TÉCNICO LISBOA

Machine Learning-based Detectors for Magnetic Signal Differentiation applied to Early Cancer Detection

Rita Susana Gomes Ramos

Thesis to obtain the Master of Science Degree in

Electronics Engineering

Supervisors: Prof. Diogo Miguel Bárbara Coroas Prista Caetano
Prof. Gonçalo Nuno Gomes Tavares

Examination Committee

Chairperson: Prof. Paulo Ferreira Godinho Flores
Supervisor: Prof. Diogo Miguel Bárbara Coroas Prista Caetano
Member of the Committee: Prof. Ana Luísa Nobre Fred

October 2022

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Here I would like to thank all those who contributed to its final form in one way or another.

I have to start, first, with a thank you to my grandparents, because without them none of this would be possible, and especially to my grandfather Simão, because it was thanks to his advice that guided me, since I was young, to follow the right paths. He was the one who helped me to complete the most difficult and important stage of my life right now. Unfortunately, even if you're not here to see it, I'm sure you'll feel this achievement, wherever you are.

Thank to my boyfriend, João, for his patience and especially for his companionship and affection over the last 6 years. It was IST that brought us together, and without you, no doubt that this journey would have been more difficult both personally and academically. Knowing that we have someone next door to share questions and solve problems makes it easier. I look forward to your turn to finish this stage too.

Thank to my family, who never doubted my abilities and always encouraged me to follow the academic path. To my mother, Florbela, who raised me and gave me the necessary foundations that made me who I am today, and who, even in the greatest difficulties she faced, always did her best so that I didn't lack for anything. To my brothers who, even without knowing, with your example have always guided me to be a better person. To my brother Pedro and Rosa who encourage me, with their work and dedication, to invest more in me to achieve a better future. And to my brother Tiago for always being present and for his way of being, transmitting security, affection and love. It is a privilege for me to be your sister.

Thanks to my friends, with whom I vented hours on end about how difficult this academic journey was and who helped me overcome and forget about some problems whenever we were together.

And finally, thank my supervisors Prof. Diogo Caetano and Prof. Gonçalo Tavares for their knowledge and guidance during this work. Especial thanks to Diogo, for his experience, has always been there to help with his councils, to clarify any doubts, and to require presentations and reports so wouldn't let the work be delayed. I would also like to thank all my colleagues for their teamwork and information exchange, and in particular, thank engineer Ruben Afonso, provided me with the first simulation script, which served as the origin of all my work and who helped me whenever I needed it. It is a privilege for me to be part of a project and a team like this, and to know that I participated and helped a little more to reach such a noble objective that I am sure it will succeed.

Abstract

This dissertation presents the implementation of a machine learning model, namely an artificial neural network for distinguishing magnetic signals from cells marked with magnetic particles, and clusters of free magnetic particles. This dissertation is part of a project that is being developed between INESC-ID and INESC-MN, whose main objective is the detection of cancer cells in a blood sample. The database that served as training for the neural network was designed based on simulations carried out using the magnetic dipole equation as a model of the behavior of the magnetic particle for the two different cases. The training of the model was implemented using the ReLu function as an activation function of the hidden layer, and an accuracy greater than 90% was obtained for a system with an RMS value of the noise signal lower than 7 μV, an ideally higher number of magnetic particles per cell/cluster at 20, and a maximum magnetic particle travel speed of 2.5 m/s.

Keywords

Manetic Flow Cytometer; Cancer Cells; Magnetic Signals; Machine Learning; Artificial Neural Networks.

Resumo

Esta dissertação apresenta a implementação de um modelo de aprendizagem automática, nomeadamente uma rede neuronal artificial para distinção de sinais magnéticos provenientes de células marcadas com partículas magnéticas, e agregados de partículas magnéticas livres. Esta dissertação está inserida num projecto que está em desenvolvimento entre o INESC-ID e o INESC-MN que tem como principal objectivo a deteção de células cancerígenas numa amostra de sangue. A base de dados que serviu de treino para a rede neuronal, foi desenhada tendo por base simulações realizadas através da equação do dipolo magnético como modelação do comportamento da partícula magnética para os dois casos distintos. O treino do modelo foi implementado utilizando como função de ativação a função ReLu e foi obtida uma precisão superior 90% para um sistema com um valor RMS do sinal de ruído inferior a 7 μV, um nº de partículas magnéticas por célula/aglomerado idealmente superior a 20, e uma velocidade máxima de viagem de partícula magnética de 2.5 m/s.

Palavras Chave

Citómetria de fluxo magnético; Células Cancerígenas; Sinais Magnéticos; Aprendizagem automática; Redes Neuronais Artificiais.

Contents

List of Figures

List of Tables

Acronyms

ANN	Artificial Neural Network
CT	Computed tomography
FC	Flow Cytometry
GMR	Giant Magnetoresistances Sensors
INESC	Instituto de Engenharia de Sistemas e Computadores
INESC-ID	Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento
INESC-MN	Instituto de Engenharia de Sistemas e Computadores - Microsistemas e Nanotecnlogias
MFC	Magnetic Flow Cytometer
ML	Machine Learning
MMP	Micro-Metric Magnetic Particles
MNP	Magnetic Nano Particle
MR	Magnetoresistive sensor
OFC	Optical Flow Cytometry
SNR	Signal-to-Noise Ratio
WHO	World Health Organization

1

Introduction

Contents

1.1 Motivation

The advancement of technology has allowed human beings to have more and more knowledge of themselves and their surroundings. With this evolution, and especially with the scientific discoveries of recent times, namely the end of the last century, medicine is one of the areas that has benefited most from the discovery of new techniques for prevention, diagnosis, and treatment. An example of this is oncological diseases.

Cancer is a disease that results when cellular changes cause the uncontrolled growth and division of cells [7]. Apoptosis is a phenomenon of the natural death of a cell. A cell receives the instruction to die so that the body can replace it with a newer cell that functions better. However, cancerous cells don't respect these instructions causing the uncontrolled growth and division mentioned before. That causes the accumulation of dead cells in the body, which can form tumors, impair the immune system, and cause other changes that prevent the body from functioning regularly.

Cancer is a major burden of disease worldwide. Each year, tens of millions of people are diagnosed with cancer around the world, and more than half of the patients eventually die from it [8]. This disease brings, direct or indirectly, many consequences and social impacts on health and the economy. That's why it is so important to prevent it.

In the last decades, there has been a great evolution both in the diagnosis and in the treatment of cancer cases. These technological advances have marked a new perspective in the fight against cancerous diseases and, currently, there are already numerous diagnostic tests for the detection of cancer, such as various types of biopsies and cytology.

According to the World Health Organization (WHO), early diagnosis improves cancer outcomes [9], and the reason this happens is because as cancer results in uncontrollable cell division and multiplication, so early detection may mean less treatment and less time spent recovering [10], leading to fewer costs to fight the disease and a better chance to survive. Due to the importance of early detection of cancer in the fight against the disease, WHO promotes and supports Member States to develop and implement cancer early diagnosis and screening programmes, according to assessed feasibility and cost-effectiveness of screening, and with adequate capacity to avoid delays in diagnosis and treatment.

Usually, to detect cancer, a Computed tomography (CT) scan is initially performed to identify the structures of cancer cells, and subsequently, a biopsy is performed to understand at what stage the disease is. These tests can identify the disease, however not always at an early stage and the costs turn out to be very high. The solution may be to use flow cytometry, as it has a much more beneficial cost and faster results. Thus, Flow Cytometry (FC) appears as a new way of detecting cancer cells in a blood sample, and it is also possible to detect bacteria.

1.2 Objectives

This main objective of this work is the study, and simulation of magnetic signals produced by Magnetic Nano Particles (MNPs) that are attached to the cells or clusters flowing in a microfluidic channel in a cytometer and the implementation of a Machine Learning (ML) model, for distinction of these signals. A Magnetic Flow Cytometer (MFC) is a laboratory technique used to count, detect, identify, and characterize cells in a solution using Magnetoresistive sensor (MR) sensors. This magnetic-based cytometer, that is represented in Figure 1.1, is part of a project that is being developed at Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID) and Instituto de Engenharia de Sistemas e Computadores - Microsistemas e Nanotecnlogias (INESC-MN), which my dissertation is part, and its main objective is the detect the presence of cancer cells in a blood sample.

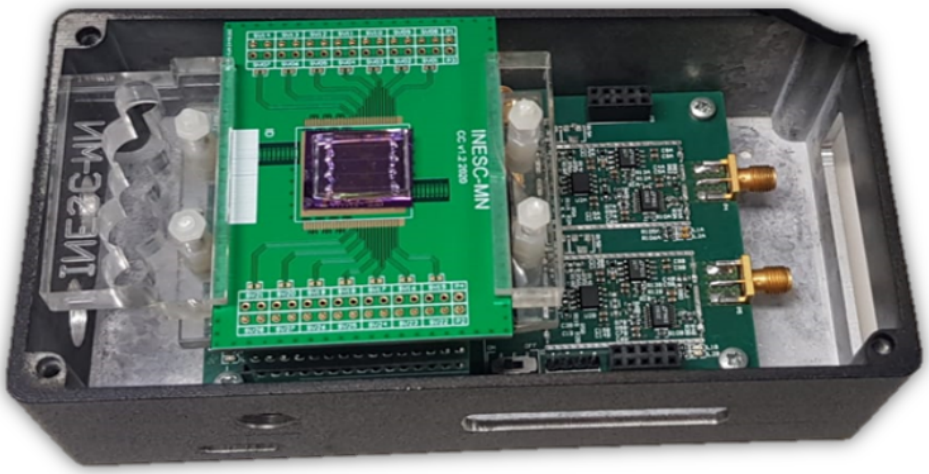

Figure 1.1: Prototype of an MFC made by engineers Rita Soares and Ruben Afonso.

The MFC can count cells in a flow if these cells are previously marked with antibodies or other probes carrying MNPs. Then, these marked cells are injected in a microfluidic channel, forcing them to come in proximity with the MR, while in a continuous flow [3], the magnetic field produced by the magnetic particles is gradually picked up by the MR sensors (Figure 1.2). The interaction between sensors and particles causes a particular signature similar to the Gaussian pulse and its derivates, which can be seen in Figure 1.3, depending on the angle of the magnetic particle to the sensor's plane.

These pulses, produced by MNP, can be identified as part of the Gaussian family due to their similarity with the Gaussian curve and its derivatives. The problem that occurs is that the reading of these magnetic signals ends up being very difficult, due to numerous factors, such as the distance from the

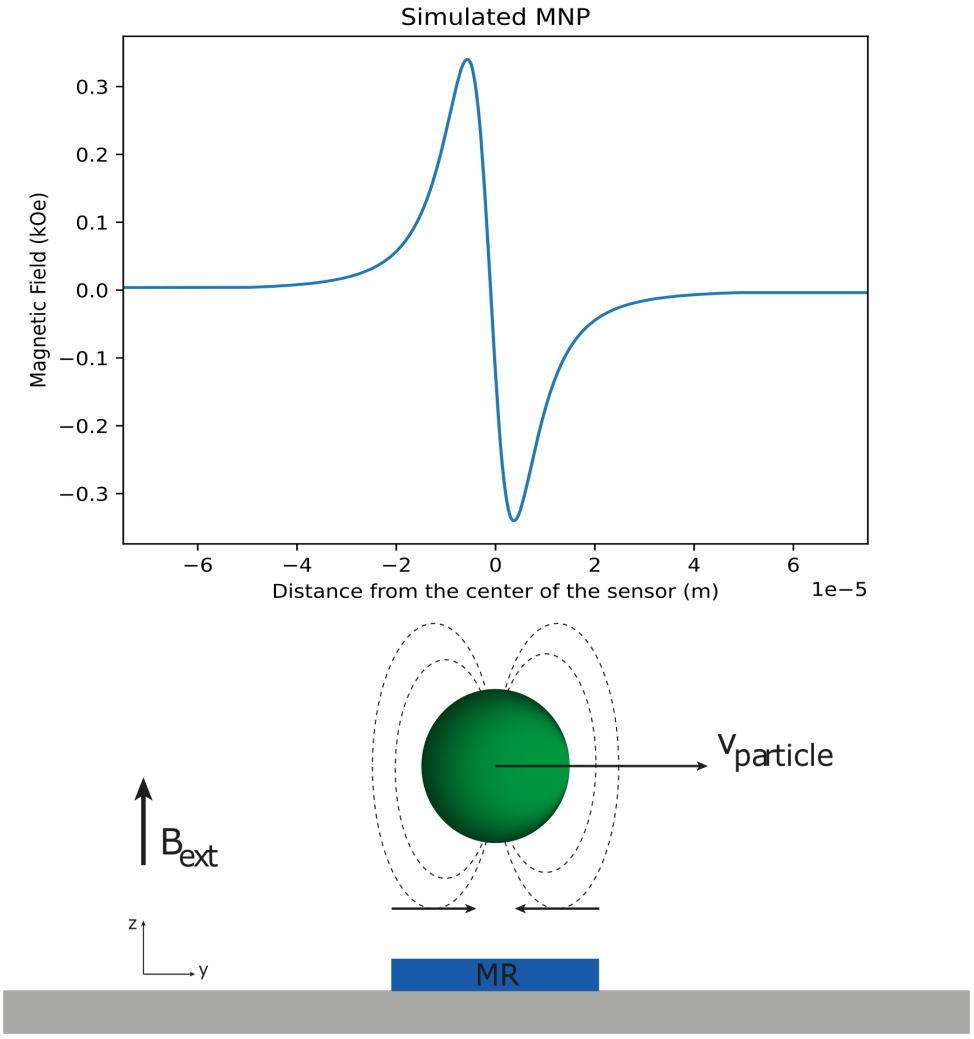

Figure 1.2: Interaction between the MR sensor and a particle that crosses the channel of a cytometer.

particles to the sensor, the signal noise, the speed of the marked cells throws the channel, the sensitivity of the sensors, etc. Thus, sometimes a signal is identified as a false positive for the observation of a magnetic signal.

So, the objective of this work is to develop a ML model, namely, an Artificial Neural Network (ANN), and through the study of some characteristics of the targets (number of particles, distance from sensors, noise signal) that influence the reading results of cytometer sensors, detect the presence of magnetically labelled cells more accurately, and distinguish between marked targets and clusters of free MNP. In the end, and if possible, the objective is also to test this implementation experimentally using cancer cells.

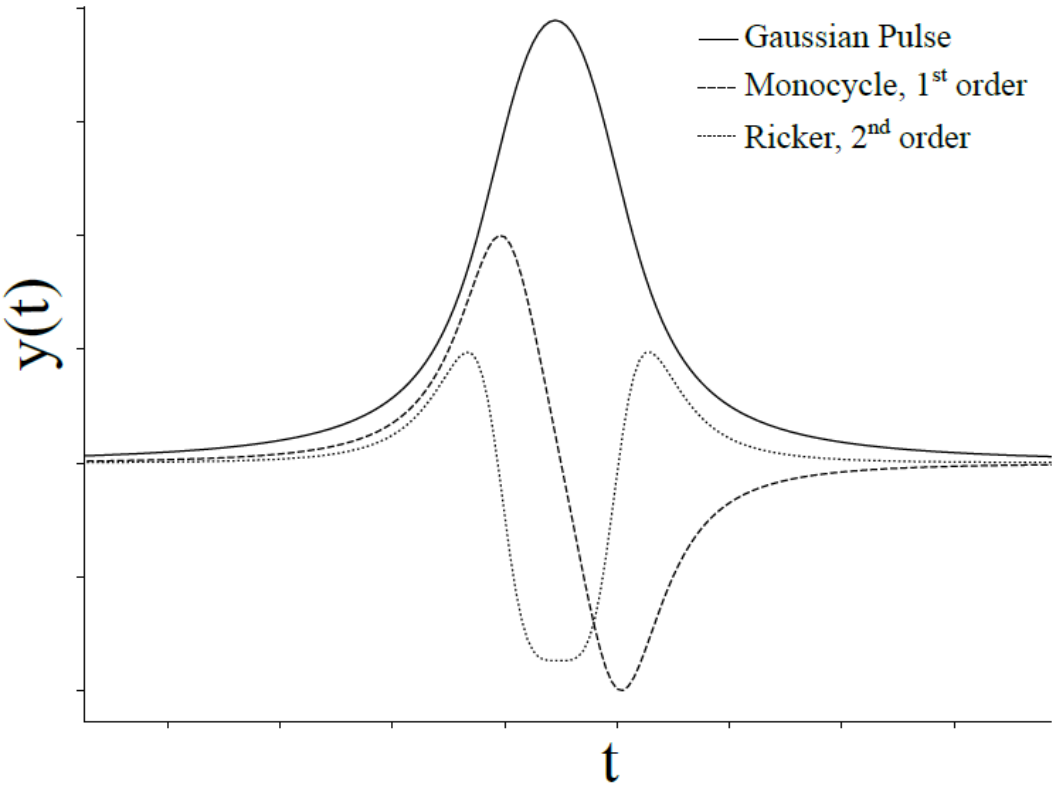

Figure 1.3: Illustration of the signature of the Gaussian pulse and its derivates.

1.3 Organization of the Document

This document reports the research and work developed to simulate an MNP in a microfluidic channel of a cytometer in many ways, and mainly uses the data of these simulations to implement an ANN capable to distinguish between a cell or a cluster of free MNP.

Chapter 2 begins by explaining what is FC in general and MFC in particular and how it works. Then some works already carried out in the same scope of this project are shown, namely classifiers used to distinguish cells in a cytometer. At the end of the chapter, an analysis of the resulting signals from magnetic sensors and how this study can help in the development and implementation of the MNP's simulations is presented.

Chapter 3 gives an introduction to what machine learning is, to better understand its role in pattern recognition, how it works and explains how it is possible to use ANN as a method to differentiate the signals from the magnetic sensors.

In chapter 4 the performed simulations are presented, namely, the simulation of a single MNP and then the combination of them that form cells and clusters, and the explanation of how they were implemented.

Chapter 5 shows how the ANN was implemented and the results obtained.

Finally, in 6, a few conclusions about this work are presented and the main achievements are revisited. Some ideas are also proposed to continue the work.

2

Magnetic Flow Cytometer

Contents

2.1 Introduction to Flow Cytometry

FC is a laboratory technique used to count, detect, identify, and characterize cells in a solution. This method allows the analysis of physical and/or chemical characteristics of the target analytes such as size, volume, cycle, and count, thus allowing researchers to get highly specific information about individual cells. Back in the 1950s, FC was first used to measure the volume of cells in a rapidly flowing fluid stream, they passed in front of a viewing aperture. Over time, the advancement of new technologies and the study of innovations by researchers and engineers have culminated in the modern FC, which is able to make measurements of cells in solution as they pass by the instrument's laser at rates of 10,000 cells per second. In Optical Flow Cytometrys (OFCs), the fluorescent intensity produced by fluorescent-labelled antibodies is measured (using lasers and photodetectors) to detect proteins, or ligands that bind to specific cell-associated molecules, and thus identify certain characteristics unique to various cell types. It has been reported that fluorochromes suffer from several limitations related to background interference, fluorescence noise, stability, and lifetime, which limit the performance of these types of flow cytometers. Also, conventional FC methods have limitations concerning the requirements for preparations of homogeneous cell suspensions, complex optics, lasers, and photodetectors. These limitations affect the speed of response, encouraging the development of new FC methods.

2.2 Working Principle

Interest in the development of point-of-care FC using magnetic sensors has been increasing thanks to its advantages, such as high sensitivity, rapid response, and the use of a small volume of reagents required per assay in handy devices. In Magnetic Flow Cytometry, the conventional and complex fluorochromes and photodetectors are replaced by magnetic particles (for labeling) and magnetic sensors (as the detectors) [2]. A flow cytometer typically comprehends some elements. One of them is an excitation source, like laser light, a magnetic field, an electric field, or radiofrequency waves). In this case of the magnetic FC, the fringe field is generated by magnetic particles. The excitation source can be any appropriate device that produces a known magnetic field, like a permanent magnet, that is usually placed below the sensors. In this case, the sensor used is a MR (e.g., Giant Magnetoresistances Sensors (GMR), which are the most common sensors in MFCs).

A MFC is used to count a specific analyte in which a sample is present. In common samples, the target analyte does not usually have a magnetic content. Thus, the analyte must manifest a specific marker on its surface to serve as an attachment for a probe that can be magnetically functionalized. These probes can be antibodies, viruses, or any particle with an affinity to the manifested marker on the target analyte. A MNP can be conjugated with a probe of choice for them to bond with the target analyte. Functionalized probes can be attached to analytes, providing them a magnetic nature.

In Figure 2.1 a), one can see a representative schematic of the labeling process. It should result in a homogeneous cover of the antigens' surface, like in Figure 2.1 b)), in which one can see fluorophores labeling a large portion of the bacteria's surface. Alternatively, it may also result in less efficient coverage of the antigen, as seen in Figure 2.1 c), in which several MNPs are aglomerated in specific regions, but most bacteria are not well covered. That's why a high labeling efficiency is desirable. After this process of functionalization, MNPs can be used to perform the assay. Figure 2.2 represents how the process of preparation of the sample is carried out before the sample is introduced in the MFC.

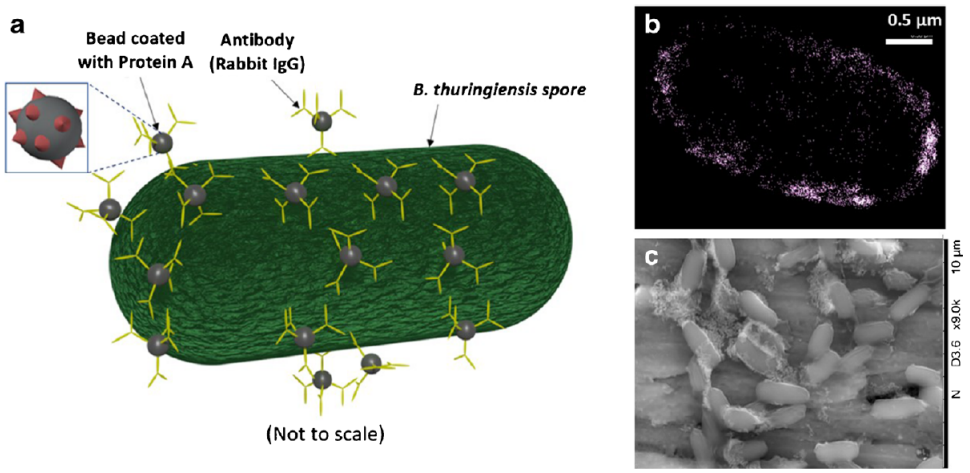

Figure 2.1: Schematic of functionalized particles attached to bacteria from [3] a) bacterial spore labelled with nanoparticles/antibody conjugate (protein A and antibody's Fc region. b) STORM (Stochastic Optical Reconstruction Microscopy) image (x 50,000 magnification); Immunofluorescence image from Bacillus cereus reacted with rabbit anti-B. cereus antibody and goat anti-rabbit secondary antibody labelled with Alexa 647, presenting high labeling efficiency. c) SEM picture of B. thuringiensis spores labelled with magnetic nanoparticles, presenting low labeling efficiency ($\times$ 20,000 magnification)

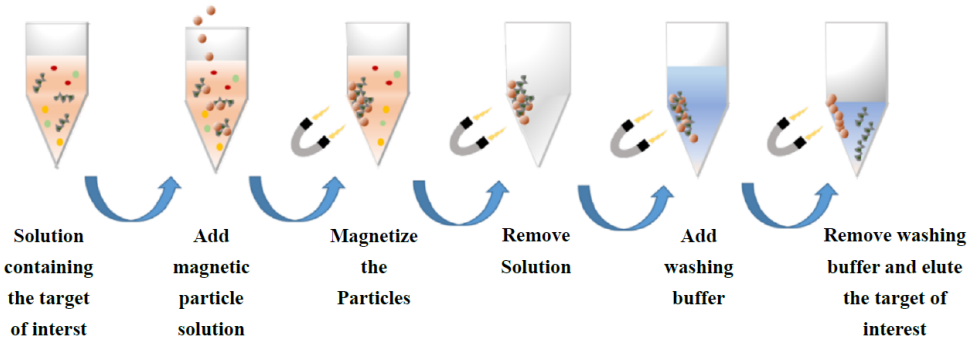

Figure 2.2: Representation of the sample preparation process before it is injected into the MFC's microfluidic channel. [4]

The magnetic particles are mixed in a solution containing the targets of interest and will bond to the specific marker after a specific amount of time. Depending on some factors, like efficiency, number of target cells contained in the sample, and others, some or several MNPs may remain free in the sample.

The resulting mixture contains the original solution with the marked targets/cells and free MNP. This final solution can be submitted to a purification process or be directly injected into the magnetic cytometer. Figure 2.3 represents a real picture of cancer cells (with 20 µm diameter) with attached MNPs flowing in the microfluidic channel, and it is possible to see that, as was said before, there are some MNPs that did not attach to the cells and remained free in the channel.

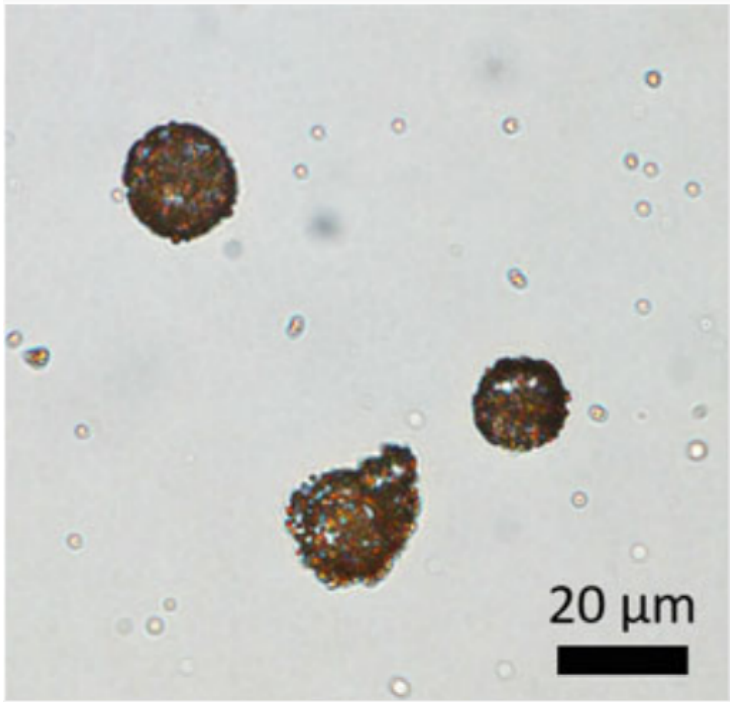

Figure 2.3: Representation of SW480 cancer cells labelled with 1 μm diameter magnetic beads. [5]

After the sample preparation described before, the sample can be introduced in the MFC. The excitation source can be any device that produces a magnetic field such as a permanent magnet and is used to magnetize the MNPs attached to the target and the biological analyte takes on a magnetic signature. When in proximity to the sensor, the magnetic field produced by the particles is gradually picked up by the MR sensors. The interaction between the marked particle/cells and the sensor is shown in Figure 2.4.

The MR sensor's resistance varies with the integrated magnetic field intensity following the Equation 2.1.

$$R(H) = R_{NOM} + \triangle R(H).[\Omega] \tag{2.1}$$

R_{NOM} is the sensor's resistance at zero fields and $\triangle R(H)$ is the field-dependent resistance variation. Depending on the position of the MNP to the sensor's plane (magnetization angle), the vector of the magnetic field could have a different direction so this variation, $\triangle R(H)$, can be positive or negative,

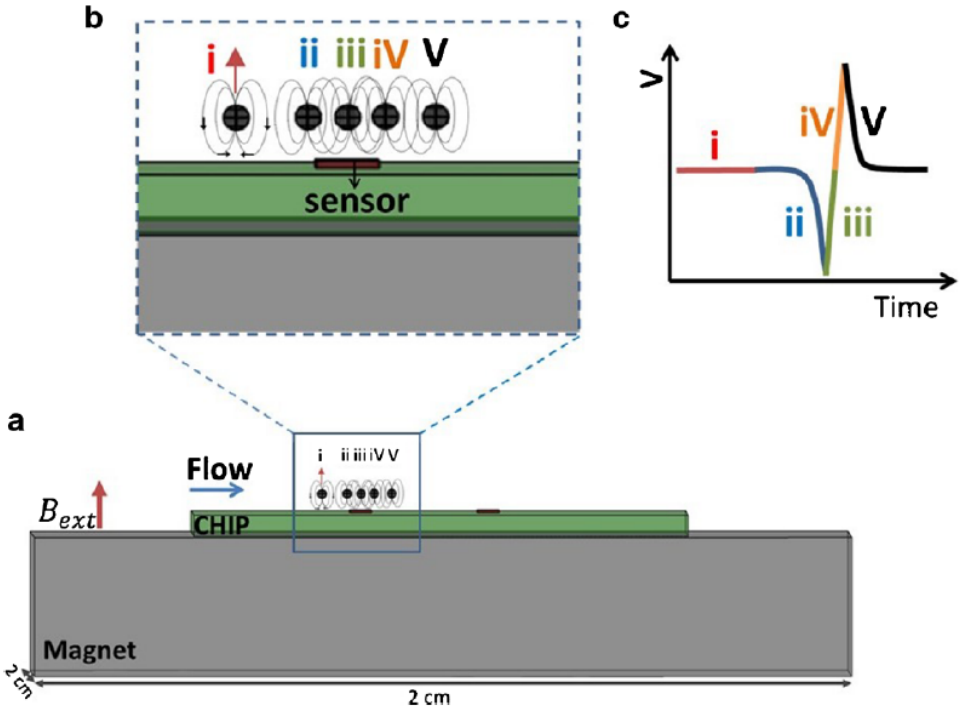

Figure 2.4: Schematic representation of the MFC concept for a magnetic particle detection. [3]

producing different pulses like the monocycle pulse, that is represented in Figure 2.4 or other gaussian pulses (Figure 1.3).

2.3 Magnetic Flow Cytometry Problems

Despite all the advantages of Magnetic FC over conventional methods, there are still some goals to be met to improve the efficiency of this method.

In the last decade, a lot of work and effort has been contributing to the advance in some areas like the definition and optimization of microfluidic channels, the biological processes involved in the binding of MNP/target complexes and the improvement of the magnetic sensors used. However, it is difficult to find some research about signal acquisition and processing in Magnetic FC. Paralleling with the continuous development of electronics and signal processing, the use of MNPs in conjunction with magnetic sensors in microfluidics for the detection of biological targets has made a significant advance. However, it is important to put more effort into developing better sensor acquisition systems, specific for Magnetic FC to improve the overall performance, namely, creating smarter and more complex signal processing algorithms.

In the preparation process described before, which is illustrated in Figure 2.2, the MNPs are mixed in the solution and will bond to a specific marker. As it was said before, depending on some factors,

like efficiency, the number of target cells contained in the sample, and others, some or several MNPs may remain free in the sample, so the resulting mixture contains the original solution with the marked cells and free MNPs. This free MNPs, which did not bind to targets of interest, end up joining, forming clusters. When the sample is injected in MFC, the sample contains, not only the labelled analytes but also clusters so the resulting magnetic field felt by the MR sensors contains the data from both signals. This interference of free magnetic labels on the detection signals of magnetically labelled targets may lead to false positive results. For this reason, it is important to try to differentiate both signals as much as possible. That is the objective of this project, to simulate both signals and try to distinguish them using the data from the simulations to train an ANN as a model of ML.

2.4 Magnetic Flow Cytometry Classifiers - State of Art

There's not a lot of work using machine learning models to detect or differentiate magnetic signals in the application of magnetic flow cytometry, because we're going to be the first to do it and we have a patent for it. However, some simple classifiers have already been implemented that try to achieve this objective here at Instituto de Engenharia de Sistemas e Computadores (INESC), and in the following subsections, I will address how they were developed.

2.4.1 Automatic System to Count and Classify Bacteria [1]

In an attempt to improve the identification of the number of magnetically labelled bacteria associated with each detection moment, a signal analysis method was created that compares the sensor output signal (pulse amplitude vs. pulse width analysis) with simulations, developing a classification method based on an algorithm developed using Matlab. To discriminate the different magnetic signatures, analytical simulations have been performed using a MNP random distribution on the cell surface, considering that overall magnetic moment is given by the combination of the individual moments (m), $9.6 \times 10^{-16} Am^2$, of each MNP. By some studies, the MNP can be approximated to the magnetic dipole [11] [12], that creates a magnetic field (H) at a position of x-axis (a), give by Equation 2.2. Using the integration of the magnetic dipole, it is possible to obtain the fringe field generated by each MNP, given by Equation 2.3.

$$H(a) = \frac{1}{4\pi}\left[\frac{3a \times ma}{|a|^5}\frac{m}{|a|^3}\right] \tag{2.2}$$

$$H(x, y, (a+h)) = \frac{m}{4\pi}\int_{-L/2}^{L/2}\int_{-W/2}^{W/2}\left[\frac{3x(a+h)}{(x^2+y^2+(a+h)^2)^{5/2}}\right]dxdy \tag{2.3}$$

The resulting simulations, represented in Figure 2.5, were performed considering different configurations of bacteria: a single bacteria and aggregates of bacteria, flowing along the x and y axis at different

heights over the sensor.

Comparing the orientation x and y from Figure 2.5, higher voltage amplitude signals were observed for bacteria flowing side by side. On the other hand, longer time length signals were observed for bacteria flowing one after another, as expected. Furthermore, increasing the number of bacteria the higher the voltage amplitude and longer time length signals were obtained, regardless the direction of the movement. Moreover, the higher the distance of the bacteria from the sensor the lower the voltage amplitude and longer the time length.

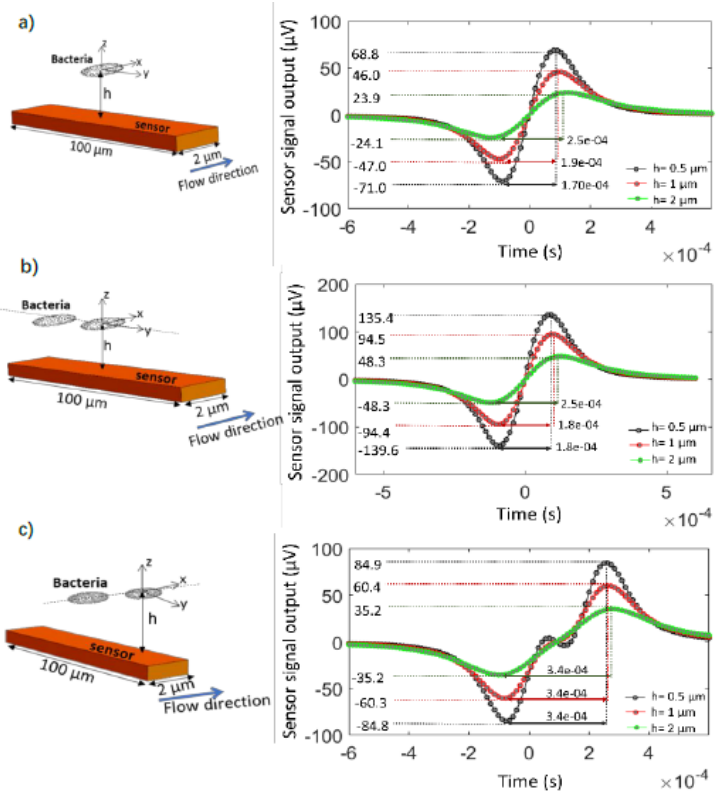

Figure 2.5: Simulated signals for three bacteria configurations at different heights from the sensor: h = 0.5 µm (grey), h = 1 µm (red), h = 2 µm (green).(a) One bacteria; (b) Two bacteria distanced 3 µm from each other, flowing side by side (y axis); (c) Two bacteria distanced 3 µm from each other, flowing one after another along the flow direction (x axis). [1]

Experimental measurements of magnetic labelled Pseudomonas aeruginosa were performed to evaluate the ability to distinguish between different peak signatures according to the number of bacteria detected. Figure 2.6 shows for each number of bacteria simulated (according to the different configurations), the range of amplitudes and time lengths represented by circles, while the real data is represented by blue stars, red hexagons, and black squares. The red hexagons and black squares represent the real signals with good fitting and no fitting with the simulated data, respectively. The real signal matched with the simulations coming from 5 bacteria moving along the sensor's length at 2 µm height.

The comparison of the experimental results with the simulated values confirmed the precision of

the method in the attempt to discriminate the number of magnetically labelled bacteria using the pulse amplitude and pulse width analysis. However, this method does not attempt to differentiate bacteria from free magnetic particles and does not take into account the distance from the x axis between magnetic particles, as their signals are completely aligned (observe Figure 2.5) and in fact, there should be a mismatch between them, as we will see later.

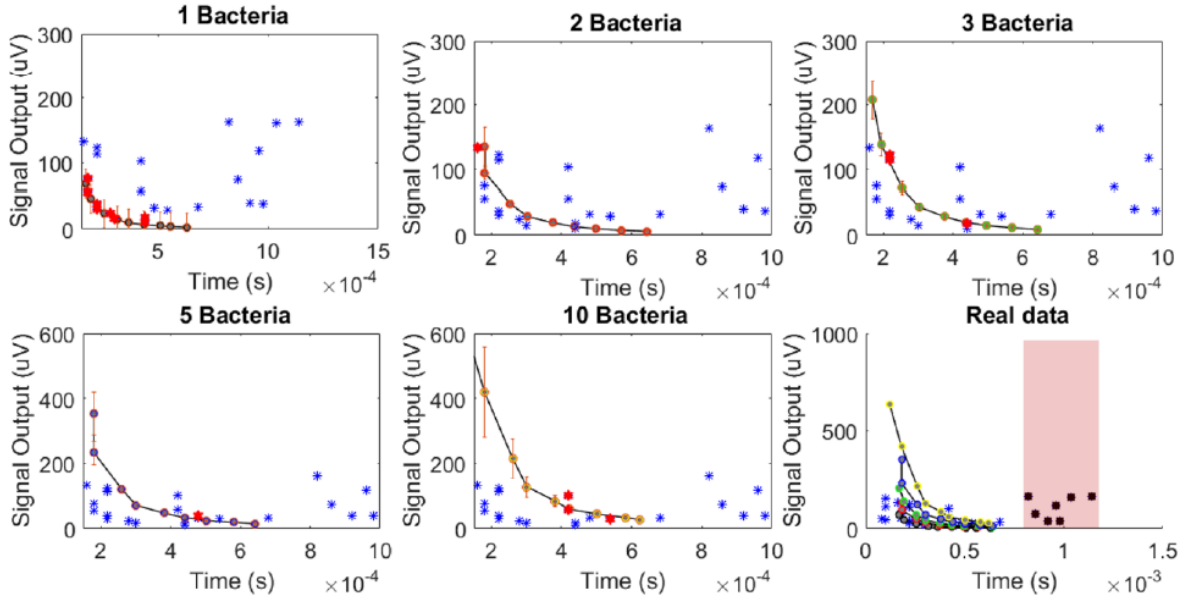

Figure 2.6: Comparison between real signals (represented by blue stars , red hexagons and black squares) and simulations (represented by circles fitted by black lines). [1]

2.4.2 Analytical Strategy [2]

In order to improve the detection system accuracy, reducing the number of false positives, a method to distinguish signatures of free or clustered magnetic particles from magnetically labelled targets was developed. This classification methodology for labelled targets, based on a custom Matlab algorithm, compares the sensor output signals (pulse amplitude vs. pulse width analysis) with simulations.

The amplitude of the peaks is related to the number of magnetic labels and the height from the sensor. On the other hand, the peak width is related to the dimensions of the flowing element over the sensor and is also dependent on the height at which the entities pass from the sensor.

The signal analysis was performed using a custom algorithm developed in Matlab. A threshold is defined according to the noise baseline, and the program will find the bipolar peaks detected above this threshold. The detected peaks are considered signals coming from magnetic entities. Then, full width at half-maximum and half-minimum values are calculated for each peak (Figure 2.7). The simulations of magnetically labelled targets were performed in Wolfram Mathematica, which allows to predict the num-

ber of magnetic particles covering the targets and can be used as an indicator of the functionalization efficiency, (Figure 2.8, dashed lines vs blue points).

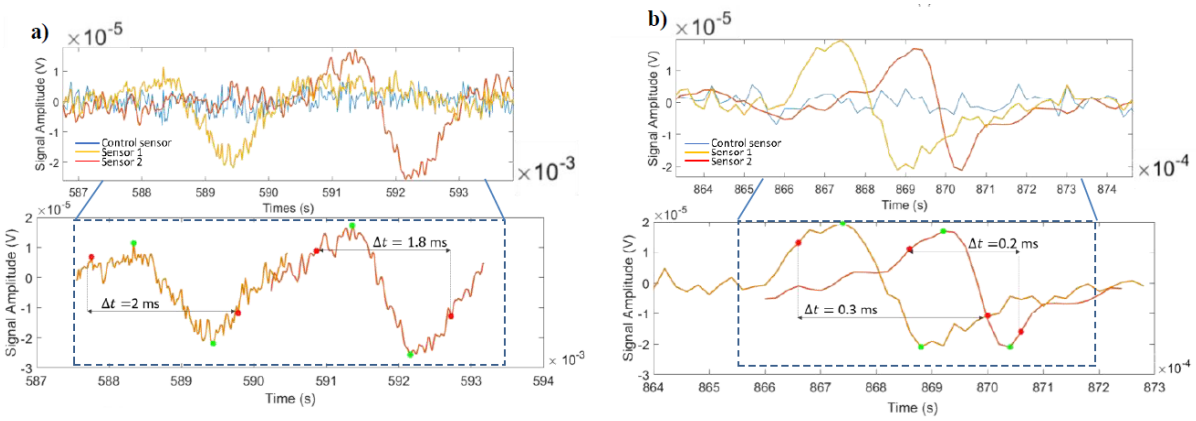

Figure 2.7: Measured data showing a) detection of conjugated MNP with antibody and peak width analysis (green points define the peak amplitude and red points define peak width). (b) detection of 5 μm labelled beads and peak width analysis (green points define the peak amplitude and red points define peak width). [2]

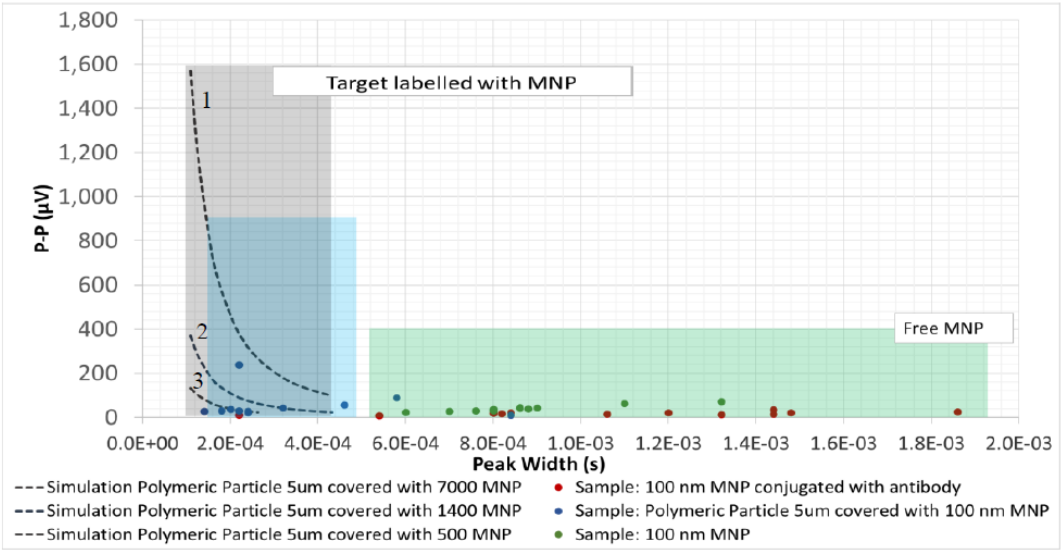

Figure 2.8: Comparison of real time signal output points. The signals coming from labelled targets are represented in blue points. The signals coming from free MNPs are represented in green (individual MNPs) and in red (MNP conjugated with antibody). Simulations are represented by the three black dashed lines, each of one representing a different coverage of MNPs. [2]

From Figure 2.8, it is possible to conclude that the signals obtained from the experiment are in accordance with the values obtained from the simulations, so it was possible the discrimination between signals, using the compilation of the parameters peak amplitude and peak width. The accuracy of the method is confirmed. However, in addition to not simulating the free MNP, this method is a classifier based on only two variables: pulse amplitude and pulse width analysis, they do not take into account

other variables such as the number of particles that are marked (since this number is unknown), the distance between particle (x-axis), and their proximity to the sensor in height (z-axis).

2.5 Signal Analysis

In order to improve the detection and differentiation of signals, namely the distinction of signals from magnetically labelled cells and free MNP, it is necessary to increase the detection rate. We can divide signal processing into two approaches. Increasing the signal quality (Signal-to-Noise Ratio (SNR)), or using the knowledge about the signal (time and frequency domain characteristics) to 'find' it in the middle of the noise.

Specifically, in very noisy or low SNR systems, knowing the signal well is very important for signal processing. Since this is the case in Magnetic FC, before explaining any signal processing or pulse detection, it is beneficial to understand the signals' nature, from a signal processing perspective.

2.5.1 Gaussian pulse and Gaussian Mono-cycle

As stated in section 1.2, an MFC can count cells in a flow if these cells are previously marked with antibodies or other probes carrying MNP. In turn, these marked cells are injected in a microfluidic channel, forcing them to come near the MR sensors, while in a continuous flow. The interaction between the sensors and the particle causes a pulse with a specific signature. These pulses can be approximated to Gaussian pulses, as the ones are shown in Figure 1.3. The first and second-order derivatives are obtained from the original Gaussian pulse. As seen by comparing Figure 1.3 with real experimental data acquired (Figure 2.9) the most relevant signal shape for this work is the Gaussian monocycle. In extreme cases, the signal will also resemble either a positive or negative Gaussian pulse or its second derivative, the Ricker wavelet.

Figure 2.10 shows how the magnetization angle can change the shape of the pulse produced by a magnetic particle. When the magnetic axis of the dipole is perfectly perpendicular to the sensing axis, the resulting signal, which can be seen in Figure 2.10 a), is bipolar and resembles a Gaussian monocycle. On the other hand, when the magnetic axis is parallel to the sensor the resulting signal is unipolar and resembles a Gaussian pulse or Ricker Wavelet, depending on if it is positive or negative, as is shown in Figure 2.10 b).

In this work, it will be assumed that the signatures produced by the MNPs and detected by the MR sensors are similar to the Gaussian monocycle pulse, since, the experiments are performed mainly by making sure the magnetization field produced by the permanent magnet is perpendicular to the sensors sensing axis (Fig. 2.4), using reduced flow rates (to avoid particle rotation). The similarity between the magnetic signal and Gaussian family pulses is important because these impulses are well characterized.

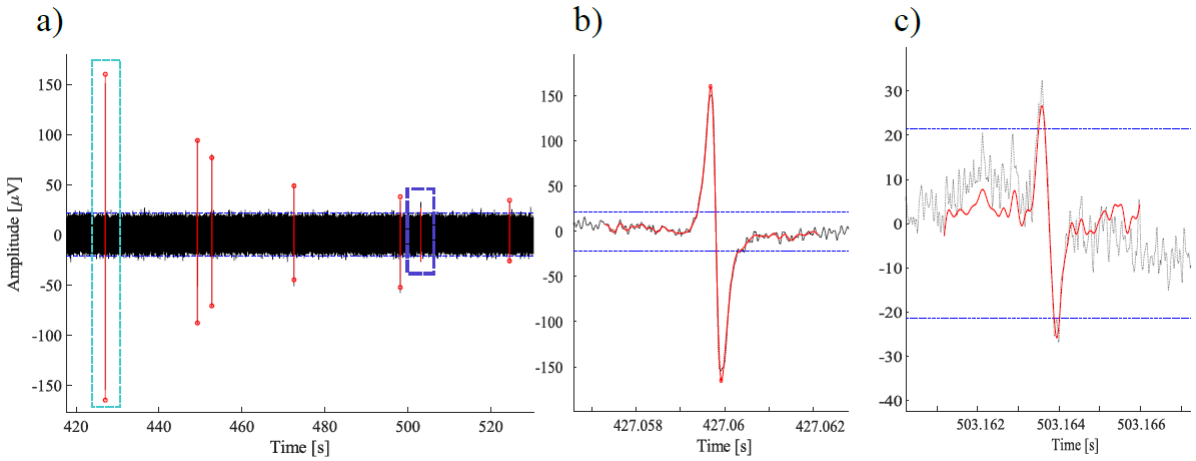

Figure 2.9: Example of experimental data acquired in the MFC. a) clip of full experiment containing seven highlighted bipolar pulses. b) zoom of a pulse with large SNR, crossing upper and lower decision boundaries. c) zoom of a pulse with low SNR, barely crossing the decision boundaries. Both b) and c) show the pulse before (black) and after filtering (red).

The Gaussian pulse (y_p) is expressed in Equation 2.4 and its derivate, the Gaussian monocycle (y_{mc}) is represented in Equation 2.5, where A is the maximum of the impulse, a is the constant that determines the slope of the pulse and t is time [13].

$$y_p(t) = Ae^{-a^2t^2} \tag{2.4}$$

$$y_{mc(t)} = -2a^2tAe^{-a^2t^2} \tag{2.5}$$

By taking the Fourier transform of Equation 2.5, the spectral response of the Gaussian monocycle described in Equation 2.6 is obtained with a maximum at a frequency f_0 given by Equation 2.7.

$$Y_{mc}(w) = \frac{iwA}{a\sqrt{2}}e^{-\frac{w^2}{4a^2}} \tag{2.6}$$

$$f_0 = \frac{a^2\sqrt{2}}{2\pi} \tag{2.7}$$

The 3-dB bandwidth between the two frequencies lower and higher than f_0 (f_L and f_H) for which the amplitude pulse is $1/\sqrt{2}$ of the maximum can be derived by Equation 2.7 where $T_p = 1/f_0$ is the pulse duration. Equation 2.8 demonstrates that the 3-dB (half power) bandwidth of a Gaussian pulse is approximately 115% of the center frequency between f_h (2.9) and f_l (2.10). Thus, to determine the effective bandwidth of the signal in MFC one must consider the maximum and minimum particle speed.

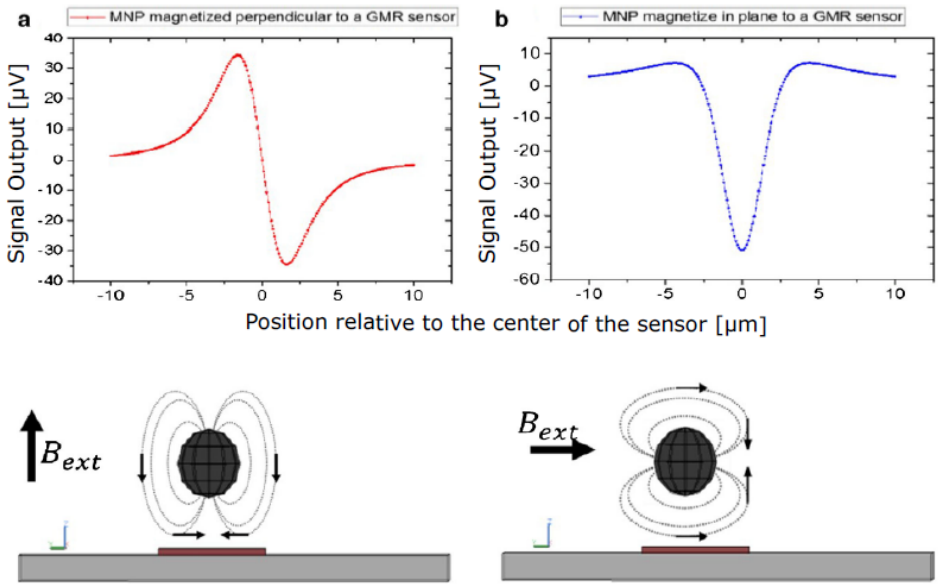

Figure 2.10: Schematic representation of signal output for an MR sensor when applying a) a perpendicular or b) an in-plane excitation field for a particle of 2.8 µm of diameter (Dynabeas M-280 streptavidin) at 2 µm [3].

$$\triangle f = f_h - f_l = \frac{1.155(a\sqrt{2})}{2\pi} = 1.155 f_0 = 1.155/T_p \tag{2.8}$$

$$f_0 + \left(\frac{1.155 f_0}{2}\right) \leq f_h \tag{2.9}$$

$$f_l \leq f_0 - \left(\frac{1.155 f_0}{2}\right) \tag{2.10}$$

Assuming that the signal's power is symmetrically distributed around f_0 and considering the maximum and minimum pulse duration, solving (2.9) and (2.10), the fasted and slowest particles produce signals with the following -3dB bandwidths: 3.3 kHz, from 2.54 kHz to 5.25 kHz, and 372 Hz, from 137 Hz to 510 Hz.

2.5.2 Signal Strength

The signal strength depends on two factors of the magnetic target, how saturated it is, and its proximity to the sensor. The first depends on the type of MNP, how good the binding process to the target is, and on the strength of the magnetic field produced by the biasing element. The sensor's distance depends on the particle's position inside the flow, limited by the channel's height.

It is possible to count the exact number of events if magnetic simulations are performed in perfect conditions (e.g no target superposition), to establish the maximum height for the channel. In Magnetic

FC, with small targets like cells and bacteria, the field intensity is maximized at the sensor's point closer to the source, and it decays rapidly as the distance increase. Thus, the size of the sensor relative to the size of the target is crucial.

The point-like behavior of an Micro-Metric Magnetic Particles (MMP) or MNP can be modelled as a magnetic dipole [11] [14], if it is assumed that the particle is spherical, and the dipole center is at the geometrical center of the bead. Considering this, D. M. Caetano performed a simulation in his work [6], that is represented in Figure 2.11, using the magnetic data from the Dynal M280 beads that carry a magnetic moment of a 1.86×10^{-13} Oe in a sphere with a diameter of a 2.8 µm.

The simulation is based on integrating the magnetic field produced by a magnetic dipole on the MR sensor's surface area, predicting the signal's amplitude as the dipole/particle crosses the sensor while flowing at a fixed height and magnetization angle. The simulation results contemplating combinations of different heights and angles generate most of the signal shapes that the sensor produces.

Equation 2.11 represents the integral of the field produced by a magnetic dipole in the sensitive axis (x) of the sensor used in this simulation. x_0 and y_0 represent the center of the sensor, l and w represent the length and width of the sensor, θ represents the magnetization angle, and x and y represent the position of the center of the dipole relative to the center of the sensor.

$$Hx_{avg} = \frac{\left(\int_{-\frac{1}{2}-y_0}^{\frac{1}{2}-y_0} \int_{-\frac{w}{2}x_0}^{\frac{w}{2}x_0} \frac{M}{4\pi} \frac{4\pi}{1000} \times \frac{3\cos\theta(x+h)+\sin\theta x^2}{(x^2+y^2+h^2)^{5/2}} - \frac{\sin\theta}{(x^2+y^2+h^2)^{3/2}} dxdy\right)}{w \times l} [Oe] \tag{2.11}$$

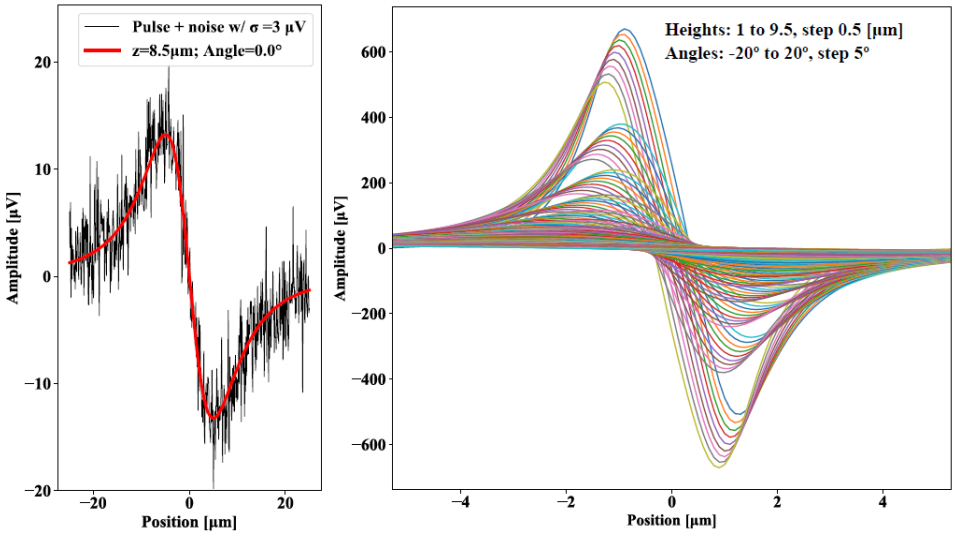

Figure 2.11: Simulation of dipole/sensor interaction for different height and magnetization angles using Equation 2.11. a) The red line represents the specific case of the largest distance possible for a 2.8 µm particle with the magnetization field perpendicular to the sensor. b) representation of the complete 180 simulations for all the different heights and angles [6].

The simulation of the dipole to the sensor for a 2.8 µm particle is represented in Figure 2.11. The

amplitude in volts is computed by assuming a GMR sensor with a sensitivity of 0.4 Ohms per Oersted, biased with a constant current of 1 mA. In Figure 2.11 a) it is possible to see the simulation for a particle at 8.5 µm and perpendicular magnetization angle producing a signature close to what is a Gaussian monocycle, as mentioned before. In Figure 2.11 b) is represented the simulations for a particle traveling in the center of the channel with the height's parametric sweep range from 1 to 10 µm, and magnetization angles between -20 and 20 degrees, relative to the plane perpendicular to the channel's base, and direction of flow.

In this thesis, I will use the same equation, ignoring the magnetization angle since, as it was said before in subsection 2.5.1, it will be assumed that the magnetization field produced by the permanent magnet is perpendicular to the sensor's sensing axis, so the magnetization angle can be ignored. Instead of doing these simulations with just one magnetic particle, these simulations will be done with numerous particles, representing two main types of signals: marked cells (with multiple MNPs around it) and clusters (set of free MNPs grouped). This simulation with several particles will be at different distances from the center of the sensor (according to z and x) resulting in signatures superimposed on each other. This will give us information about how the different kinds of particles interact with the sensor and this information will be our data to train the artificial neural network that will recognize patterns and mainly distinguish the signal between a marked cell and a cluster of free MNPs, as will be explained in the next chapter.

3

Introduction to Machine Learning

Contents

3.1 Machine Learning

Machine Learning (ML) are computer-based methods to perform a specific task that can be improved through experience, i.e., observing the many different events of the phenomenon under study [15] and that can be considered part of artificial intelligence. These ML's algorithms are built from a "training data-base", that uses this information to train and predict decisions.

ML is, in part, based on a model of brain cell interaction. The model was created in 1949 by Donald Hebb in a book titled "The Organization of Behavior" [16]. The book presents Hebb's theories on neuron excitement and communication between neurons. Some years later, in 1959, the term machine learning was coined by Arthur Samuel (IBM) with his research "Some Studies in Machine Learning Using the Game of Checkers" where he used a computer to test if a computer using machine learning could win a game of checkers versus a human being. Robert Nealey, the self-proclaimed checkers master, played the game on an IBM 7094 computer in 1962, and he lost to the computer [17].

Currently, ML models are used in internet search engines, email filters to sort out spam, websites to make personalized recommendations, banking software to detect unusual transactions, and lots of phone applications such as voice recognition. First, it is necessary to create the data-set, that is the data that will be used by ML's model to train and predict the results. There are some models that have been used and researched. Some examples are decision trees and support vector machines, but the model that will be used in this work is an Artificial Neural Network (ANN). This model is used in many applications, like pattern recognition, because of its ability to reproduce and model nonlinear processes. ANN is also known for creating very flexible and precise classifiers and producing universal regressors.

3.1.1 Training Data-Set

Every ML's model needs a data-set, it is this data that is used to train the model to predict results. A data set is a collection of data, or in other words, consists of the sample output data and the corresponding sets of input data that have an influence on the output. In fact, defining the data that the ML's model will use is not such a simple task. It is necessary to define well what the problem is and what results we want (outputs), and then what observations (inputs) are needed. Then, it is necessary to frame the data at a scale that the ML's model can train. For example, if the objective is to understand which shades of clothing are in fashion and if color of a sweater is one of the inputs, the color must be defined as a number.

For this project there are three main objectives:

- Identifying magnetic signatures in the noisy signal.
- Determining event characteristics, like, target speed/time-of-flight, volume, distance from sensors, magnetization angle, etc;

- Differentiating between marked targets and clusters of free MNP;

The first two points were already worked by D. M. Caetano in his thesis [6]. My objective is to continue that work but mainly implement the third point.

3.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains. This definition reminds us of the model created by Donald Hebb in 1949, mentioned earlier, which said that machine learning is, in part, based on a model of interaction between brain cells.

3.2.1 The neuron

An ANN is based on the connection of units or nodes called neurons. The neuron is the basic building block of an ANN, and as a biological neuron, the artificial neuron is made up of several inputs and an output as seen in Figure 3.1.

According to the human brain, the inputs could be the five senses, vision, touch, or smell, it is basically the information that the brain gets from the senses. In terms of ML models, these are the input values, the independent variables. The inputs of the neuron can be only data or can be the output of another neuron (forming a connection of neurons – ANN). Each input corresponds to an observation. For example, if we were to access the purchasing power of a bank customer, one possible input would be the salary of each customer.

There are weights that are assigned to each input and correspond to the degree of importance that that observation has for the neuron. The output result is calculated by an internal function (activation function) which is the sum of the inputs according to their weights. When the artificial neural network is trained, all the weights are adjusted, that's the reason why weights are so crucial to the ANN functioning. A single neuron can be seen as a linear classifier and it is by itself a powerful tool that can solve binary classification problems in which the classes are linearly separable. However, when several neurons are connected, (Figure 3.2), impressive results can be achieved.

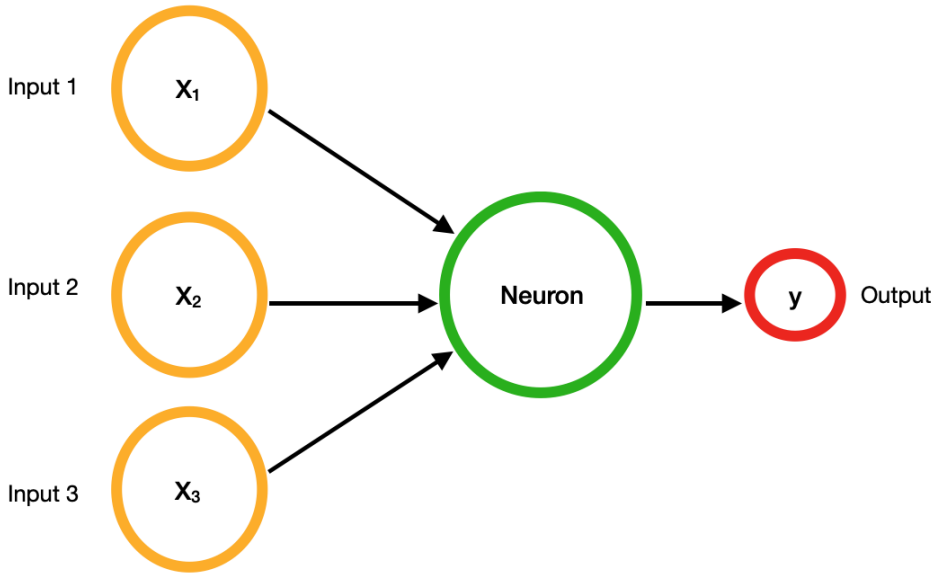

Figure 3.1: Representation of a single neuron.

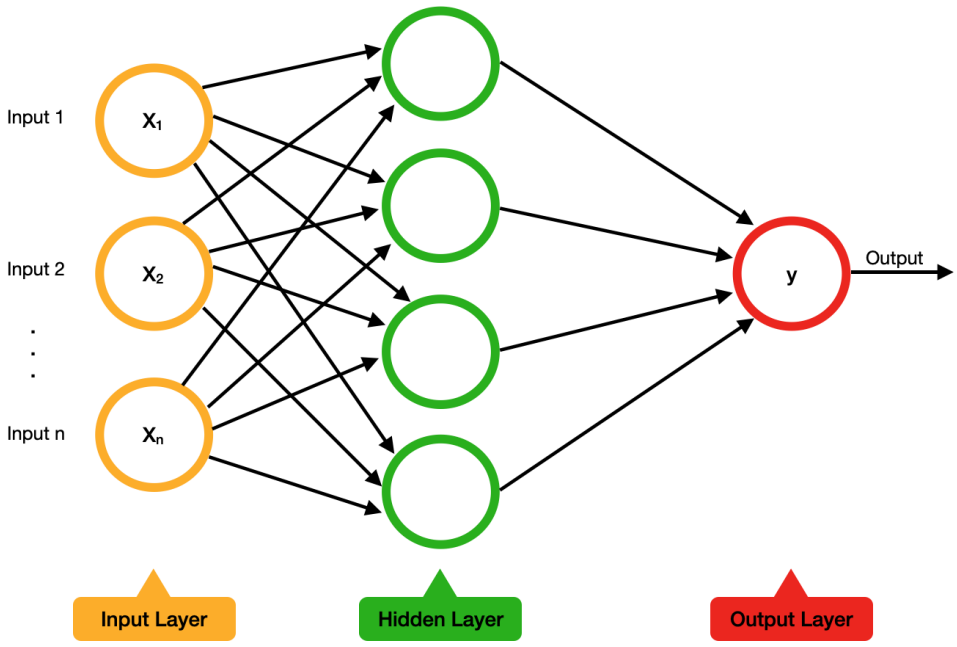

Figure 3.2: Representation of a neural network, showing the connections between each layer.

3.2.2 Activation Functions

As mentioned earlier, the activation function is the function that is applied by the neuron using their inputs and weights to calculate the output. There are many activation functions, but three of the most

common are the sigmoid (Eq. 3.1), the hyperbolic tangent (Eq. 3.2), and the rectifier linear unit – ReLu (Eq. 3.3) [18]. Figure 3.3 represents the shape of these activation functions.

$$\phi(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

$$\phi(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{3.2}$$

$$\phi(x) = max(x, 0) \tag{3.3}$$

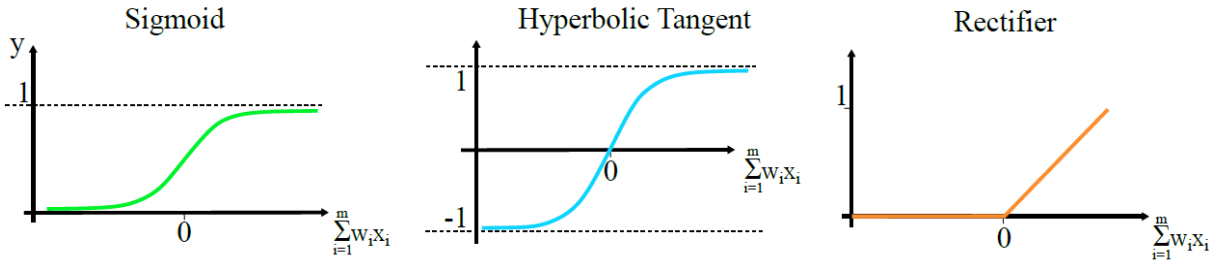

Figure 3.3: Common activation functions: sigmoid, hyperbolic tangent, rectifier or ReLu [6]

The sigmoid and hyperbolic tangent activation functions are very well defined mathematically. The same cannot be said of the rectifier linear unit. This function computes the maximum between zero and the value of x, so if the weighted sum of all the inputs is larger than zero, then the result is linearly proportional to the input. However, if the weighted sum of all the inputs and offsets results is a negative number, the neuron outputs zero. This ends up bringing simplicity to the function, furthermore, ReLu function has a strong biological motivation, and it has been demonstrated to enable better training of deeper networks than other activate functions [18]. For that reason, in the work we are developing here will use the ReLu function as an activation function in the hidden layer of the artificial neural network.

3.2.3 ANN training and Back Propagation

As was said before, the ANN uses the input values, that correspond to the observations to calculate the output prediction, using an activation function. In fact, the output prediction value ($\hat{y}$) is not necessarily equal to the value of the actual output (y), so there is a difference between these two values (even if it is small). What the ANN does is calculate the 'error' between these values based on a formula called the cost function. There are very formulas for the cost function, but the most common is represented in Equation 3.4. The goal is to minimize the cost function because the lower the cost function, because the lower the cost function, the close the prediction output value ($\hat{y}$) is to the real value (y).

$$C = \frac{1}{2}(\hat{y} - y)^2 \tag{3.4}$$

After this process, the information with this 'error' is propagated back into the neural network and the weights, mentioned before, are updated, and the cycle is repeated, decreasing the cost function, thus bringing the predicted value closer to the actual value. When the whole training set passed through the ANN, that makes an epoch. An epoch is a term used in ML and indicates the number of passes of the entire training data-set the ML algorithm has completed. Later on, I will address in more detail about how the ANN was implemented in this work, namely, how the training data-set and the network structure were defined.

4

Magnetic Particles Simulations

Contents

4.1 MNP Simulation

As mentioned before, in subsection 2.5.2, since it is possible to model the signature of an MNP using the equation of a magnetic dipole (Equation 2.11) if it is assumed that the particle is spherical, and the dipole center is at the geometrical center of the bead, a simulation was performed to determine the magnetic field, measured by a GMR sensor, through its position (in x, y, z in relation to the sensor) of a single MNP. The simulation, that is represented in Figure 4.1 a), represents the magnetic field measured at the output of the entire system (after the signal has been measured by the sensor and been amplified), in µV, produced by an MNP of 2 µm of diameter at 5 µm of height (distance from the particle relative to the plane perpendicular to the channel's base). It is possible to see that the simulation of a particle produces a signature close to what a Gaussian monocycle pulse is (Fig. 1.3), as mentioned before. Figure 4.1 b) represents the simulations for a particle traveling in the center of the channel with the height's parametric sweep range from 3 to 15 µm, relative to the plane perpendicular to the channel's base, and to direction of flow.

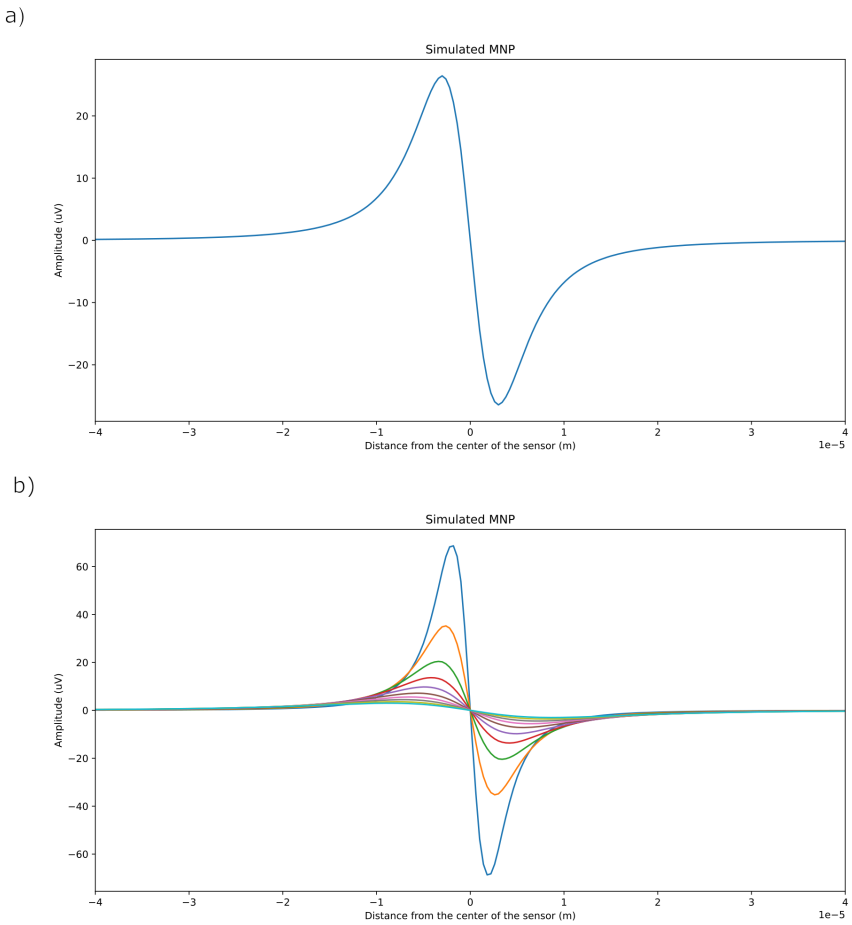

Figure 4.1: Representation of the simulation of a single MNP: a) 1 MNP at 5 µm of height; b) 1 MNP at 10 different heights from 3 µm to 15 µm.

4.2 Cell Simulation

Based on the confirmation that the dipole simulation can be used to describe the comportment of an MNP, it is now possible to simulate the MNP's around a cell. Through Figure 4.2, it is possible to observe the result of the simulation that uses a sphere to represent a cell with 10 µm of diameter, and an algorithm was performed that randomly distributes points along the spherical surface (blue dots), thus representing the MNPs distributed around the cell.

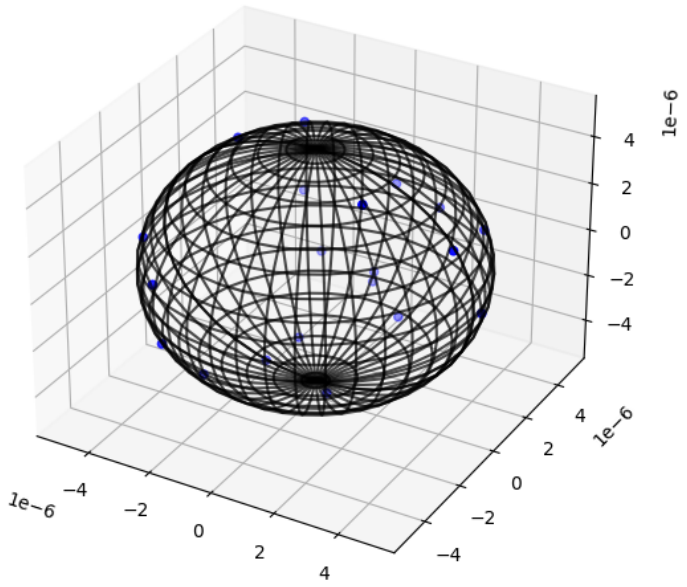

Figure 4.2: Representation of a cell with 10 µm of diameter with 20 particles (blue dots) on its spherical surface randomly distributed.

Using this random distribution of particles around the cell, it is possible to use it combined with the MNP simulation to simulate the output signal produced by the particles when they are attached to the cell. Each particle will have a different position (height and width) in relation to the center of the sensor, so the final result will be the sum of the magnetic field produced by all particles.

Figure 4.3 shows the simulation of 15 particles randomly arranged on the spherical surface of a cell, as well as the signal resulting from the sum of all particles. The cell has 10 μm of diameter and is located at a height (z) of 6 μm. It is possible to observe that the MNPs are at different distances from the center of the sensor, to the x-axis, and as they are also located at different heights to the z-axis, the MNPs that are closer to the sensor have a bigger magnetic field signal than those that are further away.

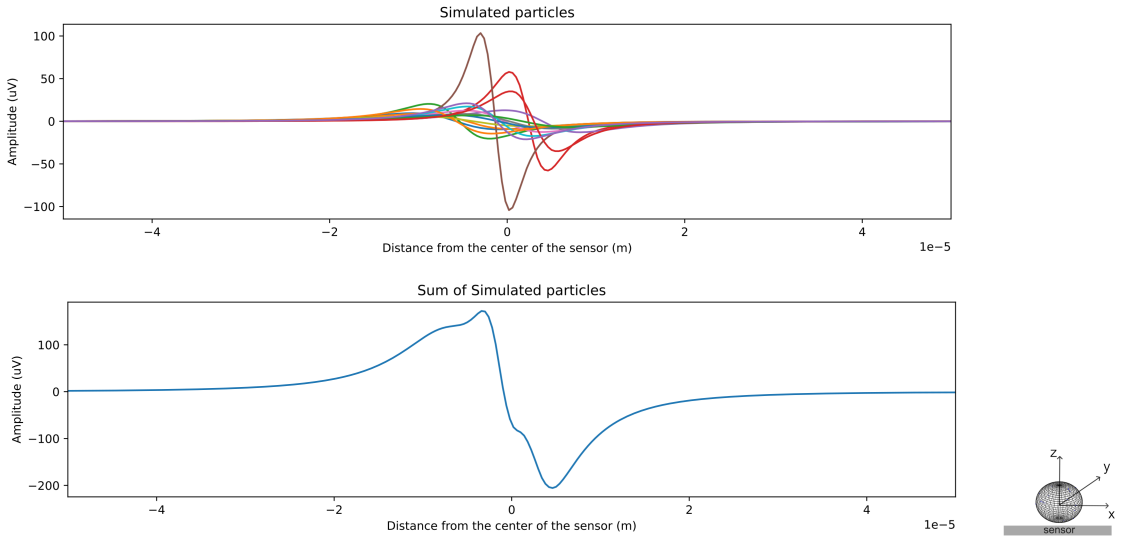

Figure 4.3: Simulation of 15 MNPs around a cell with 10 µm diameter, located at 6 µm of height. Above, the representation of each particle, below, the representation of the sum of the simulated particles.

4.3 Cluster Simulation

Clusters are sets of MNPs that never came together with the analytes/cells present in the samples, and that, therefore, end up joining each other. To simulate clusters, it is necessary to define how they are grouped. Assuming that all MNPs are spheres, it is possible to use the packing of spheres as a grouping method. In geometry, a sphere packing is an arrangement of non-overlapping spheres within a containing space. The spheres considered are usually all identical in size, and the space is usually three-dimensional Euclidean space. There are two main methods of sphere packing, square packing, and the hexagonal packing. Both are represented in Figure 4.4.

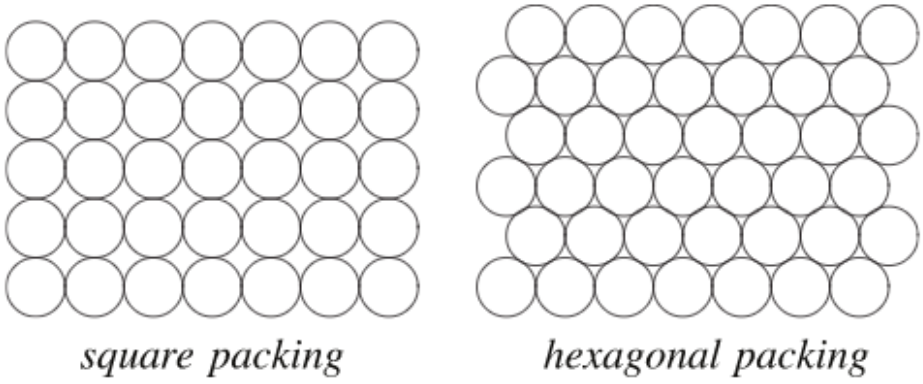

Figure 4.4: Square and hexagonal packing representation for spheres (MNPs).

The first method used to simulate clusters was the square packing, because, although it is a method that uses more free space than hexagonal packing, it is easier to implement, since if we define the main sphere/MNP, with 2 μm of diameter, placed in a certain position, all the others are situated at a distance of 2 μm in all x, y, z directions from the main one.

In Figure 4.5 is represented the simulation of a cluster with 15 MNPs using the square packing.

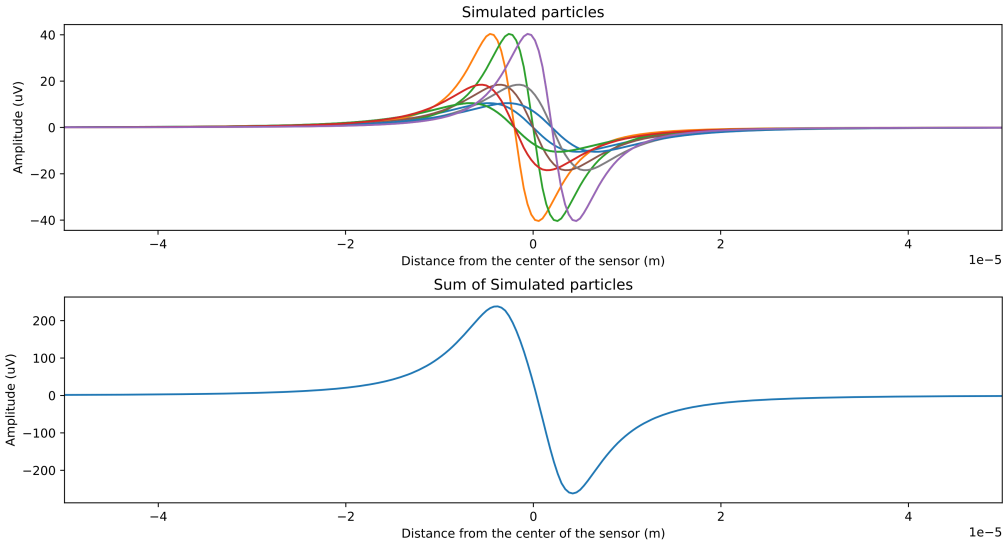

Figure 4.5: Cluster Simulation for 15 particles using square packing.

From this simulation, it is possible to conclude that, unlike cell simulation, in cluster simulation, the MNPs, from the point of view of the x-axis, are more close to each other and they have all spaced the same distance, which in this case is 2 μm which corresponds to the diameter size of the particle. This happens because in this case there are no cells between the particles, and therefore they are closer to each other. For this reason, it is natural that the signals from the cluster simulation have a greater amplitude because as the distance between the particles is smaller, the sensor detects a higher magnetic field when the particles pass through it.

4.4 Cell vs Cluster Simulation

One way to compare and differentiate the signals from the simulation of cells and clusters is to observe the two signals simultaneously when subjected to the same conditions of particle numbers and heights in relation to the sensor's plane. The following Figure 4.6 shows the simulated signals of cells and clusters for the same number of particles (15 MNP) with a height (from the center of the cell/cluster to the plane of the sensor) relatively close to the sensor (6 μm).

From this Figure 4.6 it is possible to conclude that the magnetic field signal coming from the cluster

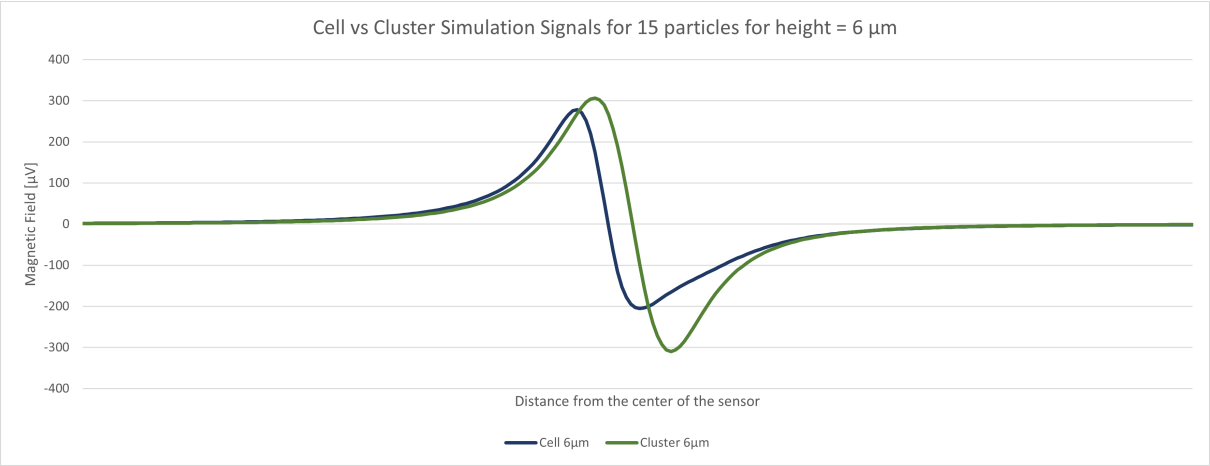

Figure 4.6: Cell vs Cluster Simulation signals for 15 MNPs located at 6 µm of height to the plane of the sensor.

simulation (green signal) has a greater amplitude in comparison to the cell simulation (blue signal). To compare both signals with a different perspective, Figure 4.7 represents the same signals but for a distance further from the sensor plane (10 μm).

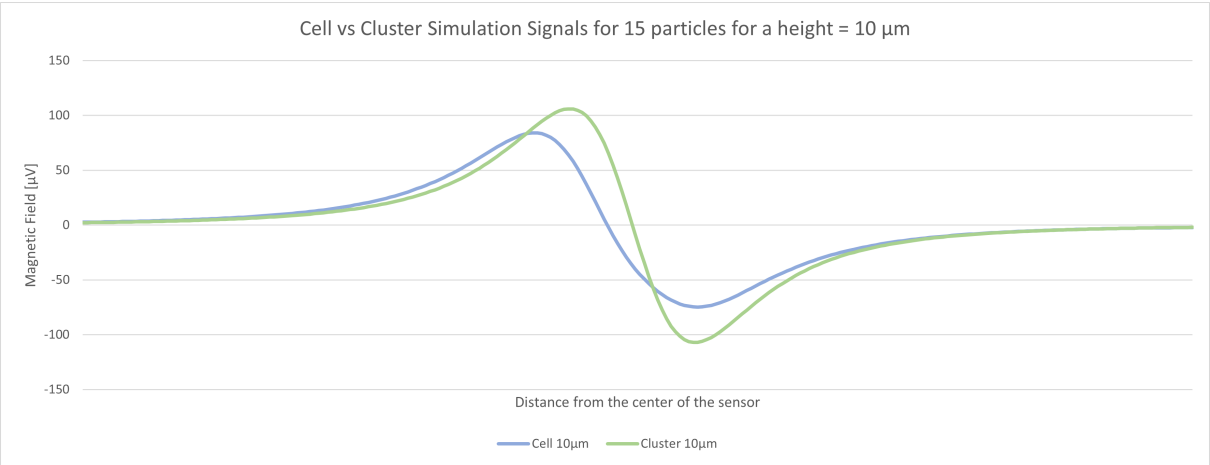

Figure 4.7: Cell vs Cluster Simulation signals for 15 MNPs located at 10 µm of height to the plane of the sensor.

It is possible to observe that the smaller the distance in height of the particles to the sensor's plane, the greater the amplitude of the signal. The reason is because the magnetic field produced by the particles, felt by the sensor, is stronger, once the magnetic particle is closer to the center of the sensor. It is possible to conclude that the simulations seem to correctly demonstrate what happens in reality.

4.5 Adding Noise to simulations

Although the simulation results seem acceptable and similar to what is expected from the real signals, the absence of the noise signal in the simulation can be a negative point for the artificial neural network's results, because the real signal resulting from the interaction of the magnetic particles with the MR sensors in the cytometer contains noise. The simulated results shown here have no noise, which makes it very simple for the neural network to distinguish the signals between cell and cluster, so the ANNs results would be inconclusive, as they would not translate to real data.

Thus, to add noise to the simulation, a value obtained randomly through a normal distribution was added to the original signal. A normal distribution (also known as Gaussian, Gauss, or Laplace–Gauss distribution) is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is described in Equation 4.1 where μ is the mean or expectation of the distribution and σ is it standard deviation. For this case, it can be assumed that μ is zero and σ will be the parameter that defines the amplitude of the noise signal in μV.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \tag{4.1}$$

Figures 4.8 and 4.9 represent the simulation for 1 MNP located at 5 μm of height in relation to the center of the sensor with a noise signal of σ=2 μV and σ=10 μV, respectively.

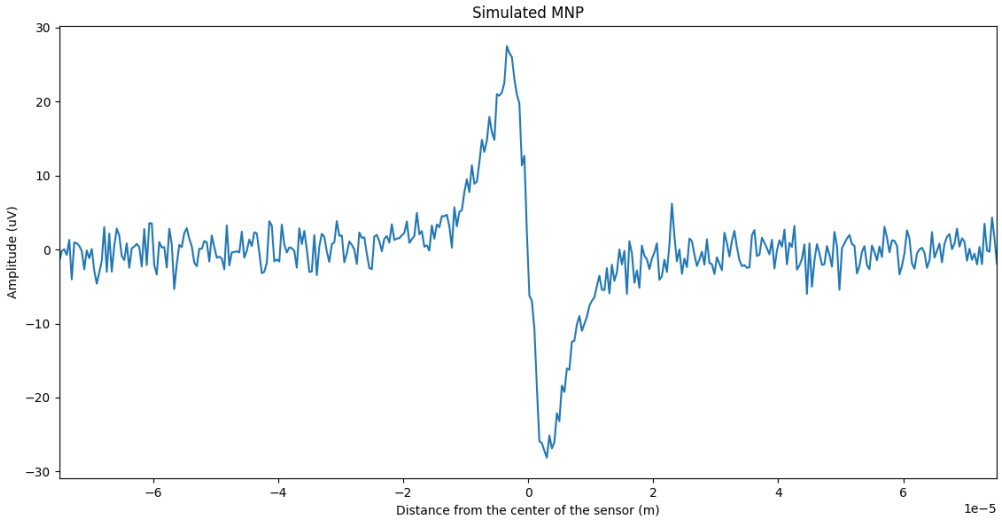

Figure 4.8: Simulation of an MNP located at 5 µm of height with a noise signal of σ = 2 μV.

By observing the figures, the greater the value of the sigma parameter, the greater the amplitude of the noise signal, and therefore, the more difficult it is to distinguish the signature from the presence of the MNP in the microfluidic channel of the cytometer. The same will also happen in relation to the

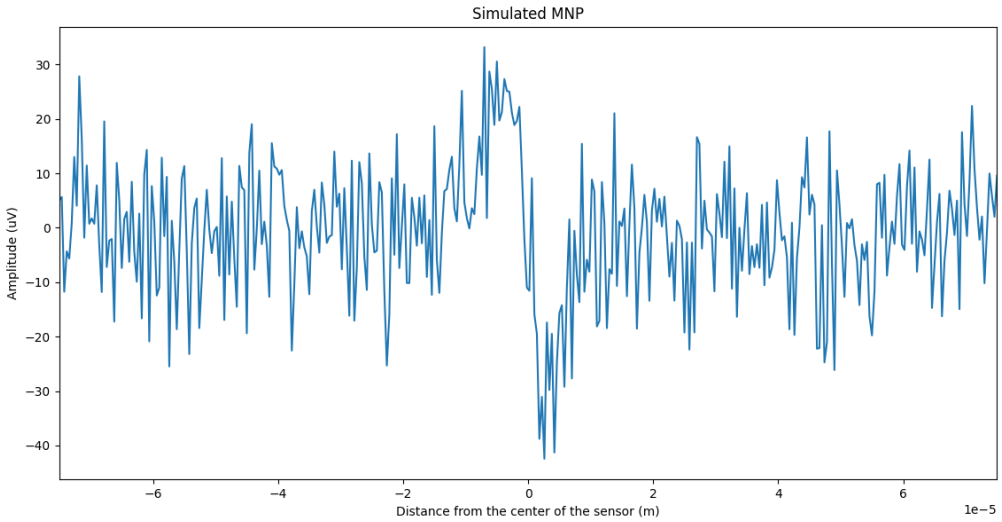

Figure 4.9: Simulation of an MNP located at 5 µm of height with a noise signal of σ = 10 μV.

height that the particle is to the sensor, because for the same sigma value, the greater the distance to the sensor's center, the more difficult it will be to find the signature. This can be observed by comparing Figures 4.8 and 4.10, the difference between them being the height of the MNP. For particles located further away from the sensor, the amplitude of the signal is smaller, as was concluded before, and it becomes more difficult to detect their presence.

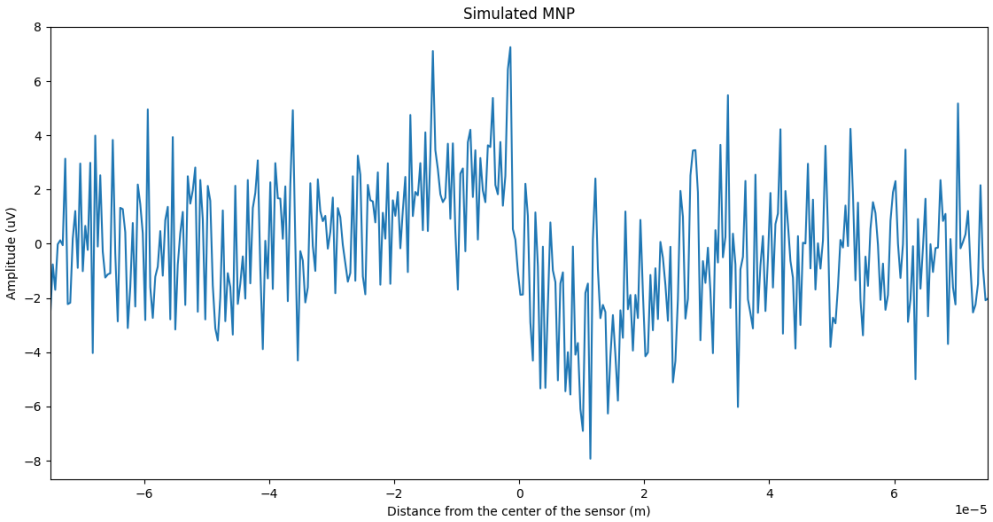

Figure 4.10: Simulation of an MNP located at 15 µm of height with a noise signal of σ = 2 μV.

Thus, concluding that the implementation of cell and cluster simulations is well defined and according

to what happens in reality, it is now possible to use the data obtained from the simulated magnetic field to train the artificial neural network as will be stated in the next chapter.

5

Artificial Neural Network Results

Contents

5.1 Data-set structure

A training data-set is the base for any machine learning application, and as was explained in Chapter 3, it contains the data that will be used by a ML model, in this case, an ANN, to train and predict the results. Thus, remembering that the objective is to distinguish between a cell and a cluster through the signal of the magnetic field, we have as input of the neural network the signal itself, that is, the signature that was shown earlier that resembles the Gaussian pulse. So this task can be approached as a classification problem, and as an output of the ANN, we have the value '1' if this signal corresponds to a cell and the value '0' if it is a cluster. Each input has 250 points that correspond to the pulse and that is translated by the sum of the magnetic fields produced by the magnetic particles that constitute the cell or the cluster. Table 5.1 is a representation of how is the data-set that will serve as training for the neural network.

Table 5.1: Representation of the data-set training.

			ANN Output	ANN Input
NºParticles	Sigma's Noise Signal [μV]	Height[m]	Cell/Cluster	Magnetic Field [μV]
17	2	$5e^{-6}$	1	[-1.94703, 1.90645, ... , 0.86984, 1.33735]
17	2	$5e^{-6}$	0	[1.05238, 3.31398, ... , 1.82329, 3.76494]
17	3	$10e^{-6}$	1	[-0.59911, 2.35205, ... , 1.44749, -0.91202]
17	3	$10e^{-6}$	0	[-2.63580, -1.91170, ... , 0.74629, 1.26064]
(...)				
70	12	$15e^{-6}$	1	[1.11419, 3.08200, ... , 0.41205, -0.05114]
70	12	$15e^{-6}$	1	[-1.15678, -1.056711, ... , 0.75632, 2.27631]

Table 5.1 only represents a few cases for information purposes, however, the dataset is made up of thousands of cases since each characteristic that affects the magnetic field will vary as follows:

- Height ranges from 5 to 15 μm with a step of 0.1 μm;
- The Sigma value of the noise signal varies between 2-12 μV with a step of 1 μV;
- The number of particles per cell/cluster will take the following values: [6, 10, 17, 22, 31, 40, 55, 62, 70].

The choice of the number of particles was taken into account that it would be very difficult to make a dataset with all possible numbers of particles, and for that reason, a random number was chosen for every ten up to a maximum of 70 particles. Furthermore, for each different case, 5 different noise vectors are generated, to take into account that each noise vector is implemented randomly, through the equation 4.1, and to have more data in the dataset.

5.2 ANN structure

As explained in Chapter 3, an ANN consists of 3 main layers: the input layer, the hidden layer, and the output layer. To implement our ML model, first, it is necessary to define its size, namely the number of neurons in the hidden and output layers.

In [19] these are a few of the rules suggested to size the hidden layers:

- The number of hidden neurons in each layer should be between the input layer's size and the size of the output layer.
- The number of hidden neurons in each layer should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons in each layer should be less than twice the input layer's size.

These methods provide guidelines and a possible starting point for the optimization of the ANN. However, trial and error are always necessary to achieve the best ANN architecture. For this reason, I used some of these methods as a test to find the best result for the neural network. Using the same number of neurons in the input layer for the hidden layer, which in this case is 250, which corresponds to 250 points/samples of the magnetic signal obtained through the simulation, the accuracy results obtained were the best compared to other values or methods. Since the input is fixed, because it always has 250 values extracted from the signal, we can assume that the number of neurons in the input and the hidden layer is always fixed and equal, as we have obtained the best accuracy results in this case.

The output layer is composed of only one neuron once this is a classification problem where there are only two classes, 'cell - 1' or 'cluster - 0'. For that reason, the activation function in this layer is the sigmoid function. As mentioned before the activation function for the hidden layer is the ReLu function because in addition to its simplicity, this function has a strong biological motivation, and it has been demonstrated to enable better training of deeper networks than other activate functions.

The data preparation, the ANN architecture, and training were performed in python language, using the Keras library with the TensorFlow 2.2.0 back-end and the sci-kit learn package.

5.3 Training

Usually, the process of training is measured in epochs. One epoch refers to the process of running the entirety of the training set through the model, only once. In practical terms, the greater the number of epochs, the greater the accuracy of the neural network, but this brings other associated problems such as over training, and the time it takes to train the data-set, so it is necessary to make the right decision.

The number of epochs can be set to an integer value between one and infinity, so, to define the number of epochs that the ANN needs to train with a good performance, I start to perform a training session with the entire data-set using 250 epochs, the results of the accuracy in the function of the number epochs are represented in Figure 5.1.

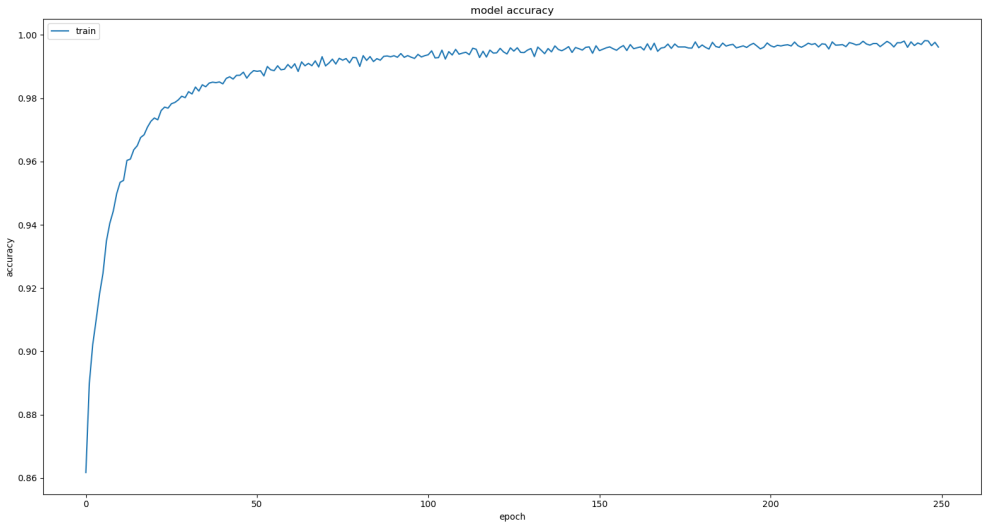

Figure 5.1: Evolution of the classification accuracy during training for the train set of examples.

It is possible to conclude that 150 could be a good value of epochs to train because it is when the accuracy becomes more stable. Accuracy means the degree to which the result (of the model to predict if the signal is a cell or a cluster) is correct. Training again the data-set with 150 epochs, the maximum value of accuracy was 99.68%, which means that, with a data-set of 77 thousand cases, 247 cases cannot be correctly classified by the model. That is, to predict whether the signals are cell or cluster for all signals, the model is tested by predicting for each signal whether it is cell (1) or cluster (0), if the prediction value matches the real value the result is correct (1). If the prediction is wrong, the result is incorrect (0). The accuracy value corresponds to the average of the results. For example, for the results represented in Table 5.2 the accuracy (Eq. 5.1) will be the mean of the result column.

NºParticle	Prediction	Real Value	Result
17	1	1	1
17	1	1	1
17	0	0	1
17	0	0	1
17	0	0	1
17	1	0	0

Table 5.2: Example of model results for nºparticles equal to 17.

$$Accuracy = \frac{1+1+1+1+1+0}{6} = 0.8(3), \tag{5.1}$$

For this example, the accuracy will be 0.8(3), which means for 6 cases the model predicted correctly 5 and failed in 1 case.

The objective of this project is to distinguish between cell and cluster magnetic signals, for a different number of particles and positions. So to interpret better our accuracy results, the model was trained using the entire dataset, and to observe the predictions of the model as a function of different parameters, I split the data-set test into a different number of particles and different heights to classify and observe how the accuracy of the model varies with the variation of these parameters. The results obtained are represented in the next section.

5.4 Results

The first results obtained are presented in Figures 5.2 and 5.3. Figure 5.2 represents the accuracy of the model as a function of 2 parameters: the number of particles per cell/cluster, and sigma of noise signal. For a bigger number of particles, the color is darker and for smaller particle numbers the signal is lighter. Figure 5.3 represents the accuracy of the model as a function of the height of the cell/cluster to the plane's sensor and also the sigma. For heights closer to the sensor (5 μm) the signal is darker and for heights more furthest (15 μm) the signal is lighter.

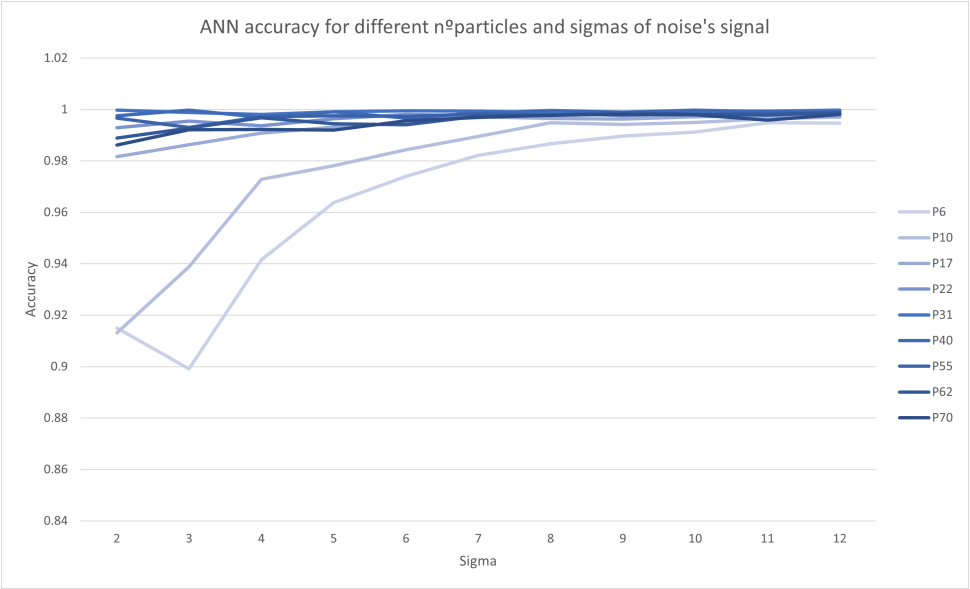

Figure 5.2: ANN accuracy for different nºparticles and sigma values.

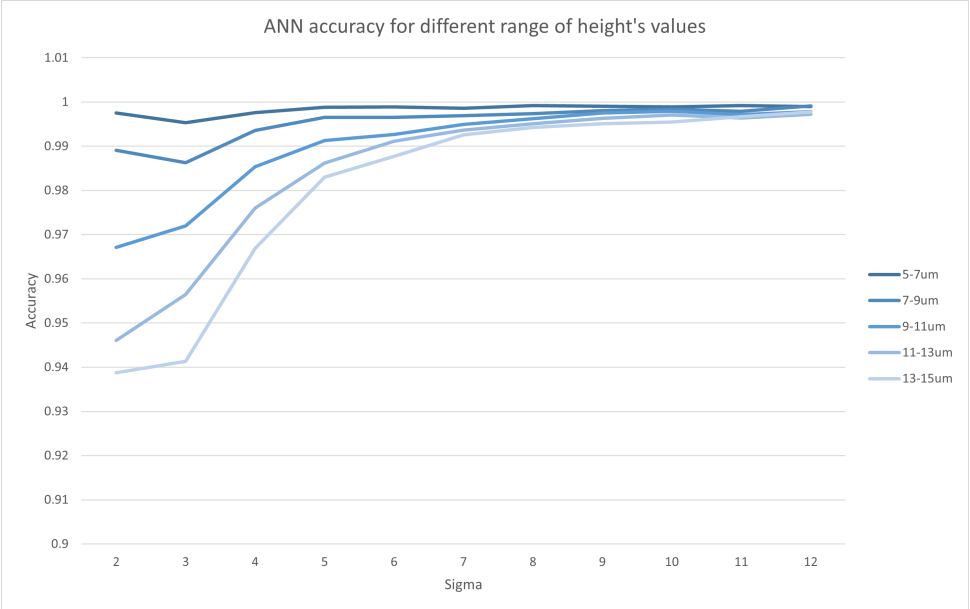

Figure 5.3: ANN accuracy for different range of height's values and sigma values.

As mentioned in the conclusions of the simulations in the previous Chapter 4, for larger particle numbers and smaller heights the magnetic field is higher since in the first case the magnetic field produced is larger because there are more particles, and in the second case because as we are closer to the sensor, the signal is felt by it more intensely. Furthermore, for higher sigma, the noise signal is also higher. Putting these two facts together, supposedly what should happen was that the accuracy should decrease over sigma values since the noise signal was higher and it should be more difficult

to distinguish between cell and cluster in the middle of the noise. However, what happens is precisely the opposite, which leads us to the conclusion that the ANN, from a certain moment on, no longer distinguishes between cell and cluster, to start distinguishing between noise signal and cluster signal (for having the largest amplitude of magnetic field), which is a bit easier to differentiate. This is a negative result because the ANN begins to misclassify with a lower SNR values (greater sigma values), which will be more common in reality.

That said, for this not to happen I decided to decrease the maximum number of sigma of the noise signal to 7 μV, since it was from this sigma value that accuracy became very high and stable close to 1, as can be seen from the previous figures. From now on, to make the results more understandable, these will be represented in SNR values in dB instead of sigma values, following Equation 5.2. Besides that, the number of particles less than 10 can be removed because the signal is not strong enough to make useful conclusions.

$$SNR\,[dB] = 20log\left(\frac{\sqrt{\left(\sum_{n=1}^{n=250} Magnetic\,Field(n)\right)^2}}{\sqrt{\left(\sum_{n=1}^{n=250} Noise\,Signal(n)\right)^2}}\right) \tag{5.2}$$

Applying the decisions made above, the result that translates the accuracy as a function of SNR value intervals by the number of particles is represented in Figure 5.4. In this case, the noise signal is higher for lower SNR values, so the results show that, on average, with increasing SNR value range, accuracy tends to improve. Furthermore, for larger numbers of particles that make up cells and clusters we have higher accuracy, to smaller numbers of particles. Figure 5.5 is represented the accuracy of the model as a function of the height of the cell/cluster to the plane's sensor and also SNR value intervals. For smaller height's values, when the position is more closer to sensor, the accuracy is greater. For bigger distances to the sensor the accuracy decreases. The figure shows that overall there is a tendency for accuracy to improve as SNR values increase.

It can be concluded that this ANN can differentiate, with good accuracy, the cell signals from cluster signals for a system with a maximum RMS noise signal of 7 μV which is the equivalent of a minimum SNR value of 9.90 dB, for cells and clusters with a minimum of 20 particles. Although there is no maximum number for particles, the maximum number used to train was 70 because it seems to be a reasonable number for particles that are attached to a cell. The actual value of RMS noise in the MFC that we are working on in this project is between 2-5 μV at 200KHz bandwidth, so this model can predict any signal that is between these values.

To have a visual perception of the signals that we are talking about, the following Figures 5.6, 5.7, 5.9 and 5.8 represent both signals (cell/cluster) in the same conditions to exemplify the magnetic signals with the maximum and minimum values of SNR that this model can predict. By observation of the

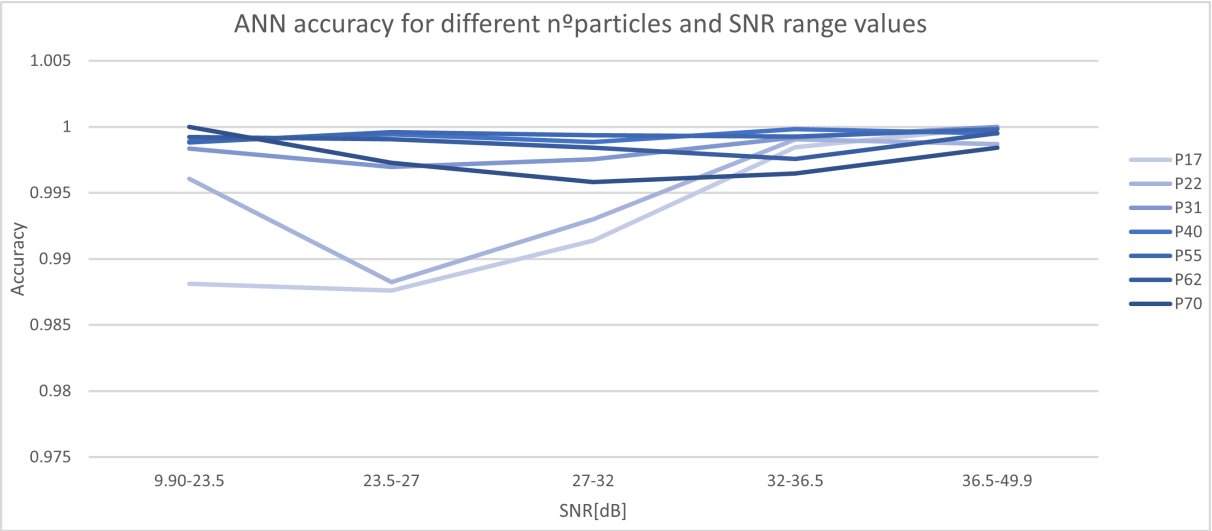

Figure 5.4: ANN accuracy for different nºparticles and SNR range values.

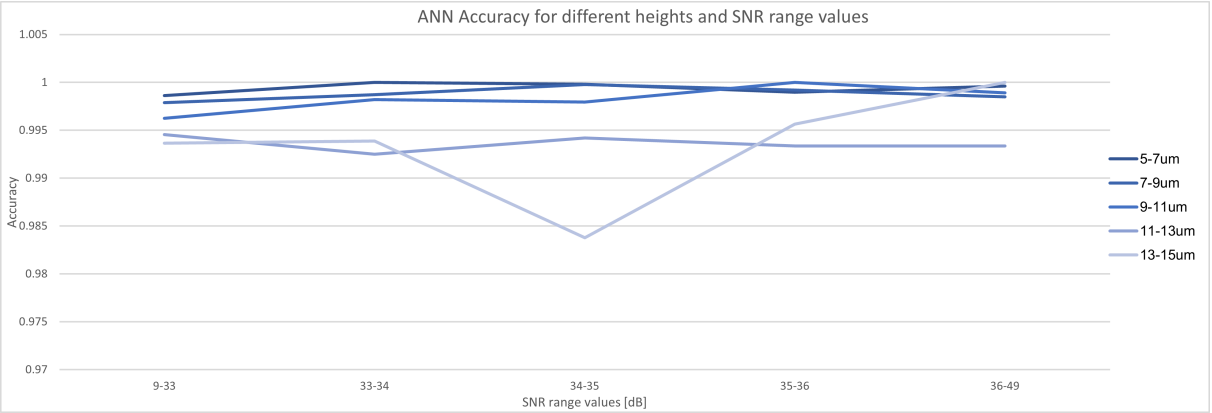

Figure 5.5: ANN accuracy for different heights and SNR range values.

first two Figures (5.6 and 5.7), it can be concluded that for a minimum SNR value, that corresponds to worst conditions, both simulations look similar which can be translated as harder for the ANN to distinguish the signals. Another characteristic that must be mentioned is that for the same conditions, the cluster simulation has always a bigger SNR value, that happens because the cluster signal has a bigger amplitude (because of all the reasons mentioned before) and the noise signal has the same amplitude for both signals so SNR values will be bigger for cluster simulations. For better conditions (Fig. 5.9 and 5.8), the cluster simulation has a greater amplitude than cell simulation so it is much easier for ANN to differentiate it.

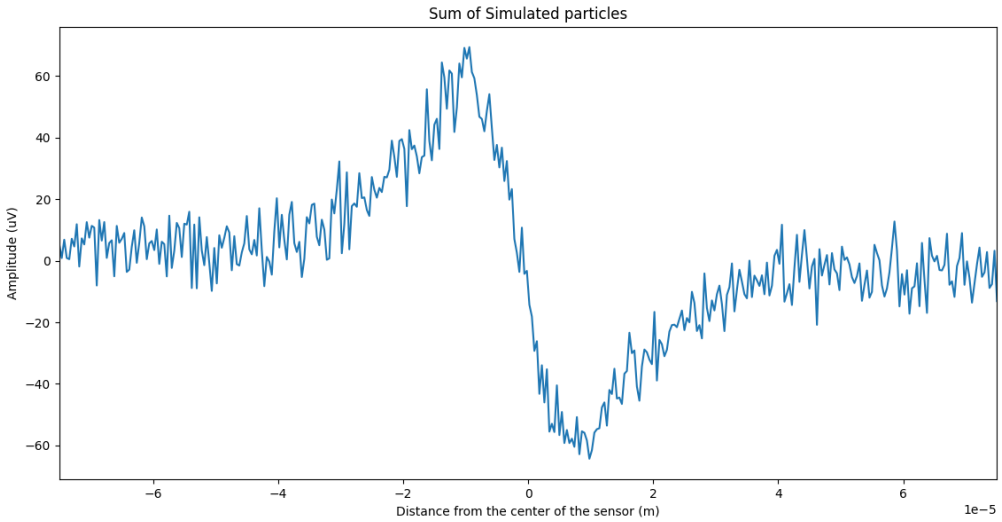

Figure 5.6: Example of a signal for minimum SNR (11.41 dB) for a cell simulation with 20 particles at 15 μm of height.

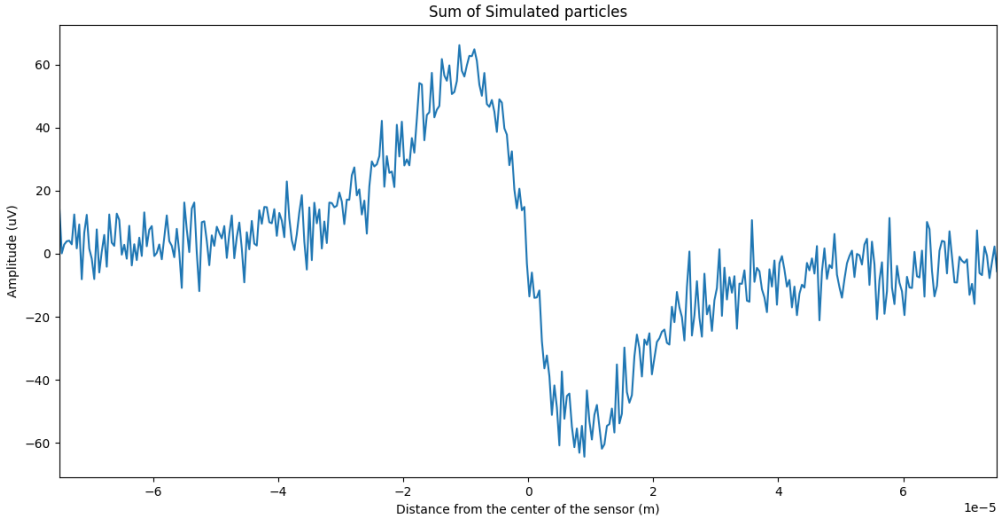

Figure 5.7: Example of a signal for minimum SNR value of 13.24 dB, for a cluster simulation with 20 particles at 15 μm of height.

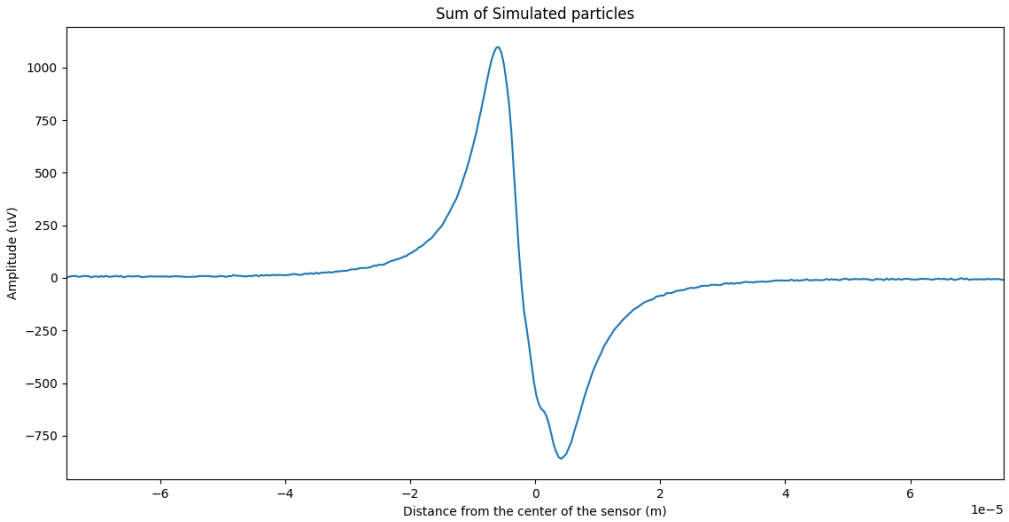

Figure 5.8: Example of a signal with a SNR value of 44.9 dB, for a cell simulation with 70 particles at 5 μm of height.

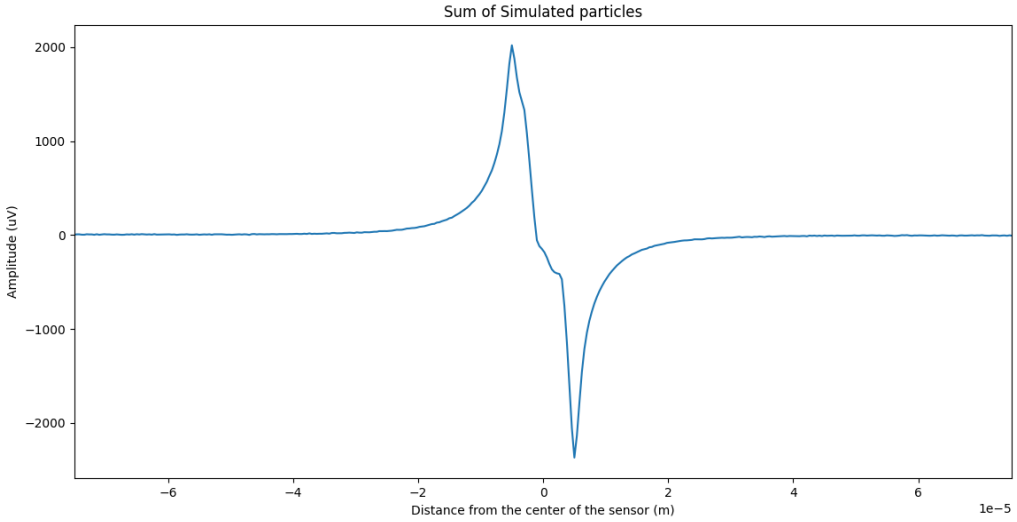

Figure 5.9: Example of a signal with a SNR value of 48.02 dB, for a cluster simulation with 70 particles at 5 μm of height.

5.5 Signal Subsampling

The results obtained so far translate the accuracy of the model that uses as input the signal that results from the sum of the magnetic field produced by the particles. This signal contains 250 samples, however, it will not always be possible to obtain this amount of samples, because if the particles travel at a higher

speed for the same sampling frequency we will have a signal with fewer samples. So the idea is to try to understand if the artificial neural network, that was implemented, can distinguish the signals when the particles travel at a higher speed. And if so, establish a limit situation by which the model fails to distinguish.

In Figures 5.10, 5.11, and 5.12 there are some results of ANN accuracy for simulations with different particle's speed and with a frequency sample of 200 kHz.

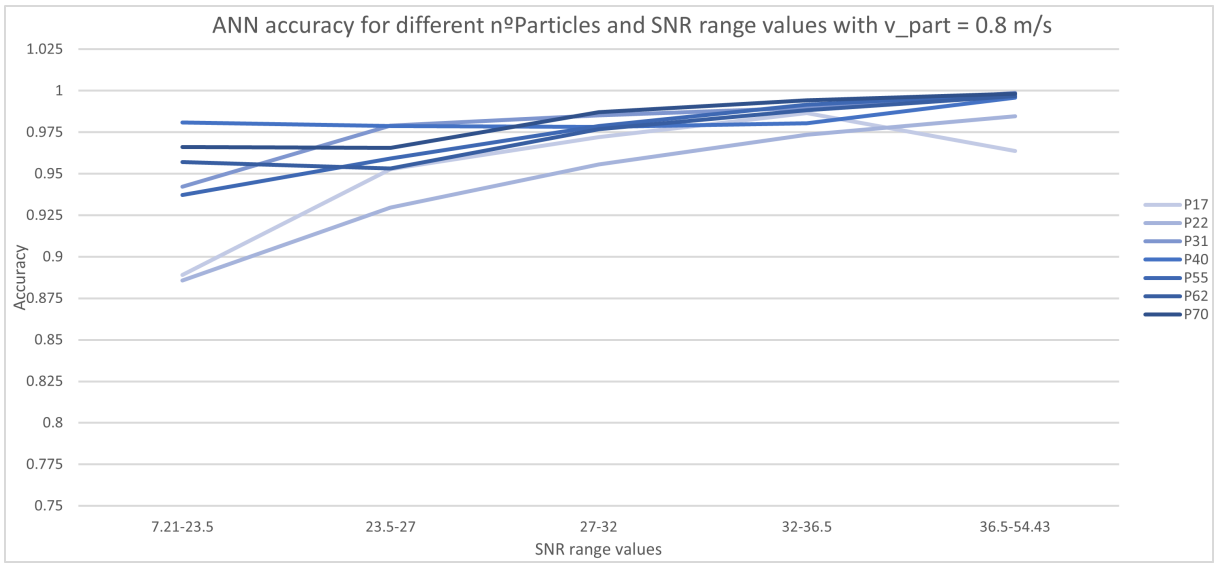

Figure 5.10: ANN accuracy for different nºparticles and SNR range values for a particle speed of 0.8 m/s, which corresponds to a signal with 30 samples.

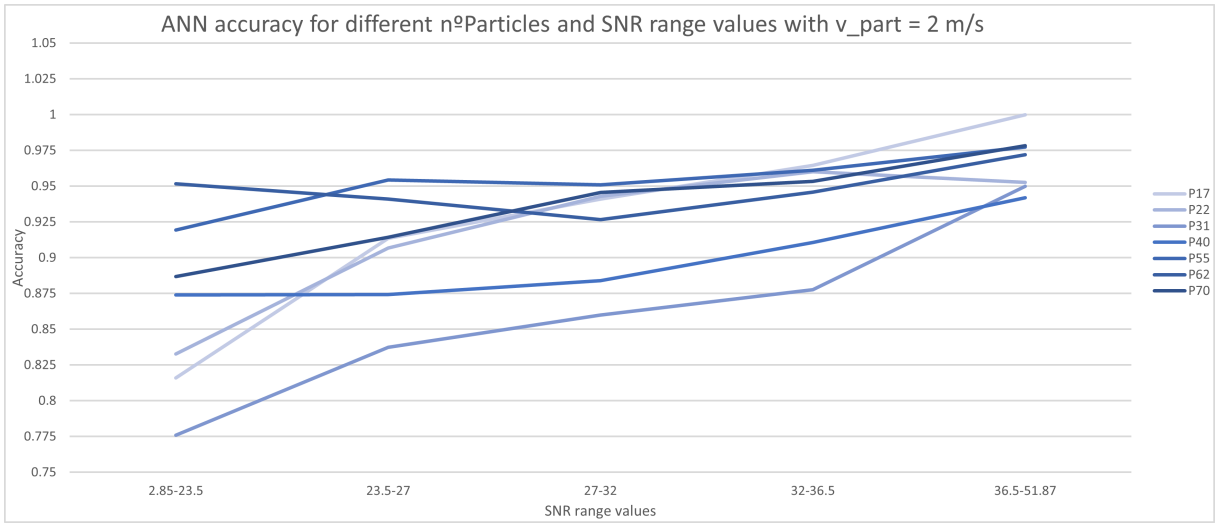

Figure 5.11: ANN accuracy for different nºparticles and SNR range values for a particle speed of 2 m/s, which corresponds to a signal with 14 samples.

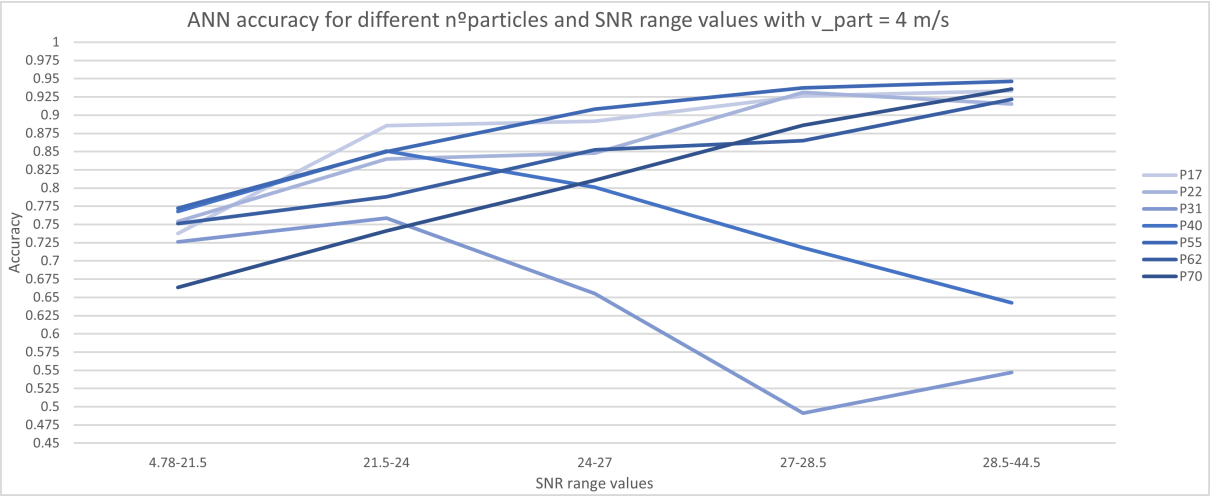

Figure 5.12: ANN accuracy for different nºparticles and SNR range values for a particle speed of 4 m/s, which corresponds to a signal with 6 samples.

For observation of the figures, it is possible to conclude that for a greater particle's speed (fewer samples), in general, the accuracy values tend to decrease, as it is possible to see from Figure 5.13, which shows how the model accuracy varies as a function of the particle's velocity.

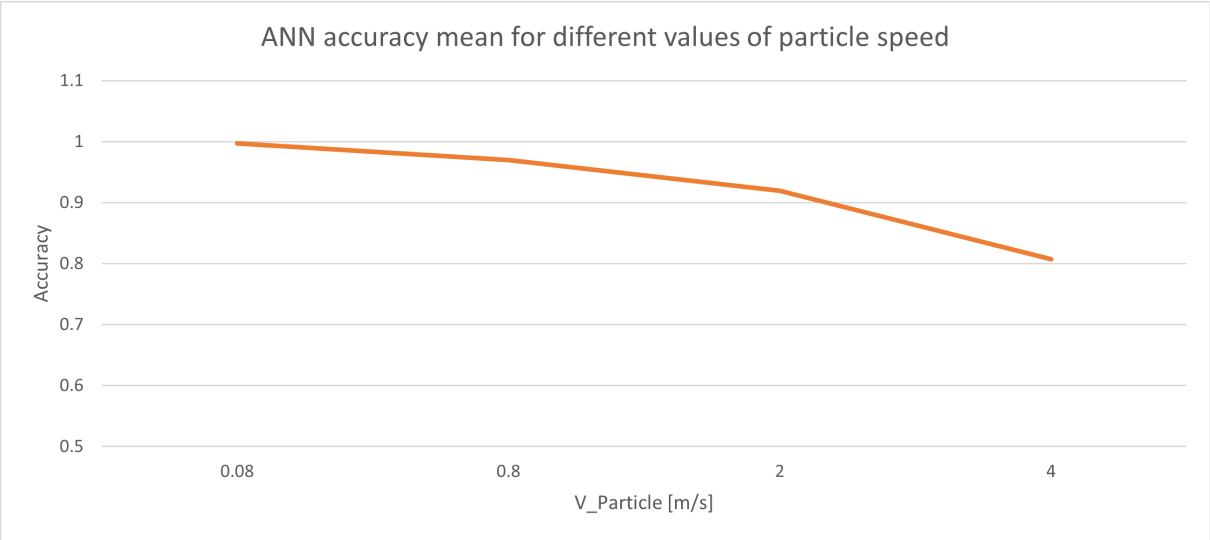

Figure 5.13: ANN accuracy mean for different values of particles speed.

According to [6] there is a estimation that is necessary at least 10 samples to fully characterise the pulse, as suggested by [20], [21]. This estimate arises through the Equation 5.3, that estimates the maximum flow rate in μl/min to use on the cytometry experiment considering: sampling rate (F_s) the ratio between pulse duration and maximum to minimum duration ($\alpha = T_{full}/T_{H-L}$) empirically estimated to be between 3-4 in the case of bipolar Gaussian mono-cycles, the sensor length (L_{sens}), the minimum desirable number of samples to reconstruct the signal (S_{min}), the microfluidic channel's height (h_c) and

width (w_c), as well as the channel section (A_{sec}).

$$Flow\,rate = \frac{L_{sens} \times \alpha \times F_s}{\left(-0.56(\frac{h_c}{w_c})^2 + 1.15(\frac{h_c}{w_c}) + 1.5\right) S_{min}} Asec \times 60 \times 10^9\ [\mu l/min] \tag{5.3}$$

Thus, according to this and by observing the results present in Figure 5.13, it is possible to conclude that using at least 10 samples (that corresponds to approximately 2.5 m/s of particle's speed) the ANN can differentiate the signals with an overall good accuracy.

6

Conclusion and Future Work

Contents

6.1 Conclusions

This work started with the simulation of various forms of signals from magnetic particles attached to cells and clusters formed from free magnetic particles (using square packing). Since these simulations have shown good results since they translate what happens in reality in a microfluidic channel of a magnetic flow cytometer, the data obtained through these simulations were used to train a machine learning model, namely, an artificial neural network to differentiate the signals coming from the magnetic particles connected to the cells from the signals coming from the magnetic particles that form clusters.

By observing the accuracy results obtained, it was concluded that this model can differentiate the signals with greater precision for a system with a maximum RMS noise signal of 7 μV which is the equivalent of a minimum SNR value of 9.90 dB, for cells and clusters with a minimum of 20 particles. By subsampling the signal it can also be concluded that the maximum speed that the particles can travel in the microfluidic channel is at a speed of 2.5 m/s.

6.2 Future Work

For future work, I can suggest some tasks that can be carried out to improve the project.

- Carry out a more in-depth study of how free magnetic particles forms clusters when subjected to magnetic excitation, through a real experience, and, if it makes sense, since it has not been tested, implement hexagonal packing and see if the results obtained are better than square packing;
- Introduce the particle's magnetization angle in the simulations because in this work it was assumed that the particles were always perpendicular to the sensor's plane;
- Test another ML models and activation functions - Since the work carried out so far using ML models has always used artificial neural networks, and the results shown were very good, I chose to implement the same model, however, exploring other models and activation functions could bring different results;
- Test the implementation of the artificial neural network experimentally using cancer cells.

Bibliography

[1] A. R. Soares, R. Afonso, J. Lampreia, J. J. V. C. Martins, M. Piedade, and S. Cardoso, "Automatic system to count and classify bacteria based on magnetic cytometry," in *IEEE MAGNETICS LETTERS,*, 2016.

[2] A. R. Soares, R. Afonso, D. Caetano, V. C. Martins, M. Piedade, and S. Cardoso, "Analytical Strategy for Magnetic Flow Cytometry Signals Classification," 2019.

[3] R. Soares, V. C. Martins, R. Macedo, F. A. Cardoso, S. A. M. Martins, D. M. Caetano, P. H. Fonseca, V. Silvério, S. Cardoso, and P. P. Freitas, "Go with the flow: advances and trends in magnetic flow cytometry," in *Analytical and Bioanalytical Chemistry*, 2019.

[4] P. P.Freitas, V. C. Martins, F. A.Cardoso, E. Fernandes, T. Sobrino, J. Castillo, A. Chicharo, M. Abal, R. Lopez-Lopez, T. S. Dias, , and S. Cardoso, "Spintronic biochips from the laboratory to pre-clinical applications," in *Nanomagnetism: Applications and Perspectives*, 2017, pp. 165–200.

[5] D. Caballero, S. C. Kundu, and R. L. Reis, "Microfluidics and biosensors in cancer research." in *Springer*, 2022.

[6] D. M. Caetano, "Ph.d. dissertation, instituto superior técnico," in *Circuits and signal processing for magnetoresistive sensor arrays*, 2021.

[7] Rachel Nall. (2020, January) What to know about cancer. [Online]. Available: https://www.medicalnewstoday.com/articles/323648

[8] Xiamei Ma and Herbert Yu. (2007, October) Global Burden of Cancer. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994799

[9] U. . h. World Health Organization. (2022) Promoting Cancer Early Diagnosis.

[10] Canadian Cancer Society. (2022) What are the benefits and limitations of regular cancer screening? [Online]. Available: https://cancer.ca/en/cancer-information/find-cancer-early/screening-for-cancer/benefits-and-limitations-of-regular-cancer-screening

[11] J. Loureiro, Z. P. Andrade, S. Cardoso, L. C. Silva, M. J. Cabral, and P. P.Freitas, “Magnetoresistive chip cytometer,” in *Lab on a Chip*, 2011, pp. 2255–2261.

[12] A. C. Fernandes, C. M. Duarte, F. A. Cardoso, R. Bexiga, S. Cardoso, , and P. P. Freitas, “Lab-on-chip cytometry based on magnetoresistive sensors for bacteria detection in milk,” in *Sensors*, vol. 14, no. 8, 2014, pp. 15 496–15 524. [Online]. Available: https://www.mdpi.com/1424-8220/14/8/15496

[13] C. Nguyen and M. Miao, “Design of cmos rfic ultra-wideband impulse transmitters and receivers.” in *Springer Briefs in Electrical and Comuputer Engineering*, 2017.

[14] J. L. andd R. Ferreira, S. Cardoso, P. P. Freitas, J. Germano, C. Fermon, G. Arrias, M. Pannetier-Lecoeur, F. Rivadulla, and J. Rivas, “Toward a magnetoresistive chip cytometer: Integrated detection of magnetic beads flowing at cm/s velocities in microfluidic channels,” in *Applied Physics Letters*, vol. 95, no. 3, 2009, pp. 34–104. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.3182791

[15] T. M. Mitchell, “Machine learning,” in *New York: McGraw-Hill*, 1997.

[16] K. D. Foote. (2021, December) A brief history of machine learning. [Online]. Available: https://www.dataversity.net/a-brief-history-of-machine-learning/#

[17] I. C. Education. (2020, July) What is machine learning. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning

[18] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[19] J. Heaton, “Introduction to neural networks for java,” January 2008.

[20] D. Issadore, J. Chung, H. Shao, M. Liong, A. A. Ghazani, C. M. Castro, R. Weissleder, and H. Lee. (2012) Ultrasensitive clinical enumeration of rare cells ex vivo using a micro-hall detector. [Online]. Available: https://www.science.org/doi/10.1126/scitranslmed.3003747

[21] D. Issadore, H. J. Chung, J. Chung, G. Budin, R. Weissleder, and H. Lee. (2013) Micro-hall chip for sensitive detection of bacteria. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adhm.201200380