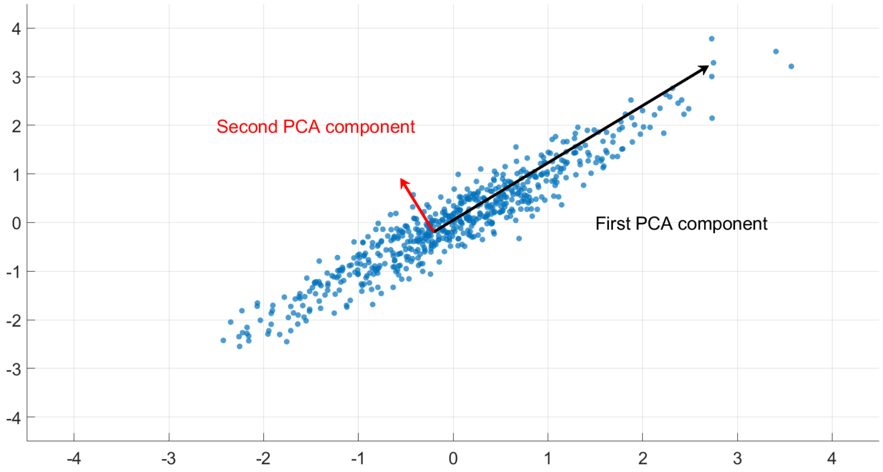


Figure 4.11: 2D Example of principal components

Figure (4.11) shows a case in which there are many points belonging to the two-dimensional space, the arrows instead show the two principal components. It is evident that most of the variability of the data develops along the direction of the first eigenvector, while in the orthogonal direction, the second component, there is less variability. The interesting thing is that generally only a few main components are needed to capture the variability of the data, so the reduction of dimensionality is very effective.

In the case of images, they are transformed into vectors by flattening them. The number of parameters corresponds to the number of pixels, so it can even be of the order of millions, in the specific case of this thesis it is generally around tens of thousands, this is the dimensionality of the space where each image is represented as a vector. It is important to note that the eigenvectors can be expressed as linear combinations of components in the original space, therefore the transformation is obtained with a matrix product, in the case of images the eigenvectors instead correspond to ”eigen-images”.

Figure (4.12) shows how the PCA is used for face recognition, from a $\mathbb{R}^{2500}$ space it is obtained a $\mathbb{R}^{2}$ space after applying the PCA. It can be noticed that groups of close points correspond to the same face, this means that only two coefficients are enough to distinguish each face from the other.

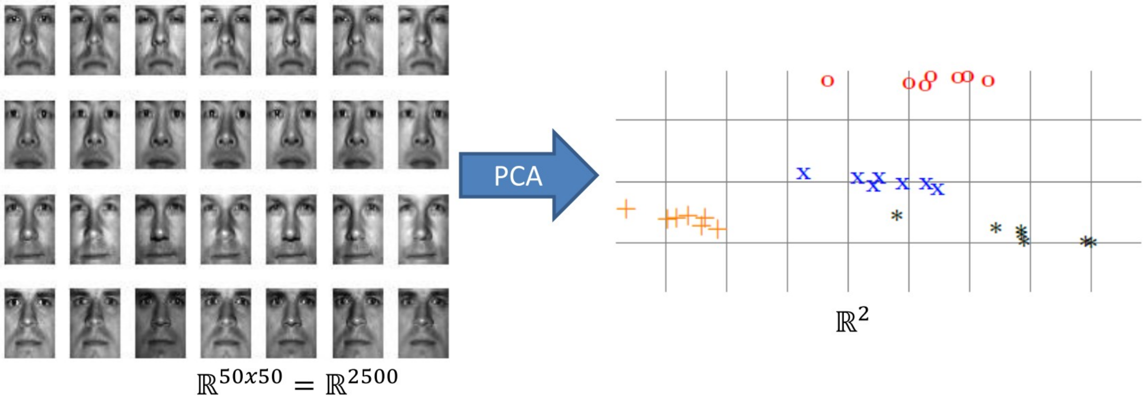


Figure 4.12: PCA for face recognition

Coefficients in the two-dimensional plane correspond to the coordinates along the two main components. Actually, the example just reported is a very simplified case as the faces are very similar to each other, but it shows the potential of the technique.

To better understand the role of self-images (or eigen-images) just think that a vector in the original space can be decomposed as a linear combination of principal components vectors, the same applies to the original image can be decomposed as a linear combination of these, as shown in Figure (4.13).

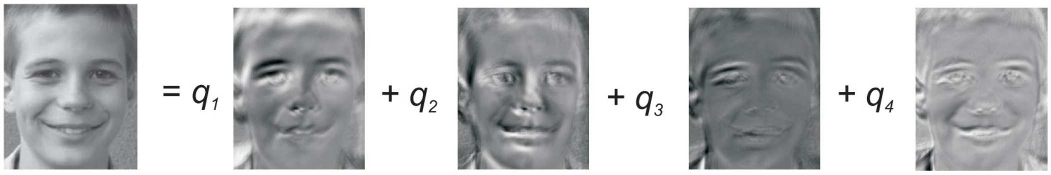


Figure 4.13: Autoimage decomposition [17]

Clearly, the approach is very simple and deep learning algorithms are currently used for face recognition, however, this procedure has been tested with the images acquired by the spectrometer.

Regarding this thesis, the objective is clearly not to classify the faces but to exploit the PCA to identify which are the areas in which the diffused light is concentrated more: given that the acquisitions are made with the same compositions of gasses (air), what characterizes one position from another is precisely the diffused light. It is expected, for what has been said above, that the first autoimage corresponds precisely to the area in which the diffused light varies the most.
The intensity of spectral lines vary slightly between different positions with respect to the diffused light zone, this phenomenon is particularly useful with PCA because this implies that the first self-image does not consider the spectral lines for the variability between the different positions, but only the zone of diffused light. Despite this, the intensity of the spectral lines varies a little even if not much and therefore the regions where the spectral lines are present have been zeroed as done for the algorithm that considers the percentage of pixels above the threshold.

PCA calculus

1. Creation of the $M \times N$ matrix A, with M equal to the number of different positions, and N the number of parameters (pixel), that has been reduced only to the useful part, to make the computation faster. The vector of length N is obtained by flattening the matrix of the acquisition.

2. Centering of data, so subtract the average evaluated through columns, in this way each parameter becomes centred to zero.

3. The $p \times p$ covariance matrix C is defined as:

   $$C = \frac{A^T A}{M-1} \tag{4.3}$$

   that is symmetric so it can be diagonalized as:

   $$C = VLV^T \tag{4.4}$$

   with V a matrix of eigenvectors and L a diagonal matrix whose eigenvalues are λ_i.

4. It's possibile to apply the singular value decomposition (SVD) to the matrix A, such that:

   $$A = USV^T \tag{4.5}$$

   with U a unitary matrix, S a diagonal matrix of values s_i

5. Combining 4.6 with 4.4 the result is:

   $$C = \frac{VSU^T USV^T}{M-1} = V\frac{S^2}{M-1}V^T \tag{4.6}$$

   This means that vectors V are the principal directions and the singular values s_i are related to eigenvalues of C through $\lambda_i = \frac{s_i^2}{M-1}$. Projections of the original data along the principals component are called scores and are obtained in the columns of XV or US.

In practice, instead of evaluating the eigendecomposition of C, the same information can be obtained by evaluating the singular value decomposition of the original data X. The calculus of the PCA via SVD is faster than the eigendecomposition's algorithm because in this case, the number of variables exceeds the number of observations [8].

PCA for position selection

Similarly to what has been done previously, the aim is to find a parameter from a single image that can then be correlated to the signal-to-noise ratio. To resume, applying the PCA is possible to recognize regions where the variability of the scattered light is high. Since the first self-image contains the region where there is the largest variability, the component of the acquisition along that direction determines a parameter related to the amount of light scattered in that region.

It is interesting to note that the PCA indicates the areas of variability and not those saturated by diffused light, this is very functional because always saturated regions, both for good and bad positions, are weighed less than those with high variability and in fact, they cannot be used to discriminate the position.

To construct the PCA, many measures are needed, in order to avoid overfitting, which in this case would result in a wrong construction of the self-images, which would present areas of high intensity where indeed, for large numbers, the variability would not be so high. The same measurements used for the analysis of Sec(4.3.1) have been used to construct the PCA matrix, also to be able to compare this procedure with the one that exploits the percentage of pixels above a threshold: twenty five repeated acquisitions of a hundred of different positions.

In the case under examination, the first component contains within itself 86% of the variability, instead, to describe more than 95% of the variability, the first three components are needed. The first three self-images are presented below, in Fig (4.14), spectral lines generally result in a lower variability as expected, but to make the procedure more resistant to possible shifts or defects in the experimental setup, they have been zeroed.

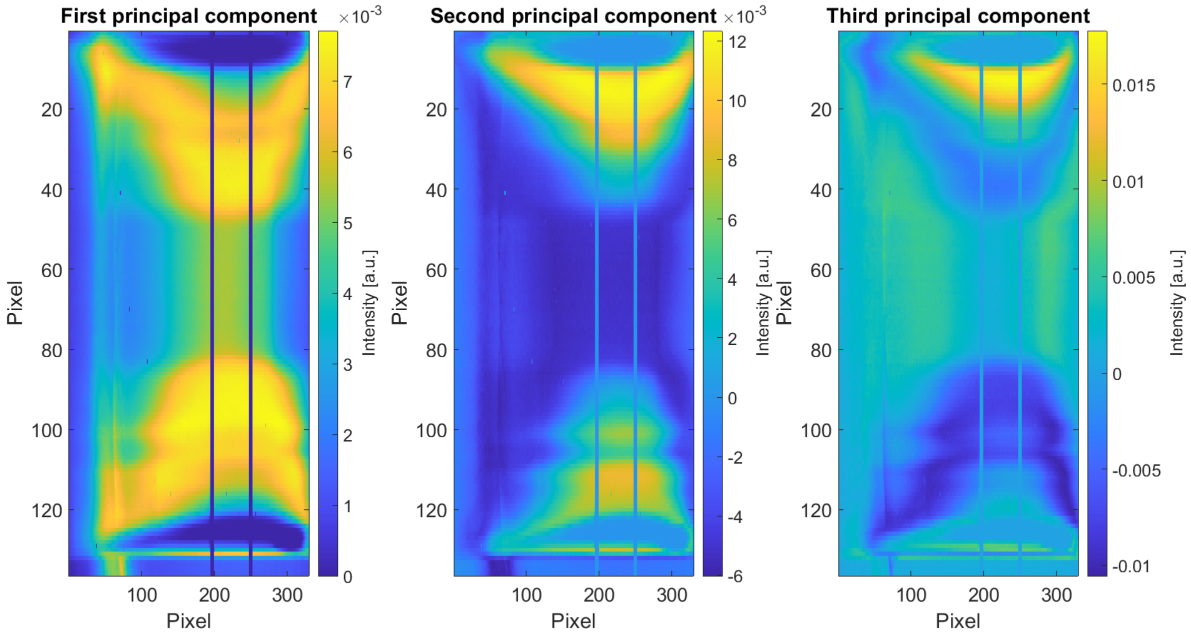


Figure 4.14: First three eigen-images

It should be noted that the first self-image clearly shows the area where the diffused light is usually found, verifiable by checking from the single acquisitions. As for the other components, previous considerations are valid, but is difficult to interpret their meaning.

To verify the correlation between the two variables, the signal to noise ratio of each position has been plotted as a function of its first principal component, the result is presented in the next figure.

Similar conclusions to what has been said for the previous parameter apply, a negative value of the first principal component does not necessarily ensure a high SNR, but a very high value of the same component certainly does not lead to a high SNR value.

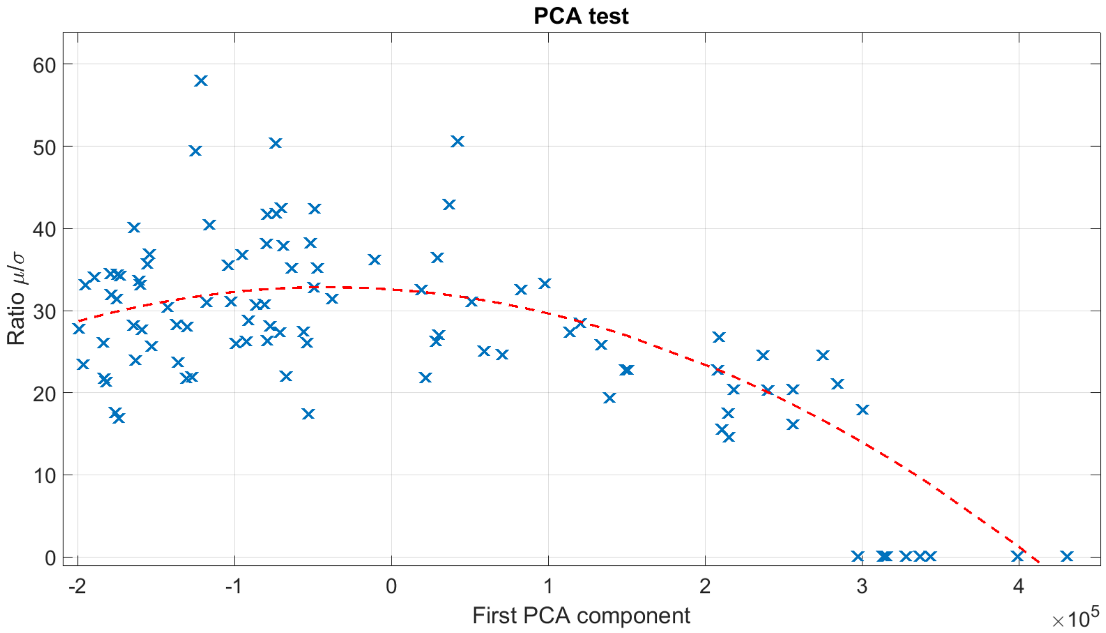


Figure 4.15: First PCA component test

Although the parameter cannot be correlated with the signal to noise ratio, the usage of PCA components has several advantages:

- Once the PCA rotation matrix is built, the calculation of the components for a new image is immediate and can be done in real time.

- The method is independent of internal thresholds, as was the case for the previously chosen parameter (% of pixels above a threshold).

- First principal component has a precise physical meaning. It captures the essence of the problem, finding regions where the scattered light is concentrated in order to discriminate the positions. Saturated areas common to all images have no weight.

4.4 Manual labeling

Given that the parameters found above have a physical sense but cannot be correlated with the signal to noise ratio, it is fair to ask whether the latter is really a parameter that indicated the goodness of the measure. Indeed, it should be considered that the presence of constant light in the signal increases the signal-to-noise ratio, but this does not correspond to a real improvement in the measurement conditions.
For this reason, another type of parameter has been investigated that can estimate the goodness of the position.

Since in the feasibility study of the instrument good positions were manually established by an operator [1], for each position of the hundred acquired, a label has been assigned manually, discriminating between good (label=1) and bad (label=0). Manual labels have been plotted as a function of the two parameters extracted from a single image, i.e. % of the pixel above a threshold and first PCA component, the result is shown in Figure (4.16).

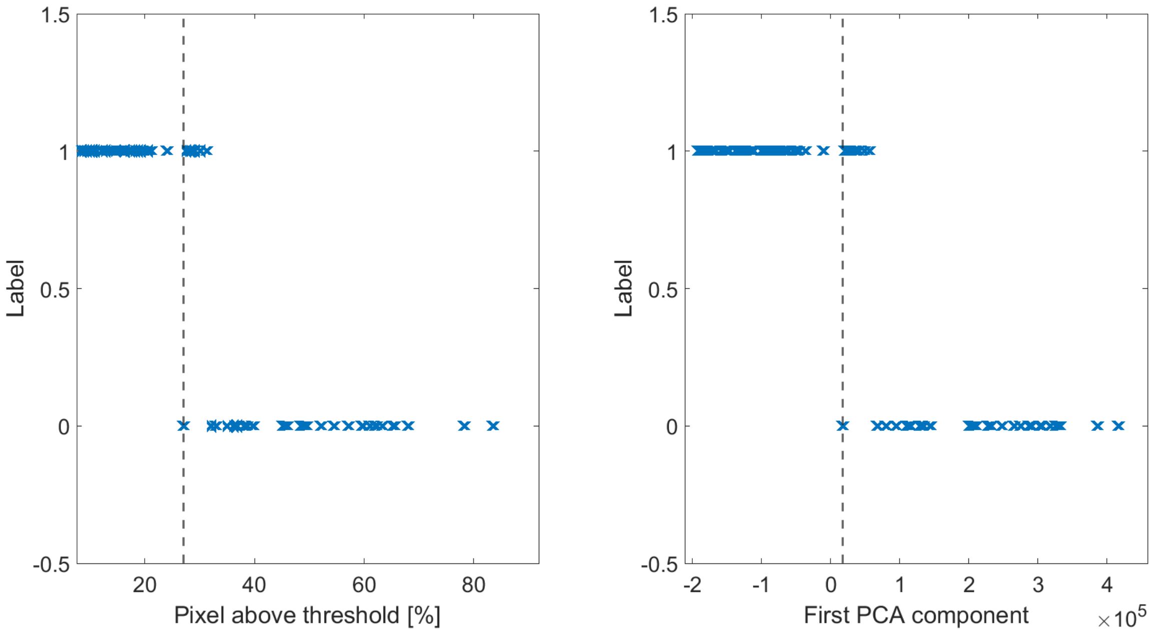


Figure 4.16: Manual labeling test

Thresholds have been set equal to the minimum value of the parameter for which over that value there are bad images, this is because it is reasonable not to want to exchange bad images as good, while the opposite is more tolerated. In both cases, there is a clear separation of the good images from the bad ones. The performances in the two cases are equivalent: with such a choice of thresholds, which is for sure not noise resistant (but is useful for testing the idea) about 12% of the good images are lost. This method is subjective, the assignment of labels manually done by a different person would probably lead to different results, but the idea remains valid that the parameters capture the problem and that good/bad labeling instead of a quantitative type could be useful. Furthermore, the positions on which the analysis was made are only one hundred, a very small sample. The defects mentioned above have been solved by creating automatic labeling of the images and testing the approach on a thousand images.

4.5 Automatic labeling

It is important to remember that the final output of the instrument is milk infected/not infected, based on the presence or not of the hydrogen. A possible approach is to manually assign a label good/bad to each image and then use a binary classifier. This procedure has been tried, but there is a lot of hysteresis on the labeling: after a lot of bad images is "easier" to assign a label "good" if an image is not so bad and vice-versa. Moreover, the manual labeling requires a lot of time and its difficult to keep the same criteria all over the procedure. Finally, the manual procedure is conditioned by the person who carries it out. To overcome all these problems, a new parameter has been proposed to characterize the goodness of a position.

Figure (4.17) shows how the acquired image and its spectrum change between a good and a bad image.

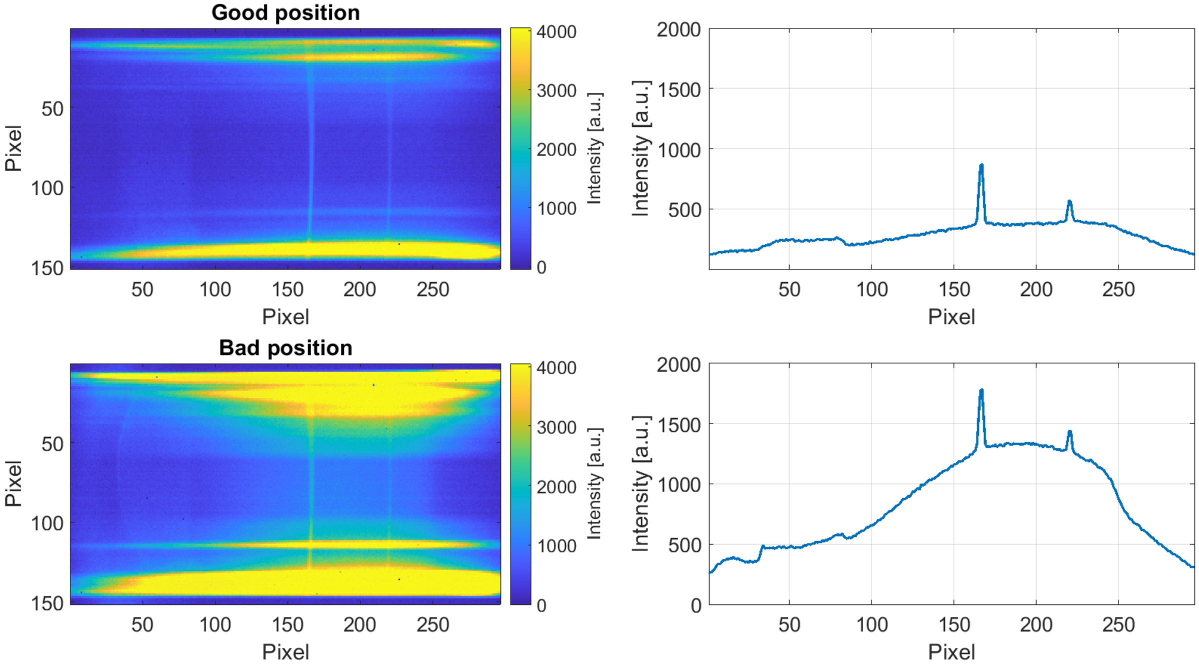


Figure 4.17: Good and bad position comparison

The previous figure clearly shows as diffuse light disrupts the spectrum and in particular the region where there are the spectral lines that is the most important. This fact implies that a "good" image is characterized by a low signal intensity between the peaks of nitrogen and that of oxygen.
For this reason, the ratio between the nitrogen peak and the average intensity between the two peaks (nitrogen and oxygen) has been proposed as a goodness parameter.
it is important to remember that the label thus defined cannot be calculated by the final software, as the gas concentrations are unknown and there may be favourable positions for the measurement but there is no oxygen inside the sample because it is exhausted by the bacteria's respiration, or if the pressure becomes very high, some of the nitrogen may escape from the test tube, as can be seen in Figure(1.1), that shows the acquired image of a sample infected by *clostridia* and there is no nitrogen.

Remains the problem of determining which threshold to use to distinguish good from bad images. Firstly, manual labeling has been used to find a first reference value, which remains purely qualitative, with this method a threshold equal to 1.6 has been found. To make the selection of the quantitative threshold, the standard deviation of the nitrogen's peak integral of the images classified as good according to the selected threshold has been analyzed. In practice, once a threshold is set, its value leads to a certain quantity of images classified as good and a spectrum is built from good images, once the baseline has been subtracted, the integral of the nitrogen's peak has been calculated. Depending on the threshold there will be more or less good images and this will vary the standard deviation of the integrals computed from spectra created from good images. All calculations for this optimization are based on the window optimization algorithm Sec(4.2.1).
To make the selection of the threshold more valid from a statistical point of view and to more consistently test the correlation between label and parameters from the single image, a thousand different images have been acquired.
The following figure shows the result obtained.

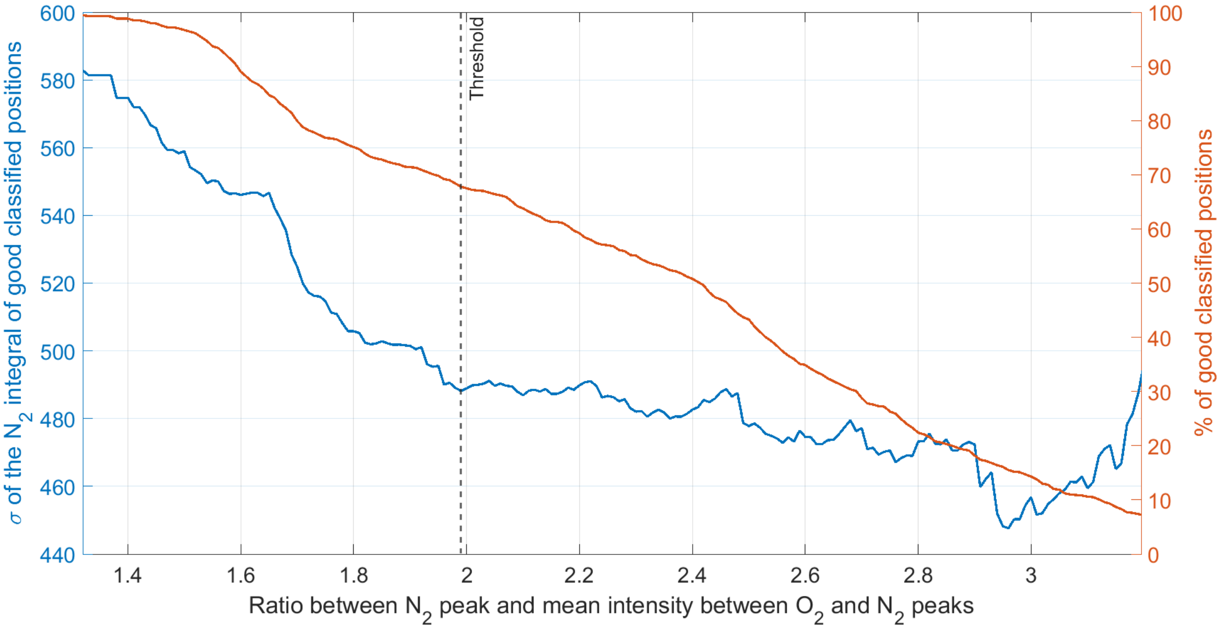


Figure 4.18: Ratio optimization for autolabeling

The trend is the expected one, with a larger and larger threshold there is less and less diffuse light in the image and therefore less and less images are classified as good, this leads to a smaller dispersion of the integral of the nitrogen peak produced by these good images. The threshold must not be too tight otherwise the majority of the images would be discarded, the human manual labeling suggests a ratio of 1.6 so the final threshold has been set equal to 1.99, this value is larger than the human one, hence, for sure, the worst images are classified as bad. With a larger threshold, there is not a sharp decrease in the dispersion and with less and less good images the values loses significance. With this selected threshold the 67.8% of the images is classified as good. The goal is now to test if the previous parameters can discriminate between good and bad images.

Regarding the percentage of pixels above a threshold, the result is presented in the next figure:

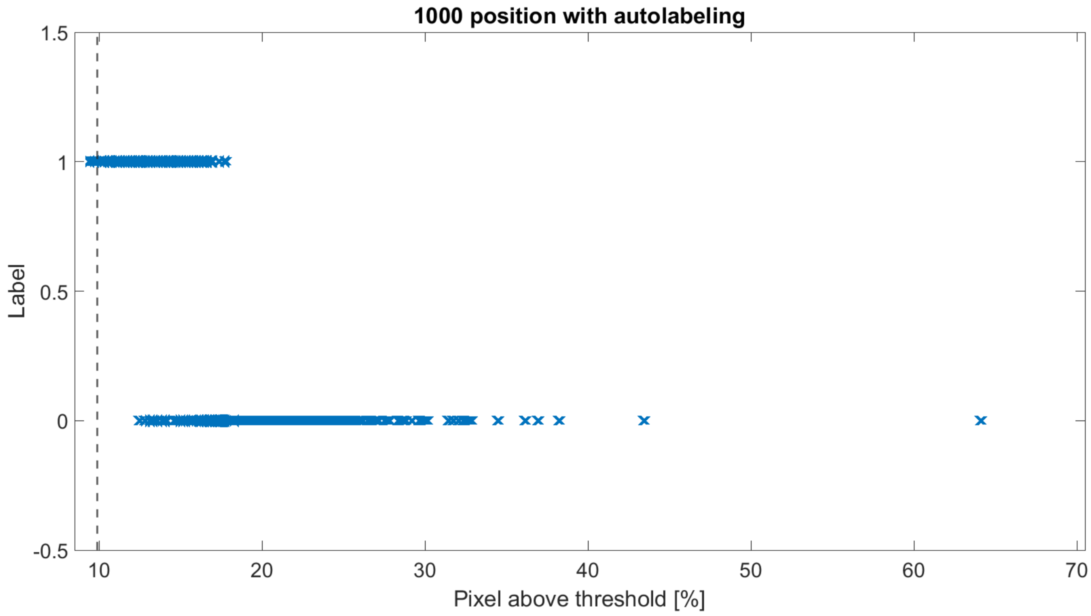


Figure 4.19: Percentage parameter vs autolabeling

Let's describe better the procedure: images whose ratio between the nitrogen peak and the mean intensity between nitrogen and oxygen peaks larger than 1.99 are classified as good. The label is compared then as a function of the percentage of pixels above half of the maximum intensity of the acquired image. The final algorithm uses a threshold, for the percentage parameter, set at the minimum percentage above which bad images start to appear, minus the five percent of the range swept by bad images. Basically, the threshold has been set in such a way to guarantee some tolerance with respect to noise.

The performance of this algorithm for the position classification is very poor, only the 1.03% of the good images is classified as good, and so the overall percentage of good images according to this procedure is 0.7% while the expected value is 67.8%.

The same has been done to check if the first PCA component is enough to classify images correctly, in this case, the result is:

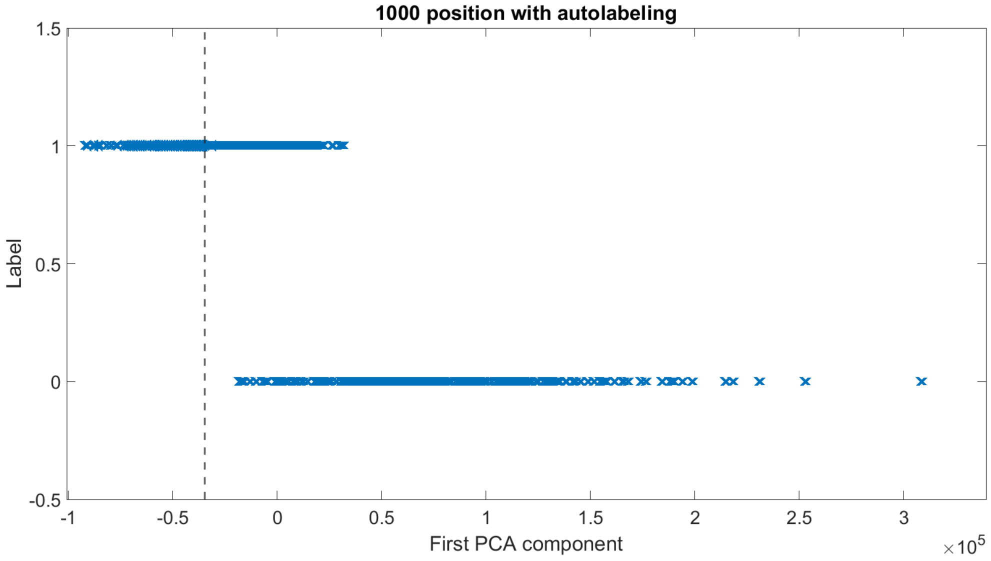


Figure 4.20: First PCA component vs autolabeling

The performance is better than the previous attempt, but not completely satisfactory. Only the 58.55% of the good images is classified as good, and so the overall percentage of good images according to this procedure is 39.7% while the expected value is 67.8%.

4.6 Machine learning approach: binary classifier

Summarizing what has been discovered so far:

- The SNR ratio is not a good parameter for characterizing the goodness of a measurement position.
- An internal standard has been defined to define a measurement position as good from a single measurement: when the ratio between nitrogen's peak and average intensity between nitrogen and oxygen peak is greater than 1.99. However, this method can be applied when there is only air in the vial.
- Both the percentage of pixels above a threshold and the first coordinate of the PCA are not enough to classify the images correctly.

Since now we have principal components of the image set and their label (valid only for air), the idea is to build a binary classifier that can work even with unknown gas concentrations.
The components of PCA arise from images with masked gas lines (even for hydrogen and carbon dioxide), so they are already independent of the contents of the tube. The label, on the other hand, is exactly what must be predicted, but, after having trained a classifier in this case, where both principal components and labels are known, it is possible to obtain a classifier that works in the most general case. The input is given by the first p components of the PCA of an image and the output is good/bad position for that acquisition.

One of the main problems when using a machine learning approach is having enough data available. Unfortunately, the acquisition of measurement takes about three seconds, even if the integration time is one second, but the camera-Matlab interface and the rotation of the tube slow down the process. In addition, the number of vials available is reduced to eighteen and these must be carefully cleaned before measurement, requiring additional time. The importance of using different vials will be described in the section below.

4.6.1 The importance of using different vials

As already mentioned, one of the main problems of machine learning is avoiding overfitting. To check if the approach was correct, a Neural Network (NN) was created using a thousand measurements taken from one vial and then it was tested on a dataset created with a different vial. The training vial has about 70% good images, as shown in Figure (4.18). The number of good positions predicted by the NN for the second vial was zero, but from the labeling rule, the percentage value of good images was actually 5%, therefore the NN was not completely wrong, even if all the good images are lost. This difference in the percentage of good positions (according to the autolabeling rule described above) between different vials is due to the different impurities presence in the vials' glass as they have all been cleaned with isopropyl alcohol before the measurement. In practice, what happened was having trained the NN with a particularly performing vial and having tested the net on a vial very rich in impurities.

This phenomenon implies that to avoid overfitting it is fundamental to train the model with as many vials as possible, this does not change so much the regions where the diffused light is found, but drastically changes the probability distribution of the two labels with which the neural network is trained.

To further investigate this fact, three hundred images of ten different vials were acquired, and the (labeling) percentage of good images varies greatly from vial to vial, as shown in the next Figure(4.21). Moreover this phenomenon makes impossible to determine a priori an estimate of how many rotations are necessary to find a good size, and so of the time required to process a vial; for vial number one, four out of five positions are fine, for vial number five one in ten is fine.

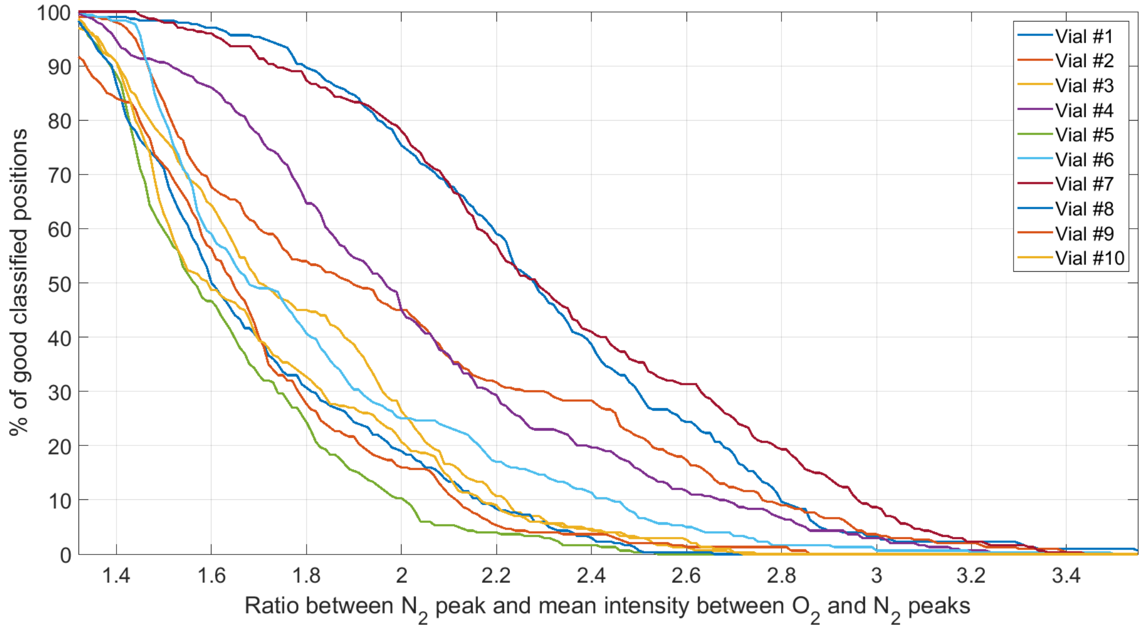


Figure 4.21: Different vials comparison

4.6.2 Data preparation

Given the need to use different vials to train machine learning models, a single image was acquired, in three hundred different positions, for eighteen vials, for a total of 5400 images. Each vial has been cleaned with isopropyl alcohol before the measurement. For each image, a 0/1 label has been associated with the procedure described in Sec(4.5), with the threshold optimized at 1.99. It is remembered that the assignment of the label is valid only with the vials containing air, but since the input of the model, the first components of the PCA, does not depend on the spectral lines, as the were masked, the trained model is independent of the contents of the tube.

The data were divided into training and test sets, since is delicate to make predictions on new vials, the data relating to the first twelve vials made up the training set, and the remaining test set. From the analysis of the PCA made on the test set, it appears that the first seven main components are needed to have more than 95% of the data variability.

The training dataset is therefore composed of three hundred vectors, for twelve different vials, with eight components, the first seven being the coordinates along the first seven principal components, the octave being the label. Since the task concerns classification, the 0-1 loss function has been used. In principle, another type of loss function could be used, to make a false negative more important than a false positive, but as will be presented later, even this simple loss function leads to good performances.

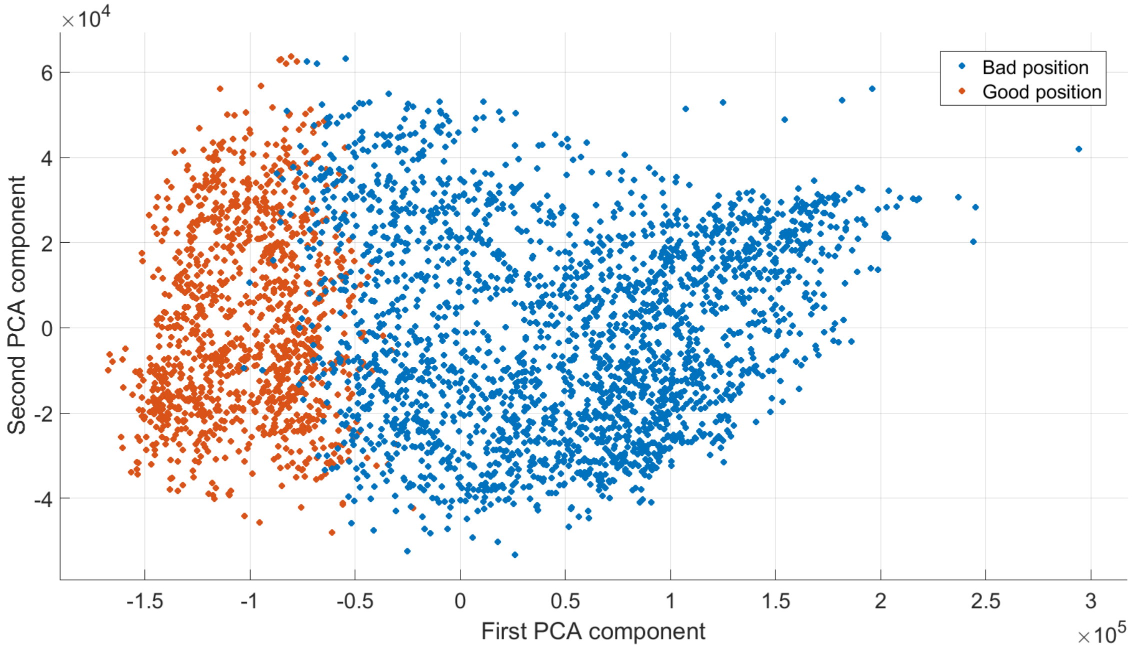


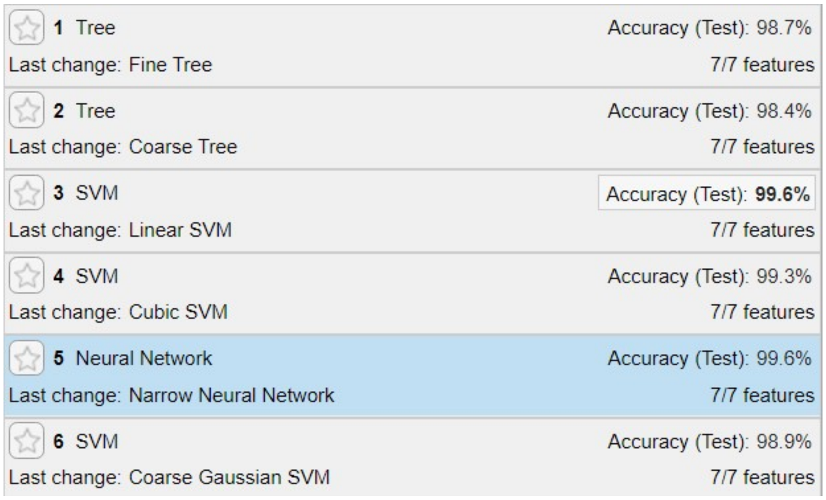
Figure 4.22: Training data scatter plot

Figure (4.22) shows the distribution of the data along the two first principal components, as can be seen, the bad positions prevail and the two regions are quite separate. Unfortunately, there is no wide margin between the two areas.

It can also be noted that most of the positions are considered bad by the labeling rule, overall, using the whole dataset made of eighteen vials, about 24% of the positions are considered as good. This leads to an estimate that it takes about four rotations to find a good position, therefore a time of about ten seconds, clearly this time depends very much on the vial as shown above in Figure (4.21).

4.6.3 Model selection

Several binary classifiers have been trained and tested with Matlab, as can be seen from Figure(4.23), the performances are all very good. Although the accuracy of the linear Support Vector Machine (SVM) is the same as that of the neural network, it was chosen to use this second model as the false negative rate is lower in this case, as shown in Table(4.1). Having a low false negative rate is important especially in cases where the vial has few good positions, so discarding a few good ones implies a faster analysis speed. However, this results in a higher false positive rate, but as will be shown later, these false positives are actually acceptable.



☆ **1** Tree — Accuracy (Test): 98.7%
Last change: Fine Tree — 7/7 features

☆ **2** Tree — Accuracy (Test): 98.4%
Last change: Coarse Tree — 7/7 features

☆ **3** SVM — Accuracy (Test): **99.6%**
Last change: Linear SVM — 7/7 features

☆ **4** SVM — Accuracy (Test): 99.3%
Last change: Cubic SVM — 7/7 features

☆ **5** Neural Network — Accuracy (Test): 99.6%
Last change: Narrow Neural Network — 7/7 features

☆ **6** SVM — Accuracy (Test): 98.9%
Last change: Coarse Gaussian SVM — 7/7 features

Figure 4.23: Training data scatter plot

Model	False positive rate [%]	False negative [%]
Linear SVM	0.1	5.9
Narrow Neural Network	0.2	1.7

Table 4.1: Selection model

4.6.4 Neural network model

The artificial neural network (NN) is a computational model inspired by the structure of the brain in humans. In a very simplified model, the brain is made up of a large number of basic computing devices linked together. Through a complex communication system between neurons, the brain is able to perform very complex calculations. NNs are based on this paradigm. [15]
An NN is structured in layers, each layer is composed of several neurons, the first layer is the input layer, the last layer is the output layer, and the layers between these two are called hidden layers.

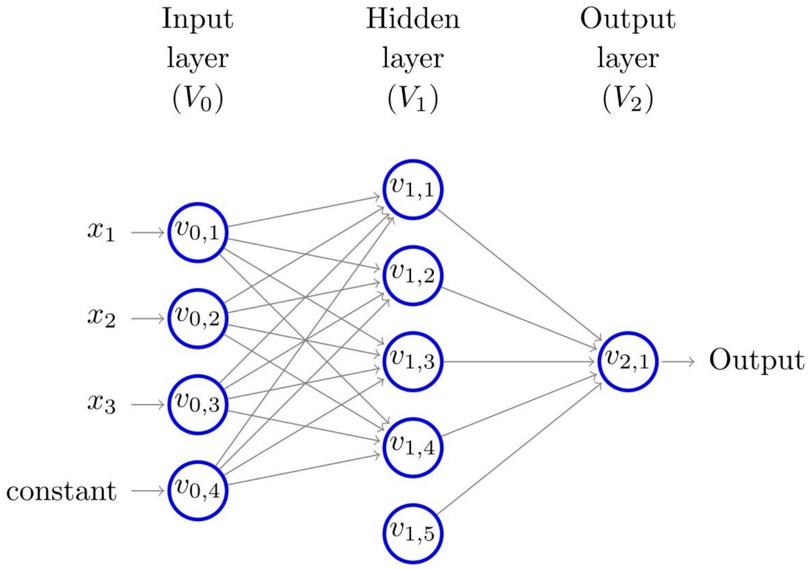


Figure 4.24: NN model [15]

A fully connected NN is a NN such that each neuron of a layer is connected to each neuron of the previous layer. The simplest NN is the feedforward, in which information is transmitted only in one direction.

From a mathematical point of view, a NN is modelled as a graph $G = \{N, E\}$ whose nodes are neurons N and whose edges E represent the connections between them and a weight function over the edges $w : E \to \mathbb{R}$.
The output of each neuron is given by the output of the previous neurons, weighted by the connection, and, once a bias is added, a non-linear function σ is applied.
Indicating con $w_j^{(t)}$ the weights from all neurons of layer $t-1$ to the neuron j of the layer t, and with $o^{(t-1)}$ all outputs of neurons in layer $t-1$, the output of the j-th neuron in layer t is:

$$o_j^t = \sigma\Big(\langle w_j^t, o^{(t-1)}\rangle\Big) \tag{4.7}$$

The σ is called the activation function, it is a scalar non-linear function, one of the most used is the ReLu function, defined as:

$$\sigma(a) = \max\{0, a\} \tag{4.8}$$

Several activation functions can be used, but ReLu is very fast and it is less expensive than other functions for example tanh or sigmoid.

Once a vector $x_i \in \mathbb{R}^d$ is given as input, with d equal to the number of neurons in the input layer, through the propagation described above an output $y_i \in \mathbb{Y} = \mathbb{R}^k$ is produced, which can be scalar or a vector.

The aim of a neural network is to create a model that can best predict the output given by a certain input, formally, the minimization of empirical risk, defined as:

$$L_s(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, (x_i, y_i)) \tag{4.9}$$

With m the size of the training set (set of data used to train the network) and with l the loss function and h the hypothesis, so the model created, which depends on all the parameters of the network, G, the weights h, σ.
The network is optimized by varying the weights between neurons in order to obtain the lowest possible error between the correct label and the one predicted by the

network. The loss function l measures how much the two values are different. The 0-1 loss function:

$$l_{0-1}(h,(x,y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases} \tag{4.10}$$

is used for classification task. L2 or squared loss function instead, is used for regression:

$$l_{sq}(h,(x,y)) = (h(x) - y)^2 \tag{4.11}$$

One of the main problems of neural network training is overfitting - creating a model that only works for training data, but when new data is used, predictions are much more wrong. To limit overfitting there are several strategies, one often used is to minimize, in addition to empirical risk, also a term called regularization, in practice, it is minimized:

$$L_s(h) = \frac{1}{m}\left(\sum_{i=1}^{m} l(h,(x_i,y_i)) + \lambda \sum_{i,j,t} (w_{ij}^t)^2\right) \tag{4.12}$$

With w_{ij}^t the weight between neuron i of layer t-1 and j of layer t, λ controls the amount of the regularization. The idea is that even minimizing the value of the weights avoids creating some that are too large and the final output depends too much on those connections.

Without going into the details of the optimization algorithm, called the backpropagation algorithm, the main steps to train the net are:

1. Weight initialization at random or with more smart strategies, for example Glorot inizializer.

2. Evaluation of the empirical loss

3. Evaluation of the gradient of the loss

4. Updating of the weights.

5. Iteration of steps 2-3-4 until a minima of the empirical risk is reached.

The computationally most expensive step is the evaluation of the gradient, which to be done requires processing all the training sets at each iteration, for this reason, the well-known Stochastic Gradient Descent algorithm is often used, which in addition to being faster also has a certain resistance against the local minima from which it can come out. The field of optimization is vast and there are many efficient algorithms for network optimization such as adaptive learning rate or others.

To conclude the section, it is recalled that parameters such as the number of neurons or the amount of regularization are called hyperparameters. For the optimization of the hyperparameters what is usually done is to divide the training set into two parts, for example, 70% -30% called training and validation set. In practice, the hyperparameters are varied, training the network with the training set, and then an error is estimated on new data, not used for training, using the validation set. The hyperparameters leading to the minor validation set are chosen. The estimation of the error of the definitive network must be made on completely new data since it will be in this condition that it will then have to work. There are two possibilities: get new data or divide the data you have into three parts, training, validation and test set.

4.7 Final position selection algorithm

The final algorithm consists of the application of the rotation matrix created by the PCA and the use of a neural network for binary classification; this section describes the details of the NN's hyperparameter optimization and the performance of the final algorithm.

4.7.1 Neural network optimization

The model has been trained and tested in Matlab which allows the possibility to optimize two hyperparameters: the number of neurons and the amount of regularization.

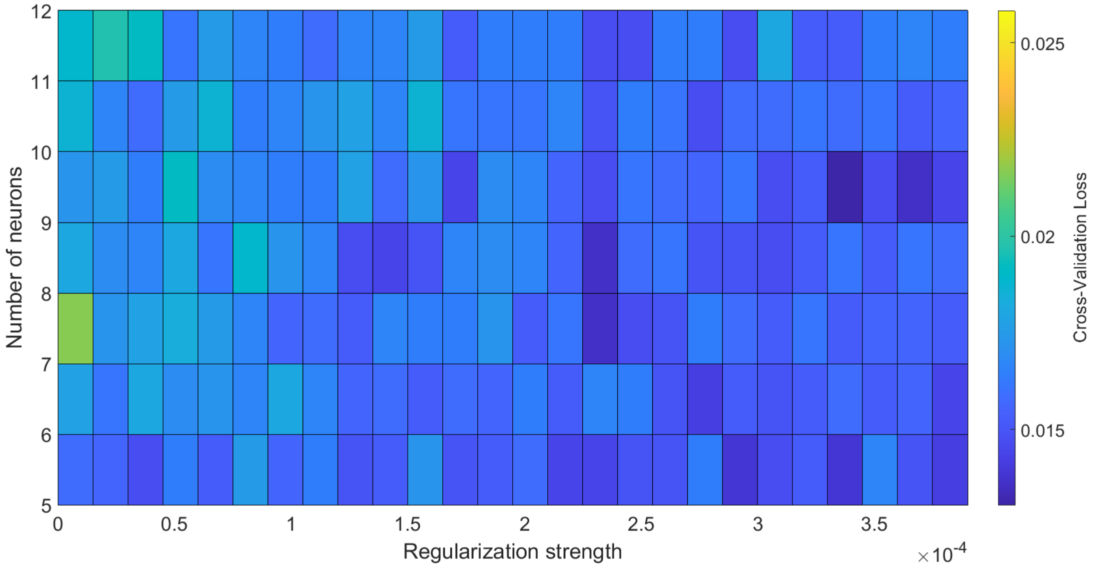


Figure 4.25: Hyperparameter tuning

Figure(4.25) shows how the optimization of hyperparameter has been done. The grid is regular and the error is the misclassified rate in decimal, calculated from a 5-fold validation set. The best model is the one with 9 neurons and a regularization strength $\lambda = 3.3 \cdot 10^{-4}$, therefore the number of total edges is 64.

It has been chosen not to use networks with multiple hidden layers to prevent the

model from becoming too complicated, as will be shown later, already two input components capture the essence of the problem, even if not completely therefore also a relatively simple model with only one middle layer will be satisfactory.
The error on the validation set is 1.3%.

4.7.2 Neural network performances

The model has been tested on the six new vials, not used in the training phase, the error on the test set was equal to 0.34%. The next figure shows the confusion matrix of the test set:

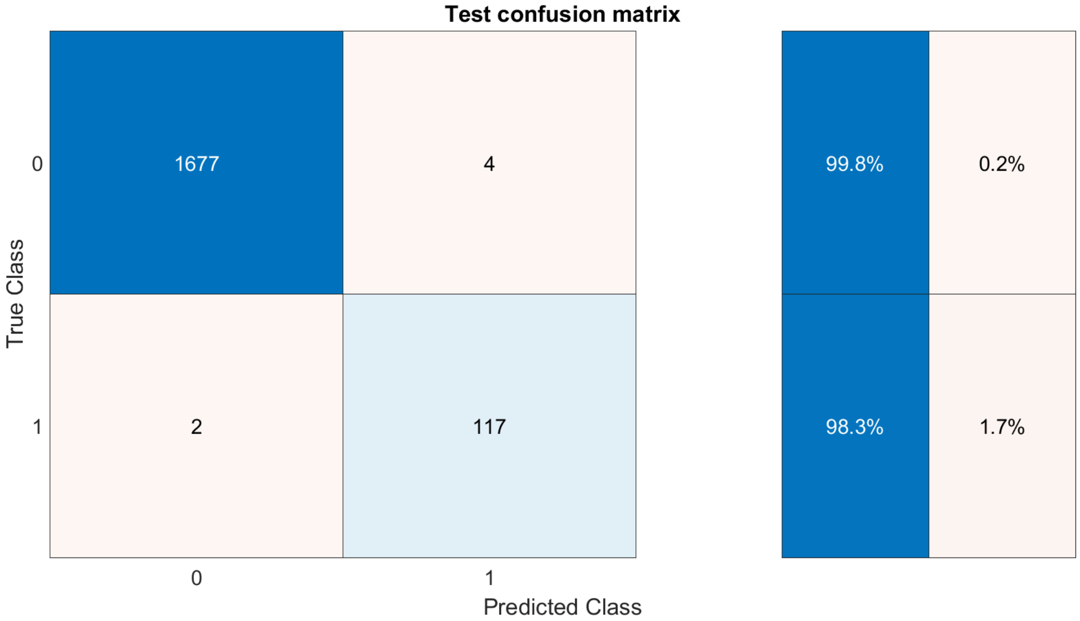


Figure 4.26: Confusion chart of the test set

As it can be seen, the performances are very good, in particular, the false positive rate is very low and this is particularly important since it would be a more serious mistake to mistake a bad position for good, rather than the other way around.

The percentage of good images expected and that predicted by the neural network for the whole set of eighteen tubes is shown below, it should be remembered that only the first twelve are used for training, and the others were used for the test. It should be noted that the input of the classifier, that is the first seven main components given an image, is calculated through the rotation matrix of the PCA created by the first twelve vials, so the test reflects what the final algorithm will have to do, that calculates the principal components and then give them as input to the classifier.

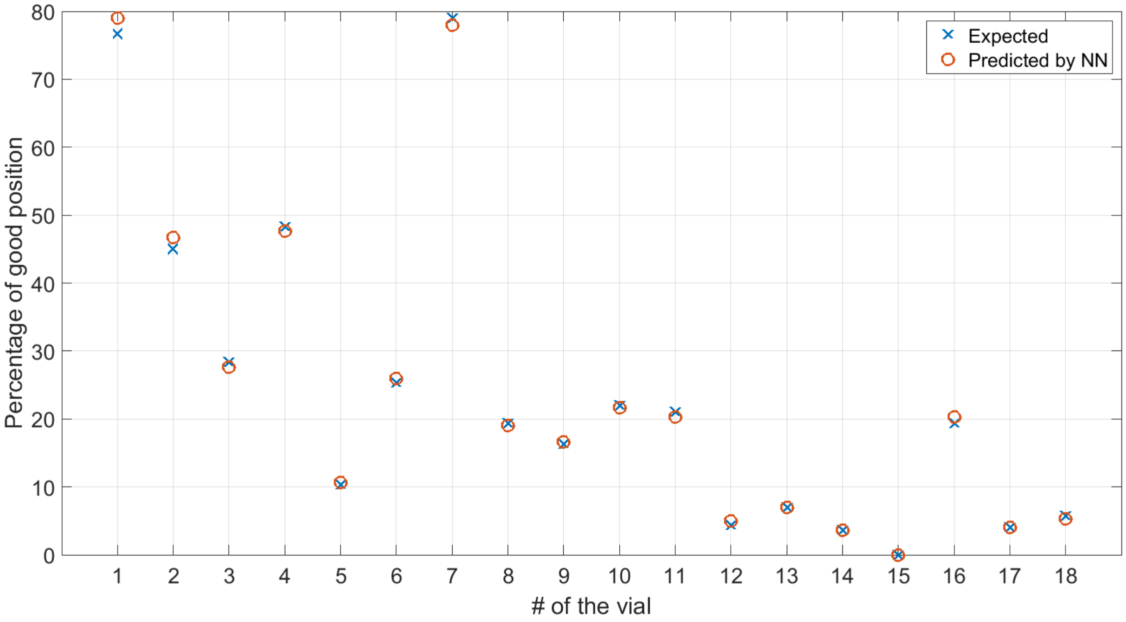


Figure 4.27: Good positions expected vs predicted

It can be noted that the fifteenth vial does not present good positions, this is interesting because it had not happened up to that moment and would involve an infinite loop of rotations if analyzed, this involves having to put a limit on the number of rotations, and in case, discard the vial without being able to obtain any useful information about the milk inside. The degree of agreement between the expected and expected values is very good, the deviation is at most a few percentage points. The performances are good, but it remains to be understood how serious the mistakes made are, that is, the positions that are classified incorrectly. Are there fully saturated images that are classified as good or vice versa?

It should be remembered that a position is classified as good if the ratio between the nitrogen peak and the average intensity between nitrogen and oxygen peak is greater than 1.99. To check how wrong the badly classified positions were, the ratio of the nitrogen peak to the average nitrogen peak to oxygen intensity of these positions was plotted, the result is shown in the next figure.

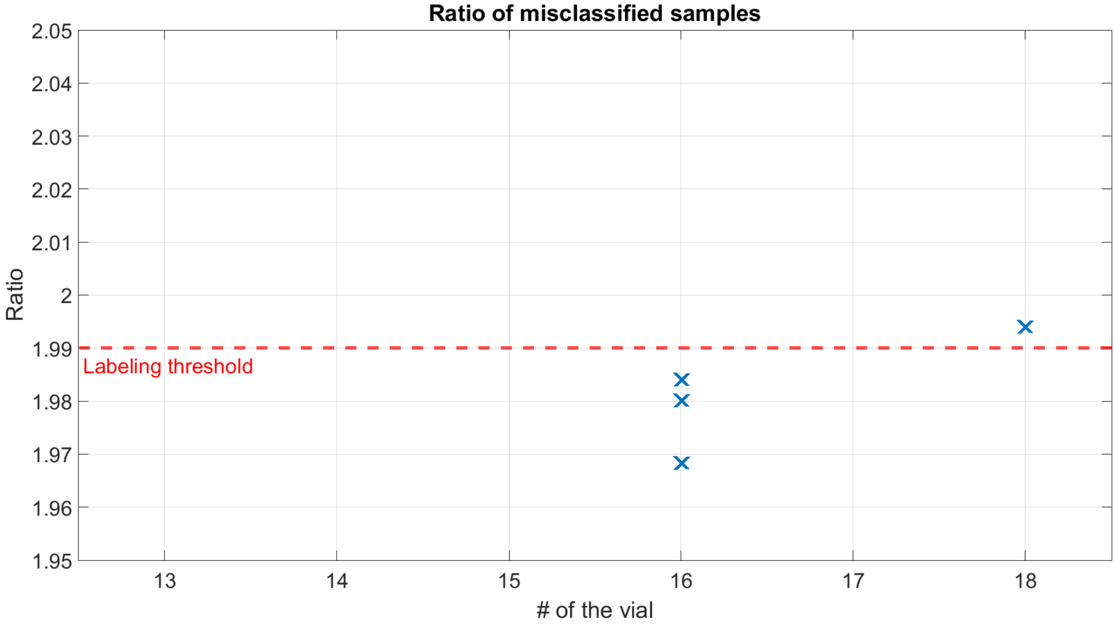


Figure 4.28: Misclassified positions

Figure(4.28) shows as the ratio between nitrogen and oxygen peak of the misclassified position is very close to the threshold. This means the error is not so serious, it would be severe if a position whose ratio is 1.4 was classified as good or with a ratio of 3 was classified as bad.

Overall, it can be said that the classifier distinguishes well the good positions from the bad ones in over 99% of the cases, moreover, in the few cases where the error is wrong, it is of little importance given that the ratio between nitrogen peak and average intensity between the nitrogen peak and the oxygen peak it is close to 1.99, that is the threshold chosen to discriminate the positions.

As the test set showed good classifier performance, the definitive neural network was eventually trained using all eighteen vials.

4.7.3 Proposed algorithm for the position selection

To conclude, the flow diagram of the proposed algorithm to find the good positions in which to carry out the measurement is shown in Figure(4.29).

The proposed algorithm complies with all the conditions required by the instrument:

1. Discriminates positions at which the image is saturated from places where it is not.

2. It does not depend on the gases concentrations as PCA rotation matrix have those lines masked.

3. It can be used in a real-time application since it requires a simply flattening of a vector, a matrix multiplication and a propagation through a very simple neural network.

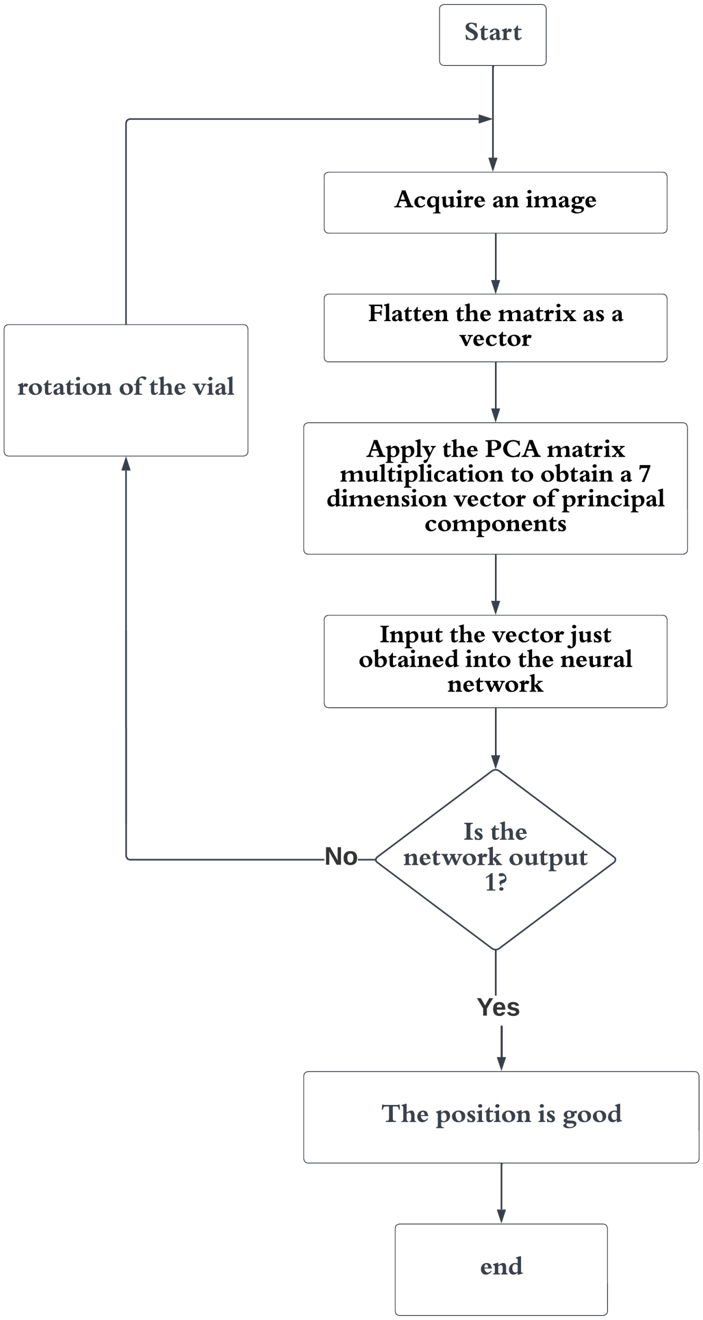

Start
Acquire an image
Flatten the matrix as a vector
Apply the PCA matrix multiplication to obtain a 7 dimension vector of principal components
Input the vector just obtained into the neural network
Is the network output 1?
No
rotation of the vial
Yes
The position is good
end

Figure 4.29: Flowchart of the position selection algorithm

Chapter 5

Conclusions

The identification of *clostridia* in milk from the analysis of gases in the vial's headspace by Raman spectroscopy is very promising, in terms of time, cost and repeatability of the test. The main problem of this technique is the interference due to the fluorescence that spoils the measurement, even completely saturating the acquisition, but fortunately, changing the measurement position can improve the situation until an acquired image is almost devoid of light generated by fluorescence. The aim of the thesis is to automate the rotation of the test tube in order to obtain only images that are not too compromised by diffused light.

First of all, a support rod for the stepper motor that rotates the test tube was initially designed. Next, it was verified whether the signal-to-noise ratio of the oxygen peak integral can be used as a quantifier of the goodness of the measurement position. However, this procedure requires multiple measurements, so it is not applicable to the final algorithm because it would slow down the analysis of samples too much. The idea was to find a parameter that can be calculated from the single acquisition and to correlate it to the SNR, so that, after applying an optimal threshold, it is possible to measure only the first and guarantee a high value of the second, without having to carry out repeated measurements. Two calculable parameters from a single measurement have been proposed: the percentage of pixels that have an intensity in the upper half of the intensity range and the the coordinate along the first principal component. The correlation between the two parameters just described and the signal to noise ratio was tested, but it was not satisfactory. The SNR of the oxygen peak integral proved not to be expressive of the goodness of the measure.

After that, a binary label (good / bad) was manually assigned to each position and it has been compared with the previously described parameters from a single image, which shows that they are effective in indicating the goodness of a measurement position. As there is a lot of hysteresis in manual label assignment, as well as being slow, it was decided to make the labeling procedure automatic. A role has been created to automatically associate a good / bad label with each image: if the ratio between the nitrogen peak and the average intensity between nitrogen and oxygen peak is greater than 1.99, the position has to be considered good, otherwise it has to be considered bad. This labeling rule is valid for samples containing only air, and cannot be used in the final application since the gas concentrations are unknown, but it is useful to get a label for each image.

A neural network binary classifier has been created that takes in input the coordinates along the first seven principal components of an acquired image and returns the good / bad label in output. Since the spectral lines were masked in the calculation of the PCA rotation matrix, from a single image, seven coordinates are obtained, which are independent of the gas concentration and therefore depend only on the diffused light in the image. The training set consists of eighteen vials and three hundred different positions for each vial.

The final algorithm proposed discriminates the good or bad positions with an accuracy greater than 98% and the few wrong positions have a ratio between nitrogen peak and average intensity between nitrogen and oxygen very close to the threshold value of 1.99, therefore the algorithm is very satisfactory.

From the hardware point of view, the instrument is still in an initial phase, it will be necessary to make a thermostat to keep the temperature of the milk at 37 degrees to help the growth of bacteria. Furthermore, a translator will have to be realized to automate the serial analysis of vials. Moreover, a software-controllable shutter must be created in order to block the laser beam when the translator changes the vial under analysis, to prevent it from catching fire. The use of a pulsed laser in coupling with the camera will be investigated, to keep the integration time in the order of nanoseconds and avoid part or all of the fluorescence, but this could imply problems that arise from non-linear optical effects of the glass, given the use of a very high intensity laser beam.

From the software point of view, on the other hand, the PCA will have to be recalculated as the use of the translator will cause the vials to position themselves in a slightly different way and this causes a different distribution of the diffused light in the acquired image. It is also conceivable to construct a rotation matrix from the PCA for each position of the vial holder. Moreover, the use of different vials is essential to obtain a training set more faithful to what will then be the proper distribution of good/bad positions and the creation of a more reliable neural network binary classifier. In addition, all the moving parts of the instrument, shutter, translator, and stepper motor must be controlled at the same time.

Finally, the presence of milk in the vial could result in the appearance of drops of condensation on the vial' surface, this may require a slower rotation of the vial and an increase in the number of rotations necessary for the each vial to find an useful acquiring position, resulting in a new tuning of the proposed algorithm.

Bibliography

[1] Cristian Andrighetto, Lorenzo Cocola, Paola De Dea, Massimo Fedel, Angiolella Lombardi, Fabio Melison, and Luca Poletto. Determination of co2 and h2 content in the headspace of spore contaminated milk by raman gas analysis. 12120:8–15, 2022.

[2] Autodesk. `https://www.autodesk.it/products/fusion-360/overview`.

[3] Basler. Basler product documentation. `https://docs.baslerweb.com/exposure-time.html?filter=Camera:acA1920-40um`, 2022.

[4] Daniela Bassi, Cecilia Fontana, Simona Gazzola, Ester Pietta, Edoardo Puglisi, Fabrizio Cappa, and Pier Sandro Cocconcelli. Draft genome sequence of clostridium tyrobutyricum strain uc7086, isolated from grana padano cheese with late-blowing defect. *Genome announcements*, 1(4):e00614–13, 2013.

[5] J Brändle, V Fraberger, K Schuller, U Zitz, W Kneifel, and KJ Domig. A critical assessment of four most probable number procedures for routine enumeration of cheese-damaging clostridia in milk. *International Dairy Journal*, 73:109–115, 2017.

[6] Vincenzina Fusco, Daniele Chieffi, Francesca Fanelli, Antonio F. Logrieco, Gyu-Sung Cho, Jan Kabisch, Christina Böhnlein, and Charles M. A. P. Franz. Microbial quality and safety of milk and milk products in the 21st century. *Comprehensive Reviews in Food Science and Food Safety*, 19(4):2013–2049, May 2020.

[7] Robert G Jensen. *Handbook of milk composition.* Food Science and Technology. Academic Press, May 2014.

[8] MathWorks. pca documentation. `https://www.mathworks.com/help/stats/pca.html#d123e694493`.

[9] Jeanne L. McHale. *Molecular Spectroscopy.* CRC Press, July 2017.

[10] Alessia Mortari and Leandro Lorenzelli. Recent sensing technologies for pathogen detection in milk: A review. *Biosensors and Bioelectronics*, 60:8–21, 2014.

[11] WHO (World Health Organization). Who's first ever global estimates of foodborne diseases find children under 5 account for almost one third of deaths, 2015.

[12] Julian Parfitt, Mark Barthel, and Sarah Macnaughton. Food waste within food supply chains: quantification and potential for change to 2050. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554):3065–3081, September 2010.

[13] Arshak Poghossian, Hanno Geissler, and Michael J. Schöning. Rapid methods and sensors for milk quality monitoring and spoilage detection. *Biosensors and Bioelectronics*, 140:111272, September 2019.

[14] Arshak Poghossian, Hanno Geissler, and Michael J. Schöning. Rapid methods and sensors for milk quality monitoring and spoilage detection. *Biosensors and Bioelectronics*, 140:111272, September 2019.

[15] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[16] E. Smith and G. Dent. *Modern Raman Spectroscopy: A Practical Approach.* Wiley, 2019.

[17] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision.* Cengage Learning, 2014.

[18] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, January 1991.

[19] Peter Vandenabeele. *Practical Raman Spectroscopy - An Introduction.* John Wiley & Sons, Ltd, July 2013.