Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

MONITORAGGIO E GESTIONE DELLE STRUTTURE E DELL'AMBIENTE - SEHM2

Ciclo 35

Settore Concorsuale: 09/E3 - ELETTRONICA

Settore Scientifico Disciplinare: ING-INF/01 - ELETTRONICA

EFFICIENT AND LOW-COMPUTATIONAL PREDICTIVE MODELS FOR SPECTRAL SENSORS

Presentata da: Leonardo Franceschelli

Coordinatore Dottorato

Luca De Marchi

Supervisore

Marco Tartagni

Co-supervisore

Annachiara Berardinelli

Esame finale anno 2023

Index

This thesis includes extracts, both in the text and in the figures, from my peer-reviewed journal publications.

1. Abstract

Spectral sensors are a wide class of devices that are extremely useful for detecting essential information of the environment and materials with high degree of selectivity. Recently, they have achieved high degrees of integration and low implementation cost to be suited for fast, small, and non-invasive monitoring systems. However, the useful information is hidden in spectra and it is difficult to decode. So, mathematical algorithms are needed to infer the value of the variables of interest from the acquired data. Between the different families of predictive modeling, Principal Component Analysis and the techniques stemmed from it can provide very good performances, as well as small computational and memory requirements. For these reasons, they allow the implementation of the prediction even in embedded and autonomous devices. In this thesis, I will present 4 practical applications of these algorithms to the prediction of different variables: moisture of soil, moisture of concrete, freshness of anchovies/sardines, and concentration of gasses. In all of these cases, the workflow will be the same. Initially, an acquisition campaign was performed to acquire both spectra and the variables of interest from samples. Then these data are used as input for the creation of the prediction models, to solve both classification and regression problems. From these models, an array of calibration coefficients is derived and used for the implementation of the prediction in an embedded system. The presented results will show that this workflow was successfully applied to very different scientific fields, obtaining autonomous and non-invasive devices able to predict the value of physical parameters of choice from new spectral acquisitions.

2. Motivation of the work

Environmental monitoring, described as the processes and activities that need to take place to characterize and monitor the quality of the environment, is a field of study that, especially in the last decade, gained a lot of traction and interest both from the academic and the industrial world. Of all the different techniques that could be used to monitor changes in physical and/or chemical variables in the environment, one of the most interesting is for sure spectral sensing, because it allows acquiring useful information not only from the surface of materials and objects (like images) but also allows to inspect changes inside the materials, that would be otherwise invisible to the human eye. The main problem behind the use of spectral sensors is that the information sought is divided between all the variables evaluated by the device (that usually are thousands of frequencies, wavelengths, etc.). Moreover, the acquired spectra are affected by all the physical and chemical variations that happen in the monitored material, not only the ones caused by our Variables of Interest (VI). So, a human user can't estimate changes directly by looking at a spectrum, and an algorithm that provides a mathematical way to automatically calculate the VI values from it is required. For these reasons, the main goal of my PhD was to study statistical techniques and apply them to experimental data acquired with different types of spectral sensors, devising in this way models that allow the prediction of the VI directly from an acquired spectrum, in real-time (or close to it) conditions. In the last century, several different types of algorithms for predictive modeling were developed, going from simple linear regression to very complex Deep Learning techniques. For our needs, *multivariate statistics* is the most suitable, being a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. In particular, I focused on a technique called *Principal Component Analysis* (PCA), developed in the chemometric field specifically for analyzing spectra, and on following evolutions of this algorithm, which allows to solve regression or classification problems. These analyses can be applied to every type of spectra, independently by the field of application and the type of sensor: in particular, I focused on 4 different applications, where the same statistical techniques are used to predict different variables, measured with different sensors in different environments. I worked on models for predicting moisture in soil and concrete, gas concentrations in gaseous mixes, and monitoring the freshness and decaying process of fish. This of course was made possible thanks to collaborations with other researchers and teams, both in the academic and industrial world: I collaborated with the Department of Agricultural and Food Sciences (DISTAL) for soil and fish applications and with the Department of Civil, Chemical, Environmental and Materials Engineering (DICAM) for concrete curing applications, both from the University of Bologna; with the Centre Agriculture Food Environment (C3A) of the University of Trento for the

predictive modelling; and with a startup called Nanotech Analysis s.r.l., based in Turin, for gas applications.

3. Workflow and organization of the thesis

Summing up, the main point behind my Ph. D. work was to obtain prediction models that allow to predict the value of one or more VIs directly from a raw spectrum acquired by a spectral sensor., obtaining a fast and non-invasive monitoring device. This approach can be applied in different practical applications, using the same workflow for all of them. As depicted in Fig. 1.it could be divided into three consequential phases, each of them necessary for the following ones. Firstly, an acquisition campaign is carried out, measuring at the same time spectra and the real value of the VI (on the same sample or different ones). The data acquisition is performed with different techniques and instruments, depending on the field of application. These data are then used as input for the chosen *Multivariate Statistical Analysis* (MSA), after an initial preprocessing phase: it could imply the application of simple preprocessing algorithms, like smoothing, first derivative, alignment between spectra, etc.; or the use of more complex techniques, like Deep Learning algorithms for object recognition and selection. The model is then refined and tested with an external set of data, not used for the calibration. In this testing phase, it is possible to compare the real VI values with the ones predicted by the model, obtaining a good understanding of the model prediction ability. My work was not focused on the implementation of algorithms in numerical code but on the analysis and optimization of MATLAB toolbox models, called PLS_Toolbox, developed by Eigenvector Inc. [1] When the model is complete, it is possible to extract from it a calibration array, used to implement the calculation of the VI directly inside the embedded electronic system. The whole procedure can be applied to every field or application that allows the acquisition of spectra containing useful information.

This workflow, as well as 4 of its applications in different fields, will be described in the thesis. In the Introduction the concepts of spectral sensing will be presented in greater detail, as well as a brief discussion of its state of the art, both for 1D and 2D spectral devices. Then, the MSAs used in the various applications will be presented, focusing on their mathematical theory and algorithm implementation Then the practical applications I worked on will be presented in detail in 4 different chapters. Finally, the Conclusions will sum up the obtained results, as well as their utility in the field of environmental monitoring.

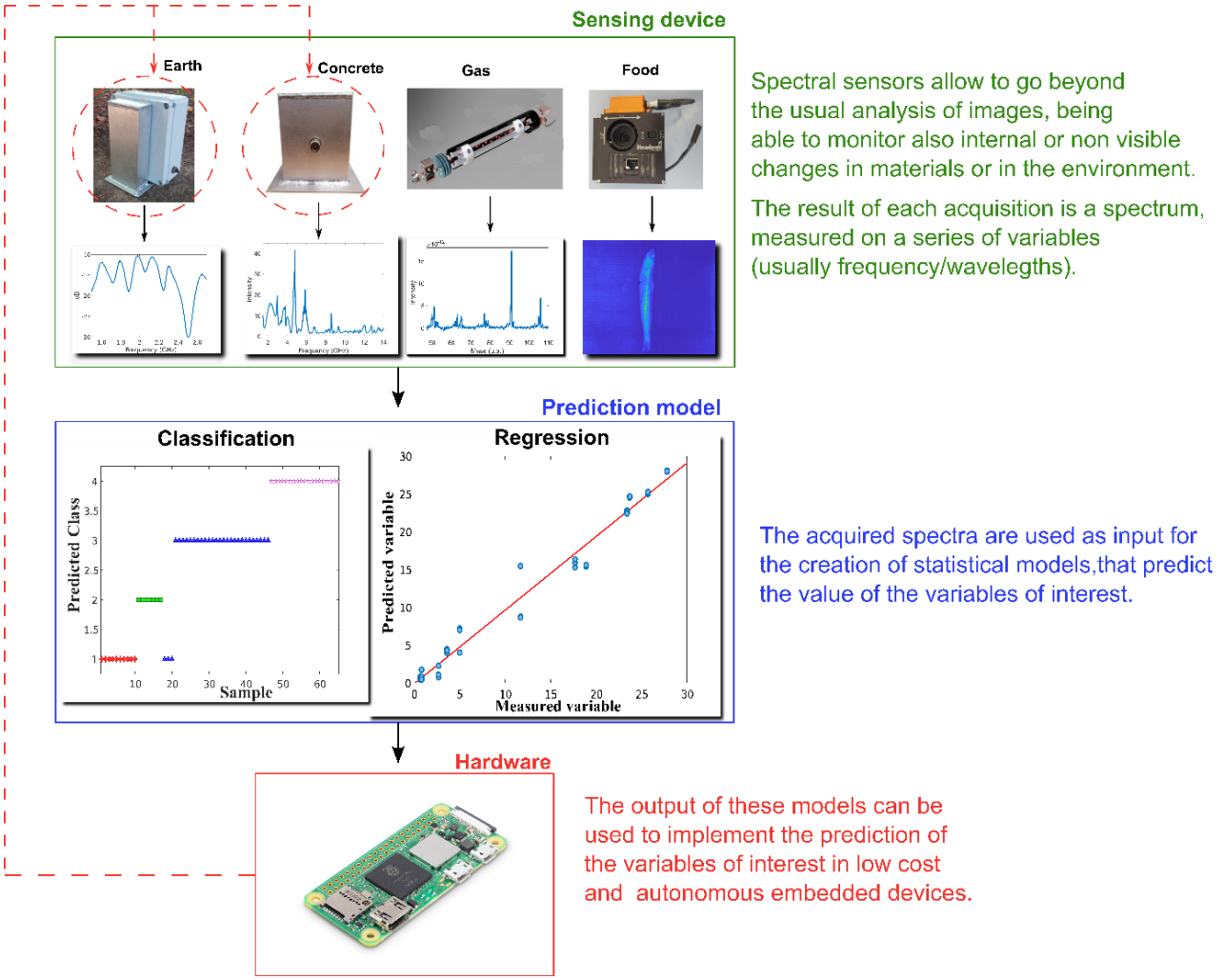

Fig.1 Flowchart of the approach used for the design of monitoring embedded devices. It can be divided into three consequential steps: acquisition of both spectral and VI data (top) using gold standard techniques based on the application; calibration and testing of the prediction model created with MSA algorithms (middle); implementation of the calculation of the VI inside an embedded system (bottom).

4. Introduction

4.1 Spectral sensing

The monitoring of changes in physical or chemical Variables of Interest (VIs) is a fundamental component of several human activities, with crucial importance in fields like safety control or environmental monitoring. This task is performed by sensors: in their broadest definition, a sensor is a device, module, machine, or subsystem that detects events or changes in its environment and sends the information to other electronics[2]. Sensors can be divided into a plethora of families, based on the physical phenomenon exploited by the device to acquire information about one or more specific VIs. Between them, image sensing is often used for the study and monitoring of the environment: the obtained images or video can be easily interpreted by the user (or even analyzed by automatic algorithms), and it is possible to integrate a lot of small and cheap cameras in basically all scientific applications. However, 2D image sensing has a major flaw: image analysis does not allow it to go beyond the surface of the material itself in its form and the material in its chemical and physical qualities. This can be overcome by using another type of analysis, called spectrometry: it is the field of study that measures and interprets the electromagnetic spectra that result from the interaction between electromagnetic radiation and matter as a function of the wavelength or frequency of the radiation. So, spectral sensors allow going beyond what is the analysis of the classical image where we can only identify the position and shape of the objects, acquiring useful information also on changes that happen also inside the monitored material or object, even at nanometric levels. Moreover, spectral sensors are also suitable for embedded in-line and/or real-time applications, given their simplicity, small dimensions, and cheapness. An example of the usefulness of spectral sensing over image sensing is depicted in Fig. 2.

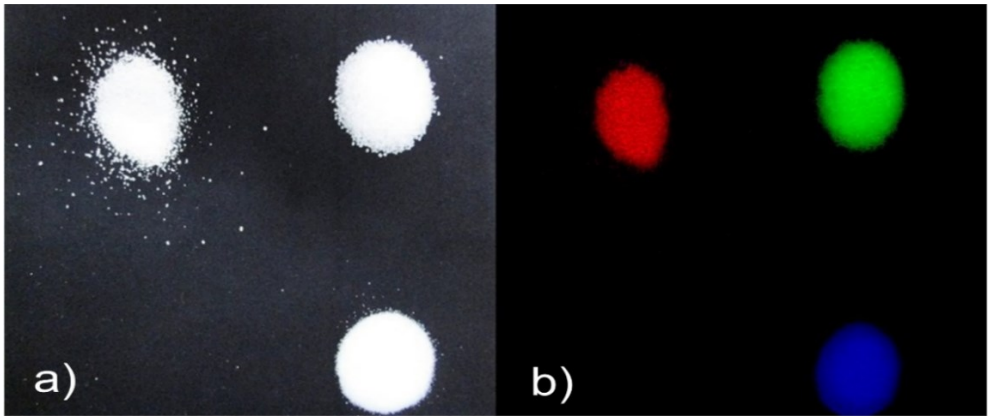

Fig. 2 **(a)** Conventional image showing three almost identical heaps of sugar, salt, and citric acid **(b)** Hyperspectral imaging of the same 3 heaps: the difference between the three substances becomes obvious due to the differences in the molecular structure and chemical properties, detected by the sensor.

In the applications I will present in the next chapters, I worked with three different techniques: Radio-Frequency (RF) sensing, Hyperspectral Imaging (Visible and Near InfraRed Spectrometry (VIS/NIR)) and Gas Chromatography–Mass Spectrometry (GC-MS).

More in detail, RF sensing is based on sending a series of waves of increasing frequency (as an example, in our applications we used the ranges 1.5-2.7 GHz (soil) and 1.5-14 GHz (concrete))and acquire the consequent waves reflected by the sample. This reflection depends mainly on the dielectric properties of the examined material, which is in turn highly influenced by its water content, making it an ideal technique for the monitoring of moisture. On the other hand, Infrared Spectrometry exploits the fact that molecules absorb frequencies that are characteristic of their structure. These absorptions occur at resonant frequencies, where the vibrational frequency matches the frequency of the absorbed radiation. These energies are affected by the shape of the molecular potential energy surfaces, the masses of the atoms, and the associated vibronic coupling. Finally, GC-MS is an analytical technique that is used to measure the mass-to-charge ratio of ions: usually, a gaseous sample is ionized thanks to an electron beam, then these ions are accelerated and subjected to a magnetic field, dividing them according to their mass. Every pure component or molecule breaks in a characteristic and unique pattern: the atoms or molecules in the sample can be identified by correlating known masses to the identified masses or through these characteristic fragmentation patterns. Finally, hyperspectral imaging mixes the best aspects of both images and sensors, blending VIS/NIR spectroscopy with computer vision. Thanks to specific instruments called “hyperspectral cameras”, it is possible to acquire a full spectrum for every pixel of an image, obtaining three-dimensional data called a “hyperspectral cube” [3], obtaining both spatial (in the pixels) and chemical (in the spectra) data at the same time.

State of the art

In the literature it is possible to find an example of applications of these 3 techniques in a lot of different fields, obtaining generally very good results for a non-invasive and fast prediction of several VI. RF sensing is often used for the monitoring of metals and other materials: as an example, the use of an open-ended waveguide was tested in 2003 by Abu-Kousa et al. [4] for the non-destructive detection of air gaps in carbon composite materials. They demonstrated that waveguides have a good sensitivity for the detection of voids, especially if they are filled with dielectric materials (like water). This result was supported by another study by Bin Sediq and Qaddoumi [5], showing that circular waveguides are better for this monitoring, penetrating more deeper into these types of material. More

recently, McClanahan et al. [6] studied the use of waveguides for the detection of superficial cracks in metals, obtaining a good accuracy and also for small defects (0.25 mm). RF sensors were also investigated for biological studies, both on humans and food. In 2019, Meaney et al. [7] developed a dielectric probe array that could be used for early diagnosis and characterization of skin burns, whereas Obol et al. [8] designed a coaxial probe technique for the microwave characterization of biological tissues, testing it on cancerous samples of both human and animal tissue. Regarding food, Shivamurthy et al. used an open-ended waveguide to differentiate between healthy and infected mangos, finding a difference of the 30%-40% between the two dielectric responses.

Regarding GC-MS, it is very effective to monitor residuals of dangerous gaseous substances in the air, as was demonstrated by several papers. In 2019, Healy et al. [9] used two GCs to measure benzene, toluene, ethylbenzene, and xylenes concentrations in air at a highway, obtaining a mean error ranging from 6% (benzene) to 30% (xylene). In the same year, Yoo et al. [10] used a particular type of GC (called barrier discharge ionization detector) to monitor the concentration of formaldehyde rapidly, also for very low values (under 1 part per million (ppm)); whereas Chang and Heinemann used spectra acquired with a particular GC, called zNose, to create models for the prediction of pleasantness for odors emanating from dairy operations, and compared the predictions to pleasantness values given by experts, obtaining a mean difference of only 1 point. Also GC-MS can be used for medical screening and/or for food monitorin: for example, Sklorz et al. [11] studied the detection of ethylene in fruits with a miniaturized GC; whereas He et al. [12] used a combination of surface acoustic wave sensor and GCs for an early screening of lung cancer. Several other applications of GC for non-invasive medical diagnostics can be found in the review written by Casas-Ferreira et al. [13] in 2019.

Finally, HSI is probably one of the sensor techniques which attracted the most interest in the last years, given the possibility to merge spatial and chemical information in “prediction maps”, in particular in the food industry. For example, in 2022 Yao et al. [14] used HSI to monitor egg freshness, in conjunction with statistical algorithms, obtaining a small mean error in the prediction; and Hu et al. [15] acquired hyperspectral images of Tibetan tea and analyze them with PCA-derivated techniques, achieving good prediction of quality. Several applications of HSI and artificial intelligence for quality assessment of fruit, vegetables, and mushrooms were reported by Wieme et al. [16]. Finally, also applications in other fields related to environmental monitoring were investigated: in 2015, Gagnon et al. [17] used HSI to monitor ships plumes in their operating environment; whereas Karaca et al. [18] explored the use of shortwave infrared HSI to sort waste materials like plastics, papers, glass, and metals, with a mean accuracy higher than 90%.

Information from spectra

The main problem related to all spectra, independently from the type of sensor used to acquire them, is that the information contained in them is extremely difficult to interpret, and, very often, it is impossible to differentiate between the useful information and the noise or the changes in the spectra due to variations of other uninteresting variables. Moreover, also the changes related to the VIs are not always easy to see: if we change some chemical content amount of a material, the resulting spectra give non-monotonic behavior at different wavelengths that are not discernable at first sight [19]. To use spectrum sensing at its full potential for monitoring purposes, we need to find an automatic way to infer the value (and the consequent changes) of the physical Variable of Interest (VI) from the acquired spectra. For example, let's say we want to monitor the soil moisture using an RF sensor, that acquires the reflection component of electromagnetic waves(this application will be discussed in detail in the next chapter). It is well known that the amount of water highly influences the reflection, but it does not exist in the literature an analytical expression that allows calculating the soil moisture starting from a spectrum. This is because the reflection is influenced also by a plethora of other physical parameters (soil and air temperature, type of soil, presence of detritus, etc.) which are not of interest to our application, and a user can't differentiate between the changes in the spectra which are caused by changes in the soil water content and the one caused by changes in other parameters. However, the prediction of the VI could be achieved easily by using a model created by statistical analyses: analyzing spectra in this way, precise and monotonic detection of the information could be obtained with simple computational effort, even if the useful part of the information is only a small part of the variations in the acquired spectra. As shown in Fig. 3, the model is created with two different inputs. One is a matrix **X** ($N \times K$) where every row N_i represents a single spectrum, measured for a number K of variables (wavelengths, frequencies, etc.), and acquired with different sensors and techniques based on the specific field or application. The other one is a **Y** array (NxJ), containing, for every one of the spectra N, the values of the J VIs (usually 1 or 2), measured with gold standard techniques. In the previous example, **X** would be made by the RF spectra acquired on soil samples (in a lab or outdoor conditions) with different moisture values, measured with a gold standard technique, that will be instead stored in the **Y** array. The predictive model created with these two inputs can find an array B that allows measuring the value $\hat{Y}$ of the soil moisture directly from the newly acquired spectrum $\hat{X}$, following the following formula [20]:

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{X}}\boldsymbol{B} \tag{1}$$

It is important to stress that eq. (1) has a very low requirement regarding computational power, and so, after the offline creation and tuning of the model, it is possible to implement the prediction of the VI in every type of embedded and/or autonomous device, minimizing the power and the time required for the calculation. Moreover, this process can be applied to every type of spectrum that shows variations due to changes of the VI, making it applicable to a plethora of fields and applications. Fig. 4 shows the whole process, already briefly described in the Workflow chapter: it starts from the offline acquisition of spectra and VI values, as well as the creation of the model (dotted line); then proceeds with the operating mode, where a new spectrum is acquired and processed by embedded hardware, obtaining the value of the VIs. Summing up, a suitable predictive model based on multivariate analysis could enhance any spectrum-based detection technique.

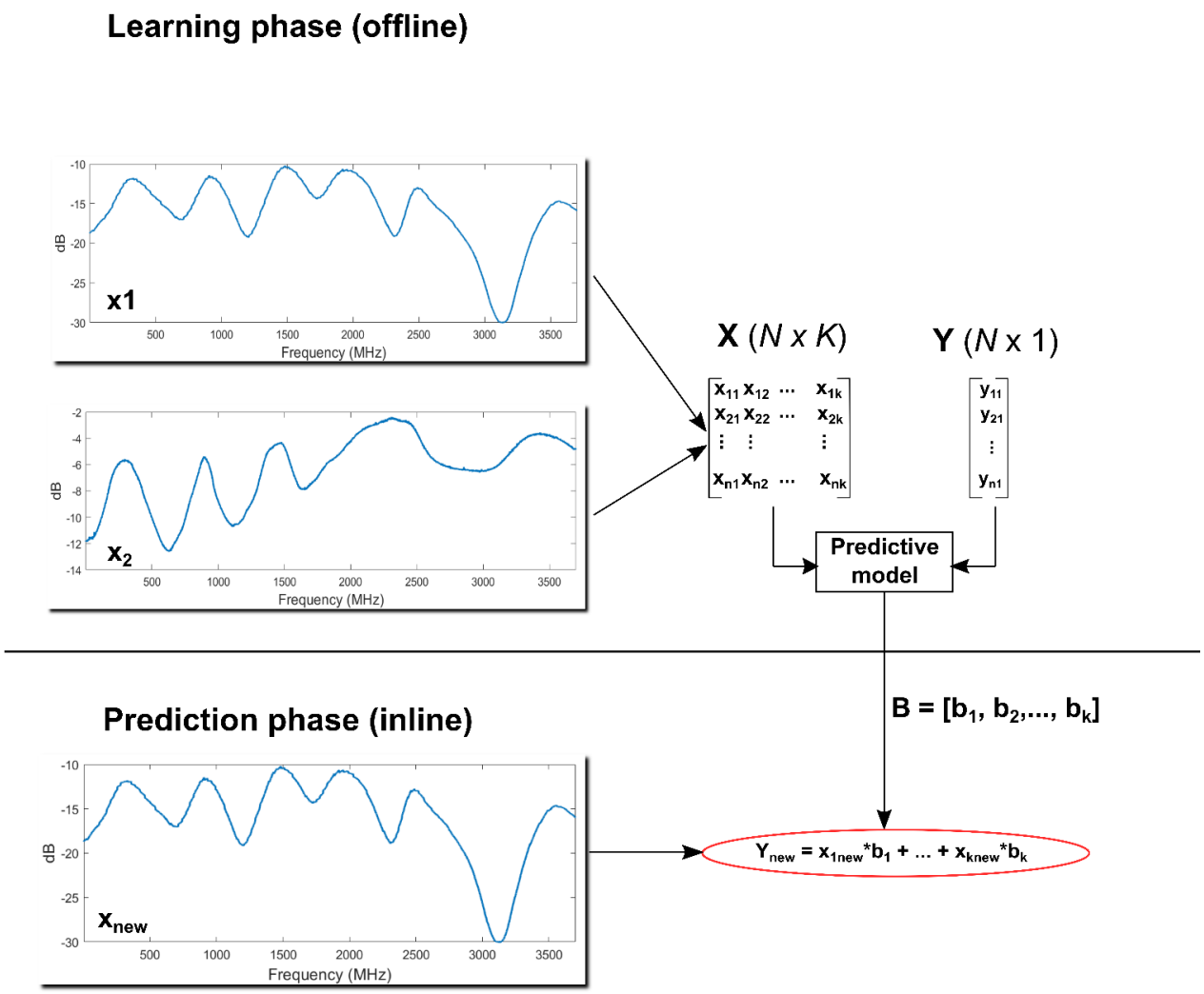

Fig. 3 Scheme of the creation of a predictive model (**top**) and its application for the calculation of the VI (**bottom**)

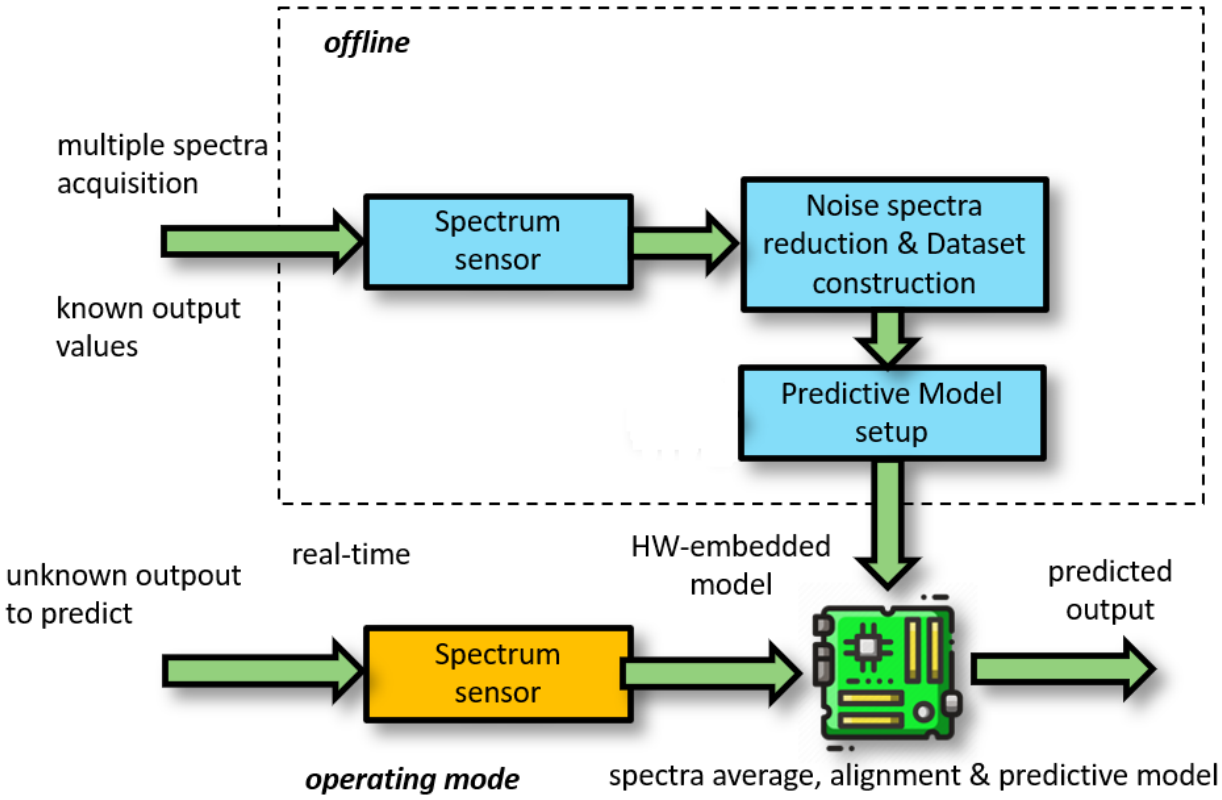

Fig. 4 Representation of the workflow used in the practical application for the prediction of the VI from the acquired spectra. The process starts with the acquisition of spectra and the corresponding VIs with gold-standard techniques, used as input for the creation of the predictive model (dotted box). Then the calibration coefficients are saved in the embedded hardware, allowing the prediction of the value of the VIs from new spectral acquisitions.

4.2 Mathematical tools

In this section, the mathematical foundations of the PCA-based MSA will be presented. In particular, the focus will be on PCA (the base of every other prediction technique, and still used alone to analyze the acquired data), Partial Least Square Regression (PLSR), used to solve regression problems, and Soft Independent Modelling of Class Analogy (SIMCA), used to solve classification problems. Then the basics of the preprocessing of the data and the evaluation of the prediction abilities of the models will be discussed. After that, the basic algorithm used for the implementation of PCA and PLSR (called NIPALS) will be presented in detail. Finally, these techniques will be compared with two other very popular families of predictive modeling techniques: Multiple Linear Regression (MLR) and Deep Learning (DL), showing why PCA-based MSAs are the most suitable for our applications.

PCA

PCA was first developed by Karl Pearson in 1901 [21] and re-developed independently by Harold Hotelling in 1936 [22], who also named it. It is based on the idea that a whole spectrum could be seen as a single point in a *K*-dimensional space, where *K* is the number of the acquired variables (frequency, wavelengths, u/e^- etc.). Generally, a group of *N* spectra could be defined as *N* observations described by a series of *K* variables or a cloud of *N* points in a *K*-dimensional space [23]. So, a spectra dataset is arranged in a matrix $\boldsymbol{X}\ (N \times K)$, also referred to as a "data matrix" containing *N* spectra, each defined by *K* variables. As stated before, in spectral applications K is usually higher than N, even by one or two magnitude order. Moreover, the useful information, carried by the changes in the spectra due to the influence of the VI, is spread between all the variables: these are the best choice for the spectra acquisition (measuring one frequency after another in ascending order), but not for the visualization of the useful information. So, we need to manipulate these data, finding a new dataspace with reduced dimensions where the variations between the spectra are clearer and easier to see and study. PCA solves this problem, by finding a subset $A < K$ of directions that still contain most of the total variance of data. From a geometrical point of view, these new directions, called Principal Components (PCs), define an *A*-dimensional subspace of the original *K*-dimensional data space, where the variance is included in descending order in the PCs: the first one will be the one explaining the higher percentage of the total variance, the second one the higher after the first, and so on. But how it is possible to find this new dataspace?

Generally, a data matrix **X** could be decomposed into two arbitrary matrices, $\boldsymbol{T}\ (N \times A)$ (score) and $\boldsymbol{W}\ (A \times K)$ (loadings), so as $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}'$, where $\boldsymbol{W}$ is an orthonormal matrix. Following classical regression theory, it could be shown that optimal $\boldsymbol{W}$ minimizes the error residual matrix $\boldsymbol{E}\ (N \times K)$, where

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{W}' + \boldsymbol{E} \tag{2}$$

so that we can find the best projection of $\boldsymbol{X}$ (prediction) along loadings directions, in a dot product fashion, as $\widehat{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{W}'$. It could be shown that loadings are the eigenvectors of the correlation matrix $\boldsymbol{X}'\boldsymbol{X}$. However, the calculation of all eigenvectors (as already stated, the original variables are usually in the order of thousands or tens of thousands) would be computationally intensive and subject to experimental noise, especially for lower eigenvalue values. PCA approach reduces to *A* the number of directions (PCs) along which calculate maximum variance, using numerical iterations (that will be presented later) and uses the residual matrix $\boldsymbol{E}$, containing the difference between the original data

matrix ***X*** and the new representation in the PCs space, as a control of the convergence to stop iterations. In this way, the original information is still contained in the new dataset, but it will have only a few dimensions (usually, from 3-4 to 10) and they will be ordered to the variance, allowing us to easily see trends and clusters in the data even with a simple plot between the first two PCs. To sum it up, the PCA algorithm allows us to find the value of the loading that minimizes the residual error, maximizing in this way the projection of the data in a simpler dataspace. Doing so, we could say that **T** is an acceptable summary of the original **X** matrix, because, even if it has a reduced dimension, it still contains most of the information. A geometrical representation of PCA is depicted in Fig. 5 (for the first 2 PCs). In the image, the two PCs, represented by orange lines, define a PCA model with 2 variables (dotted orange line).
If the error was already low enough, these 2 PCs might be enough to describe well the acquired spectra contained in **X**.

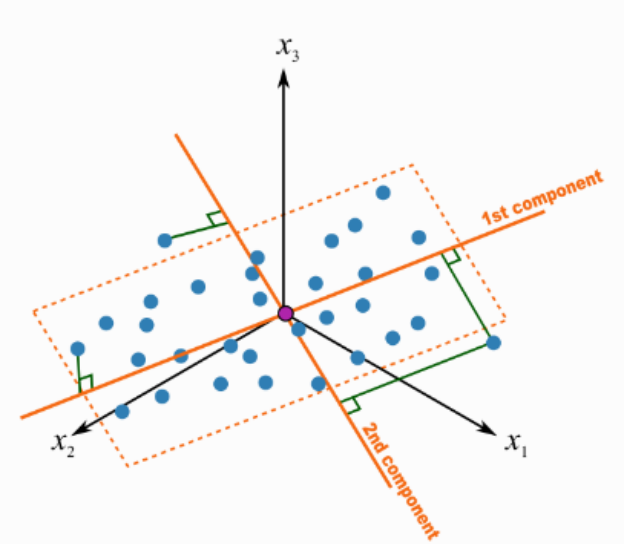

Fig. 5 Depiction of the results of a PCA analysis. The dotted line represents the plane individuated by the first 2 PCs [24].

PLSR

From an applicative point of view, PCA is very useful to easily identify trends and clusters in the data, and more in general to understand if the acquired spectra contain useful information for the prediction of the VI. However, it is not possible to directly obtain a prediction from it: we need to use one of the several techniques that were developed starting from it. For the applications that will be presented in the following chapters, I focused mainly on an algorithm called Partial Least Square Regression (PLSR), an evolution of PCA for solving regression problems, introduced by the Swedish statistician Herman O. A. Wold [20]. Whereas PCA is used to study the **X** dataset and find clusters

or trends in the data, PLSR allows to create of a predictive model for a VI, inferring in this manner an unknown variable value starting from an acquired spectrum. To do this, in addition to the **X** dataset, also a **Y** initial dataset is required, containing the value of the VI linked to every one of the *N* spectra which form the **X** matrix. In PLSR, the decomposition of the **X** matrix is very similar to the one performed in PCA and could be described with (2). However, the new directions found by the algorithm are not the previously described PCs: in fact, it is not said that the directions that contain most of the **X** variance are also the directions that best explain the **Y** dataset [25]. Therefore, PLSR identifies slightly different directions, called Latent Variables (LVs), which try to maximize at the same time the variance of **X**, the variance of **Y,** and the covariance between the two. So, PLSR finds the score and loadings for both the input matrixes, following eq. (2) and this one for **Y**:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}' + \boldsymbol{F} \tag{3}$$

Where U, C, and F are the score matrix, the loading matrix, and the error matrix, respectively, of **Y**. Finally, the covariance (between the 2 scores) could be expressed by:

$$Cov(\boldsymbol{T}, \boldsymbol{U}) = Correlation(\boldsymbol{T}, \boldsymbol{U}) \cdot \sqrt{\boldsymbol{T}'\boldsymbol{T}} \cdot \sqrt{\boldsymbol{U}'\boldsymbol{U}} \tag{4}$$

Eq. (4) unites the 3 conditions expressed before:

1) The X matrix in the new dataspace ($\sqrt{\boldsymbol{U}'\boldsymbol{U}}$)
2) The Y matrix in the new dataspace ($\sqrt{\boldsymbol{T}'\boldsymbol{T}}$)
3) The relationship between X and Y ($Correlation(\boldsymbol{T}, \boldsymbol{U})$)

Mathematically, this could be summarized by the fact that in PLSR the score matrix **T** of the **X** matrix is not only a good predictor of **X** but also of **Y** ($N \times M$):

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{C}' + \boldsymbol{F}; \tag{5}$$

This equation could also be easily interpreted from a geometrical point of view: **C** identifies the best direction in the A-dimensional subspace **T** that also shares the maximum covariance with the matrix **Y** [24]. Therefore, (2) and (3) can be merged:

$$\boldsymbol{Y = TC' + F = XWC' + F = XB + F} \tag{6}$$

From (6) is easy to see how the prediction of the VI is accomplished by the PLSR: the algorithm estimates an array of coefficients **B** ($K \times 1$), which allows us to use the linear eq. (1), easily finding the value of the VI $\widehat{\boldsymbol{Y}}$ starting from a newly acquired spectrum $\widehat{\boldsymbol{X}}$**.** A summary of the whole technique is depicted in Fig. 6.

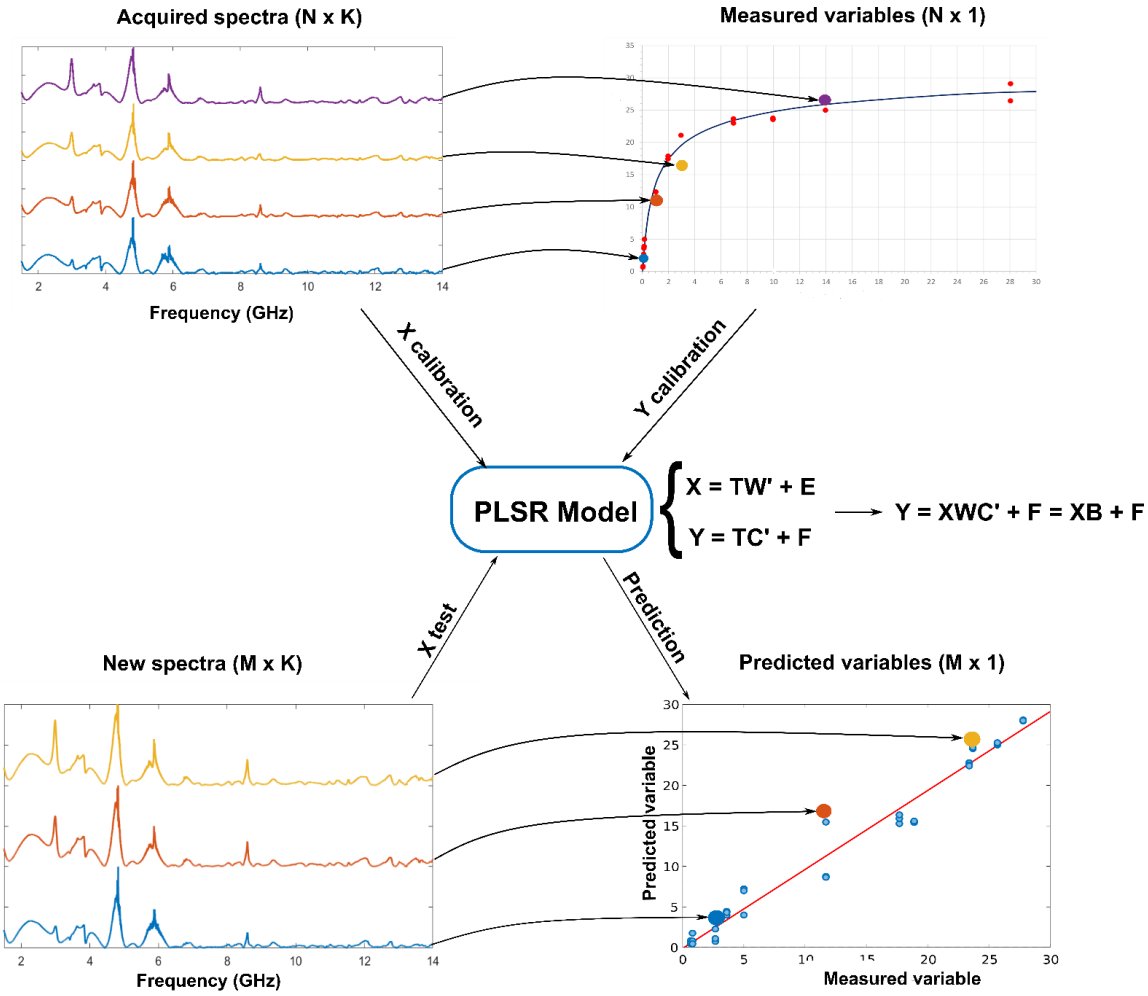

Fig. 6 Partial Least Square Regression (PLSR). The X calibration dataset (upper-left) represents the spectra obtained with a spectral sensor, each of these linked with the respective VI (Y calibration dataset, upper right), measured with gold-standard techniques. These two datasets are used as input for the creation of a PLSR model and the calculation of an array of calibration coefficients B. This allows us to give the model new spectral inputs (X test dataset, lower left) and obtain as an output the relative VI value. In the lower right, we can see a plot of the prediction of the model vs the measured values: the closer the point is to the bisector (red line), the better the prediction ability of the model.

SIMCA

The PLSR algorithm is very powerful and often allows the user to obtain a prediction of the VI's precise value. However, for some applications is not necessary to obtain quantitative predictions, but only qualitative ones, assigning every new acquisition to a class (for example over or under a certain threshold value). To solve this type of classification problem, it is possible to use another technique derived from PCA, called Soft Independent Modeling by Class Analogy (SIMCA), ideated by Wold and Sjostrom [26]. This approach could be divided into two subsequent steps. In the first one, a PCA analysis is carried out separately on the spectra of each cluster class, finding a series of A PCs for each of them, all contained in the original K-data space (A could be different for each class). From the geometrical point of view, projection subspaces are defined for each class. Then, for each new acquisition submitted to the model, SIMCA calculates for each class two statistical parameters, called Q residuals and Hotelling's T^2, describing how well the new sample belongs to these classes [27]. In particular, the Q residual, also called Squared Prediction Error (SPE), measures the squared error of the projection of the new sample in the PC spaces [28]. Thus, it indicates the amount of information lost in the process, measured as the distance from the PC's subspace [29]. In general, for an i-th observation, the Q-residual related to a given cluster is calculated as

$$Q_i = e_i e_i', \tag{7}$$

where e_i is the residual of the projection of the observation in the PCA models calculated by the SIMCA analysis, and it is a row of the $\boldsymbol{E}$ matrix for the related cluster. On the other side, Hotelling T^2 measures the difference between the new observation and the mean value of the model. From a geometrical point of view, it measures the Euclidean distance between the center of the PCs subspace and the projection of the observation, and it is calculated as

$$T_i^2 = \frac{t_i t_i'}{\lambda}, \tag{8}$$

where t_i is the i-th observation *score*, that is, the projection value of the observation along all the considered PC variables (it is a row of the $\boldsymbol{T}$ matrix); λ represents the distribution variance [28]. Note that T_i and Q_i are orthogonal contributions. For a given new observation, SIMCA combines these two parameters in a final parameter d_i, with the following formula

$$d_i = \sqrt{(Q_i^2 + T_i^2)}. \quad (9)$$

The observation is finally assigned to the class with the lowest d_i value: it is the class that better describes the new spectrum. Fig. 7 shows a geometrical visualization of the prediction process for a new sample between two classes.

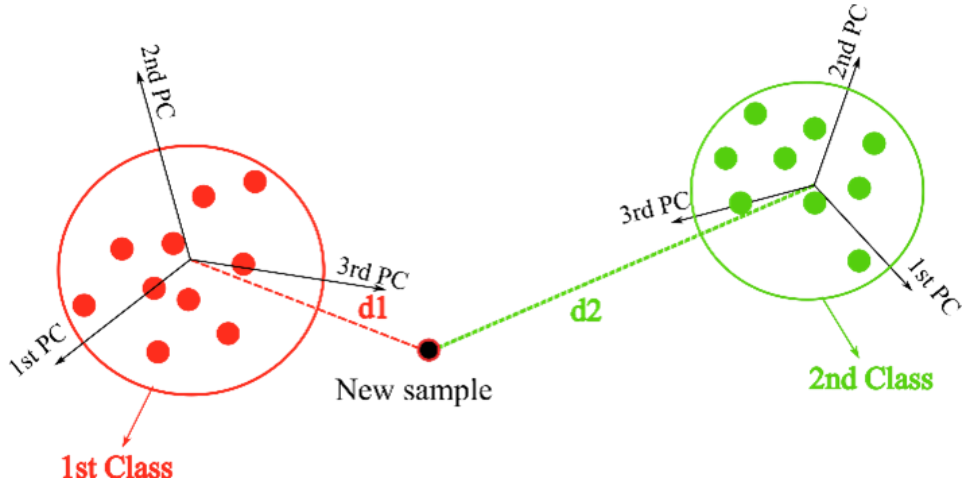

Fig. 7 Visual representation of the prediction of a new sample with a SIMCA algorithm. Every point represents a spectrum, whereas PCs create a subspace in the k-dimensional space of the original variables (frequencies). In the present case, *d* is defined by eq. 8.

Pre-processing

It is important to point out that, for all types of analyses, the **X** dataset needs to be pre-processed before being used. Performing an MSA on raw data is often useless: these types of analyses are focused on the variations of data, while the actual values are not important. So, the **X** dataset is usually mean-centered, subtracting from every variable the corresponding mean. Sometimes is also important to change the scale of data, especially if measurements acquired with different instruments are mixed: if a type of measurement has values much higher than the others, it could "cover" the useful information contained in the smaller ones. This is avoided by dividing the spectra by the standard deviation. These two pre-processing techniques are fairly common, and the application of both of them is called autoscaling [30]. After that, other preprocessing algorithms can be applied depending on the particular application and type of spectrum, to reduce the noise or enhance the PCA results: the most common ones are Savitsky-Golay smoothing and derivatives[31], Multiplicative Scatter Correction [32], variable alignment, etc.

Choice of the variables number and model validation

One of the main features of MSA is the huge reduction of the problem complexity, moving from thousands of variables to a few of them, aiding both the interpretation and the visualization of the data, and aiding at the same time also the statistical analysis itself [33]. The choice of the right number is up to the user, and the correct answer often depends on the scope of the performed MSA. If it is performed to enhance the visualization and clarity of the data (PCA is often used for "explorative study", as will be shown in chapter 3), is not necessary to set a specific number of new variables, but limit them to the first few: the majority of the data variance is often contained by the first 2-3 PCs, and each extra component after them do not add a lot [34]. However, for other applications, especially if the model will be used to predict the VI from new spectra (so PLSR/SIMCA model, etc.), it is necessary to have a stricter selection of the number of variables, to divide the useful information and the noise in the best way possible. This is done thanks to a process called Cross-Validation (CV), first introduced by Wold for PCA [35], that allows testing the prediction ability of the model without needing other acquisitions. The main idea behind the CV approach is to divide the input datasets into a series of subsets, then set aside them and create a model with the rest. Finally, the left-out subset will be given as input to the model, and the prediction results compared to the real values, obtaining an initial estimation of the error thanks to a parameter called Predicted REsidual Sums of Squares (PRESS), calculated as:

$$PRESS = \sum_{i=1}^{I} \sum_{j=1}^{I} (e_{ij}^{(r)})^2 \tag{10}$$

Where $e_{ij}^{(r)}$ is the difference between the prediction and the real value for the variable *j* of sample *I*, considering *r* components. From that, the Root Mean Square Error for Cross Validation (RMSECV) is calculated:

$$RMSECV = \sqrt{\frac{PRESS}{IJ}} \tag{11}$$

This process is repeated for every subset, and a mean RMSECV trend vs the number of variables is plotted at the end, making it possible to test the model on itself [11]. RMSECV is a very useful parameter for the selection of the number of variables because it will show an increase when the variables containing noise start to be added, clearly indicating the end of the useful information. It is also possible that its trending is always negative: in this case, the number of variables can be chosen by looking at how much each variable adds to the overall reduction of the error (as shown in Fig. 8). Finally, it is important to point out that for the practical applications, it is always recommended to test the prediction ability of the model also with an external test set, not used for the calibration or the CV. The obtained parameter is called Root Mean Square Error in Prediction (RMSEP), and, if the test set is chosen carefully, is the parameter that gives the most accurate indications about the model prediction ability.

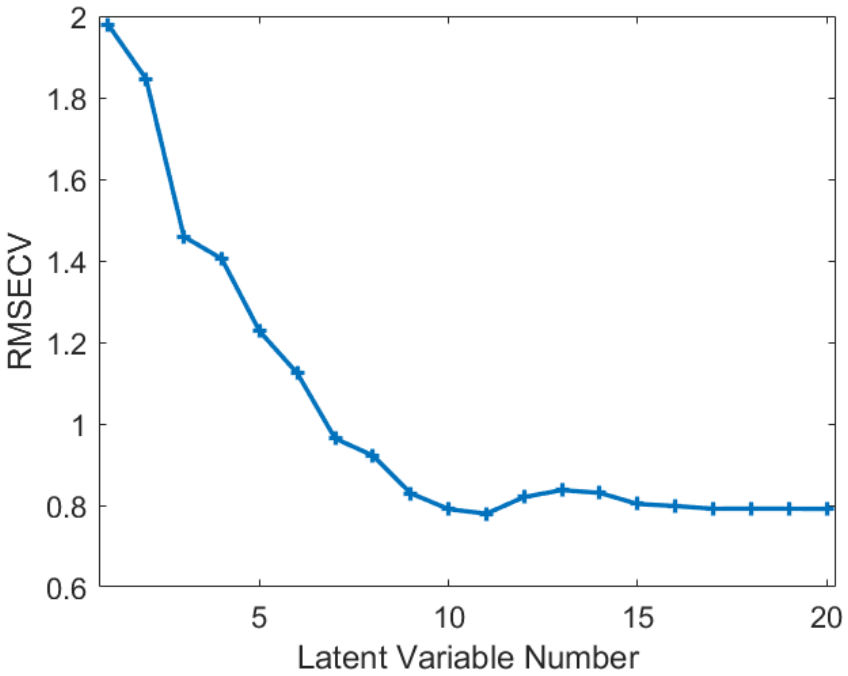

Fig. 8 Example of an RMSECV plot for the selection of the number of variables. It shows clearly that the variables after the 9th do not add a sufficient amount of useful variance. In fact, after the 11th variable, the RMSECV starts to increase, highlighting the presence of noise in the following variables.

NIPALS

Several different algorithms were developed in the last 50 years to implement both PCA and PLSR. Of them, the most used and well-known is called Non-linear Iterative PArtial Least Squares, which presents two main advantages: it handles missing data, and calculates the components sequentially, allowing the user to stop the calculation when the first (and the most valuable) directions are computed.

The actual simpler way to implement the PCA algorithm is to use a procedure called Eigenvalue Decomposition [24]: the loadings that form the **W** matrix are the eigenvectors of the matrix **X**'**X**, and it is possible to perform a selection of the variables by calculating the corresponding eigenvalues and ordering them from largest to smallest, and then choose arbitrarily a subset of them. However, this algorithm is never used, because it requires the calculation of all the eigenvectors, that working with spectra can reach a number in the order of ten of thousands, requiring excessive computational power and time.

The steps necessary for the calculation of a single PC with the NIPALS algorithm (in this example the first one) are 4 [36]:

1) A first representation of the scores t_1 is created, it could be random numbers or one of the columns of the **X** matrix.
2) Every column of **X** is then regressed onto t_1, and the regression coefficients are saved in the array p_1. This is done with an ordinary least square regression ($y = \beta x$), where the x variable is t_1 and the y variable is the k column of **X.** Considering β as $w_{k,1}$, the solution is given by the formula:

$$w'_1 = \frac{t_1'X}{t_1't_1} \tag{12}$$

Where t_1 is a $N \times 1$ column vector, and X a $N \times K$ matrix. This process is represented in Fig. 9(a).

3) The vector w'_1 is rescaled to have a magnitude of 1.0.

$$w'_1 = \frac{w_1'}{\sqrt{w_1'w_1}} \tag{13}$$

4) Every row of **X** is regressed onto w_1, calculating in this way new values for the starting vector t_1:

$$t_1 = \frac{Xw_1}{w_1'w_1} \tag{14}$$

where x_1 is a $K \times 1$ vector. This process is represented in Fig. 9 (b).

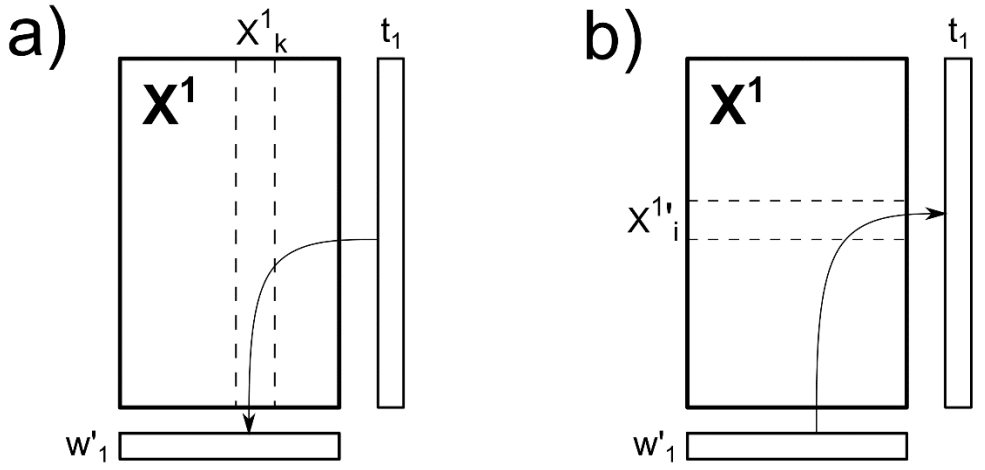

Fig. 9 **(a)** Regression of one of the columns k of X onto t1, finding the k-th value of p1 **(b)** Regression of row i of X onto t1, finding the i-th value of t1

Steps 2,3,4 are then repeated until the changes of t_1 are in the order of $10^{-6} – 10^{-9}$. Finally, the vectors t_1 and p_1 are saved as the first column of the matrices **T** and **P**, respectively. To calculate the other PCs, it is necessary to deflate the matrix **X**, eliminating from it the variability related to the first PC, already calculated. The new X matrix is calculated as

$$\boldsymbol{X}_2 = \boldsymbol{X}_1 - t_1'w_1 \tag{15}$$

This also assures the orthogonality between the different components: the second PC is calculated on data that not contains the variability of the first one, assuring that they can not explain the same variability.

NIPALS (PLS)

It is possible to use NIPALS also for the implementation of PLSR, with some adjustments that also account for the presence of an input dataset Y, containing the values of the VI. The starting point is now the selection of one of the columns of Y^1 as the initial u_1. Then the algorithm is developed with the following steps, reported also in Fig. 10 for the calculation of the first PC [25]:

1) Every column of X^1 is regressed onto the vector u_1, and the results are saved in the array p_1:

$$p_1 = \frac{X'^1 u_1}{u'_1 u_1} \tag{16}$$

2) The vector w_1 is normalized:

$$p_1 = \frac{p_1}{\sqrt{p'_1 p_1}} \tag{17}$$

3) Every row of X^1 is regressed onto the vector p_1 and the results are saved in the array t_1

$$t_1 = \frac{X'^1 p_1}{p'_1 p_1} \tag{18}$$

4) Every column of Y^1 is regressed onto t_1, and the data are saved in the array c_1

$$c_1 = \frac{Y'^1 t_1}{t'_1 t_1} \tag{19}$$

5) Finally, every column of Y^1 is regressed onto c_1, and the data are saved as the new values of the array u_1

$$u_1 = \frac{Y'^1 c_1}{c'_1 c_1} \tag{20}$$

As before, the steps are repeated until convergence. Then the loadings w_1 of the X matrix for the array t1 are calculated. They are not part of the final PLSR model, but they are used for the deflation of the **X** matrix. The t_1 scores are calculated with eq. (14), and the X^1 matrix is deflated using eq. (15). Also Y^2 is calculated in a similar way, using the loading c_1 instead of w_1:

$$Y_2 = Y_1 - t_a c'_a \tag{21}$$

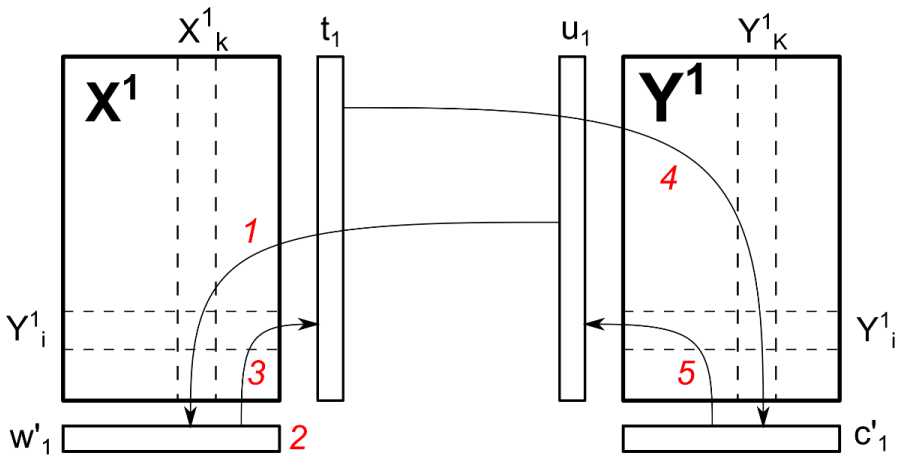

Fig. 10 Representation of the NIPALS algorithm for PLSR

Comparison with MLR and Deep Learning

In this section, the PCA-based technique presented in the previous sections will be compared to two of the most well-known and diffused families of MSA techniques, Multiple Linear Regression (MLR) and Deep Learning (DL), explaining also the problems in the use of the latter two with embedded systems that acquire spectral data.

At first glance, MLR may seem ideal for the creation of a predictive model for spectra, because it only requires as input the two datasets **X** and **Y**, and derive the array **B** directly from them, following the formula:

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{22}$$

Moreover, (22) is also the matrix notation of the Ordinary Least Squares [37], another well-known and simple method for the creation of linear regression models. So, MLR these methods may seem simple and fast ways to obtain the prediction array B. However, to have this algorithm work properly, the input dataset **X** needs to respect a series of conditions. In particular:

1) The *K* columns of **X** (representing the measured variables, in our case usually frequencies or wavelengths) need to be uncorrelated

2) The number of observations (the *N* rows of **X**) needs to be bigger than *K*

3) It needs to be noise-free

4) It cannot have missing values

In spectral applications, it is practically impossible to obtain these conditions, especially the first two: spectra are usually highly correlated and during the acquisition campaign is easier to acquire few measurements on a greater range of variables (so N<<K) with respect to the opposite. Moreover, it is difficult to have spectra with a very high Signal Noise Ratio (SNR), especially monitoring small concentrations or changes in the VI. For these reasons, using spectral data on MLR will result in a model prone to high variability in the results: small changes in the **X** data may cause great variations in **B**, even changing the sign of the elements, and the obtained results will have wide confidence intervals for the coefficients, undermining the usefulness and reliability of the whole model [24]. This is visualized in Fig. 11, which represents two variables, x1 and x2, that are strongly and positively correlated. If these two variables are used to create a prediction model with MLR, the resulting $\widehat{\boldsymbol{Y}}$ would be represented by the plane defined by $b_0 + b_1x_1 + b_2x_2$, which will minimize the residual error. However, very small variations in the **X** data will cause huge variations in the **B** values, and consequently in the solution: the plane will rotate around its axis (dashed line). So, we will obtain

very different values of b_1 and b_2 depending on the rotation, whereas the real function minimum is presenting only small changes. This will result in large confidence intervals for the coefficients since the off-diagonal elements in $\boldsymbol{X'X}$ will be large [24].

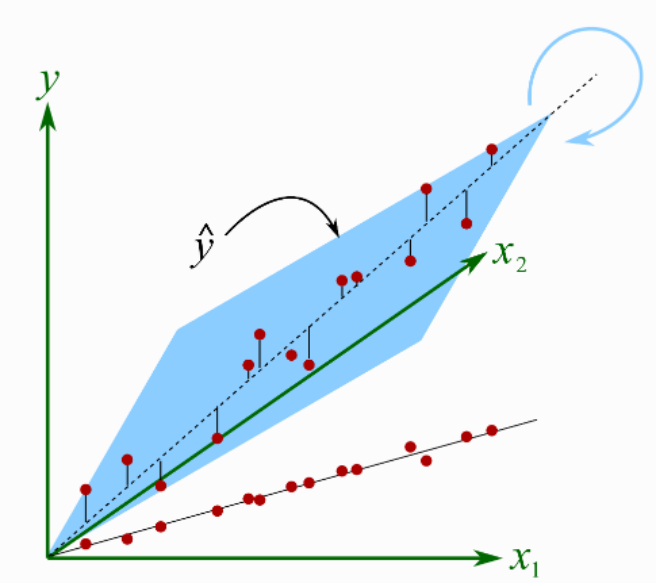

Fig. 11 Depiction of the effect of small changes in **X** values, causing a rotation of the solution plane. [24]

The usual solution used for MLR is to select only the useful variables, obtaining an uncorrelated small fraction of the original **X** matrix. However, in spectral applications, *K* is usually very large, and the computational burden requested for the variable selection is too high, considering also the fact that it is not always easy to understand how many and which columns need to be selected. PCA-based algorithm solve these problems: they find a subset of the original variables, reducing the problem dimension from thousands to usually under 10. Moreover, the selected variables are the one that contains the majority of the data variance, selecting in this way only the useful part of the spectral data, and eliminating what we consider noise for our prediction model [34].

Regarding the comparison with Deep Learning (DL) techniques, the previous discussion about the implementation algorithms of PCA and PLSR shows clearly that their mathematical foundations are clear and not too complex to understand. This clarity of the basic concepts is for sure one of the reasons that made us prefer these techniques over more powerful but complex DL ones. In particular, Artificial Neural Networks (ANN) have attracted great interest in the last 5-10 years: these algorithms use multiple layers to progressively extract higher-level features from the raw input until the value of the chosen Variable of Interest (VI) is obtained as output. This type of technology can solve complex problems, like the automatic selection of objects in images or vocal recognition, and usually obtains excellent results in prediction, but present some technical issues for the implementation in embedded and low-power systems. They require a lot of computational power for convolutional calculations, especially for the analysis of images. For this, it is not possible to train an ANN algorithm on simple computers, but a set of dedicated hardware, for example, GPUs, is required to do it in an acceptable amount of time. Moreover, they need a great amount of memory to store the weights used for the prediction (up to hundreds of Mbytes) and a huge input dataset for the training of the ANN and the calculation of the prediction weights. The typical DL database contains several samples ranging from 10000 to millions. For these reasons, it is not always possible to implement these algorithms in an embedded device: Internet of Things (IoT) applications require low computational resources due to the lack of power supply reserve, and it is not always possible to create a huge input dataset in an acceptable period during the testing phase of these algorithms in a specific application.

MSA techniques are less powerful compared to the ANNs, but they have a series of great benefits: their mathematical foundation is simpler but solid and easier to understand, to the “black-box” effect of DL algorithms, where it is very difficult, or impossible, to understand what happens to the weights of the inner layers, and only see the output result. Moreover, MSA needs smaller input datasets, usually in the range of 100-2000 samples, and the prediction weights occupy far less memory (4-8

Kbytes). Finally, the MSA has significantly lesser computational requirements, and it is worth stressing that its use is based on two phases, the first of which is performed offline and consists of processing reference datasets to set up the predictive model before the actual measurements, whereas the second consists in using the predictive model with newly acquired datasets. Once the predictive model is completed, we obtain the weight matrix B, which is then implemented in the chosen hardware for digital processing. More precisely, for a given acquired input spectrum X, a basic output is simply given by (1). The computational requirements for the matrix product are the execution of K sums of products, either in fixed- or floating-point notations, resulting in linear complexity (O(K) complexity). Such linear complexity is very much treatable for real-time analyses even by single-core low-cost CPUs, microcontrollers, or DSPs. If N input spectra are acquired and M output variables are required, computation times for alignment increase by N, and matrix products by N*M. Then, with N and M in the order of unity, computation times are still expected in the order of seconds [38]. To summarize, once the predictive model is set up, the multivariate approach allows for the execution of challenging achievements with relatively low computational cost and on low-cost hardware compared to other approaches.

4.3 Practical applications examples

In the next chapters, 4 different applications of PCA-based algorithms to spectral data will be presented. Despite the very different scientific fields and the electronic device used, the workflow will be always the same, already presented and depicted in Fig. 1 and Fig. 3. In particular, these applications will be the following ones:

1) Prediction of soil moisture, acquiring reflection electromagnetic spectra with an open-ended waveguide (RF spectrometry)
2) Prediction of concrete moisture, acquiring reflection electromagnetic spectra with an open-ended waveguide (RF spectrometry)
3) Prediction of fish freshness (in days), acquiring hyperspectral images on sardines and anchovies samples with a hyperspectral camera (Hyperspectral Imaging)
4) Prediction of gas concentrations in mixes, acquiring mass spectra with an innovative gas chromatograph (Mass spectrometry).

For all of them, a brief introduction and a description of the state of the art will be given at the start of the chapter. Then, the sensing device used will be presented in the detail, as well as the acquisition campaign that was performed and the consequent application of MSA to these spectra. Finally, the statistical results will be discussed, and a brief conclusion will sum up the work already performed and possible future developments, as well as discuss their limits and criticisms.

5. Soil moisture content detection

The purpose of this chapter is to show a compact and fast sensing architecture approach for soil moisture, based on RF spectroscopy and MSA, applied to a real soil environment. Therefore, the measurements are made using a compact electronic architecture in a practical environment and not by calibrated instruments in laboratory conditions. The starting point for this application was an article previously published by our team [39], in which the potentiality of a non-invasive technique based on an open-ended waveguide was investigated, starting from "gain" (defined as the ratio between the power of the received and the emitted waves) and "phase" data (defined as the difference of phase between received and the emitted waves). Spectral data were acquired on samples of different types of soils in controlled lab conditions. The results of that work, obtained in combination with PLSR and multi-way PLS tools, showed that a different moisture content (%) in the soil involves changes in both "gain" and "phase" waveforms. In this chapter it will be shown that with PLSR it is possible to create a calibration model from the "gain" and "phase" data measured by the system, then use the calibration coefficients to implement the calculation of the moisture value directly in the device microprocessor. The obtained system can be considered an innovative approach compared with the traditional techniques, avoiding invasive probes, being independent of data post-processing, and having good accuracy, thanks to the proposed statistical tools.

5.1 State of the art

In the last decade, indirect techniques for soil moisture assessment have become increasingly important as alternative tools to the standard time-consuming thermo-gravimetric method [40]. These techniques are based on the assessment of physical and chemical soil properties and are crucial in the new "precision agriculture" application field. Among these, a considerable part of the literature is focused on the inference of moisture on soil dielectric properties [41], variables that describe the electric polarization of the matter when subject to an external electric field. The complex (or apparent) relative dielectric permittivity is defined by real and imaginary components. The real component accounts mainly for energy stored in the system owing to the alignment of dipoles with the electromagnetic field. Usually, applications in active and passive remote sensing use the change of this part of the dielectric constant because it contains most of the information [42]. The imaginary component (related to the dielectric loss factor) is mainly due to molecular relaxation and accounts

for energy dissipation effects [40]. Even if a smaller part of the information resides there, it could be utilized to boost the sensing performance [39]. In addition to the temperature and frequency of the electromagnetic field [39], [43], dielectric properties are also a function of parameters such as moisture content [44], bulk density [45], and soil constituents [43], which could thus be described thanks to a dielectric characterization, making it a fundamental technique for indirect soil sensing methods. Several attempts were made to develop non-invasive systems based on microwave reflection, especially in the spectral range of near-infrared (NIR), exploiting multivariate statistical algorithms to predict moisture values. Some applications were created to be used in movement, attached to the subsoiler chisel or shank of a tractor [46], [47], whereas the majority of them were immobile, directly in contact with the soil or just above it. They are based on different technologies: LEDs [48], optical units [49], antennas [50], [51], bistatic scatterometers [52], and waveguides [39], [53]. Finally, some works focus more on the application and comparison of different statistical analyses, using existing spectrometers to acquire data [54].

5.2 Device Description

Physical Sensing Approach

The main components of the system are shown in Fig. 11. It could be divided into two fundamental parts: an open rectangular waveguide (96 mm x 245 mm x 46 mm), containing Tx and Rx antennas, and a plastic box (200 mm x 250 mm x 100 mm) containing the electronic circuit. Fig. 12 also highlights the main components involved in the system's general functioning: a series of electromagnetic waves are created by an RF source, and then transmitted to the soil through the open-ended waveguide. The reflected waves are read by the Rx antenna and used as input for a data analysis circuit to obtain “gain” and “phase” data. These are multiplied in the microcontroller unit (MCU) with calibration coefficients calculated using multivariate analysis to obtain soil moisture estimation. To measure this parameter, the system exploits the concept of impedance spectroscopy, an indirect method based on physical interactions between the matter under test and an electromagnetic excitation. At the interface between the two different materials, air and soil, rapid variations of physical and electric properties cause refraction and reflection phenomena in the wave.

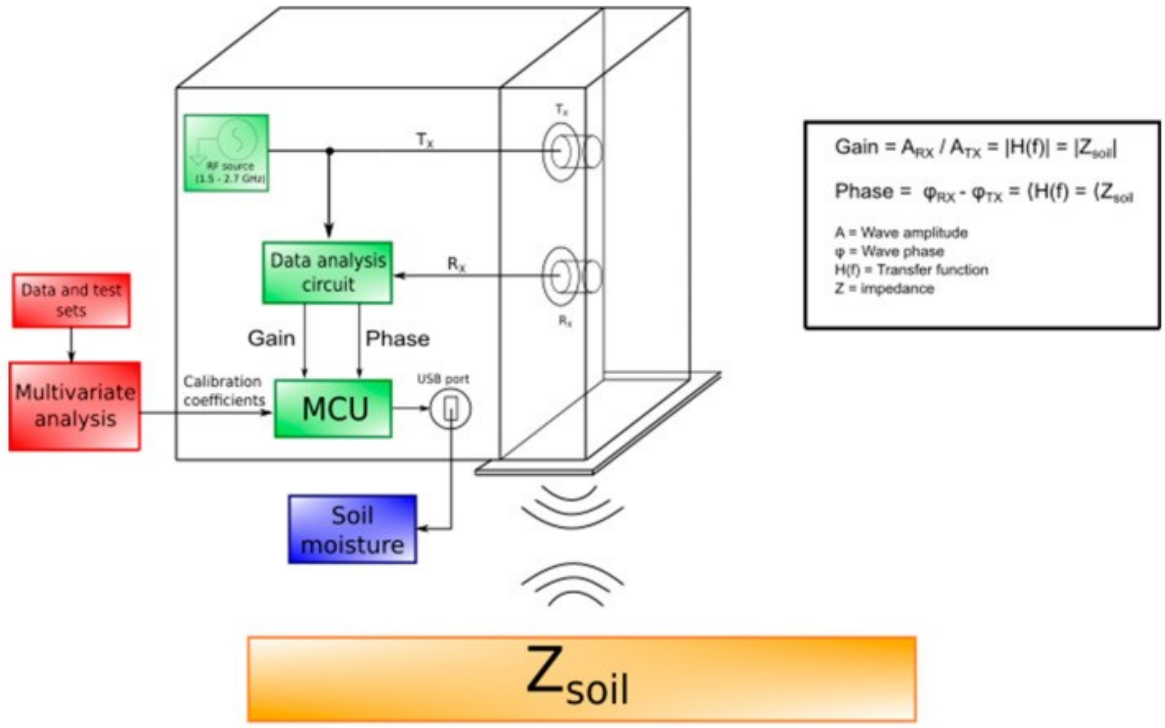

Fig. 12 System general functioning. A data analysis circuit calculates "gain" and "phase" data, using as input transmitted and received electromagnetic waves. These data are used to calculate the soil moisture directly in the microcontroller unit (MCU), thanks to calibration coefficients obtained from a multivariate statistical analysis.

The transmitted and reflected waves can be described, respectively, as follows:

$$T_x(t) = A_{Tx} e^{j\varphi Tx} e^{j2\pi ft} \tag{22}$$

$$R_x(t) = H(f) e^{j\varphi Rx} e^{j2\pi ft} = A_{Rx} e^{j\varphi Rx} e^{j2\pi ft} \tag{23}$$

where A is the wave amplitude; j is the imaginary unit; f is the frequency; t is the time; φ is the phase; and $H(f)$ is the transfer function between x and y, represented by the soil impedance [55]. It is possible to see that the relationship between the amplitudes and the phases of the two signals is as follows:

$$A_{Rx} = |H(f)| A_{Tx} \rightarrow \frac{A_{Rx}}{A_{Tx}} = |H(f)| = |Z_{soil} \tag{24}$$

$$\varphi_{Rx} = \langle H(f) + \varphi_{Tx} \rightarrow \varphi_{Rx} - \varphi_{Tx} = \langle H(f) = \langle Z_{soil} \quad (25)$$

These equations explain that the reflection of an electromagnetic wave is still a wave, with the same frequency as the transmitted one, but with a different amplitude value and phase shift. H(f) is determined by the ratio between A_{Rx} and A_{Tx}, called "gain", and the difference between φ_{Rx} and φ_{Tx}, called "phase". The frequency range of electromagnetic waves was chosen to be 1.5–2.7 GHz, maintaining the same range that gave very good results in previous work [39]; in this spectral region, it is possible to obtain information about soil moisture and, at the same time, design the whole system practically and compactly. The waveguide dimensions (96 mm x 245 mm x 46 mm) were chosen to obtain a cut-off frequency equal to 1.56 GHz.

Electronic Architecture

The measurement system is composed of three principal components: a data-control and elaboration system, an RF source, and a gain-phase detector. The first consists of a microcontroller, a D/A converter, and a serial–USB converter. The selected microcontroller is a MICROCHIP PIC24FJ256GB606. It manages the measurement process and the communication with the serial interface (UART/USB converter cable). This microcontroller has a few interesting features, like SRAM data storage with a capacity of 32 Kbytes and 16-bit addresses, making PIC24F the ideal microcontroller for data-logging applications of significant amounts of data. The D/A converter is a 16-bit Analog Devices AD5761R. It connects the MCU with the RF source, translating the digital value provided by the microcontroller into an analog voltage thanks to a voltage ramp from 0 to 20 V, approximately. It has a resolution of 16-bit, a unipolar or bipolar output, an output noise equal to 35 nV/Hz, a maximum integral nonlinearity (INL) of ±2 Least Significant Bit (LSB), and a maximum output settling time of 12.5 µs (with a step equal to 20 V). The RF source consists of a VCO MiniCircuits ZX95-2700A, that translates the output voltage of the DAC into an ideally sinusoidal wave at a frequency dependent on the input voltage. Its most interesting features are the frequency of the wave generated from 1.5 GHz to 2.7 GHz (compatible with the project specifications), the typical output power of 3.3 dBm, the low phase noise, the tuning voltage from0.15 V to 25 V, the supply voltage of 5 V, and the maximum power supply current of 35 mA. The signal coming out of the VCO has a power of around 4 dBm at maximum, which is too low compared with the design specifications, so it is necessary to perform an amplification to reach the minimum transmitted power through an RF amplifier circuit. The chosen component is the QORVO TQL9092, an ultra-low noise amplifier (LNA) with an operating band from 0.6 to 4.2 GHz. Finally, the signal amplified by the LNA is

supplied to the rectangular guide, which generates waves and collects the reflected ones, thanks to a Tx and an Rx antenna. The gain phase detector compares the transmitted and reflected waves and provides the measured information to the microcontroller. The characteristics, measured by the instrument, are in the range of -30/+30 [dB], with a scale of 30 mV/dB for the gain, and 0–180° with a scale of 10 mV/° for the phase. The gain and phase output voltages vary in a range from 0 V to 1.8 V. In addition, the component provides a reference voltage of 1.8 V, which serves as a full scale for the output voltages. A layout of the electronic systems is shown in Figure 13.

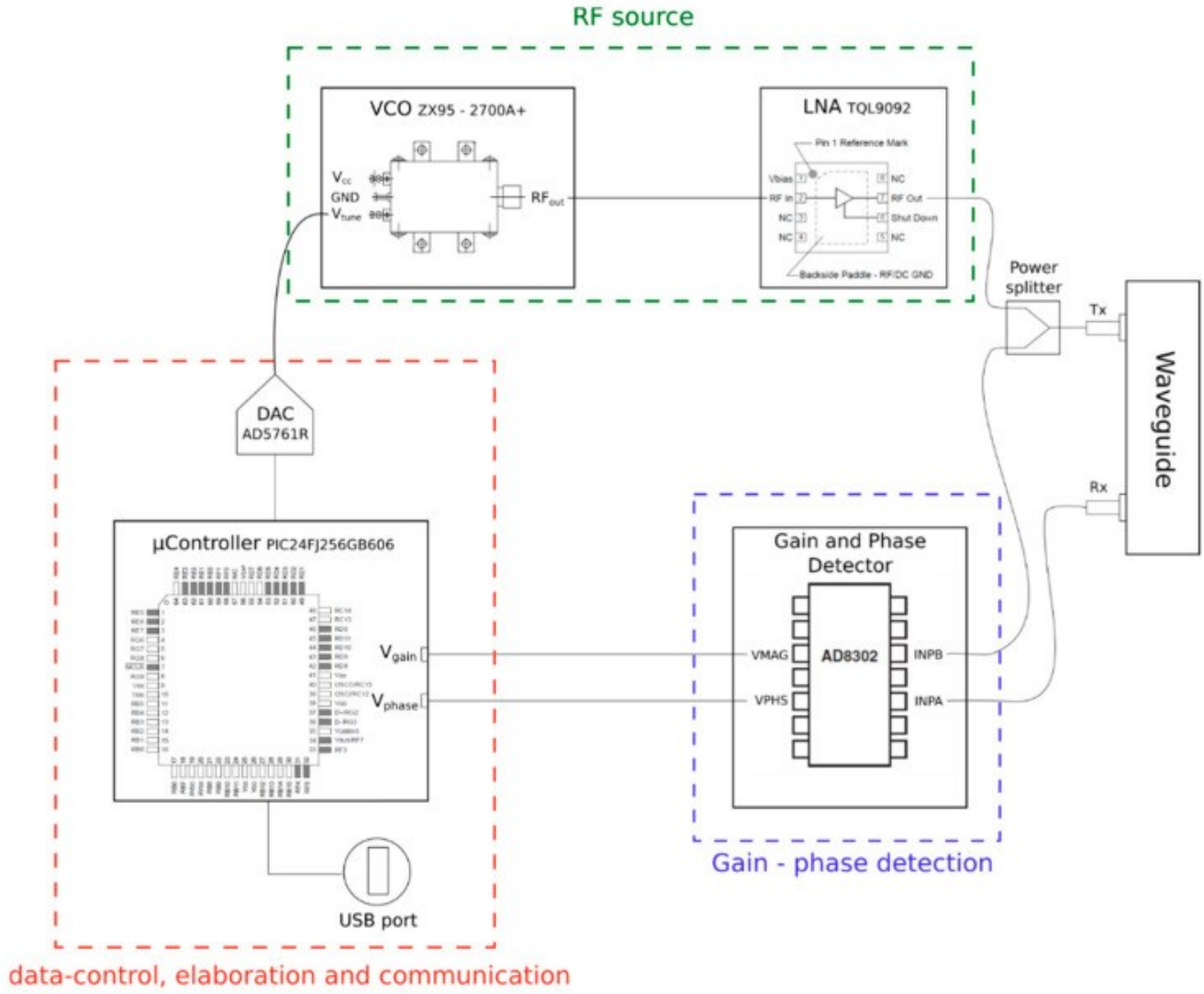

Fig. 13 Electronic system layout.

5.3 Experimental setup

Data acquisition

Spectral acquisitions were conducted on different areas of silty clay loam soil located in the Romagna region (Italy), described by an average value of the soil bulk density of 1.40 ± 0.09 g/cm^3, calculated starting from nine measurements conducted on different soil areas. Measurements were done with a core drill machine, with a diameter of about 48 mm and a height of about 56 mm. For each chosen location, three measurements were acquired, rotating the container to its axis at an approximately constant angle (40°). For each acquisition, the system was placed in such a way that the longitudinal axis of the waveguide was normally oriented to the soil surface; a good adhesion of the waveguide aperture to the soil was ensured after a brief cleaning of fallen leaves and other organic detritus that could alter the measure. The time requested to perform a single measurement is about 45 s. The acquisitions were taken from October to November 2018, with a soil temperature range of 8.1–17.5 °C and an air temperature range of 7.7–20.9 °C.

A total of 345 measurements (115 soil areas x 3 acquisitions per area) were obtained over two months. After each triplet of acquisitions, a section of metallic tube was used to perform a coring process on the soil, obtaining cylindrical samples of about 10 cm and a diameter of about 0.8 cm. Example photos of an acquisition and a sample coring are displayed in Figure 14.

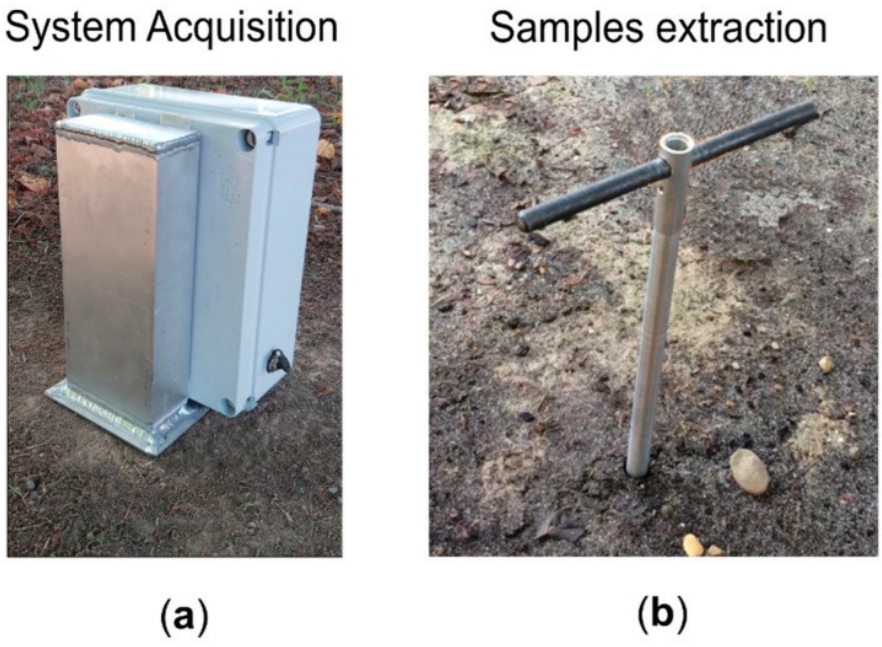

Fig 14 **(a)** System acquisition and **(b)** samples extraction on real soil

For each sample, the moisture value (%) was evaluated with the thermo-gravimetrical method; each sample was weighed, then put in an oven at 105 °C for about 24 h and weighed again. The weights were used to find the gravimetric moisture values θ_M, thanks to the following formula:

$$\theta_M = \frac{(w_w + t_a) - (w_d + t_a)}{(w_d + t_a) - t_a} \times 100 \quad (26)$$

Where the parameters w_w and w_d represent the masses of the soil before and after the day in the oven, respectively, and t_a represents the tare mass [40].

Spectral results

Samples of soil were characterized by a moisture content ranging from 9.3% to 31.7% and by a temperature from 8.1 °C to 17.5 °C. The acquisitions were distributed to the temperature as follows: 14% in the range of 7 °C–11 °C, 18% in the range of 11 °C–14 °C, and 68% in the range of 15 °C–17.5 °C. Examples of "gain" and "phase" waveforms acquired at different soil moisture levels (%) are shown in Fig. 15.

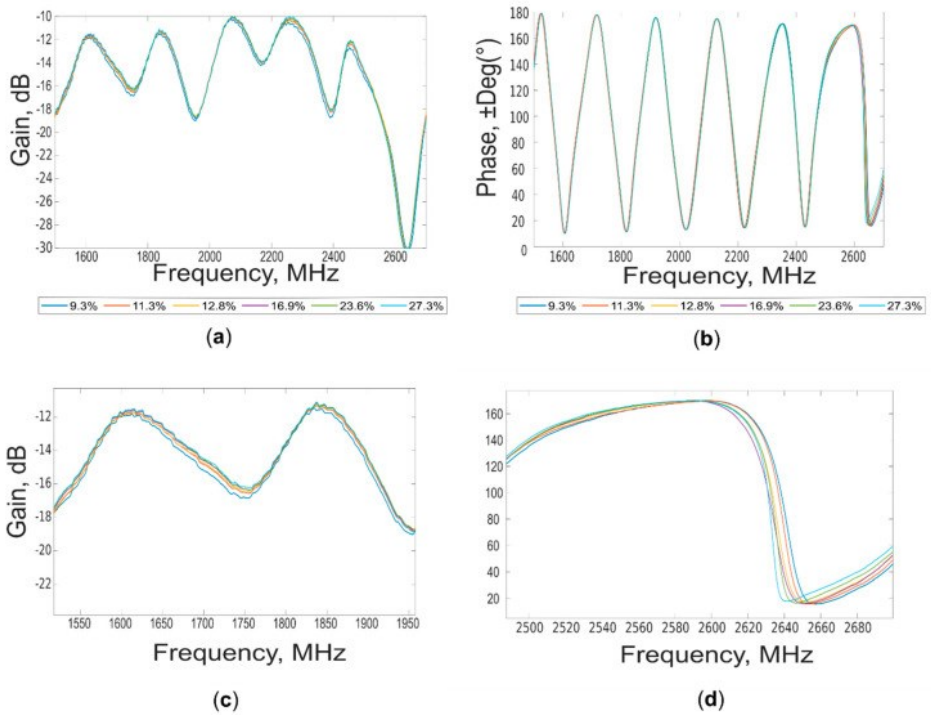

Fig. 15 **(a)** "Gain" and (**b**) "phase" waveforms acquired for different moisture content (%). (**c**,**d**) Magnifications of the waveforms, in the regions where the spectral differences are more pronounced.

Spectral differences can be appreciated during the entire range of the explored frequencies, especially for “gain” spectra (1.5 GHz–2.7 GHz), according to the moisture content. It appears evident the influence in the spectra of the complex water–soil chemical-physical interactions. These changes can be better visualized in the lower portion of the figure showing a magnification of both “gain” and “phase” in two different spectral ranges. For these acquisitions, a temperature of the soil ranging from 16.5 °C to 17.5 °C was measured. No evident loss of continuity in the waveforms was detected for the cut-off sub-frequency range (from 1.50 GHz to 1.56 GHz), where the emitted power appears enough to return information related to moisture content.

To understand the influence of the temperature on spectral acquisitions, other examples of “gain” and “phase” waveforms are shown in Fig. 16.

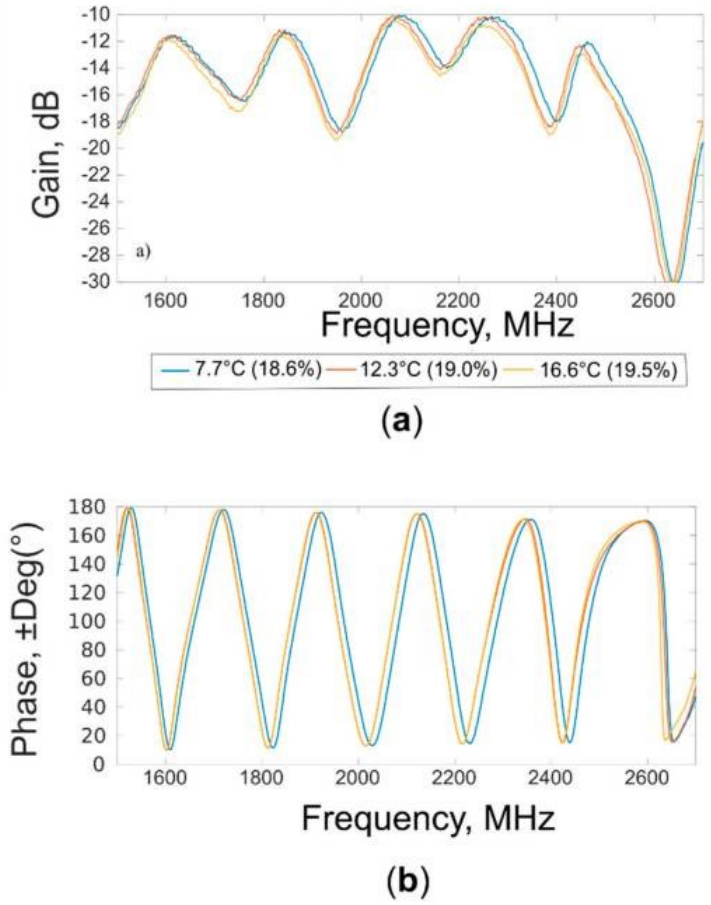

Fig. 16 **(a)** “Gain” and **(b)** “phase” waveforms acquired at different soil temperatures (°C), with a moisture value of about 19.0%.

The waveforms were acquired at soil temperatures of 7.7 °C, 12.3 °C, and 16.6 °C on soil samples characterized by a small variation of moisture contents (18.6%, 19.0%, and 19.5%, respectively). Shifts of the waveforms are evident in different parts of the spectrum for both “gain” and “phase”. These differences reflect the known dependence from the temperature of the loss factor and dielectric constant.

Moreover, the mutual coupling between the antennas was characterized, calculating the power emitted by the Tx antenna (which varies by the wave frequency, in a range of 20.4–22.5 dB) and subtracting from it the measured gain values, for the different frequencies, to obtain the R_X power. The coupling was obtained as the ratio between Tx and Rx power, for acquisition taken on both air and soil, to compare the response when open-end waveguide radiated in open space and soil. The results are shown in Fig. 17; as expected, different trends were obtained for acquisition on air and soil, evidencing the ability of the systems to well differentiate the two media, both remaining in an acceptable range of power ratio values.

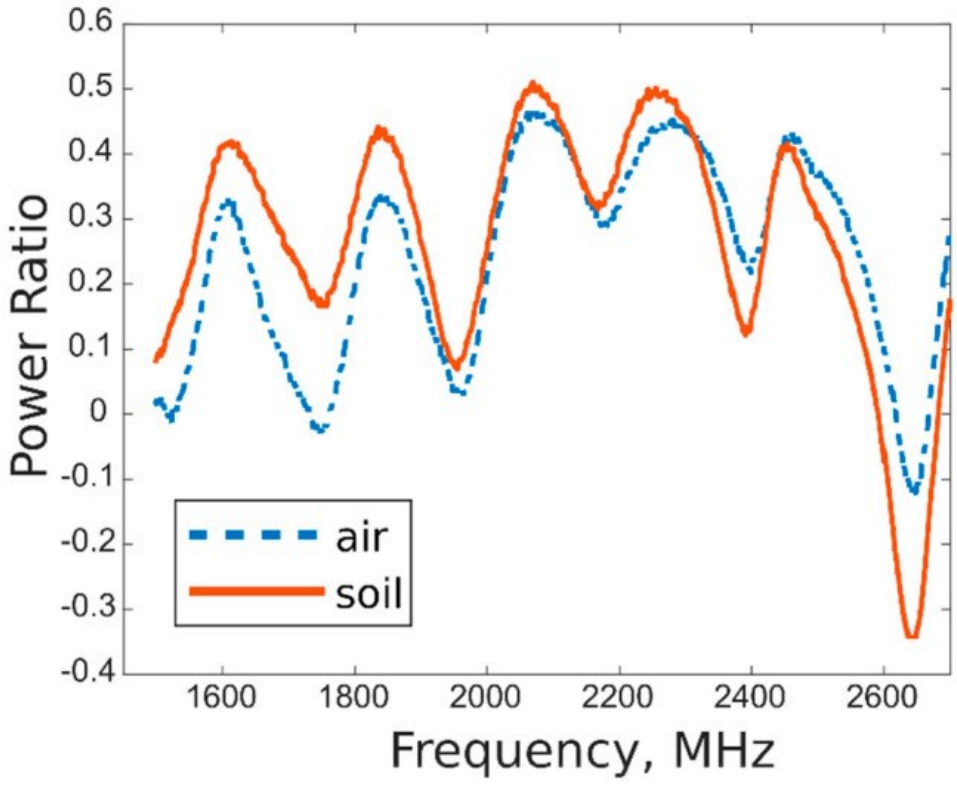

Fig. 17 Characterization of mutual coupling between Tx and Rx antennas, for acquisitions on air and soil.

5.4. Embedded model approach

In the end, the independent variables (X dataset) were arranged in a K = 3700 (frequencies) × N = 345 (spectral measurements) matrix, whereas an N=345 (measurements) × M = 1 (moisture content %) vector column was created for the dependent variable (Y dataset). Then, both datasets are split between calibration, with NC = 285 measurements, and test, with NT = 60 measurements.

These two datasets were used as input for the creation of a prediction model with PLSR. Before the model creation, autoscale pre-processing was applied to both X and Y data sets. For cross-validation, a method called "Venetian blinds" was used: each subset is determined by the selection of every nth object in the data set, starting at objects numbered 1 through s. This method is simple and easy to implement, and generally safe to use if there are relatively many objects already in a random order[56]. The RMSECV parameter was used for the selection of the number of latent variables, done automatically by the software. Another algorithm, already implemented in the software, was used to select only the useful X variables, to improve the regression model. According to this algorithm, in the first run, the variables with the lowest Variable Importance in Projection (VIP) values are eliminated. If the model improves, this is repeated until convergence [57]. Finally, new measurements were used to create X and Y test sets.

PLSR results

The values of these parameters for PLSR analysis conducted on "gain" and "phase" spectra for the prediction of the moisture content (%) are summarized in Table I, for calibration, segmented cross-validation, and test set validation. The optimal numbers of latent variables were 6 for the "gain" model and 5 for the "phase" one. The test sets were created with measurements and moisture values chosen from X and Y calibration sets (and thus not included in model creation). For every chosen sample, all three acquisitions were put in the test set, to avoid the presence of the same sample's acquisitions in both data sets. The samples (60 measurements, about 20% of the total) were randomly selected to cover the moisture range of the data used for the calibration.

Table I - Partial least squares (PLS) regression models for the prediction of soil moisture content (%) from "gain" and "phase" spectra. RMSE, root mean square error; LV, latent variable.

Data	LVs	R^2	RMSE (%)
Gain – Cal.	6	0.958	0.7
Gain - CV	6	0.949	0.7
Gain – Pred.	6	0.892	1.0
Phase – Cal.	5	0.764	1.8
Phase - CV	5	0.741	1.9
Phase – Pred.	5	0.732	1.8

As expected from the waveform's visual exploration, the best predictions were obtained using "gain" spectra. In test set validation, the moisture content can be predicted with an R^2 value of 0.892 and an RMSE of 1.0% with the gain model. For the "phase", an R^2 value of 0.781 and an RMSE of 1.8%, were obtained. The scores plot for the model obtained from "gain" spectra (in calibration) is reported in Figure 15 for the first two latent variables (LV1 and LV2). This plot shows how much each measurement is influenced by the two LVs, allowing us to understand which physical parameters they represent; that is, sample scores are distributed according to the moisture content (%) along the first latent variable, which accounts for 50.84% of the variance. The second latent variable (38.6% of the variance) appears to describe differences in the spectra due to temperature changes. As possible to observe from Table I and Fig. 18, the model created with "gain" data is better than the "phase" one, obtaining higher values for R^2 and lower values for RMSE, in calibration, cross-validation, and prediction.

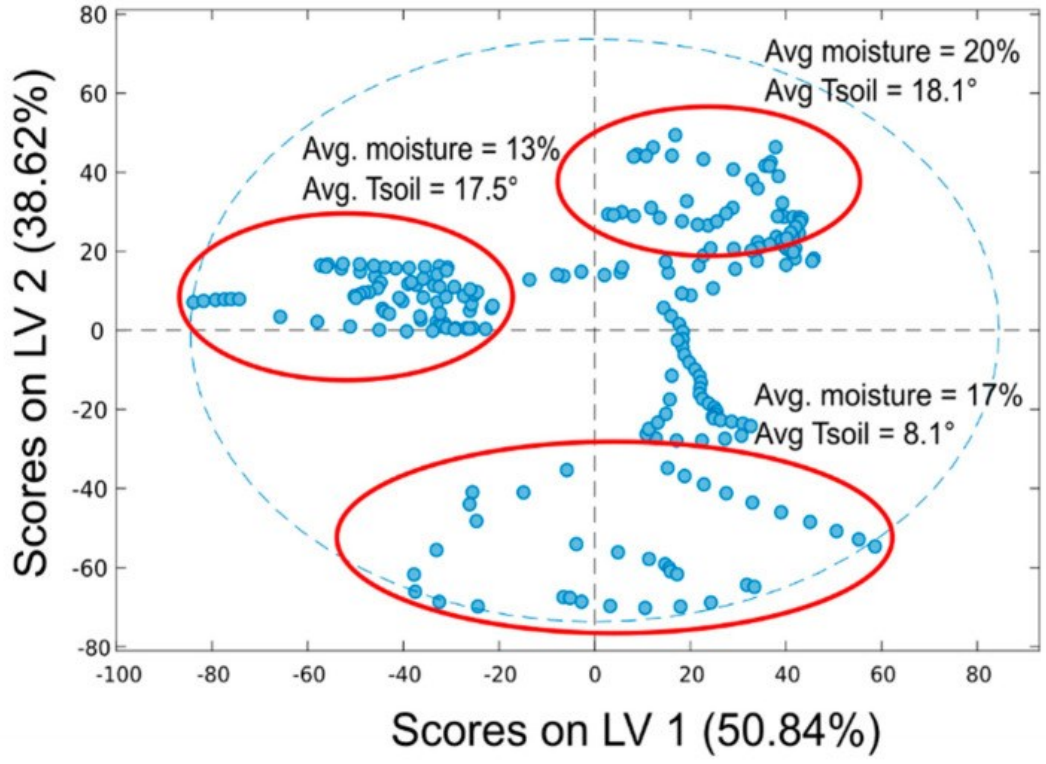

Fig. 18 Scores plot for the PLS model in calibration obtained for "gain" spectra (first two latent variables (LVs)).

5.5. Implementation in the microcontroller

Calibration coefficients were obtained from the final model, resulting in an array **A** and a single set value **B**, with which it is possible to calculate the soil moisture value $\widehat{\boldsymbol{Y}}$ directly from a new measurement $\widehat{\boldsymbol{X}}$, with a formula similar to (1):

$$\widehat{\boldsymbol{Y}} = A\widehat{\boldsymbol{X}} + B \tag{27}$$

Using (27), the calculation of the soil moisture was directly implemented in the microcontroller of the system: for each frequency, the newly acquired spectral values are multiplied by the corresponding coefficients and summed up with the previous multiplication in a new variable. At the end of the whole acquisition, an offset B (1 x 1), also calculated by PLSR, is summed to this variable, obtaining the final moisture value, sent to the PC through a serial port. The whole process is shown in Fig. 19.

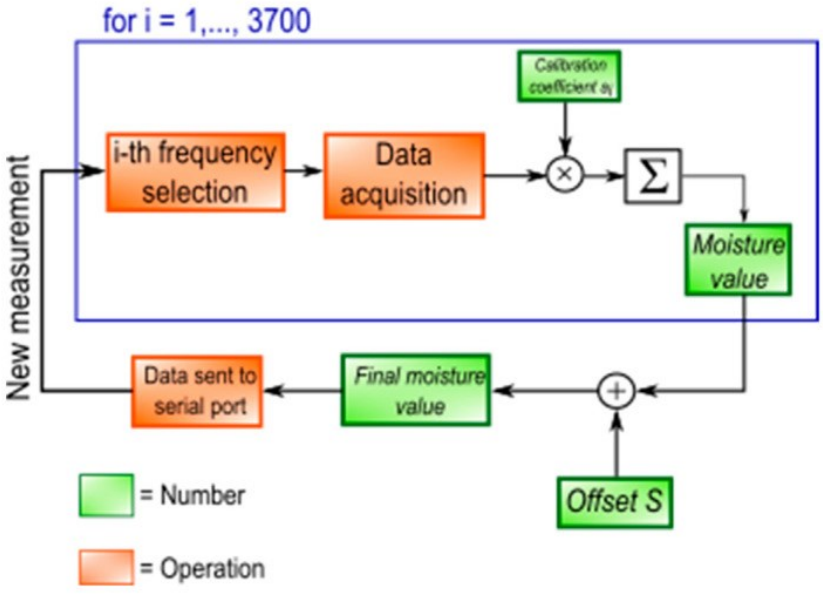

Fig. 19 Moisture calculation in the MCU.

5.6. Conclusions on soil moisture detection

A proof of concept of a contactless open-ended waveguide, designed and developed for soil moisture rapid evaluations, was assessed on real silt-clay loam soil. “Gain” and “phase” spectra acquired from 1.5 GHz to 2.7 GHz on soil characterized by different moisture contents (from about 9% to 32%) were used to build PLSR models. The best predictive model was obtained starting from “gain” spectra (R^2 = 0.892, RMSE = 1.0 for an external test set validation), whereas “phase” spectra did not produce accurate results in terms of moisture content prediction (R^2 = 0.781, RMSE = 1.8%). As expected, soil temperature can affect both “gain” and “phase” spectral waveforms. However, the multivariate tool appeared to evaluate the variable “moisture content” as the most influential for waveform variance (50.84%, first LV), granting good results in the moisture prediction.

These results are very promising and, in the future, further research could be conducted with the suggested instrumentation and data modeling, trying to refine a more general model based on “gain” data. However, there are some criticism and limits to the system and the model that needs to be considered and further investigated. The main one is the fact that all the samples used to acquire both the spectra and the real moisture values were obtained from the same type of soil, and they were acquired in a relatively small period of time (2 months). So, there could be a lack of generalization in the model predictions: they would probably be very accurate for new spectra acquired on the same type of soil and in the same range of temperature, but they could be less precise for predictions in different environmental conditions. This problem can be solved increasing the input dataset, acquiring new spectra and moisture values on soil samples with a different composition in different periods of the year. This increased dataset would for sure help the statistical analysis to generalize further the moisture prediction, making it more accurate for a larger number of environments. Another limit of the proposed technology is the fact that the acquired spectrum is referred to a small portion of the soil, the one immediately below the instrument. So, if we want to monitor the moisture trend of a whole agricultural field, a network of devices would be necessary.Summing up, future developments could be achieved by exploring the possibility of predicting moisture at different soil levels, developing a model based on values derived from a combination of “gain” and “phase”, and extending the application of the technique to other kinds of soils and/or chemical soil components, using similar PLSR algorithms, to obtain models for specific soil. Moreover, the performance of the system in different seasons (different temperatures and moisture levels) needs to be investigated and developed, to extend the conditions where the acquisition is not distorted. Finally, the electrical and mechanical design could be improved, decreasing the measuring time, implementing instrumental calibration functions and procedures, and testing the resistance to weather conditions during prolonged field use.

6. Concrete moisture content detection

This chapter presents a sensing system based on an open-ended waveguide, used for monitoring concrete compressive strength in real-time, implementing in this way a non-intrusive monitoring sensor. The proposed approach uses a Vector Network Analyzer (VNA) connected with a rectangular open cavity antenna emitting an RF signal towards the concrete material. The VNA allowed us to grab scattering coefficients (S-parameters) and derive them in a spectrum between 1.5 GHz and 14 GHz, a range where the water component of concrete presents a higher dielectric constant (from 88 to 27) versus other constituents. Many concrete cubic samples were cast simultaneously using the same material (sample twins) from which spectra were acquired for 28 days. Part of these samples was used in destructive tests to determine compressive strength. Then, the acquired spectra and the compressive strength data were used as input for two MSA, SIMCA, and PLSR. This algorithm allows for resolving classification problems, where new samples are assigned to existing groups of similar samples, modeling the common properties of the classes [26]. Finally, a subset of spectra was used to create a preliminary regression model with PLSR, which allows for predicting compressive strength values from newly acquired spectra. Since these predictive models could be easily implemented in low computationally demanding digital platforms such as a microcontroller, this chapter aims at exploring a compact, non-destructive and cost-effective instrumental chain for concrete hardening monitoring.

6.1. State of the art

Concrete is the most used material in the civil infrastructure and construction industry. It is composed of a mix of cement powder, sand (fine aggregates), gravel (coarse aggregates), and water. The ratio of these components influences the mechanical properties of concrete. In particular, the water/cement (w/c) ratio is critical for the strength and durability of concrete due to the role of water in the hydration process and the microstructure evolution [58]. So, it is crucial to monitor the conditions and characteristics of concrete during the hydration process, especially in the first hours. Nowadays, several tests are used as the gold standard for the evaluation of hardened concrete properties, like the water absorption test (ASTM C: 642–81), rapid chloride ion penetration test (ASTM C1202), impact strength test (ACI 544.2R-89), and compressive strength test (ASTM C39/C39M). However, even if they allow a satisfactory evaluation during construction, all these methods are time-consuming and expensive. Moreover, they require an invasive approach and the creation of additional samples [59]

with different mechanical characteristics for the on-site batch due to other manufacturing processes and materials affecting the results [60], [61].

For these reasons, non-destructive techniques (NDT) have been recently studied with several different technologies: acoustic waves [62], [63], ultrasonic wave propagation [64], Ground Penetrating Radar (GPR) wave attenuation [65], [66], piezoelectric materials [67], [68], gamma scattering [69], electrical resistivity [70], [71], and many others. Among them, a particular interest is aroused by microwaves, which allow penetrating the material and gathering information on the internal water content. The microwave spectra are highly influenced by the hydration of concrete and the consequent changes in the water content because of its significant influence on the dielectric properties. Different techniques were used to determine these dielectric properties of concrete and their changes: the first tests were performed by Bhargava and Lundberg [72] in 1972 with a microwave resonant cavity, obtaining a linear relationship between the output of the instrument and the moisture content of concrete. Three years later, Wittman and Schlude [73] studied microwave absorption using concrete disk samples with a free wave technique. The results showed that the wave attenuation decreases and then reaches a plateau due to the absorption of the free water contained in the sample. In 1982 Gorur et al. [74] introduced a different measurement method, focusing on using a two-port (a waveguide section full of concrete) and using the scattering parameters S to compute the complex dielectric constant. Results showed that the dielectric permittivity decreases with time and increases with the w/c ratio of the concrete. More recently, the same results were obtained by Haddad and Al-Quady [75], that used a coaxial transmission line in the range of 100 MHz to 1 GHz, whereas Buyukozturk et al. [76] determined the complex permittivity and the loss factor of several materials from the transmission coefficient and the Time Difference of Arrival (TDOA), using a network analyzer in the range of 8-18 GHz, obtaining similar results. More in general, in the last two decades, several different techniques have been used to study and monitor the dielectric properties of concrete, like horn lens antennas [77]–[79], Ground Penetrating Radar (GPR) [80], [81], time-domain reflectometry [82], and several types of waveguides [83]–[86]. However, all the above techniques only focused on the variations of dielectric properties without trying to correlate them with the changes in the concrete physical properties. Furthermore, their outputs are often hard to interpret and interfaced with quantitative measurements and lack robust predictive models. Different from conventional spectra-based approaches, the main goal of our work was to introduce a sensing technique that operates by extracting hidden information from raw data, even if this is not evident at first sight [19].

6.2. Design of the system

Preliminary investigations

To correctly design the sensing system for concrete monitoring, a series of preliminary tests were performed, with a twofold task. Firstly, we had to determine the best microwave spectrum that copes with the application. Secondly, we needed to understand the most suitable spectrum for building the dataset of the predictive model.

As far as the first objective is concerned, we needed to investigate the penetration depth of microwaves inside the concrete to understand the best operating spectra bandwidth. This data is important since the proposed technique aims to extract material information from a sufficiently large volume where the surface can differ from the inner material conditions. Therefore, the radio frequency penetration depth was estimated from experimental permittivity measurements. The penetration depth (d_p) is defined as the distance where an incident electromagnetic wave penetrates the material under test with an intensity (power) falling to $1/e$ [87] and it can be calculated by permittivity as follows

$$d_p = \frac{c}{2\sqrt{2}\,\pi f \left\{ \varepsilon_r' \left[\sqrt{1+\left(\frac{\varepsilon_r''}{\varepsilon_r'}\right)^2} - 1 \right] \right\}^{\frac{1}{2}}}, \tag{28}$$

where c is the speed of light in vacuum, 2.998×10^8 m/s, f is the frequency (Hz), ε_r' and ε_r'' are, respectively, the real and imaginary parts of the relative complex permittivity $\varepsilon_r^* = \varepsilon_r' + j\varepsilon_r''$, depending on the frequency. It should be pointed out that equation (28) is related to the matter interaction with plane waves and in far-field conditions [88], which is only an approximation in our case. However, the calculation is useful to get a rough estimation of the spectrum range used for this application. Preliminary EM simulations considering the true geometrical features of the entire system showed penetration depths longer by a factor between two and three. The complex parameters were assessed using an open-ended coaxial instrumental chain (Speag, DAK 3.5, Swiss) and reflectometer (R140, Copper Mountain, USA), depicted in Fig. 20.

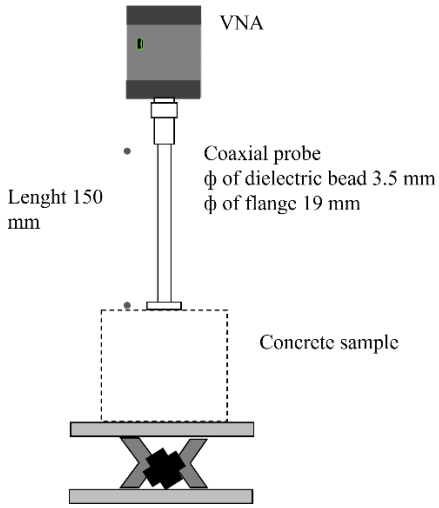

Fig. 20 Layout of the instrumental setup used for the preliminary calculation of penetration depth.

The frequency range investigated was 1.5-14 GHz, and each spectrum was composed of 10000 spectral points. Measurements were conducted after 2.2 hours of the concrete preparation (wet), after 23 hours (semi-dried), and at advanced hardening (104 hours). The latter time is considered sufficient to achieve the drying condition of the concrete uppermost layer that can be assumed (from the dielectric point of view) as representative of the hardening state of the inner part of the sample after about 28 days. On the other hand, due to its conformation, the coaxial probe performs only superficial dielectric measurements (in the order of a few mm), so the measurements can be taken as indicative of the sample's inner dielectric properties. As a reference, the real and imaginary values of the complex electric permittivity (ε_r' and ε_r'') for a frequency of about 2 GHz were the following: 27.3+j7.1 (wet), 11.2+j1.7 (semi-dried), and 7.3+j0.89 (dried). The penetration depth (d_p) calculated using (15) within the instrument spectrum bandwidth is shown in Fig. 21.

In conclusion, we can say that the spectra frequency range and the concrete's hydration level directly affect the penetration depth. Using the bandwidth we investigated, it is possible to grab information from the concrete under test with a d_p in the order of centimeters scale, especially for frequencies below 7 GHz. The penetration depth is a reference parameter that indicates the distance at which the surface intensity decays of a factor 1/e. However, the reading depth, i.e., the depth at which the VNA can detect the reflected signal with an acceptable signal-to-noise ratio, can be much greater than the penetration depth. This mainly depends on the output power and dynamic range of the instrument. In our application, previous tests conducted by placing a reflecting surface under the concrete sample found a reading depth up to about three times greater than the penetration depth. Therefore, we can still detect reflected signals at further greater depths using higher incident power and/or instrumental

chains with lower noise floors. So, data from the calculation and preliminary tests were used to design the microwave apparatus.

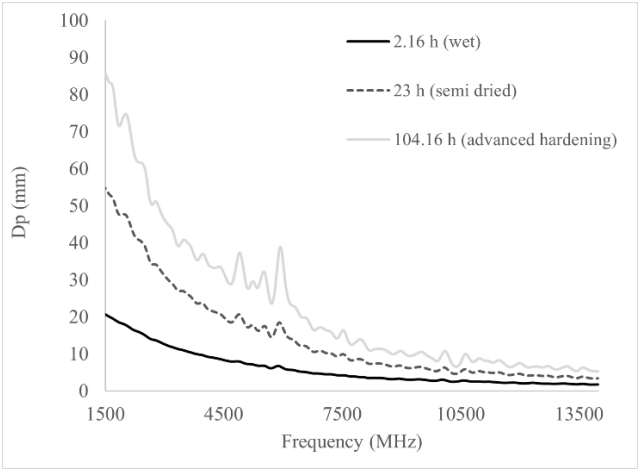

Fig. 21 Penetration depth (d_p) as a function of frequency (MHz) for wet, semi-dried, and dried concrete.

Description of the device

It was chosen as a probe, a short open section of a waveguide optimized for a TE10 mode at 2GHz with a cutoff frequency of 1.56GHz (Fig. 21a) with the antenna inserted at a ¼ of the guided wavelength.

As far as the second objective of preliminary investigations is concerned, i.e., the choice of the kind of spectrum, we proceeded as follows. We connected the designed open waveguide, as shown in Fig. 22 **(a)**, and we acquired several spectra based on S-parameters, e.g., real and imaginary parts of S_{11} and VSWR spectra, in the frequency range from 1.5 to 14 GHz. The relationship between S_{11} and VSWR is given by:

$$VSWR = \frac{1+|S_{11}|}{1-|S_{11}|}, \tag{29}$$

Where

$$|S_{11}| = \sqrt{\left(\mathrm{Re}(S_{11})\right)^2 + \left(\mathrm{Im}(S_{11})\right)^2}, \tag{30}$$

We noticed that they all show variations versus the compressive strength state due to the evolution of the chemical process. However, to understand what spectrum is better suited for building the dataset, we compared the Root Mean Square Error in Cross-Validation (RMSECV) values of the predictive

model for the datasets built on different spectra (see Section 4.2.5). Since we did not get any appreciable differences, we decided to use VSWR mainly for its better simplicity of implementation [89] in a future prototype. An example of a VSWR spectrum grabbed with the designed setup is shown in Fig. 22 **(b)**.

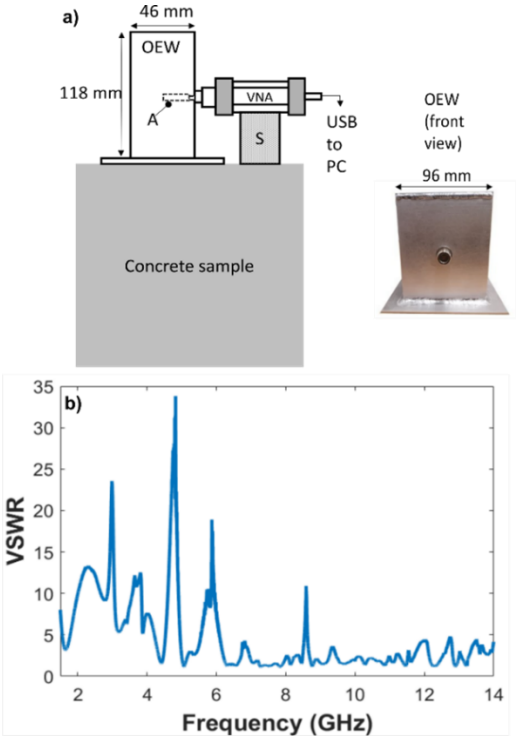

Fig. 22 **a)** Schematic layout for the acquisition of electromagnetic spectra. Legend: OEW, open-ended waveguide; A, antenna; VNA, vector network analyzer; S, NVA support. The drawing is not to scale. **b)** Example of a VSWR spectrum

6.3. Experimental setup

The data acquisition procedure was conducted as follows. Sixteen equal concrete samples were cast at the same time. Among them, two were used for acquiring VSWR spectra and temperature measurements at increasing curing time, while others were subject to destructive compressive tests.

Destructive tests

During the first 4 hours, when concrete was starting to develop its mechanical properties, its penetration resistance was measured by a Concrete pocket dial penetrometer Matest (diameter of the penetration plunger 6.4 mm, load 0-50 kg/cm^2), providing the load necessary to plunge a probe of the known area into concrete to a fixed depth. From this parameter, a proper calibration coefficient

provided the corresponding compressive strength. Two cubic concrete samples were used for this purpose, and the penetration resistance was determined as the average of three different plungings. After the first hardening, the strength was measured at significant curing times by compression tests (2 samples for each condition) in a universal testing machine with 4000 kN capacity. Tab. II reports all the compressive strength values, including both those coming from the early penetration resistance values and those measured by the destructive tests. As shown in Fig. 23, the compression strength values closely follow a predictive model available in the literature [90] described by the following equation:

$$Res = e^{s*\left(1-\sqrt{\frac{28}{t}}\right)} * R_{c,28}, \tag{31}$$

where s = 0.2, t is the time from cast expressed in days, and $R_{c,28}$ = 27.8 MPa is the 28 days compressive strength.

Table II. Time of acquisition and results of the compression strength of concrete.

Time [day]	*Strength [MPa]*	*Instrument*
0.045	0.67	Concrete penetrometer Matest
0.062	0.81	
0.083	2.13	
0.104	2.67	
0.125	3.60	
0.146	3.90	
0.167	5.00	
1	11.7	Universal Testing machine
2	17.7	
3	18.9	
7	23.4	
10	23.7	
14	25.7	
28	27.8	

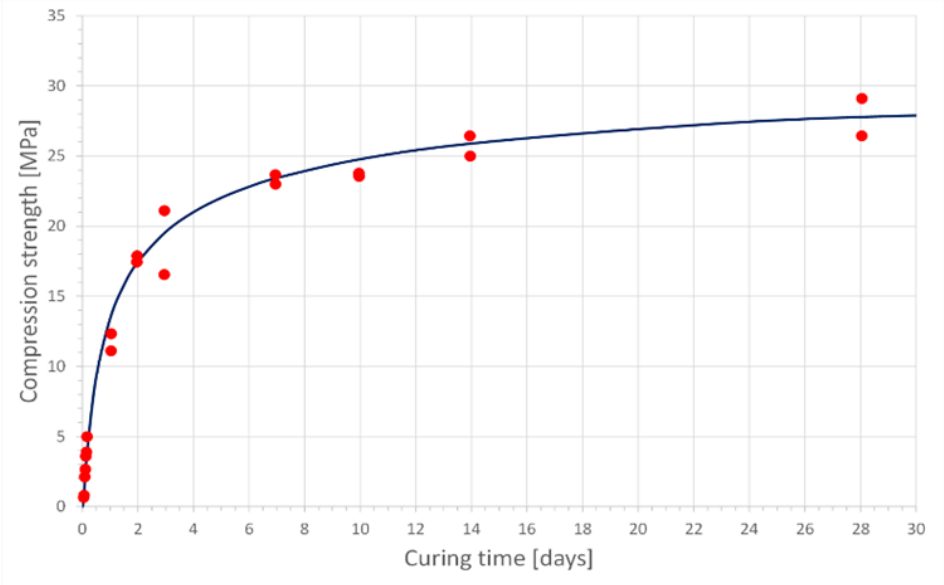

Fig. 23 Measured resistance (dots) and theoretical value according to (18) (line).

Spectra acquisition

Electromagnetic spectra were acquired with the designed cavity antenna (internal dimensions, in mm: 96 × 46 × 118) connected to a VNA (R140, Copper Mountain, USA), as shown in Fig. 19. The VNA was calibrated using a calibration kit (N1801, Copper Mountain, USA). The instrument provides spectra of the real and imaginary parts of the scattering parameters. At the end of the acquisition period, which lasted 28 days, a total of 409 VSWR spectra were acquired, each calculated as the mean of 5 consecutive acquisitions. To take the maximum information on the evolution of the process, spectra acquisition intervals should be set to have an equal increment of the compressive strength. Therefore, sampling intervals should not be equally spaced since the curing process is fast at the beginning and slows down to an asymptotic value at the final state. For the above reason, acquisition times were set as summarized in Tab.III, where intervals are denser at the beginning compared to later times.

Table III. Time interval and number of acquisitions of the electromagnetic spectra on concrete

Interval of time [days]	*Time between acquisitions [min]*	*Number of acquisitions*
0 - 0.09	10	14
0.09 - 0.96	20	59
0.96 - 3.24	30	102
3.24 – 8.98	60	127
8.98 – 14.2	180	50
14.2 - 28	300	62

6.4. Spectral results

Fig. 24 shows all the acquired spectra, colored by the order of acquisition, from blue (first measure) to yellow (last measure). It is easy to see that in certain regions there is a clear trend in the spectra, where the VSWR values gradually change over time in a monotonous way. This trend is a good indication for the creation of statistical models: it is already plausible to assume that the spectra contain useful information about the dielectric properties of the concrete and, consequently, about its free water content and compressive strength. Moreover, this phenomenon is more pronounced in the first part of the spectrum (1.5-6 GHz), which is conveniently also the part of the spectrum with the higher penetration depth, as reported previously.

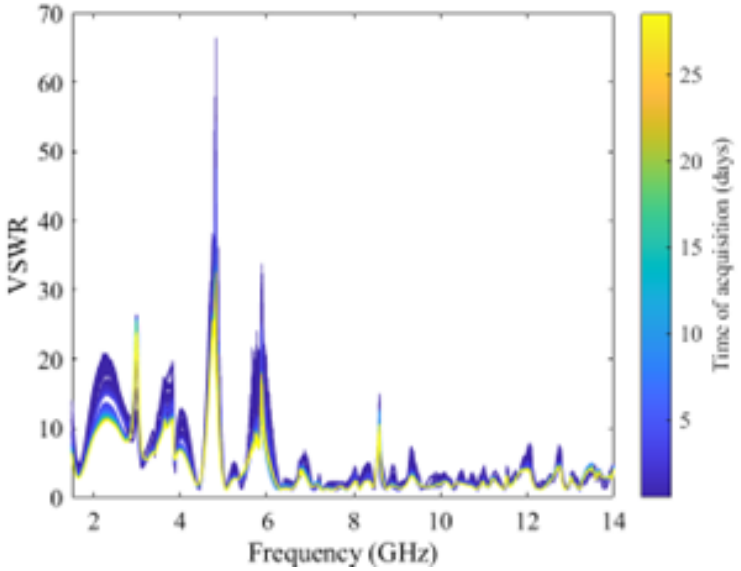

Fig. 24 Visualization of all the acquired spectra, colored following the acquisition time, from blue (day 1) to yellow (day 29).

Explorative PCA

All the VSWR spectra were preprocessed with two standard techniques: the Savitsky-Golay smoothing and autoscale, which performs a mean centering and scales each variable to unit standard deviation.

The first statistical analysis performed on these preprocessed spectra was an explorative PCA (with $A=5$ PCs) to visualize clusters in the dataset and divide them into the classes mentioned above, which is necessary for the following statistical analyses. We empirically divide the data into 4 classes based on the obtained score plot, shown in Fig. 25. This plot shows the data in the space defined by the first

2 of the 5 total PCs, which contains the majority of data variance (62.95 % the first PC and 17.37 % the second one, for a total of 80.32%), allowing us to easily visualize the data trends and clusters. Then, we linked these classes to ranges of compressive strength. To do so, we evaluated the time of acquisition of the first and last acquisition of each subset: if it is close to one of the destructive tests reported in Table II, we used the experimental value as a reference; in the other case, we evaluated the theoretical strength with (18). With this method, we obtained these references for the compressive strength: <12 MPa for class 1, 12 - 18.5 MPa for class 2, 18.5 – 23.4 MPa for class 3, and 23.4-28 MPa for class 4. These classes were used for the creation of a classification model using SIMCA (see Section 4.2.3).

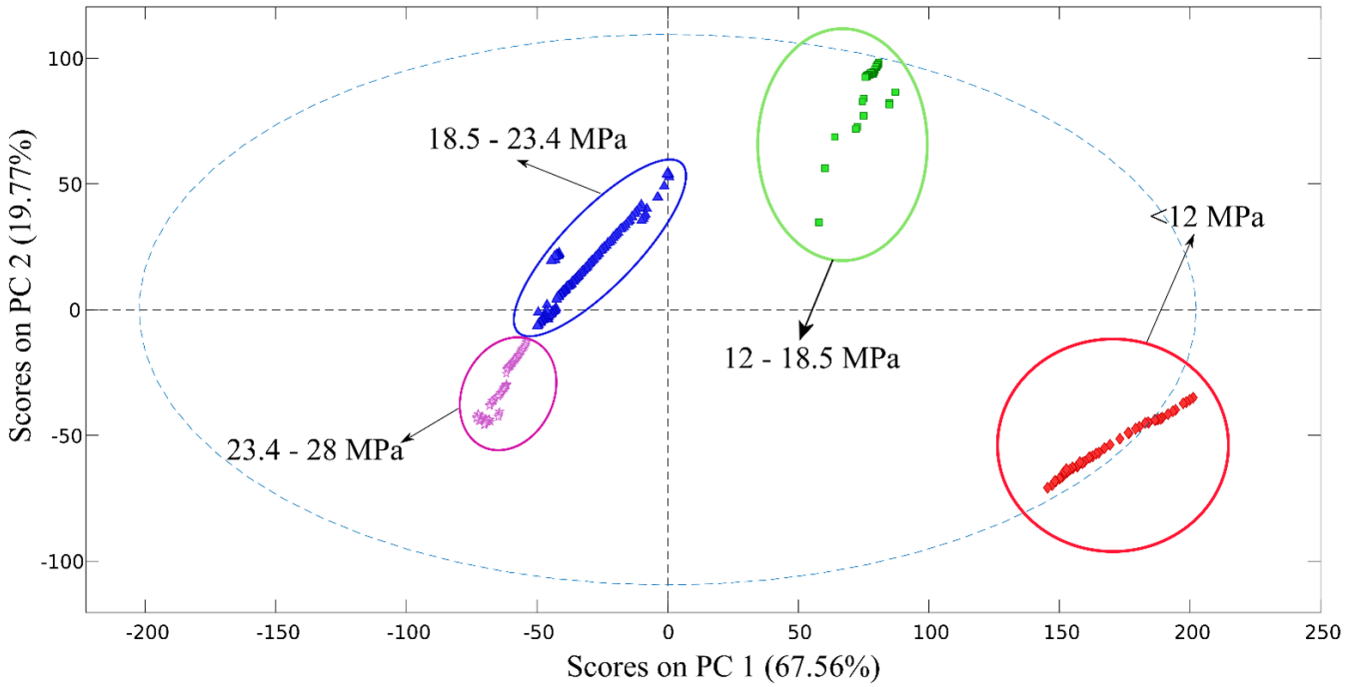

Fig. 25 PCA score plot. The 4 classes used as input for the classification algorithms are highlighted.

SIMCA Model

For the SIMCA model creation, 80% of the original spectral data (266) were used for the creation of the calibration test. The remaining 20% (66), randomly chosen and, in percentage, equally distributed between all the classes, were used as a test set to perform the validation phase: these spectra were assigned to the different classes by the model, and the results compared to the actual classes, obtaining a confusion table and a confusion matrix, useful to compare the predictive robustness to different SIMCA models.

To obtain more informative results, 10 SIMCA models were created, with different test sets (always chosen at random), and a mean value of the obtained statistical parameters was calculated. As reported in the last section, the SIMCA analysis utilizes two parameters, Q and T^2, to assign new samples to the more representative class (the one with the lowest value of both parameters). Fig. 26 shows the Q-T^2 plots of one among the 10 SIMCA models as an example. It can be observed that for the first two classes (<12 and 12-18.5 MPa), there is a considerable difference between the samples of the correct class to the other ones, while for class 3 (18.5 – 23.4 MPa) and 4 (23.4 – 28 MPa), the difference is less noticeable. However, as shown in Figs 26 (e) and 26 (f), there are still enough differences between the values of Q and T^2 to correctly classify the test samples in most cases.

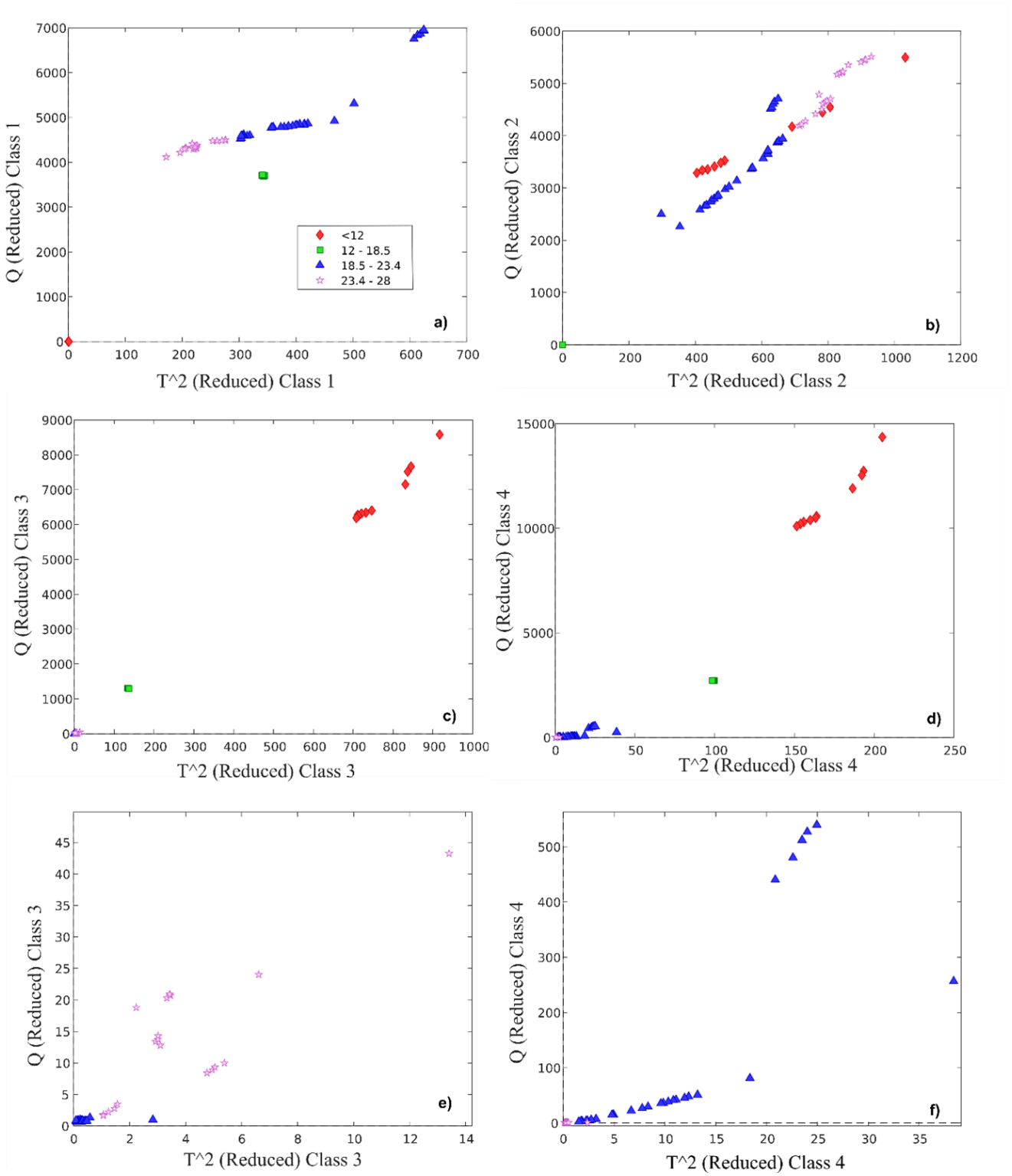

Fig. 26 **a)** – **d)** Plots of T2 vs. Q for each of the 4 classes. **e)-f)** Zoom for classes 3 and 4 to better visualize the differences between the test samples.

Table IV reports the mean confusion matrix, whereas Table V is the mean of the classification parameters, both obtained from the mean of the 10 SIMCA models. *N* is the number of spectra used as a test set for every class. The values *TPR* (sensitivity), *FPR* (false positive rate), *TNR* (specificity), *FNR* (false negative rate), Err (misclassification error), *PPV* (precision), and *F1* (F1 score) are given by the following relationships:

$$TPR = \frac{TP}{TP+FN} \tag{32}$$

$$FPR = \frac{FP}{FP+TN} \tag{33}$$

$$TNR = \frac{TN}{TN+FP} \tag{34}$$

$$FNR = \frac{FN}{FN+TP} \tag{35}$$

$$Err = \frac{FP+FN}{TP+TN+FP+FN} \tag{36}$$

$$PPV = \frac{TP}{TP+FP} \tag{37}$$

$$F1 = \frac{2}{(TPR)^{-1}+(PPV)^{-1}} \tag{38}$$

where *TP* = true positive, *FP* = false positive, *TN* =true negative and *FN* =false negative. Note that one of the most significant parameters for the classification performance is *F1*, which is the harmonic mean of sensitivity and precision, thus emphasizing the smaller of the two. Since precision is related to the displacement of the decision threshold towards the *negative result* distribution and sensitivity is related to threshold displacement for the *positive results*, *F1* is a representative indicator of the threshold placement between classes.

Table IV. Mean confusion matrix

		Observed class			
Predicted class		*<12*	*12-18.5*	*18.5-23.4*	*23.4-28*
	<12	10	0.2	1.4	0.4
	12-18.5	0	6.7	0	0
	18.5-23.4	0	0	27.3	0
	23.4-28	0	0	0	18.3

Table V. Mean statistical parameters, calculated from the confusion matrix

Class	**N**	**TPR**	**FPR**	**TNR**	**FNR**	**Err**	**PPV**	**F1**
<12	10	1.00	0.04	0.96	0.00	0.03	0.84	0.91
12-18.5	7	0.97	0.00	1.00	0.03	0.003	1	0.98
18.5-23.4	30	0.95	0.00	1.00	0.05	0.02	1	0.97
23.4-28	19	0.97	0.00	1.00	0.03	0.01	1	0.99

TPR = True Positive Ratio, FPR = False Positive Ratio, TNR = True Negative Ratio, FNR = False Negative Ratio, Err = Error, PPV = Precision, F1 = F1 score

The most interesting and promising results are the low misclassification error *Err* and the high *F1* score obtained for each class: the SIMCA models have a good prediction ability concerning the four classes. The "limit" values (0 and 1) obtained in some cases from TPR, FPR, TNR, and PPV could be explained by the fact that, given the low number of test samples and the good general clustering of data, no test samples were classified to a wrong class.

PLSR

Regarding PLSR, we created a series of models with different datasets as input, as already stated in the "Preliminary investigations" section. We inspected the parameter Root Mean Square Error in Cross-Validation (RMSECV), which indicates how closely the models predict the measured values: it can have any values starting from 0, and the smaller, the better: 0.846 (1.5-14 GHZ) –0.743 (1.4-6 GHz) for the real component of S_{11}, 0.814 (1.5-14 GHZ) - 0.835 (1.4-6 GHz) for the imaginary component of S11, 0.964 (1.5-14 GHZ) - 0.984 (1.4-6 GHz). Therefore there is no real advantage of using a kind of spectrum versus others, and we chose VSWR for future implementation ease reasons. Fig. 27 shows the results of the Cross-Validation for the chosen dataset (VSWR, 1.5-14 GHz), where

for the purpose, we used $N = 33$, which are only the spectra acquired closer in time to the destructive tests. For all the $\boldsymbol{X}$ spectra (represented by the blue dots), the prediction of the compressive strength was close to the $\boldsymbol{Y}$ values (the red line represents the ideal fit between the prediction and the observed values). This result is also supported by the excellent value of the parameters RMSECV, 0.986, and R^2_{CV}, equal to 0.997.

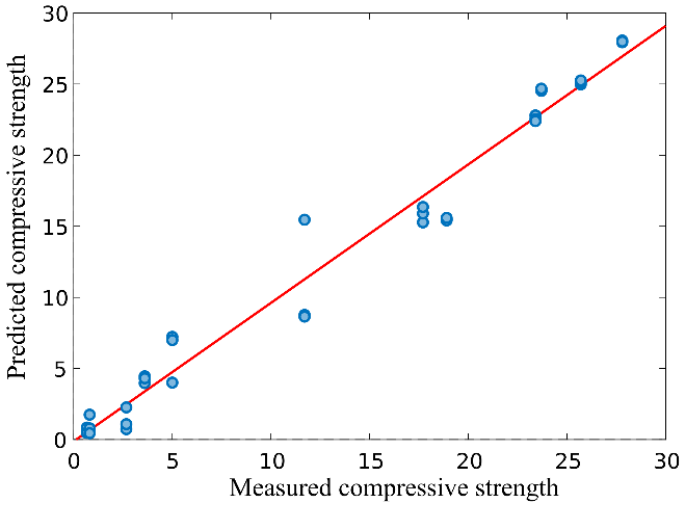

Fig. 27 Predicted vs. observed plot of the PLSR model. The blue dots are the compressive strength predicted by the model (Y-axis) versus the observed values (X-axis). The red line is the bisector, representing the ideal prediction

6.5. Conclusion on concrete moisture detection

A novel non-invasive approach for concrete resistance monitoring during curing time was explored; it consists of the acquisition of reflected microwave spectra coupled with a statistical predictive model. The technique is based on a VNA governed by a PC-based graphical user interface coupled with an open waveguide. This system was used to acquire a total of 409 VSWR spectra in the frequency range from 1.5 to 14 GHz over 28 days: in several frequency sub-ranges, these spectra present a clear descending trend concerning the acquisition time, showing that they contain good information about the amount of free water in the concrete samples. First, we divided measurements into 4 classes thanks to an explorative PCA. At the same time, a series of 14 tests for compression resistance was performed (7 with the penetrometer and 7 with the universal testing machine), and the obtained values were used to link the classes to 4 different concrete resistance ranges, up to 28 MPa. These classified data were the input for two types of predictive models: the first one, SIMCA, was able to classify new acquisitions in one of the 4 resistance classes, whereas the PLSR, applied to a subset of data, predicted the precise values of resistance. The statistical results obtained from the

SIMCA models show a relatively low error of misclassification and a high *F1* score. Also, the regression prediction ability of the PLSR model was good, with a very high value of the parameter R^2_{CV}. The limits of this approach are very similar to the ones already described for the soil moisture predictions (section 5.6): due to time limitations, both the spectra and the resistance values were acquired on the same type of concrete. So, it will surely be necessary to test the model accuracy with spectra acquired on concrete samples with different parameters (water content ratios, components, etc.) and under different environmental conditions. Also the parameters of the acquisition process could be refined in future studies: the results show that the most informative region is a subset of the one we explored (from 1.5 to 6 GHz). Moreover, other reflection parameters than the VSWR, like the phase, the impedance, or the simpler S11 parameter (real or imaginary), could contain information to predict the concrete compressive strength; whereas, regarding classification analyses, different classes and resistance ranges could be tested to find any accuracy improvements. Finally, also the mechanical design of the device needs some improvements, ensuring a stable but non-invasive contact with the samples even in the first phases of the curing process, when the concrete has not yet completely solidified. In conclusion, this initial approach is promising for developing future autonomous and non-invasive monitoring devices for curing time monitoring since the predictive models, once developed offline on a dataset, could be easily implemented in a microcontroller in real time due to its low level of computation.

7. Gas concentration detection in mixes

This chapter aims to demonstrate the effectiveness of the MSA approach, along with a novel sensing device, to achieve real-time GS column-free detection of gaseous compounds. Early and significant results for a real-time, column-free miniaturized gas Mass Spectrometer (MS) in detecting target species with partial overlapping spectra are reported. The achievements have been possible using an innovative gas chromatograph developed by a startup based in Turin, called Nanotech, based on the use of nanoscale holes as a nanofluidic sampling inlet system. Even if the presented physical implementation could be used with Gas Chromatography (GC) columns, the aim of high miniaturization requires investigating its detection performance with no GC aid. For this reason, suitable analytical models were studied to get a semi-quantitative evaluation with very low computational resources. As a study case, dichloromethane (CH_2Cl_2) and cyclohexane (C_6H_{12}) with concentrations in the 6-93 ppm range in single and compound mixtures were used. The nano-orifice approach was able to acquire raw spectra in 60 seconds with correlation coefficients of 0.525 and 0.578 to the NIST reference database, respectively. Then, a calibration dataset on 320 raw spectra of 10 known different mixtures was acquired and used as input for PLSR. The model showed a Normalized full-scale Root-Mean Square Deviation (NRMSD) accuracy of 10.9% and 18.4% for each species, respectively, even in combined mixtures. A second experiment was conducted on mixes containing two other gasses, Xylene and Limonene, acting as interferents. Further 256 spectra were acquired on 8 new mixes, from which two models were developed to predict CH2Cl2 and C6H12, obtaining NRMSD values of 6.4% and 13.9%, respectively.

7.1. State of the art

Novel generations of analytical instruments that use developments in the field of MEMS and NEMS open perspectives for devices with a very high level of miniaturization for GC. Recent advances in GC-MS analytical techniques and more targeted technologies, such as IMS (Ion Mobility Spectroscopy), SAW-MS (Surface Acoustic Waves-Mass spectrometry), GC-SAW (Gas Chromatography - Surface Acoustic Waves-Mass spectrometry), show a clear trend at reducing the size, the analysis time as well as the costs of installation and deployment. Therefore, stringent vacuum conditions should be satisfied, requiring complex differential vacuum systems, bulky connections, and expensive vacuum pumps. However, it is challenging to eliminate the need for relatively large gas inlet flows for the instruments. Examples of these efforts could be found in several recent publications, where the various gasses were injected with an sccm value in the range of 10-200 sccm

[91]–[96]. To further reduce flows, a reduction of the whole system's dimensions was studied by several researchers, obtaining the first consequential reduction of the needed inlet throughputs: for example, in 2007, Kim et al. [97] reported the first integration of a micro GC, where a 4-stage gas micropump was connected to a microcolumn with a length of 25 cm. This system obtained the best vapor separation between 0.2 and 0.3 sccm. More recently, Hsieh and Kim [98] developed a microcirculatory gas chromatography system and tested it successfully on the separation of different isomers, working at a fixed flow rate of 0.5 sccm. Similar results were reported using a particular technology, called Knudsen Pump (KP), based on parallel channels created with nano orifice membranes. On that topic, Qin et al. wrote several papers [99]–[101], developing small systems with a flow of 0.4, 0.82, and 0.15 sccm, respectively. In general, nanotechnology devices can drastically change how these measures could be carried out, allowing radical and extremely relevant system dimensional and power supply reductions. A significant improvement in system simplification [102], [103] is possible by using nanometer-scale orifices[104] as sampling points and smart gas interfaces toward atmospheric pressure.

As long as MS spectra data processing is concerned, in the literature, several papers have applied multivariate techniques (especially PCA) to spectra measured with GC-MS, focusing mainly on classification problems. These works cover several fields, where food is one of the most active: for example, in 2013 Welke et al. [105] used mass spectrometry detection in conjunction with PCA and Stepwise Linear Discriminant Analysis (SLDA) to discriminate between 5 different types of wine, with a success rate of 100%. Lv et al. [106] used GC-MS to acquire fingerprint spectra of Puerh green tea and other six green teas and then used Cluster Analysis (CA) and PCA to evaluate the difference between the Puerh variant and the other ones. More recently, Mogollon et al.[107] performed GC-MS acquisitions on Ecuadorian spirits beverage, whose samples were prepared with a particular technique called Headspace Solid-phase microextraction (HS-SPME). This pretreatment, in conjunction with PCA, allowed them to use mass spectrometry for a valuable quality inspection of these alcoholics. Other studies were conducted on humans: for example, Jha et al.[108] analyzed human body odor data acquired with GC-MS with Kernel PCA (KPCA). This technique allowed them to find volatile compounds that could act as biomarkers, obtaining a good classification between different subjects. In 2019, Stark et al. [109] applied more complex deep learning techniques to MS data acquired on melanoma samples, trying to classify them between a melanoma or non-melanoma mole. Three different deep learning algorithms were explored: Single Layer Perceptron, 1-Hidden Layer Multilayer Perceptron, and 5-Hidden Layer Multilayer Perceptron, with the second one giving better results (63.3% of correct classifications). Jajin et al. [110] considered techniques such as Soft Independent Modelling of Class Analogy (SIMCA) and Orthogonal Partial

Least Squares-Discriminant Analysis (OPLS-DA) to discriminate between subjects affected by medullary thyroid cancer and healthy ones, using spectra from GC-MS on plasma samples. By using OPLS-DA, an R^2 parameter (coefficient of determination, with maximum 1) value of 0.925 was obtained.

7.2. Device description

Physical enabling approach

In this section the critical aspects of the innovative GC will be shown, allowing it to acquire mass spectra in close to real-time conditions. In particular, it will be discussed why this technology could save orders of magnitude in power consumption and response time.

In standard MS systems, molecular ion beams can be generated in several ways and through different techniques, for instance, electronic ionization, discharge ion source, photoionization, etc.[104]. Once generated, the ion beams must fly into a mass filter first to be selected (through a single mass filter or with a tandem mass spectrometer), then through a couple of mass filters with a scattering cell down to a detector to measure the intensity of the selected ions. As well known, to reduce severe scattering effects and consequential losses of the ion beam, it is critical to reaching a gas regime for the MS system where the mean free path of the ion is comparable with the geometrical dimensions of the analytical system *D*, flying from the ion source to the detector, alias, achieving a Knudsen number $K = \lambda/D > 1$, where *D* is the dimension of the vessel and λ is the mean free path

$$\lambda = \frac{k \cdot T}{\sqrt{2} \cdot \sigma^2 \cdot p} \tag{39}$$

where *T* is the temperature (in Kelvin degrees), σ is the scattering cross section, *p* is the pressure (in Pascals), and *k* is Boltzmann's constant. Therefore, to reduce MS losses of the ion beam, it is required to reach a gas regime where the mean free path λ of charged particles is comparable with the geometrical dimensions of the analytical system D. Thus, ions should fly from the source to the detector, requiring pressures of a standard analytical system (whose length is about tens of cm) in the range of $10^{-6} \div 10^{-7}$ mbar. In this case, ions collide mainly with the inner chamber walls rather than with each other.

To understand the critical features of nanodevice-based MS, it is needed to look at the continuity equation of a single vessel having an inlet throughput or gas flow/rate Q (in $mbar \cdot L \cdot s^{-1} \equiv W$) and outlet effective pumping speed (or volumetric flow rate) $S = dV/dt$ (in $L \cdot s^{-1}$) as

$$V \cdot dp = Q \cdot dt - S \cdot p \cdot dt \rightarrow -V\frac{dp}{dt} = S \cdot p - Q \tag{40}$$

where again, *V* is the volume of the vessel.

It is easy to show that the differential equation could be solved with boundary conditions as an inverse exponential decay of pressure:

$$p(t) = \frac{Q}{S}\Big|_{\infty} - \left(\frac{Q}{S}\Big|_{\infty} - p_0\right) e^{-\frac{S}{V}t} \tag{41}$$

where $p_0$is the initial pressure, $(Q/S)_{\infty}$ is the pressure at stationary regime, and $\tau = V/S$ is the time constant of the system. When the system achieves a steady state, we have:

$$Q = p \cdot S = p\frac{dV}{dt} \tag{42}$$

Also, at the stationary regime, another equation relates pressures across an orifice through the conductance *C* (in $L \cdot s^{-1}$)

$$Q = C \cdot (p_2 - p_1) \tag{43}$$

where p_2 and p_1 are the pressures across the orifice.

When multiple vessels are interconnected by pumps and orifices, the constant mass flow constraint sets(by using (42) and (43) for each vessel) an *N-th*-order differential equation that gives pressures at each point under initial conditions. An electric equivalent model is usually defined to better understand the behavior, where electric potential, capacitance, and current are equivalent to pressure, volume, and throughput, respectively, as shown in Fig. 28.

In standard MS systems, the gas sample is eluted into a gas carrier (contained in a tank) to flow into a chromatography column for separation at constant throughput. To achieve molecular regimes, a system of multiple chambers and pumps is used, as shown in Fig. 28 (A), where $Q = S_1p_1 + S_2p_2 + S_3p_3$ (we used a simplified 3-chamber system to show the concept). As the electrical model shows, it can drop the pressure (electric potential) at constant throughput (electric current). In industrial systems, a standard unit for Q is the SCCM@1bar (in $cm^3 \cdot min \cdot 1bar \equiv 1.66W$), using 10 SCCM@1bar as a typical value. With this value, it is easy to show that for such a flow and using a unique chamber the pumping speed would be technically impractical, requiring multiple chambers. However, also in this case the power effort is very relevant. As a rough example, pump speeds of S_1~*700 L/s*, S_2~*300L/s*, and S_3~*300L/s* would require electric power of about *700 W* using 61 turbomolecular and rotary oil vane pumps.

Conversely, using a nanometric orifice by sampling the gas to be analyzed at ambient pressure allows for achieving molecular regimes with a reduced number of chambers and power requirements, as shown in Fig. 28 (B). Moreover, the dominant time-constant $\tau = V_1/S_x$ (since the inter-chamber conductivities are much lower than output effective conductivities) is dramatically decreased. More specifically, in standard systems, the typical values of V_1 are ~1000 cm^3 and S_x ~ 0.1-1 L/s, while for the proposed technique, it could be V_1 ~ 1 mm^3 and S_x ~ 0.1-1 L/s, thus reducing the time constant of orders of magnitude.

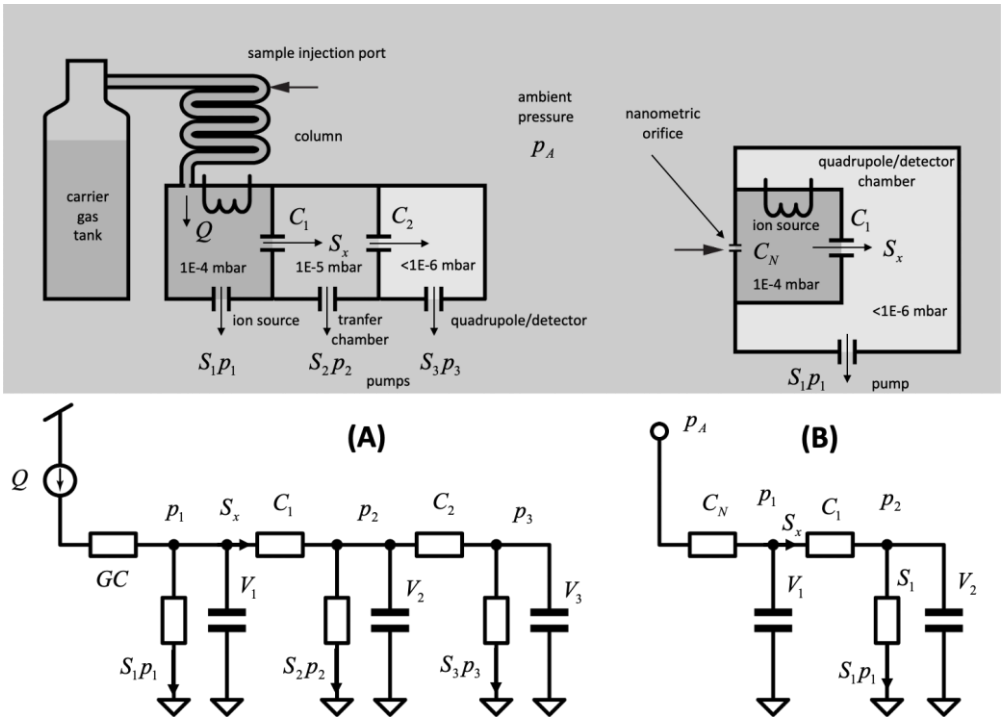

Fig. 28 Comparison between the conventional approach, **(A)** and nanodevice approach, **(B).** The electric model lets us understand the pressure behavior between chambers at transient and steady-state regimes. The model is based on the analogy: throughput - current: $Q \leftrightarrow Sp$, volume – electrical capacitance: $V \leftrightarrow C$, and pressure – electric potential: $p \leftrightarrow V$, since both systems obey to the same differential equations.

In conclusion, using nanodevices allows for a dramatic decrease in the complexity of the overall system using a single or multiple (array) orifices at a nanometric level in molecular flow, achieving very low conductances. In the molecular flow regime, rather than considering the collective motion of the fluid, we can focus on the motion of the single molecule, flying "practically alone" from one end to another of a pipe and only on a statistical base. In this case, a conductance C (in $L \cdot s^{-1}$) does not depend anymore on the pressure (as in Poiseuille's equation) at its ends but only on its geometry, average molecule velocity (or temperature), and molecular mass as:

$$C = \frac{1}{4} \cdot \left(\frac{8kT}{\pi m}\right)^{\frac{1}{2}} \cdot A = \frac{1}{4} \cdot \left(\frac{8kT}{\pi m}\right)^{\frac{1}{2}} \cdot \frac{V}{l} \tag{44}$$

where A is the surface of the aperture, T (in ${}^0K$) is the temperature of the gas, V and l are the nano-orifice's volume and depth, and m is the molecular mass (in kg) of the gas analyte. As an example, using (44) for a 490nm round hole diameter and 250nm depth, a conductance of about 21.7 nL/s for air particles is obtained. With this value, eq. (42) gives as a result a throughput at ambient pressure of about 1.3×10^{-3} SCCM@1bar, thus several orders of magnitude lesser than a standard mass spectrometer that is using about 10 SCCM@1bar. The dependence of the conductivity vs. the molecular mass (lighter gases enter at a higher rate because C is higher for the weightier gases) should not be of concern because the same effect occurs at the exit flow. Therefore, it is feasible to have the same gas concertation sampled at atmospheric pressure but a much lower pressure level under the mass balance equation.

To summarize, the nanometric orifice technique achieves the following advantages towards standard MS systems:

1) simplified mechanical implementation and reduced power consumption;
2) reduced sampling time constant;
3) reduced throughput.

These characteristics allow measurements in real-time and drastically simplify the analytical device

Mechanical setup

An exploded conceptual view drawing of the miniaturized mass spectrometer is shown in 29 (A) and is based on a combination of micro and nano technologies (MEMS and NEMS) with techniques currently used for analytical measurements. The analytical prototype is equipped with a nano gas sampling device realized through nanoscale orifices directly interfaced with standard components such as an ion source, ion lenses, mass filter, and a detector to directly sample targets at atmospheric pressure. The experimented prototype is shown in 29 (B). An encapsulated nanomembrane interface uses nanometric orifices and acts as a smart sampling device operating directly at atmospheric pressure in the molecular regime. The sampled inlet gas flows directly into an ion source, where an ion beam is generated. Then, a single quad mass filter can select defined ions detected through a Faraday cup or a second channel electron multiplier (CEM), simplifying the vacuum system and consequently realizing measurements in real time. The holes' typical diameter is 500 nm, even if this technology allows high versatility around specific needs. Depending on the application, it is possible to realize membranes with single or arrays of orifices tailored to compounds in a complex matrix at reduced concentration [111]. A quadrupole mass filter (CIS 300 by Research Systems) is equipped with a closed ion source and with a second channel electron multiplier (model 4220 Stanford Research Systems) as a detector; the spectrum could be recorded by setting the CEM Voltage to amplify the signal also for less concentrated samples. An SEM microphotograph of the nanometric orifice is shown in Fig. 30 where a membrane-in-membrane structure was adopted, and smaller sub-membranes are realized where nanoscopic holes are created [111]. Several devices were fabricated and tested, having hole diameters from 300 to 600 nm on a membrane side of 80 μm.

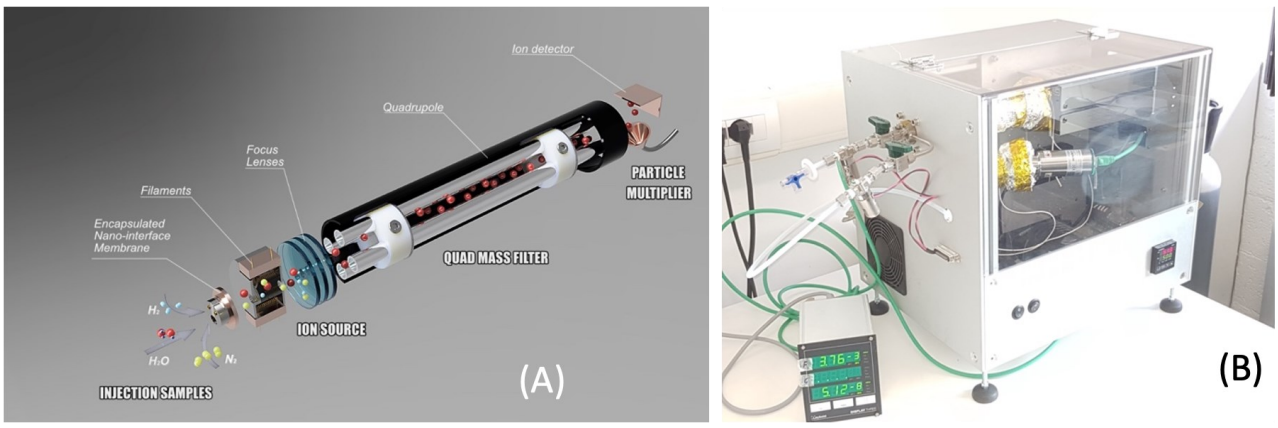

Fig. 29 Mechanical conceptual structure of the nano interface integrated into an ion source (A) and prototype implementation (B).

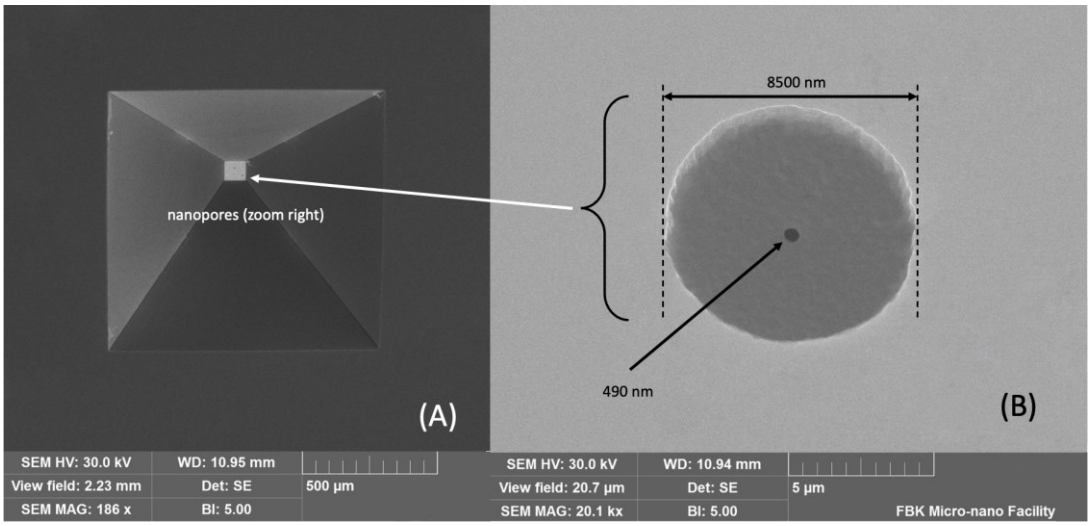

Fig. 30 Mechanical conceptual structure of the nano interface integrated into an ion source (A) and prototype implementation (B).

7.3. Experimental setup

Two common solvents, usually used in analytical applications, were chosen as prediction targets: dichloromethane (CH_2Cl_2) with a molecular weight of 85 atomic mass units (a.m.u), and cyclohexane (C_6H_{12}) with a molecular weight of 84 a.m.u. More specifically, they are both volatiles showing relatively large fragmentation patterns ranging between 40 a.m.u. and 85 a.m.u: there are regions of the spectrum where the fragments generated during the electron ionization process overlap between their relative mass peaks. CH_2Cl_2 and C_6H_{12} compounds were diluted in a gas matrix of argon prepared using three-liter bags and a polypropylene valve with a septum that was used to inject the liquid standards into the bag. Liquid compounds used for sample preparation were obtained from Merck (Darmstadt, DE). All the gaseous solutions were prepared on the day of use and stored in environmental conditions of temperature and pressure. For the preparation of the gas samples, a matrix of argon (99,9999%) (Nippon Gases), Tedlar® bags (Restek, PA, US), and a flowmeter (Brooks Scientific, DE) were used to fill bags.

Samples preparation

For the preparation of Tedlar bags standards, the Full Evaporation Technique (FET) was used [112]. This technique used a small amount of pure sample (a few µL), reducing the operator's exposure to toxic substances. The FET was based on a transfer of analytes from a condensed matrix, liquid or solid, into a confined vapor phase[112]: the analytes were induced to evaporate into the Tedlar bag until a condition of equilibrium in a short time was reached under the condition $P<P^o$, where P is the

pressure of the moles of analytes in the volume of the bags at a temperature of work, and P^o is the saturated vapor pressure of the sample. The injection tube of the Tedlar® bag was connected to the flowmeter, and the bag was flushed with argon for 18 min at ambient temperature and a primary pressure of 2 bar until 2.4 liters using a flowmeter. Then, a small volume of the liquid matrix (in the order of μL) was injected into the septum of the bag using a gas-tight syringe, and they were left to evaporate in the gas matrix to obtain the stock solution. Using a 25 mL gas-tight syringe, the diluted solutions were prepared and injected a few mL of the stock solution into the Tedlar® bag filled with 2.4 mL of argon. This method was used to prepare 6 ppm, 30 ppm, and 58 ppm solutions for CH_2Cl_2, and 20 ppm, 50 ppm, and 93 ppm solutions for C_6H_{12}, respectively. In addition to the bags with only CH_2Cl_2 and C_6H_{12}, other bags with both were prepared using the same method. The concentrations of CH_2Cl_2 and C_6H_{12} used for the dataset for a total of 10 mixture combinations are reported in IV.

Table VI: Concentration of CH_2Cl_2 and C_6H_{12} and their mixtures analyzed in the experiment.

CH_2Cl_2 (ppm)	**C_6H_{12} (ppm)**
6	0
30	0
58	0
0	20
0	53
0	90
30	50
30	93
58	50
58	93

Spectra acquisition

The chip membrane hosting nano-orifices was positioned between the external environment's high-pressure side (about 1013 mbar) and the low-pressure side toward the quadrupole, allowing to carry out samplings at constant pressure. No chromatographic column upstream was used. The Tedlar bag was connected to the sample holder compartment, and the gaseous sample was flushed a few minutes before recording the mass spectrum. Spectra have been recorded from 45 to 90 u/e^-, for the concentrations indicated in Tab. VI where each spectrum acquisition lasts about 60 seconds. At first,

the samples containing only CH_2Cl_2 and C_6H_{12} were examined. Fig. 31 (A) and (B) show an example of the mass spectra of the analytes in the maximum concentration of the experiment, 58 ppm for CH_2Cl_2 and 93 ppm for C_6H_{12}, respectively. Then, the mass spectrum of the gaseous mixtures has been recorded, where an example (58 ppm CH_2Cl_2 plus 93 ppm of C_6H_{12}) is shown in Fig. 31 (C). As a reference, the NIST spectra of the two species are shown in Fig. 30 (D). Due to the absence of the GC that would differentiate the analytical species based on the different retention times, it is apparent how spectra are partially overlapped, and significant fragmentation peaks (the main at 84 u/e$^-$) could be seen.

In a second acquisition campaign, we acquired, in the same way, the mixes presented in Tab. VII, in the range of 47 to 110 u.m.a. The NIST spectra of the 4 components are shown in fig. 32 (A), and an example of the spectra of one of these mixes is depicted in Fig. 32 (B).

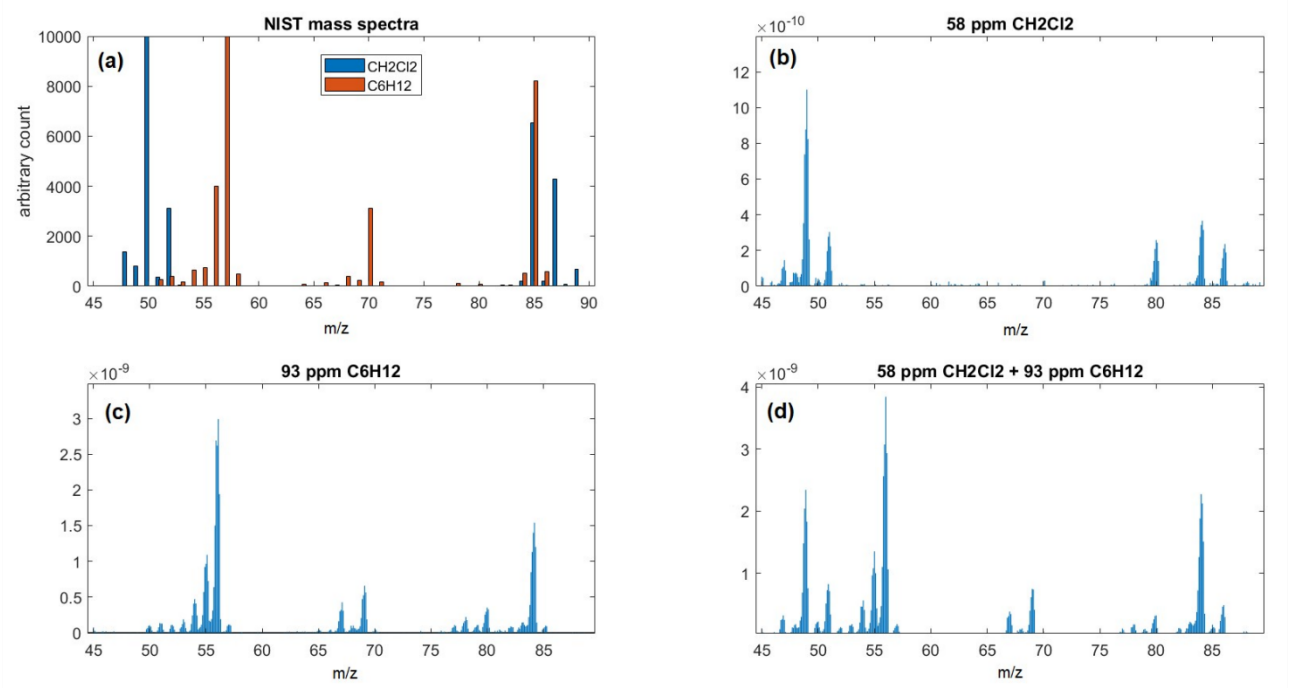

Fig. 31 Mass spectra recorded in an interval of 45-90 u/e$^-$. Each spectrum acquisition lasts about 60s. **(a)** NIST mass spectra (at 70 eV Electron ionization energy) of CH_2Cl_2 and C_6H_{12}, **(b)** Mass spectrum of 58 ppm of CH_2Cl_2, **(c)** 93 ppm of C_6H_{12}, **(d)** 93 ppm of C_6H_{12} combined with 58 ppm CH_2Cl_2.

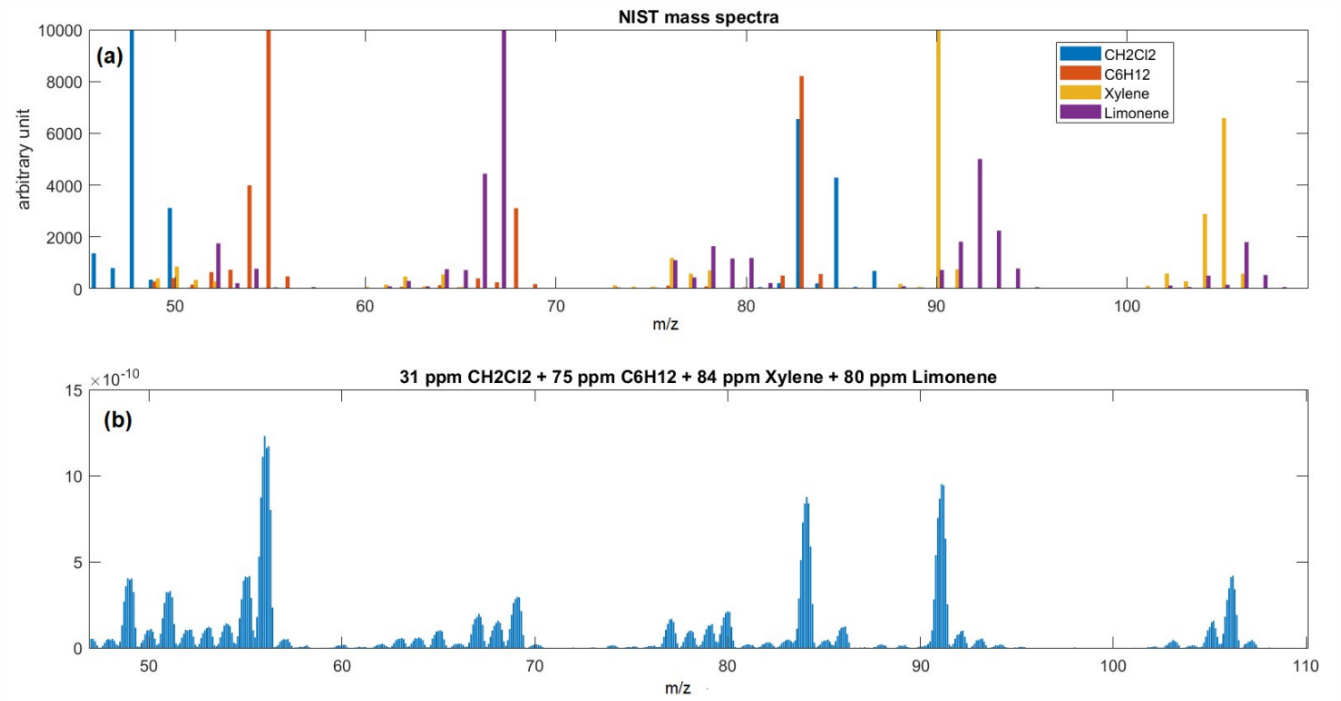

Fig. 32 Mass spectra recorded in an interval of 47-110 u/e$^-$. **(a)** NIST mass spectra (at 70 eV Electron ionization energy) of CH_2Cl_2, C_6H_{12}, Xylene, and Limonene. **(b)** Mass spectrum of a mix of 31 ppm of CH2Cl2, 75 ppm of C6H12, 84 ppm of Xylene, and 80 ppm of Limonene

7.4. Results

Spectra analysis

To evaluate the raw data consistency, the error between acquired data and NIST reference normalized data [113] was evaluated. More specifically, it is well known that MS spectra are subject to non-linearities due to physical system and electronic readout, so the global error should be considered. Therefore, the spectra were normalized to the maximum peak, as in NIST references, for singular and composed compounds. Then, the data were compared by removing the noise floor using correlation coefficient[114] after averaging 32 spectra and using alignment pre-processing (see "Data augmentation and preprocessing" section). A comparison between experimental data and NIST reference is shown in Fig. 33. The correlation coefficient is r = 0.525 for C_6H_{12}, r = 0.578 for CH_2Cl_2, and r = 0.513 for the compound C_6H_{12} + CH_2Cl_2. Therefore, it is apparent that spectra acquisition undergoes non-linear effects that alter the ratio between peaks, which is an essential feature of the spectrum fingerprint. We will see in the Discussion subsection that notwithstanding the deformation of spectra, the multivariate analysis will be able to overcome the problem, thanks to the construction of a dataset based on known references.

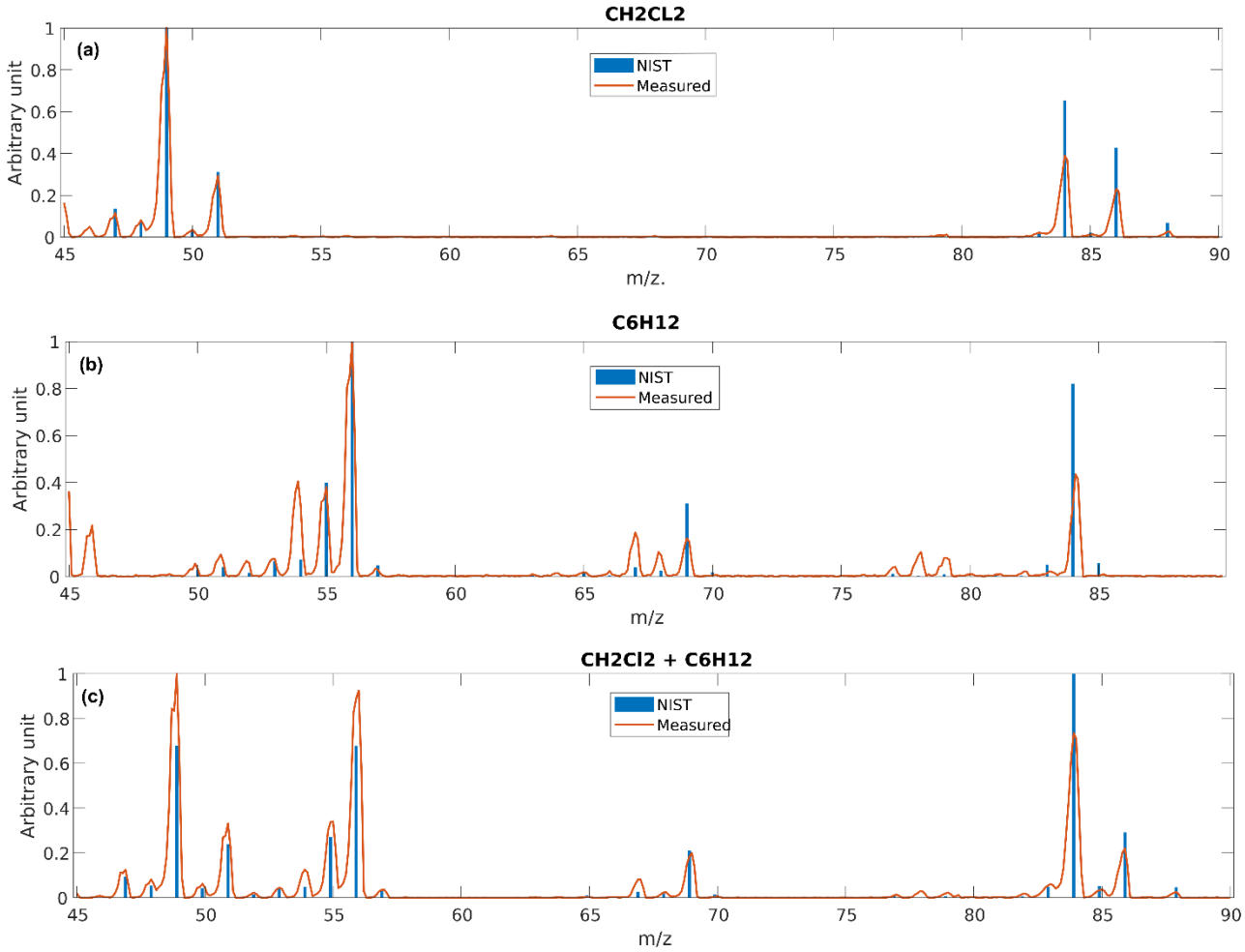

Fig. 33 Comparison between acquired spectra and NIST reference for CH2Cl2 **(a)**, C6H12 **(b)**, and CH2Cl2 + C6H12 **(c).**

For MSA, we focused on PLSR2 because we aimed at estimating semi-quantitative concentrations of CH_2Cl_2 and C_6H_{12} in a two-variable output ($M = 2$). We used 12 new spectra as a test set to investigate the model prediction ability, measured on a single gas or a combination of the two. The goodness of the model fit was assessed with the coefficient of determination (R^2) and the Normalized full-scale Root-Mean Square Deviation (NRMSD).

Data augmentation and preprocessing

The **X** calibration dataset was created with spectra acquired for each one of the CH_2Cl_2 and C_6H_{12} mixes reported in Tab VI for a total of 2277 acquisitions, collecting about 230 spectra of 427 points for each combination. To reduce noise and concurrently perform data augmentation, the mean of 20 random spectra for each combination was calculated, for a total of 100 averaged spectra for a final $\boldsymbol{X}$ dataset, where $K=427$ and $N=100$. From experimental data, it was found a shift between the peaks of spectra due to an intrinsic error of the mass spectrometer around ± 0.25 u/e$^-$, so, a Matlab function

named *icoshift*, developed by Savorani et al. [115] was applied to averaged spectra, to realign them. Finally, autoscale preprocessing was applied (mean centering and scaling of each variable to unit standard deviation).

Prediction tests

For Cross-Validation, the obtained model showed a reduced number of latent variables, LVs = 5, and appreciable values of the coefficient of determination, $R^2 = 0.886$ for CH_2Cl_2 and $R^2 = 0.900$ for C_6H_{12}. The model was tested with spectra measured on 12 unknown mixes of CH_2Cl_2 and C_6H_{12}, not present in the $\boldsymbol{X}$ calibration dataset, averaged and aligned as the calibration set. The relationship between measured and predicted concentrations is shown in Fig. 34, where prediction model error bars are also displayed. Considering all the numerical values, the estimated concentrations of gasses obtained with the PLSR are quite good: the accuracy is estimated in normalized full-scale root-mean-square deviation (NRMSD) accuracy of 10.9% for C_6H_{12} and 18.4% for CH_2Cl_2. Note that the accuracy considers the presence of both species in detection.

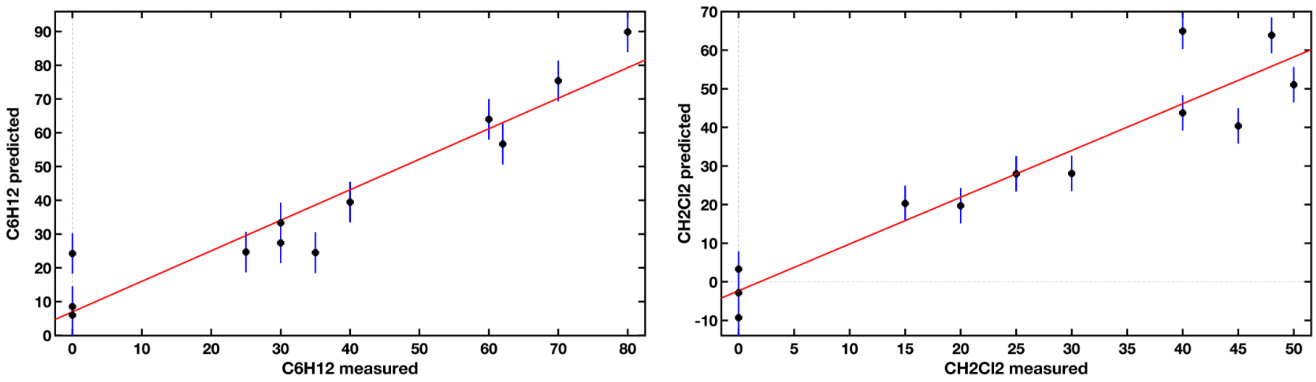

Fig. 34 Predicted and measured concentrations with prediction model error bars. Units in ppm.

With the data acquired on the second campaign, two models were created, for the prediction of CH_2CL_2 and C_6H_{12}, respectively: with the addition of Xylene and Limonene, the differences in the accuracy made us prefer two models that predict a single variable to a single model that predicts both. Given the simplicity of the calculation required for the variable prediction (see Eq. (10)), the variation in the calculation time using two models instead of one is negligible. These two models were tested with 3 unknown mixes containing all 4 elements. The prediction plots are shown in Fig. 35: the black dots represent the spectra used for the creation and the calibration of the models, and the red dots the test spectra. The results were very good: for CH2Cl2, we obtained an R^2 of 0.981 and an NRMSD of 4.7%; whereas for C6H12, we obtained an R^2 of 0.987 and an NRMSD of 17%.

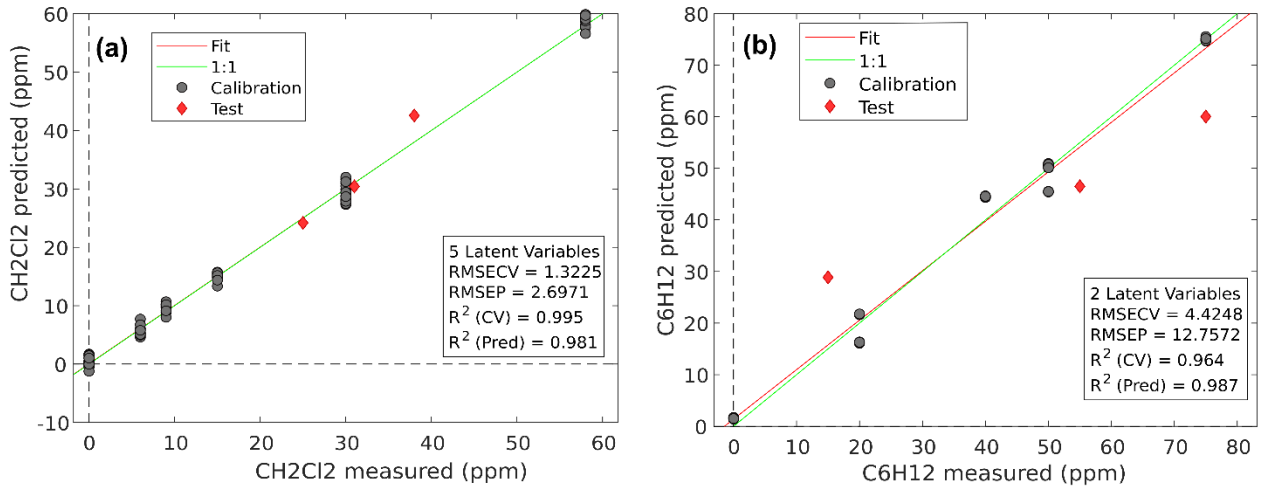

Fig. 35 Prediction plots of the two models created with the spectra acquired on mixes of 4 components. **(a)** Prediction of CH2Cl2 **(b)** Prediction of C6H12. In both plots, the black dots represent the spectra used to create and calibrate the model, and the red dots are the spectra used as the test set. The green line is the bisector of the plot, representing the ideal prediction, whereas the red line represents the linear fit of the model.

7.5. Conclusions on gas concentration detection

This section has shown semi-quantitative multivariate analysis results of experimental spectra from a nanodevice-based mass spectroscopy system where no chromatographic column upstream was used. A data set was constructed on 320 raw spectra derived from 10 different blends of two compounds with overlapped peaks acting as interferents. The model showed an accuracy with an NMRSD error of 10.9% and 18.4% for each species, respectively, even in combined mixtures. Then, a second dataset was created with 256 spectra acquired on mixes of the two compounds and another two gasses used as interferents. One model for each of the two compounds was created, and they both showed good prediction ability, with NMRSD of 4.7% and 17%, respectively.. The accuracy of the model could be increased by widening the $\boldsymbol{X}$ calibration dataset with more acquisitions in a higher number of different concentrations for the two gases, using a more extensive and time-consuming setup. However, once the model is built offline, it could be easily implemented in a real-time detection system with very low computational resources. It is also important to note that the gasses used for this preliminary test do not represent interesting targets for a real in-situ applications. So, the system needs to be tested for the prediction of more interesting gasses and/or compounds, which, however, could be more complex to monitor. A perfect example of this is the detection and monitoring of formaldehyde, whose predictions could have very important practical implications for industrial and healthcare applications, but whose characteristic peaks need to be distinguished from the ones of the nitrogen, the most present element in the atmosphere.

8. Fish freshness detection

This chapter focuses on assessing the freshness of fish samples in a non-invasive way, exploring the potentiality of the HyperSpectral Imaging (HSI) technique in sardines (*Sardina pilchardus*) freshness monitoring: HSI images of the whole fish body were acquired in the VIS-NIR range (440 – 1000 nm), and a mean spectrum was calculated from each spectrum. Then, these spectra were used as input for a Principal Component Analysis (PCA), to evaluate differences in the spectra caused by the decaying process. A second acquisition campaign of 7 days was performed on anchovies samples, focusing this time only on the fish eye region. Both techniques were able to group samples according to storage days, showing promising results for practical application, not only for determining the mechanism of fish freshness but also for developing new techniques for the non-destructive evaluation of the decaying process.

Given the fact that the best results were obtained monitoring only the eye region, an algorithm for the automatic detection of this Region Of Interest (ROI) was then developed, using an already existing Artificial Neural Network (ANN), called ResNet50. This algorithm was trained on the fish eye images acquired in the previous 2 campaigns, using a Python package called Detecto. The free platform Google Colab was used to perform all the calculations on GPUs, greatly reducing the required time for the training phase. The prediction ability of this ANN was tested on a small test set (37 images) and the results are very promising: all the eye regions were correctly identified, with a mean accuracy of 0.985 (over a maximum of 1). Section 8.4, after a brief explanation of the basis of ANN theory, will describe the creation and training of the model, as well as the obtained results.

8.1. State of the art

Consumer acceptance and food safety are key concerns for wholesalers and retailers of fresh fish and seafood products [1]. Fish and fish products are appreciated worldwide and have a very important role in balanced human nutrition, providing several different nutrients and health benefits (protein, omega-3 fatty acids, vitamins). However, fresh fish and fish products are highly perishable products and are subject to fast development of undesirable odors and flavors, and to a rapid decay process. So, freshness is the most important aspect to monitor in fish. With time dead fish show an increase in the growth and activity of microorganisms and the oxidation of lipids [116]. These two processes cause a rapid deterioration of the critical quality parameters for the consumers; appearance, odor, and taste [117], [118]. The speed and severity of these changes are highly variable and depend on

parameters such as species, fat content, storage conditions, and temperature [119]–[121]. Moreover, refrigeration does not stop microbial activity as a whole because of the presence of psychotropic bacteria [122]. The standard method to assess fish freshness is sensory inspection and this has been used by the European Union since the 1970s [123]. Nowadays, the most common and widespread sensory method is u the quality index method (QIM) which is based on evaluating the body parts of the fish changing during the decaying process, such as eyes, gill, and skin. This method is very specific because a different evaluation table is required for every fish species [124]. Regarding anchovies, several papers focus on the development of QIM schemes [124], [125], or use them to assess changes. For example, the increase of total volatile basic nitrogen (TVBN), the determination of cholesterol oxides, or the impact of natural plant extracts in the fish samples under various conditions [126]–[128]. However, these methods have some disadvantages: they are time-consuming and require highly skilled operators, making them difficult to use with in-line or online industrial applications. To overcome this, over recent decades several indirect and non-destructive techniques have been developed to assess physical and chemical parameters related to fish freshness, ranging from biosensors [129] and electronic noses to various spectroscopic analyses [130], [131]. Among the latter, techniques that operate in the visible and near-infrared range (VIS/NIR) are of particular importance, because they allow the detection of the vibrations of C–H, O–H, and N–H groups [132]. Recently, VIS/NIR devices have been used to assess several different parameters, such as the cold storage time of salmon [133], trimethylamine concentration and K-value in silver carp [134], and discrimination between fresh/thawed Atlantic mullets [135]. An evolution of these IR spectroscopy techniques is represented by hyperspectral imaging (HSI). Thanks to special cameras, a whole electromagnetic spectrum can be acquired for every pixel of an image, providing qualitative and spatial data at the same time. Good results have been obtained for a wide variety of species. Ivorra et al.[136] developed a model for shelf-life prediction of salmon, Cheng et al. [137] evaluated the K-value in grass carp and silver carp, Khoshnoudi-Nia and Moosavi-Nasab [138] assessed the values of total volatile base nitrogen, psychotropic plate count, and sensory score in rainbow trout fillets, Also, the study of more simple RBG images (usually focusing on the eyes and gills) represents an active research field, as evidenced by recent works on common carp and goldfish [139], [140] and rainbow trout [141].

8.2. First Acquisition campaign

Acquisition setup

Hyperspectral images were acquired with a Nano-HyperSpec Vis/NIR camera by Headwall. It acquires a total of 240 wavelengths, in the range of 440-1000 nm, with a pushbroom technique: the sample is in movement and the camera is stationary, acquiring all the spectra of a single line of pixels at the same time. Fish samples were placed and moved under the camera thanks to a conveyor belt, equipped with two halogen lamps sideways, with an inclination of 15°.

The whole HSI acquisition system is shown in Fig. 36. All the acquisitions were performed with the whole system covered by a cardboard case, to avoid light dispersion in the room. The images were measured on fish samples stored for 0, 1, 2, and 5 days. For each day of storage, acquisitions were conducted on 20 sardines at a temperature of about 20° C, positioning the fish on the right side. The spatial length of the resulting hyperspectral image is 640 pixels, whereas the length is variable from 500 to 700 pixels, depending on when the camera acquisition is manually stopped by the software. Black and white reference images were also acquired at the start of every day, to reduce the distortions caused by variations in illumination and sample geometry, following Eq. (45)

$$R(x,\lambda) = \frac{I - I_D}{I_W - I_D} \tag{45}$$

where R is the corrected image, I is the raw image, I_D is the dark image, and I_W is the white image.

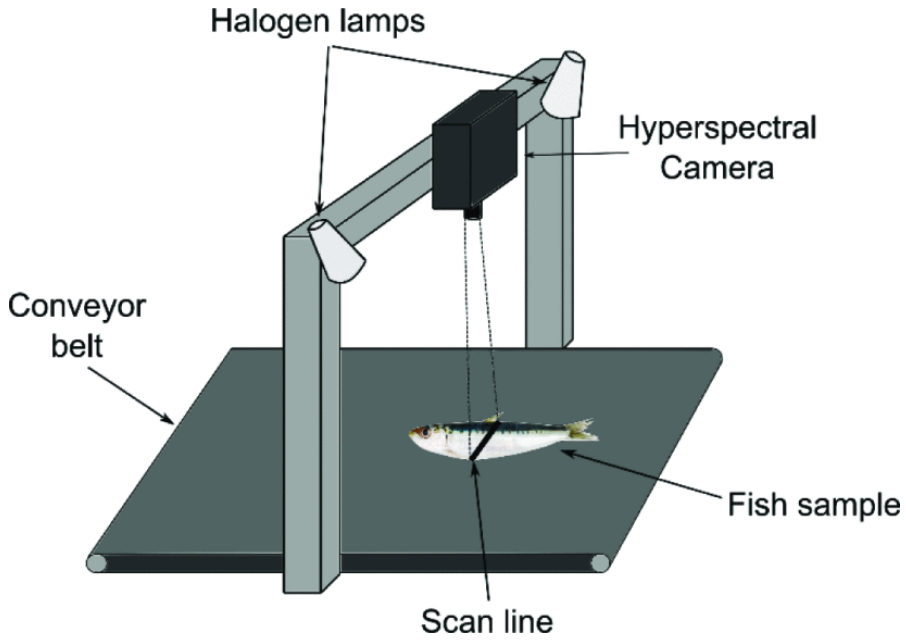

Fig. 36 Setup used for the acquisition of HyperSpectral images of sardine

First acquisition campaign

Fresh sardines (*Sardina pilchardus)* were purchased in the Romagna region (Italy) and immediately carried into the laboratory, where they were stored in ice at the temperature of 4 °C. The mean value of the fish mass and length were 9.2 g (± 1.2 g) and 95 mm (± 4 mm), respectively.

Before the creation of the PCA model, three preprocessing techniques were applied to the mean spectra: smoothing, to eliminate the noise from the data, a Savitzky-Golay second derivative algorithm with second-order polynomials and Multiplicative Scatter Correction (MSC). Derivatives and MSC are commonly used as spectroscopic preprocessing, especially in the NIR range. In particular, they are useful for scaling effects and removing baseline offsets. The raw and preprocessed data are shown in Fig. 37. It is easy to see that the original data have strong baseline shifts, and the spectra from different days are interlined between them, without any clear trend to the decaying time. On the other hand, the preprocessed data are more compact, thanks to the second derivative algorithm and MSC, and present a much clearer differentiation between the classes. These preprocessed data were used as input for the creation of a PCA model.

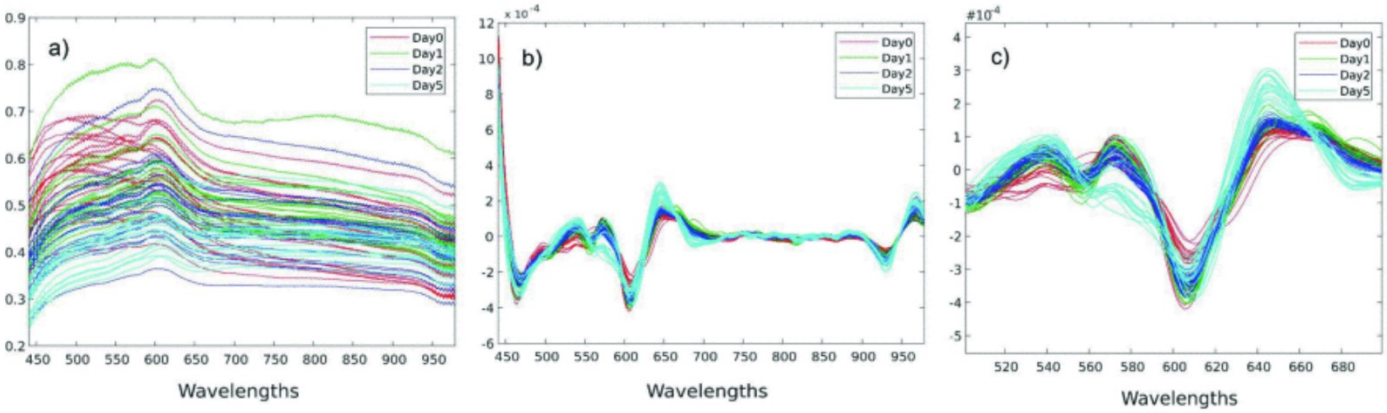

Fig. 37 **(a)** Raw and **(b)** preprocessed mean spectra of the hyperspectral images, after the selection of the ROI. **(c)** Magnification of the preprocessed mean spectra.

Results

The obtained PCA model reduces the number of variables from the original 240 wavelengths to only 3 PCs, where the vast majority of data variance is contained in the first one (94 %).

One important graph to study in a PCA model is the score plot, between the major PCs, to evaluate how the data variables can be explained by these PCs [24]. In Fig. 38 (a)-(b) the score plot of the model is compared to the score plot of a model created with raw data, showing again the benefits of the preprocessing phase. It is possible to see that, in the first plot, the sample scores are mainly distributed according to time (days) along the first PC, that alone explains almost all the variance of the system; whereas in the second plot, the distribution is along the second PC, that explains only the 3% of the variance. Moreover, in the first plot, the scores are more clustered and less spread out. Another useful plot to study the model characteristics is the loading plot, represented in Fig. 38 (c). This graph shows how much each measured frequency weight is in the calculation of a PC [33], in our case of PC1. These weights, represented on the y-axis, are called loadings. It is easy to see that in the range of 700-900 nm, the loadings have a value close to zero, meaning that these wavelengths do not carry useful information for our model. This allows us to perform a first preliminary selection of the most informative variables: the removal of the data contained in this range does not affect the new model and score plot, decreasing at the same time the number of variables and the data weight. In conclusion, the PCA analysis showed that the preprocessed data could be a good input dataset for

a classification model, allowing the evaluation and the monitoring of fish freshness and decaying state directly from HSI acquisition.

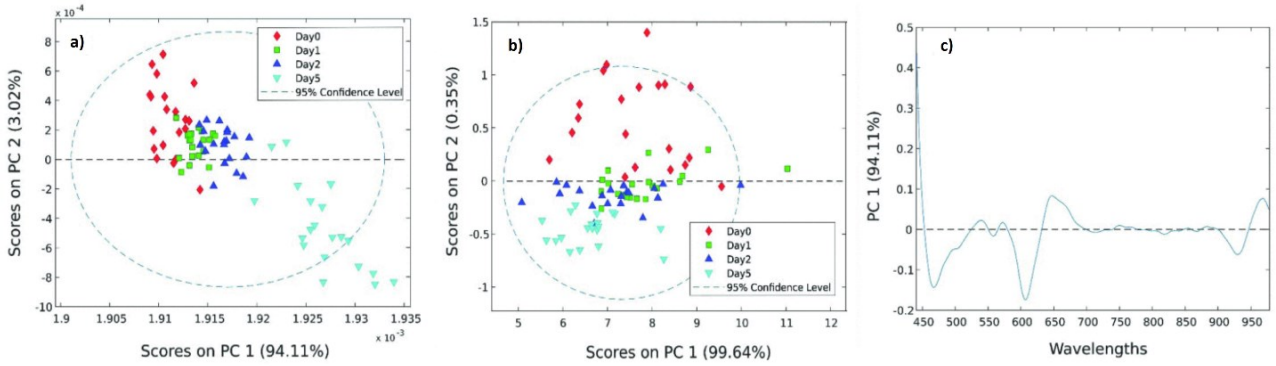

Fig. 38 Setup used for the acquisition of HyperSpectral images of sardine. Score plots of the PCA model created with (a) preprocessed and (b) raw data.

8.3. Second acquisition campaign

Anchovies (*Engraulis Encrasicolus)* were purchased soon after fishing in the Romagna region, Italy, immediately stored in a polystyrene box covered by ice, and carried to the laboratory. Fresh anchovies were stocked in a refrigerator at 0 °C (±0.5 °C). Anchovies were soon characterized by mean values and standard deviations of mass, length, and width obtaining 8.3 ± 0.98 g, 94.7 ± 3.4 mm, and 14.8 ± 0.8 mm, respectively. Measurements were conducted after 1, 2, 3, 4, and 7 d on different fish batches according to the assessment type. On each storage day, different batches of fish were investigated to consider the variability of the samples in the freshness assessment ability of the proposed non-destructive techniques. Measurement day and final storage day were chosen based on several studies evidencing the endpoint of anchovy edibility, stored at 0 °C under ice, between 6 and 8 d [124], [125], [142]. All the measurements were conducted at room temperature.

This time, a region of interest (ROI) was obtained from each image, manually selecting the fish eye and the mean spectra were calculated by averaging the spectra of this region. The spectral band between 400 and 450 nm was excluded due to the low signal-to-noise-ratio produced by the camera sensor, according to [143]. A total of 252 spectral data points were used as X-variables for the multivariate data analysis approach. The spectra were smoothed (Savitzkye-Golay method) to reduce noise from the spectra, then pre-treated by the standard normal variate (SNV) and first derivative, and finally mean-centered. SNV is a common pre-processing method implemented to correct spectra for light scattering, while the derivatives can remove both additive and multiplicative

spectral effects. Subsequently, principal component analysis (PCA), was used as an explorative technique to visualize the data according to fish storage time.

Results

The 3D score plot of the PCA is reported in Fig. 39 (a). Good separation between the samples, according to the storage time, was achieved according to storage time. Especially along the 1st PC (day1 vs day2 vs day3) and the 3rd PC (days1, 2, and3 vs day7). The loading plot in Fig. 39 (b) helps understand which physical changes are described by the PCs, showing how much each measured wavelength contributes to the variance explained by these new directions. It indicates that all the PCs had high loading values in the range of the visible (500–700 nm), especially the first PC, which alone explained the 58.55% of the data variance, pointing to changes in the sample colors during the decaying process. The second and third PCs presented maximum loadings values in the range of 900–1000 nm, in proximity to one of the absorption bands of the water (970 nm), as already described for the dielectric spectra, the moisture and water content of anchovies changed greatly during the 7 days.

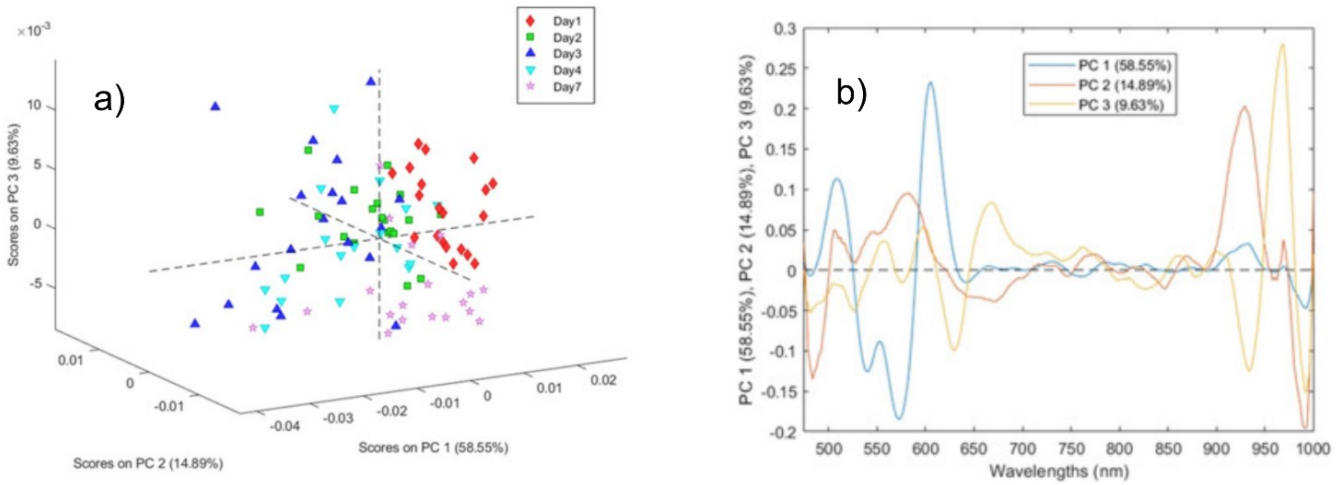

Fig. 39 **(a)** Score plot of PCA model created with hyperspectral data, showing the first three PCs. **(b)** Loading plot of the PCA model created with hyperspectral data, showing the first three PCs.

8.4. Automatic detection of fish eye

The results presented in the previous section shows clearly that the eye is an ideal region for the monitoring of the fish freshness and decaying process, due to the differences obtained between acquisition in different periods and the fact that the sample does not require a previous treatment for a correct acquisition of the HSI. However, a new problem arose: for the study presented above the eye region was manually selected from every one of the acquired HSI. For a practical application of monitoring sensors based on this approach, it is critical to develop an automatic way to divide the eye region from the rest of the body, with a fast and reliable method. This was obtained successfully thanks to a deep learning technique already cited in the Introduction, the Artificial Neural Network (ANN). In particular, I used a python package called Detecto, which allows solving object recognition problems in a fast and easy way.

ANN theory

ANN are complex computational algorithms, that try to mimic the processing power and speed of the human brain [144], thanks to the use of a very large number of processing units connected between them, divided into a series of levels, called "layers". These layers can work in parallel to solve complex problems, like classification by pattern recognition and optimizations. In fact, as with the other ML and DL techniques, ANN algorithms can model nonlinear relationships directly from the input data, and then apply the model to new data, solving both classification and prediction problems [145].

To understand the basics of ANN functioning, it is helpful to analyze the most basic family of these algorithms, called MultiLayer Perceptrons (MLP), whose scheme is shown in Fig. 40 (top), as well as the single constitutive elements of these networks, called "neuron" (bottom). These neurons accept a series of inputs X_m and process them to produce a single out, that will be sent to other ones [146]. To calculate this value, the inputs are initially multiplied by weights, and then summed between them, often with a bias[147], obtaining a numeric value that is often called the activation. Finally, the sum is used as input for a function (often nonlinear, like hyperbolic tangent or rectified linear unit) that produces the output of the neuron. Neurons are typically organized in multiple layers, that could be divided into three different types. Two of them are single: the input layer, which receives the external data used for training and prediction, and the output layer, which produce the value of the VI. Between them, there are a variable number of hidden layers, that process the input values to obtain the output one. Every layer usually accepts inputs from the previous one and gives output to the following one,

creating what is called a *feedforward network*[148]. If connections between neurons of the same layer, or from a layer to the previous one, are present, the ANN is a *recurrent network*[149].

Independently from the type of network, the output is calculated by the ANN thanks to a process called learning, which is an iterative adaptation of the weight values (and the thresholds of the nonlinear functions) to improve the accuracy and better handle the task. In our case, using a supervised approach, a cost function is calculated by comparing the predicted values to the real ones given as input. The weights are adjusted for a series of iterations (called "epochs") until the cost function is not declining anymore and reaches a plateau with a small error[150]. The usual method used to perform the adjustment of the weights is called "backpropagation", which, by calculating the gradient of the cost function to the weights, effectively divides the error between all the connections[151].

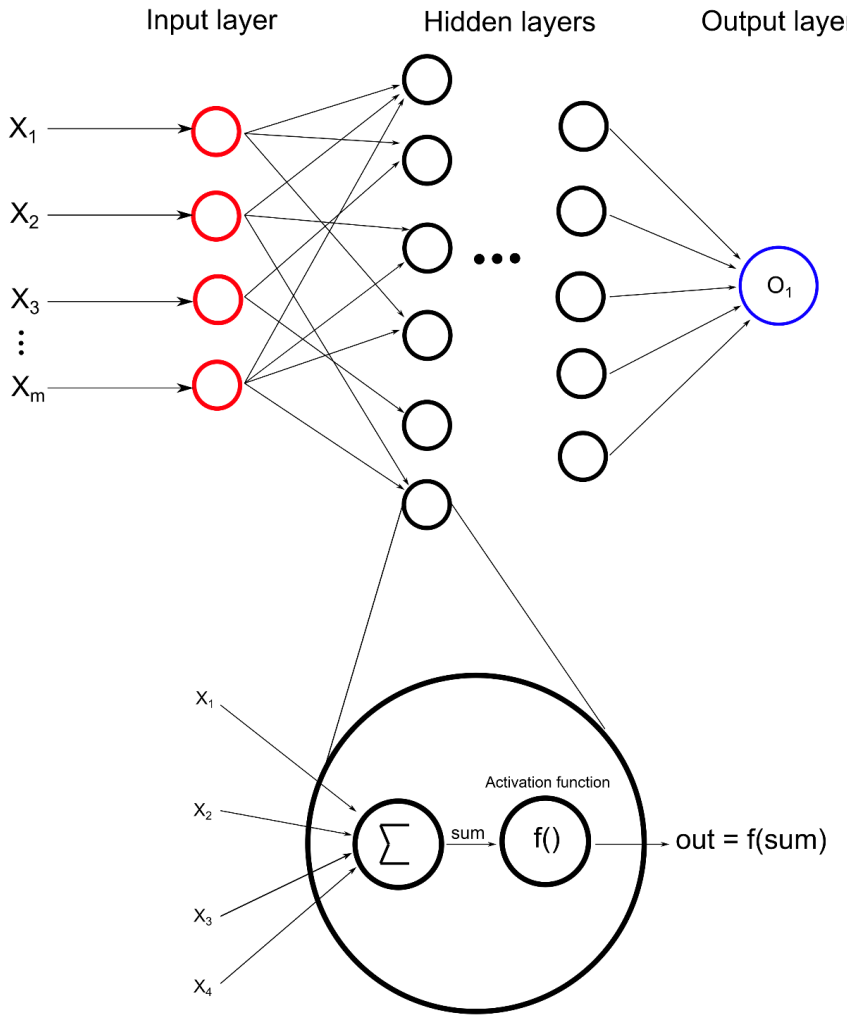

Fig. 40 **(top)** Scheme of a simple feedforward network, where the different types of layers are highlighted with different colors: red for the input layer, black for the hidden layers, and blue for the output layer. **(bottom)** Scheme of a single neuron

Database creation and data augmentation

As stated in the previous section, the ResNet50 used by Detecto is a supervised network, so a labeled dataset is needed as input. I created it with an open-source online tool called MakeSense, which allows one to load a series of images and label them manually, drawing a box on each image to select the Region of Interest and linking it to a label (in this case “eye”). Fig 41 shows an example of the labeling process.

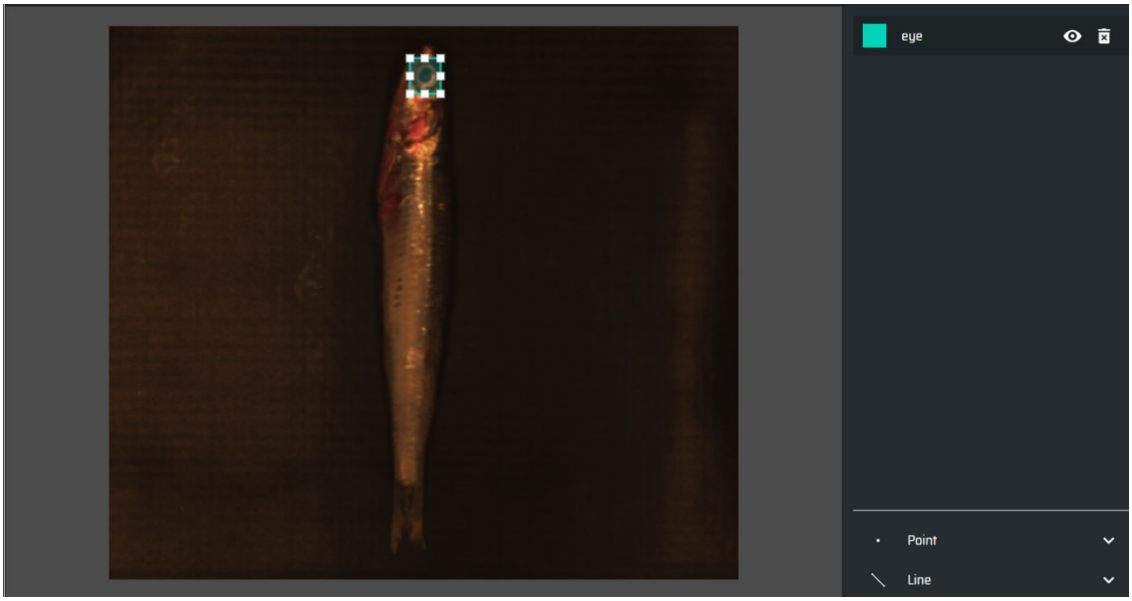

Fig. 41 Example of a fish image labeled with MakeSense. The box was drawn manually, and the label “eye” was linked to it.

Thanks to this process, I obtained a total of 191 labeled images, that were subsequently and randomly divided into the training set (154 images) and the test set (37 images). However, these data are too few to represent a good dataset for the training of a complex ANN like ResNet50, which usually requires input datasets containing 5000-10000 images. To overcome this, a data augmentation process was performed during the training of the model: for every iteration of the algorithm over the input dataset, a series of transformations (including horizontal flip, the addition of gaussian blur, changes in the contrast or saturation of the image..) was randomly added to the original image, each of them with a probability of 50%. In this way, a different combination of transformations was performed every time, making the algorithm iterate over mostly different images. An example of new images created by the application of these transformations is depicted in Fig. 42. The dimension of this augmented dataset is estimable from the multiplication of the number of images contained in the training dataset (154) and the number of epochs (usually 20), for a maximum of 3080 images. This

was still a fairly low number, but I did not augment it further because the ANN trained with it already gave very good results regarding the detection of the eye region.

Fig. 42 Example of a new image obtained with the data augmentation process

Training and testing of the ANN

The training process of an ANN, especially with 50 layers like the one I used, is knowingly very computationally heavy, and in normal computers requires hours, or even days, to complete a single training attempt. To overcome this problem, I performed the creation and training of the model on an online and free platform called Google Colab, developed by Google Research. Thanks to this site, it is possible to use a GPU for the calculation process, reducing greatly the training time, as far as 20-30 minutes.

Fig. 43 (a) shows the plot of the loss (the cost function) to the number of epochs. It is easy to see that the function presents a steep descent for the first 3-4 iterations and then reaches a plateau, with a very low loss value, with a minimum of 0.085 for epoch 5. The prediction ability of the model was then evaluated using the test set as input, and comparing the differences between the ROI selected by the ANN and the one selected manually. The obtained results were excellent, with a mean accuracy between the 37 images of 0.985 (with a maximum of 1) and a very fast prediction time for a single image, in the range of 1-2 seconds. An example of the comparison between an ROI selected manually (in blue) and by the ANN (in red) is shown in Fig. 43 (b).

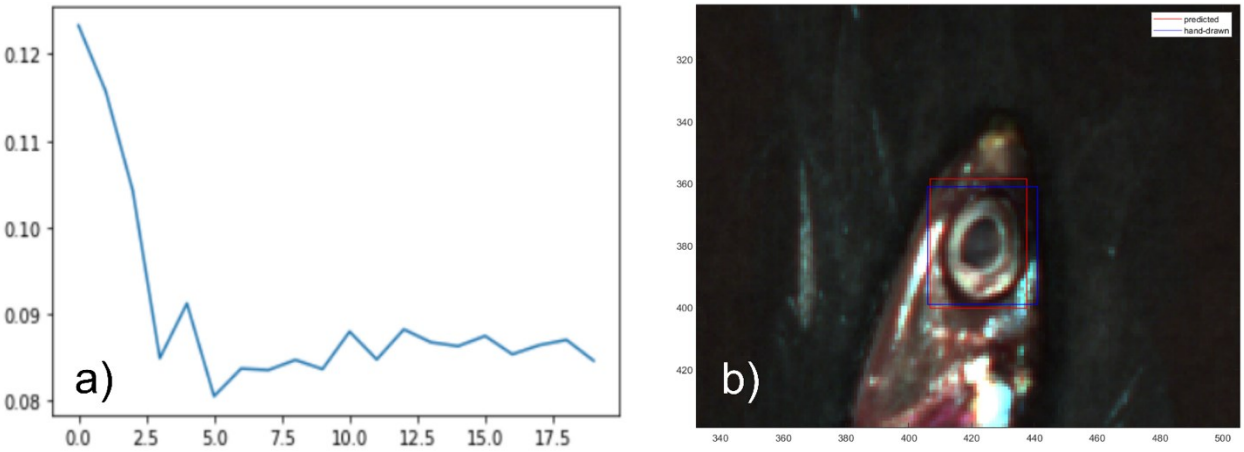

Fig. 43 **(a)** Trend of the cost function during the learning process of the ResNet50 **(b)** Comparison between the ROI selected manually (in blue) and the one selected by the ANN (in red)

8.5. Conclusions on fish freshness detection

The use of hyperspectral images to monitor the decaying process and the consequent freshness of anchovies and sardines was investigated. Images of the whole body of fishes were acquired on two different acquisition campaigns, then a mean spectrum was calculated from them and used as input for PCA analyses. For the first study the whole fish body was considered, whereas for the second one I focused on the eye region. The obtained results showed that hyperspectral images contain useful information regarding the freshness of fishes, and could be used to monitor the decaying process of them. In particular, the region of the eye is particularly informative in this sense. For the automatic selection of that region from the whole images, I focused on the use of a Python package, called Detecto. The obtained results showed that it allows to easily create and train an ANN for object recognition, with few lines of Python code, and using a free online application for the calculations. The resulting model was able to correctly predict the eye ROI in every one of the test images, obtaining a high value of the overall accuracy and a fast prediction time. The downside of the use of ANN is the requirement for a lot of convolutional calculations, both for the learning and the prediction phase, with consequent high power consumption and a slower prediction. This makes it impossible to implement the selection of the ROI directly in the embedded systems we considered for the other applications already considered because it will require too much computational power, as well as too much memory to save all the weights used by the model (that could be over hundreds of Mbytes). This problem can be solved using embedded hardware that can be connected to the Internet through wi-fi (like a simple Raspberry) and do the prediction directly on platforms like Google Colab.

Also, recent innovations in the field of electronics, in particular in the design of dedicated hardware for Deep Learning, called Convolutional Neural Networks (CNN) accelerators, could be helpful to solve this problem. These devices combine both MCU and FPGA technology [152], allowing them to both save the weights and perform the convolutional calculation inside the embedded system. Another limitation is the speed of the hyperspectral camera acquisition: the one we used for the creation of the database scans each line of the image separately (a technique called “pushbroom”), thus requiring some seconds for each fish sample. This slow acquisition time is of course incompatible with and in-line industrial applications: to overcome this, the use of a “snapshot” camera, able to acquire a true hyperspectral video, can be investigated.

9. Electronic implementations of models

As reported several times in the previous chapters, PCA-based MSAs give as output an array **B** of calibration coefficients, that allows the prediction of the value of the VIs directly from a new spectral acquisition, following Eq. (1) (i.e **Y** = **XB**). In practical applications, there is often also an offset value that is summed to the multiplication (as reported in Eq. (27)). In every case, the computational power required for this calculation is very low, needing only a series of multiplication and sums (one for every acquired frequency/wavelength of the spectrum, usually from 1000 to 10000). So, these models are an excellent choice for the implementation of the VIs prediction in embedded devices, where the requirements for power consumption and memory availability can be very strict. However, for the vast majority of spectral sensors commercially available, is not easy (or impossible), to make changes within the hardware or firmware, to implement the prediction directly on the sensing device. However, it is often very easy to connect these sensors to Single Board Computers (SBC), like Arduino or Raspberry, with different types of interfaces protocol: USB, SPI, I2C, etc. By doing so, the SBC can easily receive the raw spectrum from the sensor, apply the required preprocessing and multiply it by the array B, obtaining as output the value of the researched variables. This result could then be saved in the memory of the device or on Cloud services, if a wi-fi connection is available. Moreover, multiple sensors can be interfaced with a single SBC, allowing the simultaneous acquisition of parameters like temperature and humidity, useful for monitoring the correct functioning of the device or eventually adjusting the prediction.

This chapter will present the schematic of a PCB that allows the interface between a sensor and an SBC, as well as power up both of them with the same battery and automatically acquire spectra. In particular, I worked with a Raspberry Pi Zero W as the SBC and a NanoVNA, an RF sensor that, when in contact with the material, emits electromagnetic waves in the range of 50 KHz to 4.4 GHz, acquiring both the consequents transmitted and reflected waves. Two modes of acquisition will be implemented by this PCB: “manual”, where a single acquisition is performed by pushing a button, and “automatic”, where the systems perform a timed series of acquisitions, with the period selectable by the user. After a brief presentation of the two devices, the schematic will be presented in detail, as well as the two modes of functioning. The two devices are depicted in Fig. 44.

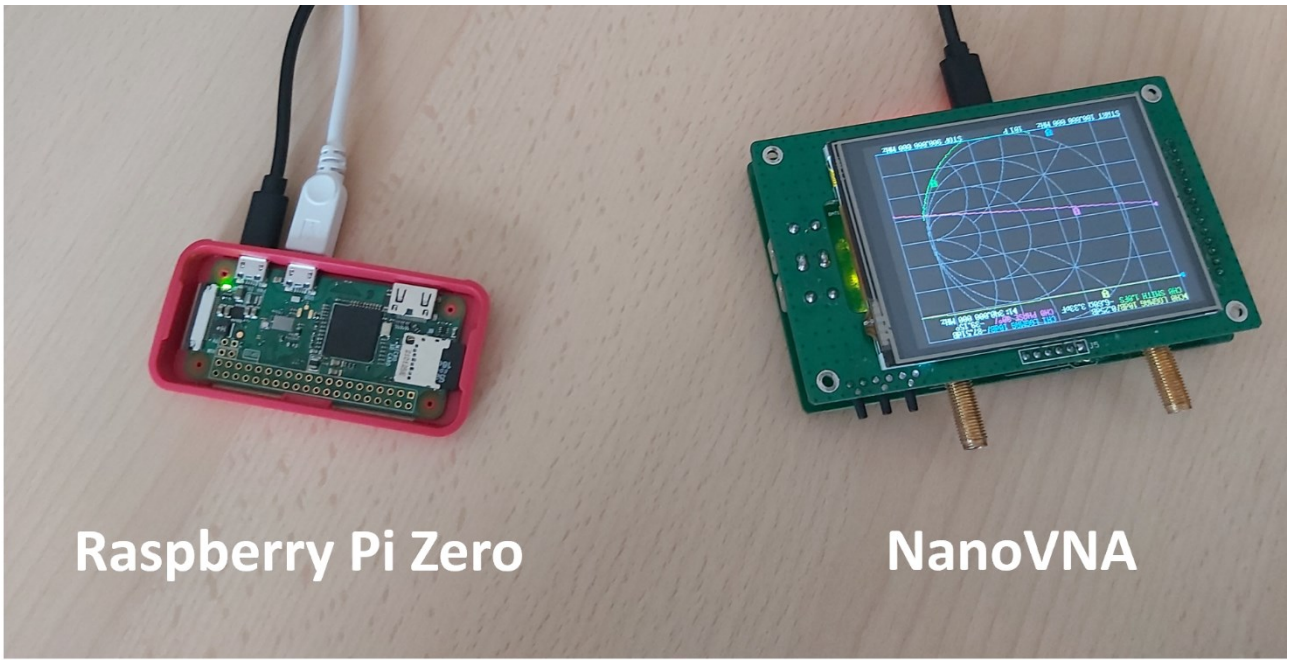

Fig. 44 Rasberry Pi Zero W (**left**) and NanoVNA (**right**), connected through a micro USB cable.

9.1. Devices

Raspberry Pi Zero W

Raspberry Pi is a series of small SBCs developed in the United Kingdom by the Raspberry Pi Foundation in association with Broadcom [153]. They are widely used in many areas, because of their low cost, modularity, and open design. In particular, I worked with a model called Raspberry Pi Zero W. It is an evolution of the Raspberry Pi Zero, a very cheap model (only 5$) released in 2015, to which a wi-fi and a Bluetooth connection were added. The main features of this model are the following ones: a 1 GHz single-core GPU, 512 MB of RAM, a Mini HDMI port and micro USB On-The-Go (OTG) port, and a HAT-compatible 40-pin header, that can be used to power up the Raspberry, or to connect it to other sensors and devices. In our application, the Raspberry performs the majority of the tasks: it runs a Python code that controls the acquisition of data through the NanoVNA, then calculates the value of the VI thanks to the calibration coefficients saved in its memory, and saved the results as txt files. Moreover, two temperature-humidity sensors are connected to it through I2C, allowing the monitoring of both the external and internal environment conditions. Despite the single-core CPU, the prediction takes little time: experimental tests show that, once a spectrum is acquired, the whole calculation requires about 3 seconds. An image of a Raspberry Pi Zero W with its principal components highlighted is depicted in Fig. 45.

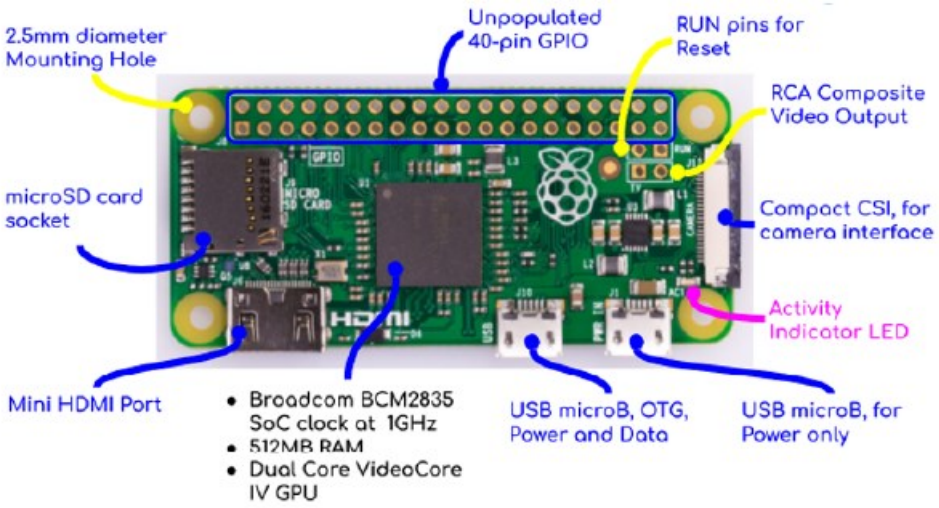

Fig. 45 Rasberry Pi Zero W schematic [154].

NanoVNA

NanoVNA V2 (S-A-A-2) is a 4GHz vector network analyzer (VNA). As it is possible to see from Fig. 46, it presents two antennae (in the upper part). From the right one, it can emit a series of electromagnetic waves in the range of 500 KHz-4.4 GHz, where the initial/ending frequency and the spacing between them can be selected by the user through the LCD screen. Then, the right antenna will acquire the component of these waves reflected by the sample, whereas the left one will acquire the components transmitted through the sample. With these two antennae, it is possible to calculate all 4 scattering parameters (S-parameters), a series of values that describe the electrical behavior of linear electrical networks when undergoing various steady-state stimuli by electrical signals. As already reported in section 5.1, the water content highly influences the electrical characteristics of the materials, making the NanoVNA highly sensitive to changes in the water content of materials. Acquisitions can be performed by connecting the NanoVNA to the computer and using the dedicated software; or, in our case, connecting it to the Raspberry with a USB cable and using a Python function to set the parameters, perform a calibration (or load an already existing one) and acquire a new spectrum. When connected to other devices, it is powered by them through the USB cable; if used alone, it can be powered by a 3.7 V LiPo battery.

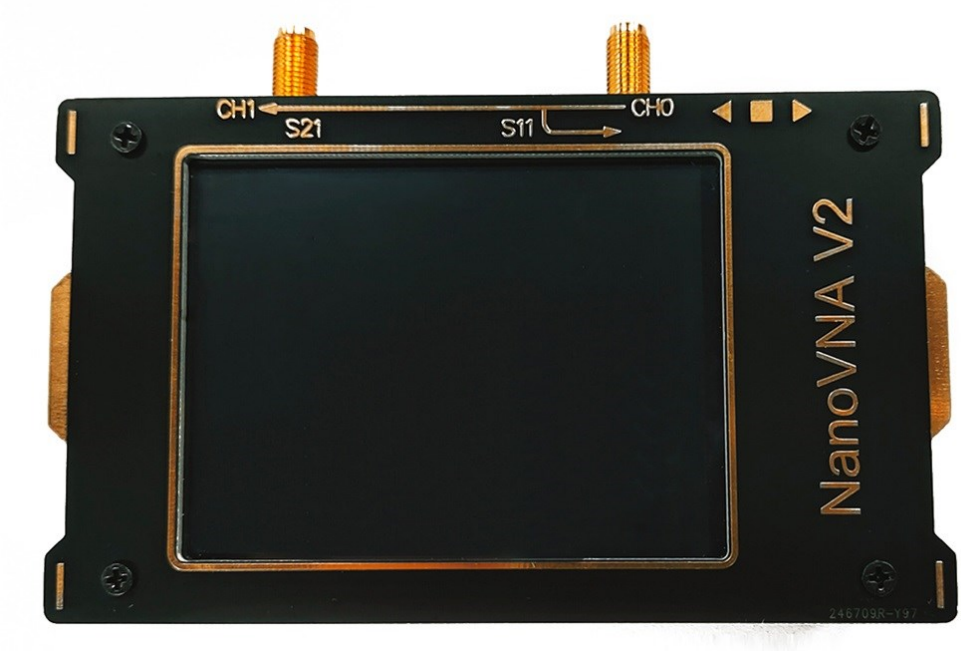

Fig. 46 NanoVNA depiction.

Sensing system

As previously stated, it is possible to connect the NanoVNA to the Raspberry thanks to a USB cable, and use a simple Python function to acquire new data. However, some enhancements were needed to obtain an autonomous monitoring sensor, that can be useful for in-line applications. In particular, the Raspberry needs to be powered with a battery, and not with the standard cable current; moreover, a way to automatically perform an acquisition every period (selected by the user) is needed. To solve these problems, I started to design a PCB that can give power to the two devices from the same battery and allow the Raspberry to automatically acquire data from the NanoVNA. Moreover, it will mechanically connect the two devices: it is designed to be mountable on top of the NanoVNA, and the Raspberry can be soldered to it using a 40 Pin header, allowing us to obtain a truly compact and self-sufficient device, which can also be connected to the Internet through the Raspberry wi-fi. Fig. 47 shows a functional scheme of the resulting sensing system: the NanoVNA is used to acquire electromagnetic spectra, then these data are sent to the Raspberry, where the calibration coefficients of an already created predictive model are saved. The Pi Zero performs the calculation of the VIs, and then the output can be either saved in its memory or on a Cloud service. To date, I completed the schematic of the PCB, using the software Kicad. In the near future, I will work on the layout and the practical realization of the board.

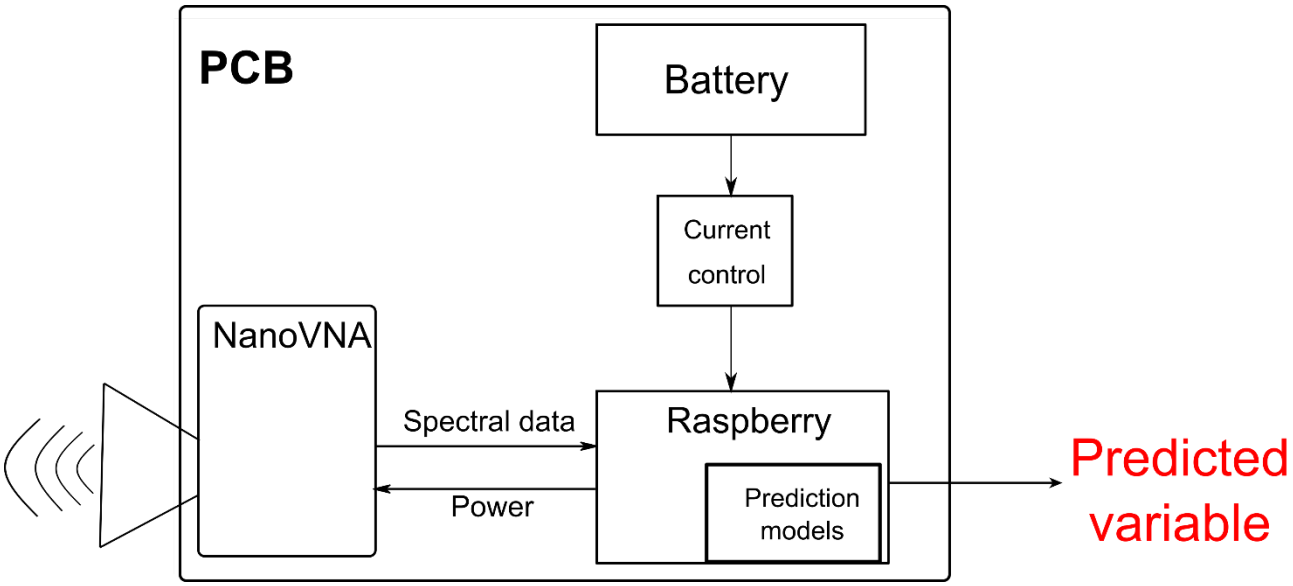

Fig. 47 Sensing system functional scheme.

9.2. PCB design

The PCB components can be divided into 3 different groups, based on their function:

1) Battery management
2) Temperature and humidity monitoring
3) Current control

These three groups, all connected and controlled by the Raspberry Pi, will be described in detail in this section.

Battery management

Fig. 48 shows the electronic component used to power up all the components, thanks to the 3.7 V LiPo battery already present in the NanoVNA. This task is mainly performed by the System On Chip IP5305, by Injoinic Technology [155]: it takes as input the 3.7 battery (pins 6-7) and gives as output 5 V (pin 8), necessary to power up the Raspberry and all the others components. Moreover, it allows the charge of the battery through a micro USB port (pin 1). Finally, 4 LEDs (pins 2-3-4) give a visual indication about the battery charge (from 4 LEDs on when higher than 75%, up to only one when lower than 25%). The whole system is powered on/off thanks to a manual switch, positioned between the battery connector and the IP5305.

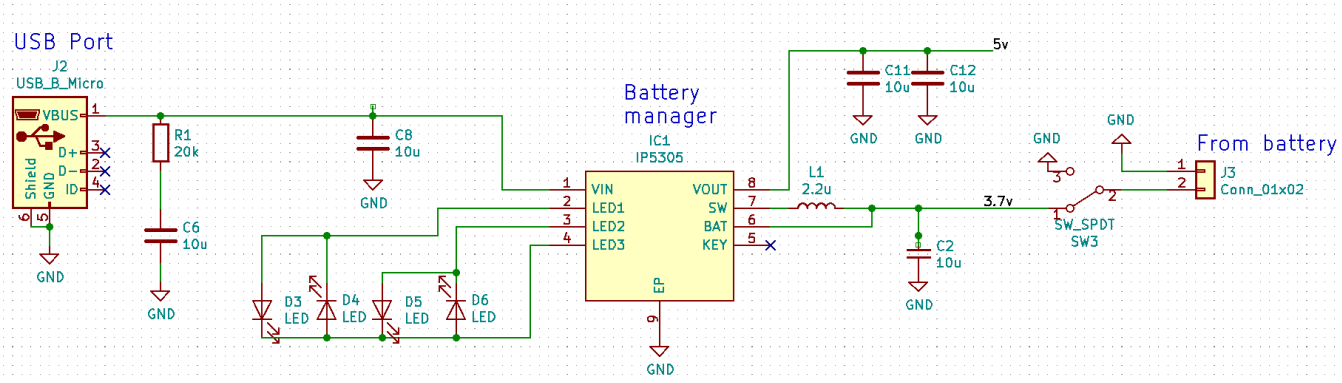

Fig. 48 Battery management.

Temperature and humidity monitoring

Fig. 49 shows the components used for monitoring both the internal and external temperature and humidity. This is implemented thanks to the use of two SHT40 by Sensirion [156]. These sensors have a humidity accuracy of ±1.8 %RH and a temperature accuracy of ±0.2 °C, as well as low current consumption: 350 µA during the measurement and 0.08 µA in an idle state. A single acquisition takes 4 to 7 ms. They are connected directly to the Raspberry Pi with the I2C protocol, using different addresses to allow the contemporary use of two of them. Moreover, the Raspberry provides them with the required power supply at 3.3V.

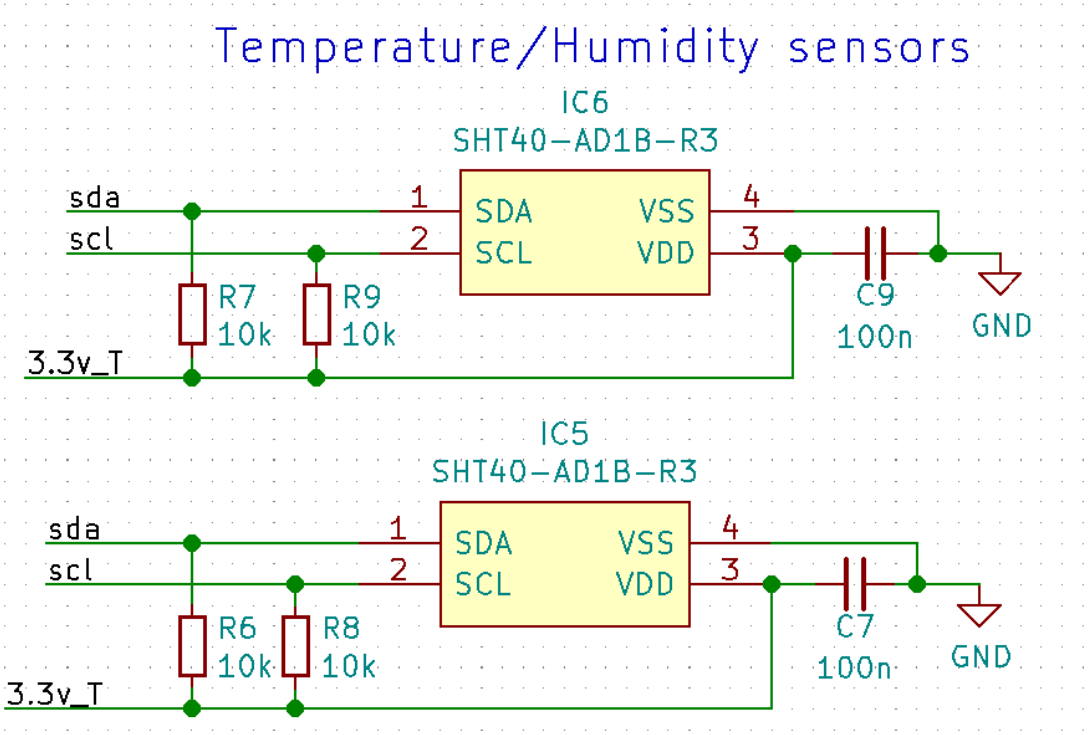

Fig. 49 Temperature and humidity monitoring

Current control system

Fig. 50 shows the components used for the implementation of the current control system, as well as the Raspberry Pi (on the left). This system is used to implement two modes of functioning (that will be presented in detail in the next section): in the "manual" one, the Raspberry is kept on until a manual shutdown, and it is possible to acquire single spectra pushing a button; whereas in the "automatic" mode, a Real Time Clock (RTC), the TPL5110 by Texas Instruments [157], drive a MOSFET connected to the power via of the Raspberry. With this, it is possible to turn on the Raspberry only for the time necessary to perform and save an automatic acquisition, enhancing greatly the duration of a single battery charge. A second manual switch allows to turn on the RTC only when the "automatic" acquisition mode is required. The period between two acquisitions is defined by the value of the resistor connected to the RTC (pin 3). I used a digital potentiometer, the MAX5161 by Maxim Integrated [158], which can take on 32 different resistance values, in a range of 0 to 200 kΩ. In this way, the period between two consecutive acquisitions can be selected by the user through the Raspberry Pi, before starting the acquisition phase. Two level shifters are used for the communications between the Raspberry, that have 3.6 V output pins, and the RTC and the digital potentiometer, that have 5V pins. 3 LEDs are driven by the Raspberry, to show clearly when the Raspberry is powered on (LED 1 on), when an acquisition is performed (LED 2 blinking), and when

it is safe to power off manually the Raspberry (LED 3 on). Finally, 2 buttons connected to the Raspberry allow to acquire a single spectrum and shutdown the Pi Zero, respectively.

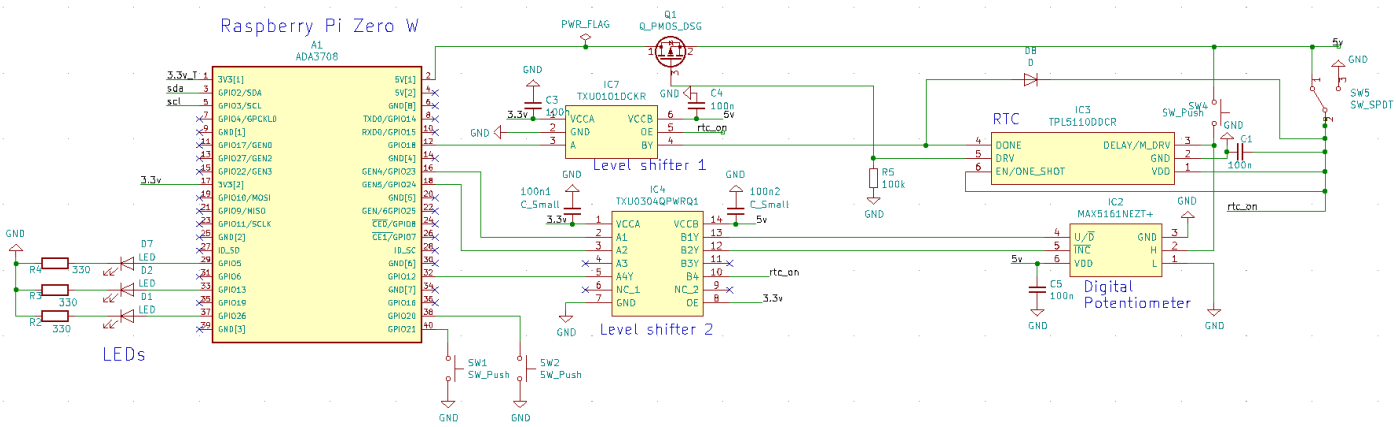

Fig. 50 Current control system (right) and Raspberry Pi (left)

9.3. Modes of operation

As already stated, the current control system included in the PCB allows to switch between two methods of acquisition, whose flowcharts are depicted in Fig. 50. The main one is the “manual”, whereas the “automatic” one can be selected through a switch.

In “manual” mode, the system is turned on with the power manual switch, and every component is kept on until the power switch is turned off. When LED 1 is on, showing that the Raspberry is ready, it is possible to push Button 1 to acquire a single spectrum and predict the VIs, saving the results as a txt file in the Pi memory. LED 2 will blink during the acquisition and save time (usually 1-2 seconds). After that brief waiting time, it is possible to acquire a new spectrum, always by pressing Button 1.

The “automatic” mode allows the automatic acquisition of spectra every period, selectable by the user, and it is necessary for the use of the device for monitoring purposes. When powered with the secondary power switch, the RTC will turn on the Raspberry, which in turn will perform an acquisition immediately after the setup time, save the data, and finally shutdown itself. The RTC will recognize the shutdown of the PI (through pin 4 of the RTC) and it will drive the MOSFET to disconnect the Raspberry to the 5V, shutting it off completely. This is necessary because the Raspberry Pi Zero W consumes a good amount of power (around 200 mW) also when is in shutdown mode, and maintaining it in this condition the whole time will greatly reduce the battery duration. After powering down the Raspberry, the RTC will wait for the time selected by the user (thanks to the digital potentiometer connected to the RTC), and then it will provide again the 5V to the

Raspberry, which will perform a second acquisition. To sum it up, spectra will be continuously acquired and saved in the Raspberry memory until the "automatic" mode is stopped by the secondary power switch.

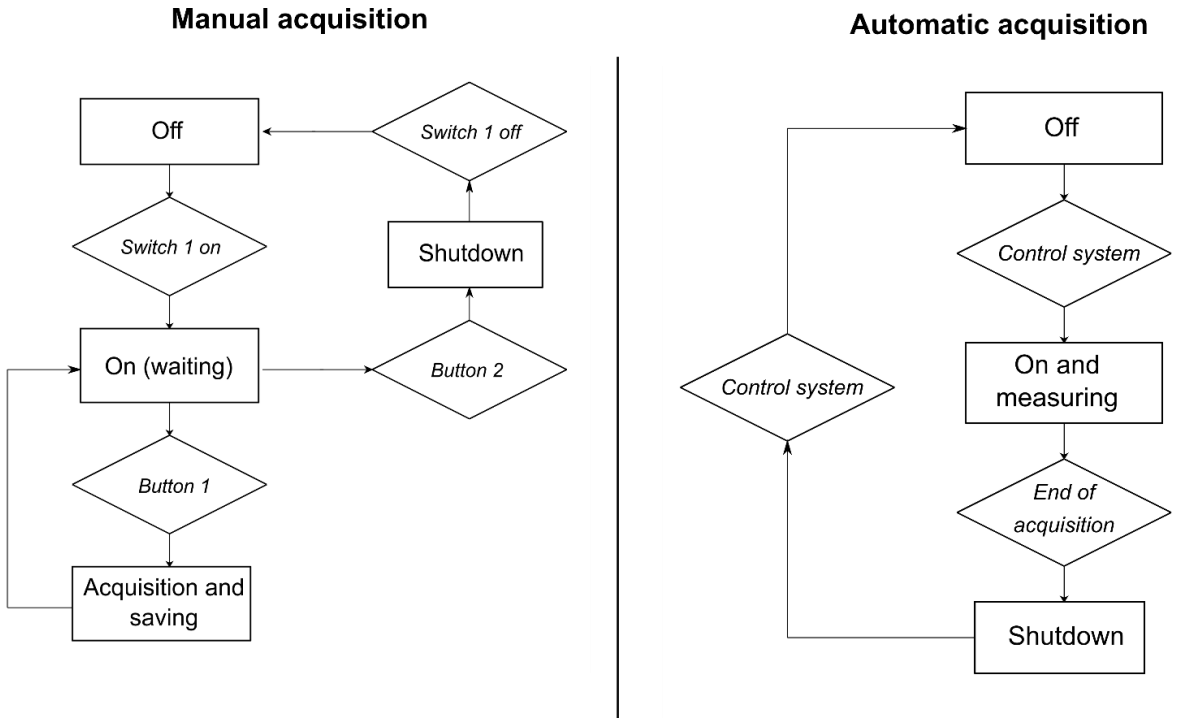

Fig. 51 Flowchart of the two modes of acquisition: manual (left) and automatic (right)

9.4. Conclusion on electronic implementation

The design of a PCB schematic for the connection between a Raspberry Pi Zero W and a NanoVNA was presented. The sensing device that can be obtained with this board will be able to acquire spectra, predict the value of one or more VIs, and save the results, offering the possibility to perform both manual and automatic acquisition. In particular, the PCB allows to power up both systems with the same battery, recharge the battery through a micro USB port, and turn off the current from the Raspberry when not used during the automatic acquisition mode. Moreover, 2 temperature and humidity sensors are connected to the Raspberry to monitor environmental conditions; as well as a series of LEDs and buttons to ensure the correct functioning of the system. In the future, the layout will be designed, focusing on a rectangular board that can be mounted on top of the NanoVNA, and the system will be tested in practical applications. If the results will be acceptable, other PCBs could be designed in the future, to allow the connection between Raspberry SBCs and other sensing devices.

10. Conclusions and future perspectives

My work was mainly focused on the design of predictive models for one or more variables of interest, starting from spectral data. The same workflow was used to create and test these models, even if the field of application will greatly change between them: spectral data and values of the VI were acquired on samples, using gold standard techniques; these data are then used as input for PCA-based analyses, that I used to create and refine the predictive models. The application of these algorithms, especially evolutions of the PCA like the PLSR and SIMCA, is quite innovative in the field of environmental monitoring. In fact, the literature contains a lot of examples where only PCA is used to analyze the acquired data, but the prediction models are not developed, or more complex Deep Learning techniques are preferred. Despite their lesser prediction power, PCA-based techniques may often be more useful, given the fact that they can offer similar predictive results, but require smaller input datasets, as well as less computational power both for the creation of the model and the prediction of the variables, making them perfect for the design of embedded sensing devices.

10.1. Lessons learned and future developments

We tested the workflow on 4 very different applications: prediction of soil and concrete moisture (spectra acquired with an RF sensor), prediction of concentrations of gasses in a mix (spectra acquired with a gas chromatograph), and prediction of the freshness of anchovies and sardines (spectra acquired with a hyperspectral camera). The predictive results were excellent for every one of the applications, as demonstrated by the low values of the RMSE and the high values of the R^2 parameters between all the cases, and demonstrated that these techniques can be used effectively for the design of monitoring systems in several different applicative fields. The main goal of my PhD period, to use the same workflow for the creation of predictive model for different application (as stated in chapter 3), was tested successfully.

The various applications showed us that, to create successful models, it is not necessary to acquire thousands of input data, like in Deep Learning, but it is possible to obtain accurate predictions with as few as 33 input spectra (the PLSR model for concrete resistance), whereas the ideal number is around 100-200. Another lesson we learned is the fact that the best results are obtained when the values of the VI used as input (the Y calibration dataset) cover its range equally, respect to have lots of acquisition linked to few different values. In the first case, we were able to obtaine models that are in general more precise and reliable in the prediction of new values of the VI. Regarding instead the

length of the spectra and the pace between the acquired frequencies, we worked with very different ranges and numbers of spectral point, with no combination which proved itself better than the others. In general, the better approach is to cover the greatest range possible during the first (and offline) creation and calibration of the model, to make sure to acquire all the useful information. Once created the first model, it is easy to visualize the most informative frequency ranges and calibrate the model only on them, reducing greatly the acquisition time of the sensor and the computational power required for the prediction of the VIs. It is worth noting that, from an implementation point of view, if the useful frequencies are few and scattered between them, it can be easier and faster to still acquire all the spectrum, and consequently multiply per 0 all the values corresponding to the non informative frequencies.Future studies can further enhance this approach, focusing on testing the created model with new data, acquired in different environmental conditions and for different ranges of the VIs, to better assess the prediction ability of the systems. Another focus will be the design of other integrated devices, to obtain embedded sensing systems that will allow non-invasive and precise environmental monitoring.

In conclusion, I think that, in the next years, the request for non-invasive and autonomous spectra sensors with embedded prediction models will increase, especially in fields related to environmental monitoring. For these purposes, spectral systems in conjunction with PCA-based predictive models could represent the ideal middle ground between the need for accurate and fast predictions and the increasingly stringent demands in power consumption, memory availability, and low cost.

11. References

[1] Eigenvector Research, Inc., “PLS_Toolbox 8.9.2 (2021).” Manson, WA USA 98831.

[2] R. Priemer, *Introductory signal processing*, vol. 6. World Scientific, 1991.

[3] B. Jia *et al.*, “Essential processing methods of hyperspectral images of agricultural and food products,” *Chemom. Intell. Lab. Syst.*, vol. 198, no. 17, p. 103936, 2020, doi: 10.1016/j.chemolab.2020.103936.

[4] W. Saleh, N. Qaddoumi, and M. Abu-Khousa, “Preliminary investigation of near-field nondestructive testing of carbon-loaded composites using loaded open-ended waveguides,” *Compos. Struct.*, vol. 62, no. 3–4, pp. 403–407, 2003, doi: 10.1016/j.compstruct.2003.09.012.

[5] A. S. Bin Sediq and N. Qaddoumi, “Near-field microwave image formation of defective composites utilizing open-ended waveguides with arbitrary cross sections,” *Compos. Struct.*, vol. 71, no. 3–4, pp. 343–348, 2005, doi: 10.1016/j.compstruct.2005.09.031.

[6] A. McClanahan, S. Kharkovsky, A. R. Maxon, R. Zoughi, and D. D. Palmer, “Depth evaluation of shallow surface cracks in metals using rectangular waveguides at millimeter-wave frequencies,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 6, pp. 1693–1704, 2010, doi: 10.1109/TIM.2009.2027780.

[7] P. M. Meaney, S. D. Geimer, R. Augustine, and K. D. Paulsen, “Quasi- Open-Ended Coaxial Dielectric Probe Array for Skin Burn Characterization,” *13th Eur. Conf. Antennas Propagation, EuCAP 2019*, no. EuCAP, pp. 13–15, 2019.

[8] M. Obol, N. Al-Moayed, S. P. Naber, and M. N. Afsar, “Using coaxial probe for broadband microwave characterization of biological tissues,” *Proc. 38th Eur. Microw. Conf. EuMC 2008*, no. October, pp. 416–419, 2008, doi: 10.1109/EUMC.2008.4751477.

[9] R. M. Healy *et al.*, “Assessment of a passive sampling method and two on-line gas chromatographs for the measurement of benzene, toluene, ethylbenzene and xylenes in ambient air at a highway site,” *Atmos. Pollut. Res.*, vol. 10, no. 4, pp. 1123–1127, 2019, doi: 10.1016/j.apr.2019.01.017.

[10] M. J. Yoo, S. H. Jo, and K. H. Kim, “An advanced technique for rapid and accurate monitoring of gaseous formaldehyde using large-volume injection interfaced with gas chromatograph/barrier discharge ionization detector (LVI/GC/BID),” *Microchem. J.*, vol. 147, no. February, pp. 806–812, 2019, doi: 10.1016/j.microc.2019.03.096.

[11] W. J. Krzanowski and P. Kline, “Cross-Validation for Choosing the Number of Important Components in Principal Component Analysis,” *Multivariate Behav. Res.*, vol. 30, no. 2, pp. 149–165, Apr. 1995, doi: 10.1207/s15327906mbr3002_2.

[12] S. T. He, Y. B. Gao, J. Y. Shao, and Y. Y. Lu, “Application of SAW gas chromatography in the early screening of lung cancer,” *Proc. 2015 Symp. Piezoelectricity, Acoust. Waves Device Appl. SPAWDA 2015*, pp. 22–25, 2015, doi: 10.1109/SPAWDA.2015.7364432.

[13] A. M. Casas-Ferreira, M. del Nogal-Sánchez, J. L. Pérez-Pavón, and B. Moreno-Cordero, “Non-separative mass spectrometry methods for non-invasive medical diagnostics based on volatile organic compounds: A review,” *Anal. Chim. Acta*, vol. 1045, pp. 10–22, 2019, doi: 10.1016/j.aca.2018.07.005.

[14] K. Yao *et al.*, “Visualization research of egg freshness based on hyperspectral imaging and binary competitive adaptive reweighted sampling,” *Infrared Phys. Technol.*, vol. 127, no. November 2021, p. 104414, 2022, doi: 10.1016/j.infrared.2022.104414.

[15] Y. Hu, P. Huang, Y. Wang, J. Sun, Y. Wu, and Z. Kang, “Determination of Tibetan tea quality by

hyperspectral imaging technology and multivariate analysis," *J. Food Compos. Anal.*, vol. 117, no. December 2022, p. 105136, 2023, doi: 10.1016/j.jfca.2023.105136.

[16] J. Wieme *et al.*, "Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review," *Biosyst. Eng.*, vol. 222, pp. 156–176, 2022, doi: 10.1016/j.biosystemseng.2022.07.013.

[17] M. A. Gagnon *et al.*, "Standoff midwave infrared hyperspectral imaging of ship plumes," *Work. Hyperspectral Image Signal Process. Evol. Remote Sens.*, vol. 2015-June, pp. 2–5, 2015, doi: 10.1109/WHISPERS.2015.8075384.

[18] A. C. Karaca, A. Ertürk, M. K. Güllü, M. Elmas, and S. Ertürk, "AUTOMATIC WASTE SORTING USING SHORTWAVE INFRARED HYPERSPECTRAL IMAGING SYSTEM Kocaeli University Laboratory of Image and Signal Processing (KULIS), MS MacroSystem Nederland ," *2013 5th Work. Hyperspectral Image Signal Process. Evol. Remote Sens.*, pp. 2–5, 2013.

[19] M. Tartagni, *Electronic Sensor Design Principles*. Cambridge (UK): Cambridge University Press, 2022.

[20] S. Wold and M. Sjostrom, "PLS-regression : a basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, pp. 109–130, 2001.

[21] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[22] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, p. 321, Dec. 1936, doi: 10.2307/2333955.

[23] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.

[24] K. Dunn, *Process Improvement Using Data*. 2021.

[25] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001, doi: 10.1016/S0169-7439(01)00155-1.

[26] S. WOLD and M. SJÖSTRÖM, "SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy," 1977, pp. 243–282.

[27] I. Eigenvector Research, "T-Squared Q residuals and Contributions," 2012. https://www.wiki.eigenvector.com/index.php?title=T-Squared_Q_residuals_and_Contributions (accessed Sep. 08, 2021).

[28] T. Kourti, "Application of latent variable methods to process control and multivariate statistical process control in industry," *Int. J. Adapt. Control Signal Process.*, vol. 19, no. 4, pp. 213–246, 2005, doi: 10.1002/acs.859.

[29] H. Oddan, "Multivariate Statistical Condition Monitoring," Norwegian University of Science and Technology, 2017.

[30] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7, pp. 1–15, 2006, doi: 10.1186/1471-2164-7-142.

[31] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, doi: 10.1021/ac60214a047.

[32] D. MacDougall, H. Martens, and P. Geladi, "Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat," *Appl. Spectrosc.*, vol. 39, no. 3, pp. 491–500, 1985.

[33] R. Bro, "Multivariate calibration," *Anal. Chim. Acta*, vol. 500, no. 1–2, pp. 185–194, Dec. 2003, doi:

10.1016/S0003-2670(03)00681-0.

[34] R. Bro and A. K. Smilde, “Principal component analysis,” *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, 2014, doi: 10.1039/c3ay41907j.

[35] S. Wold, “Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models,” *Technometrics*, vol. 20, no. 4, p. 397, Nov. 1978, doi: 10.2307/1267639.

[36] H. WOLD, “Path Models with Latent Variables: The NIPALS Approach,” in *Quantitative Sociology*, Elsevier, 1975, pp. 307–357.

[37] A. S. Goldberger, *Econometric Theory*. New York: John Wiley & Sons, 1964.

[38] A. Peleg and U. Weiser, “MMX technology extension to the Intel architecture,” *IEEE Micro*, vol. 16, no. 4, pp. 42–50, 1996, doi: 10.1109/40.526924.

[39] G. Luciani, A. Berardinelli, M. Crescentini, A. Romani, M. Tartagni, and L. Ragni, “Non-invasive soil moisture sensing based on open-ended waveguide and multivariate analysis,” *Sensors Actuators A Phys.*, vol. 265, pp. 236–245, Oct. 2017, doi: 10.1016/j.sna.2017.08.034.

[40] N. Romano, “Soil moisture at local scale: Measurements and simulations,” *J. Hydrol.*, vol. 516, pp. 6–20, Aug. 2014, doi: 10.1016/j.jhydrol.2014.01.026.

[41] M. Bittelli, “Measuring Soil Water Content: A Review,” *Horttechnology*, vol. 21, no. 3, pp. 293–300, Jun. 2011, doi: 10.21273/HORTTECH.21.3.293.

[42] Q. Chen, J. Zeng, and P. Zhang, “The simplified model of soil dielectric constant and soil moisture at the main frequency points of microwave band,” in *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, Jul. 2013, pp. 2712–2715, doi: 10.1109/IGARSS.2013.6723383.

[43] W. G. Fano and V. Trainotti, “Dielectric properties of soils,” in *2001 Annual Report Conference on Electrical Insulation and Dielectric Phenomena (Cat. No.01CH37225)*, pp. 75–78, doi: 10.1109/CEIDP.2001.963492.

[44] H. Kabir *et al.*, “Measurement and modelling of soil dielectric properties as a function of soil class and moisture content,” *J. Microw. Power Electromagn. Energy*, vol. 54, no. 1, pp. 3–18, Jan. 2020, doi: 10.1080/08327823.2020.1714103.

[45] M. J. Campbell and J. Ulrichs, “Electrical properties of rocks and their significance for lunar radar observations,” *J. Geophys. Res.*, vol. 74, no. 25, pp. 5867–5881, Nov. 1969, doi: 10.1029/JB074i025p05867.

[46] A. M. Mouazen, J. De Baerdemaeker, and H. Ramon, “Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer,” *Soil Tillage Res.*, vol. 80, no. 1–2, pp. 171–183, Jan. 2005, doi: 10.1016/j.still.2004.03.022.

[47] C. D. Christy, “Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy,” *Comput. Electron. Agric.*, vol. 61, no. 1, pp. 10–19, Apr. 2008, doi: 10.1016/j.compag.2007.02.010.

[48] Z. Yin, T. Lei, Q. Yan, Z. Chen, and Y. Dong, “A near-infrared reflectance sensor for soil surface moisture measurement,” *Comput. Electron. Agric.*, vol. 99, pp. 101–107, Nov. 2013, doi: 10.1016/j.compag.2013.08.029.

[49] P. Zhou, Y. Zhang, W. Yang, M. Li, Z. Liu, and X. Liu, “Development and performance test of an in-situ soil total nitrogen-soil moisture detector based on near-infrared spectroscopy,” *Comput. Electron. Agric.*, vol. 160, pp. 51–58, May 2019, doi: 10.1016/j.compag.2019.03.016.

[50] A. S. P. Priyaa *et al.*, “Microwave Sensor Antenna for Soil Moisture Measurement,” in *2015 Fifth*

International Conference on Advances in Computing and Communications (ICACC), Sep. 2015, pp. 258–262, doi: 10.1109/ICACC.2015.92.

[51] A. Berardinelli, G. Luciani, M. Crescentini, A. Romani, M. Tartagni, and L. Ragni, "Application of non-linear statistical tools to a novel microwave dipole antenna moisture soil sensor," *Sensors Actuators A Phys.*, vol. 282, pp. 1–8, Oct. 2018, doi: 10.1016/j.sna.2018.09.008.

[52] D. K. Gupta, R. Prasad, P. K. Srivastava, and T. Islam, "Nonparametric Model for the Retrieval of Soil Moisture by Microwave Remote Sensing," in *Satellite Soil Moisture Retrieval*, Elsevier, 2016, pp. 159–168.

[53] S. K. Dargar and V. M. Srivastava, "Moisture content investigation in the soil samples using microwave dielectric constant measurement method," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, p. 704, Feb. 2020, doi: 10.11591/ijece.v10i1.pp704-710.

[54] A. Morellos *et al.*, "Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy," *Biosyst. Eng.*, vol. 152, pp. 104–116, Dec. 2016, doi: 10.1016/j.biosystemseng.2016.04.018.

[55] A. B. Carlson and P. A. Crilly, *COMMUNICATION SYSTEMS An Introduction to Signals and Noise in Electrical Communication*, Fifth. McGraw-Hill, 2010.

[56] Eigenvector Research Inc., "Using Cross-Validation," 2020. https://wiki.eigenvector.com/index.php?title=Using_Cross-Validation.

[57] Eigenvector Research Inch., "Selectvars." https://www.wiki.eigenvector.com/index.php?title=Selectvars.

[58] S. Popovics, *Concrete Materials - Properties, Specification and Testing*. 1992.

[59] A. Wahab, M. M. A. Aziz, A. R. M. Sam, K. Y. You, A. Q. Bhatti, and K. A. Kassim, "Review on microwave nondestructive testing techniques and its applications in concrete technology," *Constr. Build. Mater.*, vol. 209, pp. 135–146, 2019, doi: 10.1016/j.conbuildmat.2019.03.110.

[60] X. Shi, N. Xie, K. Fortune, and J. Gong, "Durability of steel reinforced concrete in chloride environments: An overview," *Constr. Build. Mater.*, vol. 30, pp. 125–138, 2012, doi: 10.1016/j.conbuildmat.2011.12.038.

[61] S. Ismail, W. H. Kwan, and M. Ramli, "Mechanical strength and durability properties of concrete containing treated recycled concrete aggregates under different curing conditions," *Constr. Build. Mater.*, vol. 155, pp. 296–306, 2017, doi: 10.1016/j.conbuildmat.2017.08.076.

[62] K. Van Den Abeele, W. Desadeleer, G. De Schutter, and M. Wevers, "Active and passive monitoring of the early hydration process in concrete using linear and nonlinear acoustics," *Cem. Concr. Res.*, vol. 39, no. 5, pp. 426–432, 2009, doi: 10.1016/j.cemconres.2009.01.016.

[63] Z. Song, T. Frühwirt, and H. Konietzky, "Fatigue characteristics of concrete subjected to indirect cyclic tensile loading: Insights from deformation behavior, acoustic emissions and ultrasonic wave propagation," *Constr. Build. Mater.*, vol. 302, no. December 2020, p. 124386, 2021, doi: 10.1016/j.conbuildmat.2021.124386.

[64] N. Tareen, J. Kim, W. K. Kim, and S. Park, "Comparative analysis and strength estimation of fresh concrete based on ultrasonicwave propagation and maturity using smart temperature and PZT sensors," *Micromachines*, vol. 10, no. 9, pp. 1–17, 2019, doi: 10.3390/mi10090559.

[65] Kaplanvural, E. Pekşen, and K. Özkap, "Volumetric water content estimation of C-30 concrete using GPR," *Constr. Build. Mater.*, vol. 166, pp. 141–146, 2018, doi: 10.1016/j.conbuildmat.2018.01.132.

[66] İ. Kaplanvural, K. Özkap, and E. Pekşen, "Influence of water content investigation on GPR wave

attenuation for early age concrete in natural air-drying condition," *Constr. Build. Mater.*, vol. 297, 2021, doi: 10.1016/j.conbuildmat.2021.123783.

[67] Y. Y. Lim, K. Z. Kwong, W. Y. H. Liew, and C. K. Soh, "Practical issues related to the application of piezoelectric based wave propagation technique in monitoring of concrete curing," *Constr. Build. Mater.*, vol. 152, pp. 506–519, 2017, doi: 10.1016/j.conbuildmat.2017.06.163.

[68] Y. Y. Lim, S. T. Smith, and C. K. Soh, "Wave propagation based monitoring of concrete curing using piezoelectric materials: Review and path forward," *NDT E Int.*, vol. 99, no. June, pp. 50–63, 2018, doi: 10.1016/j.ndteint.2018.06.002.

[69] P. Priyada, R. Ramar, and Shivaramu, "Determining the water content in concrete by gamma scattering method," *Ann. Nucl. Energy*, vol. 63, pp. 565–570, 2014, doi: 10.1016/j.anucene.2013.07.049.

[70] N. Sabbağ and O. Uyanık, "Determination of the reinforced concrete strength by apparent resistivity depending on the curing conditions," *J. Appl. Geophys.*, vol. 155, pp. 13–25, 2018, doi: 10.1016/j.jappgeo.2018.03.007.

[71] L. Xiao and Z. Li, "Early-age hydration of fresh concrete monitored by non-contact electrical resistivity measurement," *Cem. Concr. Res.*, vol. 38, no. 3, pp. 312–319, 2008, doi: 10.1016/j.cemconres.2007.09.027.

[72] J. Bhargava and K. Lundberg, "Determination of moisture content of concrete by microwave-resonance method," *Matériaux Constr.*, vol. 5, no. 3, pp. 165–168, 1972, doi: 10.1007/BF02539259.

[73] F. H. Wittman and F. Schlude, "MICROWAVE ABSORPTION OF HARDENED CEMENT PASTE," *Cem. Concr. Res.*, vol. 5, pp. 63–71, 1975.

[74] K. Gorur, M. K. Smit, and F. H. Wittmann, "Microwave study of hydrating cement paste at early age," *Cem. Concr. Res.*, vol. 12, no. 4, pp. 447–454, 1982, doi: 10.1016/0008-8846(82)90059-X.

[75] R. Haddad and I. Al-Qadi, "Characterization of Portland Cement Concrete Using," *Cem. Concr. Res.*, vol. 28, no. 10, pp. 1379–1391, 1998.

[76] O. Büyüköztürk, T. Y. Yu, and J. A. Ortega, "A methodology for determining complex permittivity of construction materials based on transmission-only coherent, wide-bandwidth free-space measurements," *Cem. Concr. Compos.*, vol. 28, no. 4, pp. 349–359, 2006, doi: 10.1016/j.cemconcomp.2006.02.004.

[77] S. N. Kharkovsky, M. F. Akay, U. C. Hasar, and C. D. Atis, "Measurement and monitoring of microwave reflection and transmission properties of cement-based specimens," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1210–1217, 2002, doi: 10.1109/TIM.2002.808081.

[78] M. Jamil, M. K. Hassan, H. M. A. Al-Mattarneh, and M. F. M. Zain, "Concrete dielectric properties investigation using microwave nondestructive techniques," *Mater. Struct. Constr.*, vol. 46, no. 1–2, pp. 77–87, 2013, doi: 10.1617/s11527-012-9886-2.

[79] T. T. Dinh *et al.*, "Dielectric material characterization of concrete in GHz range in dependence on pore volume and water content," *Constr. Build. Mater.*, vol. 311, no. October, 2021, doi: 10.1016/j.conbuildmat.2021.125234.

[80] G. Klysz, J. P. Balayssac, and X. Ferrières, "Evaluation of dielectric properties of concrete by a numerical FDTD model of a GPR coupled antenna-Parametric study," *NDT E Int.*, vol. 41, no. 8, pp. 621–631, 2008, doi: 10.1016/j.ndteint.2008.03.011.

[81] P. Shen and Z. Liu, "Study on the hydration of young concrete based on dielectric property measurement," *Constr. Build. Mater.*, vol. 196, pp. 354–361, 2019, doi:

10.1016/j.conbuildmat.2018.11.150.

[82] N. E. Hager and R. C. Domszy, “Monitoring of cement hydration by broadband time-domain-reflectometry dielectric spectroscopy,” *J. Appl. Phys.*, vol. 96, no. 9, pp. 5117–5128, 2004, doi: 10.1063/1.1797549.

[83] X. Zhang, X. Z. Ding, T. H. Lim, C. K. Ong, B. T. G. Tan, and J. Yang, “Microwave study of hydration of slag cement blends in early period,” *Cem. Concr. Res.*, vol. 25, no. 5, pp. 1086–1094, 1995, doi: 10.1016/0008-8846(95)00103-J.

[84] P. Juan-García and J. M. Torrents, “Measurement of mortar permittivity during setting using a coplanar waveguide,” *Meas. Sci. Technol.*, vol. 21, no. 4, 2010, doi: 10.1088/0957-0233/21/4/045702.

[85] K. Chung and S. Kharkovsky, “Measurements of microwave reflection properties of early-Age concrete and mortar specimens,” *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, pp. 1295–1300, 2014, doi: 10.1109/I2MTC.2014.6860954.

[86] A. Wahab, M. M. A. Aziz, A. R. MohdSam, and K. Y. You, “Application of microwave waveguide techniques for investigating the effect of concrete dielectric and reflection properties during curing,” *J. Build. Eng.*, vol. 38, no. January, p. 102209, 2021, doi: 10.1016/j.jobe.2021.102209.

[87] Y. Wang, T. D. Wig, J. Tang, and L. M. Hallberg, “Dielectric properties of foods relevant to RF and microwave pasteurization and sterilization,” *J. Food Eng.*, vol. 57, no. 3, pp. 257–268, 2003, doi: 10.1016/S0260-8774(02)00306-0.

[88] M. P. Robinson, J. Clegg, and D. A. Stone, “A novel method of studying total body water content using a resonant cavity: Experiments and numerical simulation,” *Phys. Med. Biol.*, vol. 48, no. 1, pp. 113–125, 2003, doi: 10.1088/0031-9155/48/1/308.

[89] EMC Technology Inc., “A VSWR Meter Using a Power Sensing Termination,” *Microw. J.*, 2000.

[90] International Federation for Structural Concrete, *Model Code 2010 - Final draft, Volume 1*. 2012.

[91] A. F. Khalizov, F. J. Guzman, M. Cooper, N. Mao, J. Antley, and J. Bozzelli, “Direct detection of gas-phase mercuric chloride by ion drift - Chemical ionization mass spectrometry,” *Atmos. Environ.*, vol. 238, no. May, p. 117687, 2020, doi: 10.1016/j.atmosenv.2020.117687.

[92] Q. Niu *et al.*, “Exploring catalytic pyrolysis of Palm Shell over HZSM-5 by gas Chromatography/mass spectrometry and photoionization mass spectrometry,” *J. Anal. Appl. Pyrolysis*, vol. 152, no. March, p. 104946, 2020, doi: 10.1016/j.jaap.2020.104946.

[93] C. Drees, A. Schütz, G. Niu, J. Franzke, W. Vautz, and S. Brandt, “Stepwise optimization of a Flexible Microtube Plasma (FμTP) as an ionization source for Ion Mobility Spectrometry,” *Anal. Chim. Acta*, vol. 1127, pp. 89–97, 2020, doi: 10.1016/j.aca.2020.06.018.

[94] J. K. Jung, I. G. Kim, K. S. Chung, and U. B. Baek, “Analyses of permeation characteristics of hydrogen in nitrile butadiene rubber using gas chromatography,” *Mater. Chem. Phys.*, vol. 267, no. February, p. 124653, 2021, doi: 10.1016/j.matchemphys.2021.124653.

[95] J. K. Jung, I. G. Kim, K. S. Chung, and U. B. Baek, “Gas chromatography techniques to evaluate the hydrogen permeation characteristics in rubber: ethylene propylene diene monomer,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-83692-1.

[96] H. Jung *et al.*, “Selective detection of sub-1-ppb level isoprene using Pd-coated In2O3 thin film integrated in portable gas chromatography,” *Appl. Surf. Sci.*, vol. 586, no. July 2021, p. 152827, 2022, doi: 10.1016/j.apsusc.2022.152827.

[97] H. Kim *et al.*, “A micropump-driven high-speed mems gas chromatography system,” in

TRANSDUCERS and EUROSENSORS '07 - 4th International Conference on Solid-State Sensors, Actuators and Microsystems, 2007, pp. 1505–1508, doi: 10.1109/SENSOR.2007.4300430.

[98] H. C. Hsieh and H. Kim, "Isomer separation enabled by a micro circulatory gas chromatography system," *J. Chromatogr. A*, vol. 1629, p. 461484, 2020, doi: 10.1016/j.chroma.2020.461484.

[99] Y. Qin and Y. B. Gianchandani, "A facile, standardized fabrication approach and scalable architecture for a micro gas chromatography system with integrated pump," *2013 Transducers Eurosensors XXVII 17th Int. Conf. Solid-State Sensors, Actuators Microsystems, TRANSDUCERS EUROSENSORS 2013*, no. June, pp. 2755–2758, 2013, doi: 10.1109/Transducers.2013.6627376.

[100] Q. Cheng, Y. Qin, and Y. B. Gianchandani, "A bidirectional Knudsen pump with superior thermal management for micro-gas chromatography applications," *Proc. IEEE Int. Conf. Micro Electro Mech. Syst.*, pp. 167–170, 2017, doi: 10.1109/MEMSYS.2017.7863367.

[101] T. Byambadorj, Y. Qin, and Y. B. Gianchandani, "Blocking Pressure Enhancement in SOI Through-Wafer Monolithic Knudsen PUMPs," *IEEE Symp. Mass Storage Syst. Technol.*, vol. 2022-Janua, no. January, pp. 43–46, 2022, doi: 10.1109/MEMS51670.2022.9699567.

[102] US10229809B2, "Device for generating a composition-controlled and intensity-controlled ionic flow and related method," US10229809B2, 2015.

[103] US20170133212A1, "Portable electronic device for the analysis of a gaseous composition," US20170133212A1, 2017.

[104] J. M. Lafferty, *Foundation of Vacuum Science and Technology*. NJ, USA: John Wiler & Sons, 1998.

[105] J. E. Welke, V. Manfroi, M. Zanus, M. Lazzarotto, and C. A. Zini, "Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data," *Food Chem.*, vol. 141, no. 4, pp. 3897–3905, 2013, doi: 10.1016/j.foodchem.2013.06.100.

[106] S. D. Lv *et al.*, "Multivariate Analysis Based on GC-MS Fingerprint and Volatile Composition for the Quality Evaluation of Pu-Erh Green Tea," *Food Anal. Methods*, vol. 8, no. 2, pp. 321–333, 2015, doi: 10.1007/s12161-014-9900-0.

[107] N. G. S. Mogollón *et al.*, "Comprehensive two-dimensional gas chromatography-mass spectrometry combined with multivariate data analysis for pattern recognition in Ecuadorian spirits," *Chem. Cent. J.*, vol. 12, no. 1, pp. 1–10, 2018, doi: 10.1186/s13065-018-0470-x.

[108] S. K. Jha, M. Imahashi, K. Hayashi, and T. Takamizawa, "Data fusion approach for human body odor discrimination using GC-MS spectra," *IEEE ISSNIP 2014 - 2014 IEEE 9th Int. Conf. Intell. Sensors, Sens. Networks Inf. Process. Conf. Proc.*, no. April, pp. 21–24, 2014, doi: 10.1109/ISSNIP.2014.6827592.

[109] E. N. Stark, J. A. Covington, S. Agbroko, C. Peng, W. E. Hahn, and E. Barenholtz, "Deep Learning Investigation of Mass Spectrometry Analysis from Melanoma Samples," *ISOEN 2019 - 18th Int. Symp. Olfaction Electron. Nose, Proc.*, pp. 1–4, 2019, doi: 10.1109/ISOEN.2019.8823194.

[110] M. G. Jajin *et al.*, "Gas chromatography-mass spectrometry-based untargeted metabolomics reveals metabolic perturbations in medullary thyroid carcinoma," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-12590-x.

[111] A. Bagolini, R. Correale, A. Picciotto, M. Di Lorenzo, and M. Scapinello, "MEMS Membranes with Nanoscale Holes for Analytical Applications," *Membranes (Basel).*, vol. 11, no. 2, p. 74, Jan. 2021, doi: 10.3390/membranes11020074.

[112] M. Markelov and J. P. Guzowski, "Matrix independent headspace gas chromatographic analysis. This full evaporation technique," *Anal. Chim. Acta*, vol. 276, no. 2, pp. 235–245, 1993, doi: 10.1016/0003-

2670(93)80390-7.

[113] W. E. Wallace, “Mass Spectra,” in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, P. J. Linstrom and W. . Mallard, Eds. Gaithersburg MD: National Institute of Standards and Technology, 2022.

[114] D. L. Stoll, S. C. Rutan, and C. J. Venkatramani, “Peak Purity in Liquid Chromatography, Part I: Basic Concepts, Commercial Software, and Limitations,” *LCGC North Am.*, vol. 36, no. 2, pp. 110–118, 2018, [Online]. Available: https://www.chromatographyonline.com/view/peak-purity-liquid-chromatography-part-i-basic-concepts-commercial-software-and-limitations.

[115] F. Savorani, G. Tomasi, and S. B. Engelsen, “icoshift: A versatile tool for the rapid alignment of 1D NMR spectra,” *J. Magn. Reson.*, vol. 202, no. 2, pp. 190–202, 2010, doi: 10.1016/j.jmr.2009.11.012.

[116] N. Erkan, Ö. Özden, D. Ü. Alakavuk, Ş. Y. Yildirim, and M. Inuğur, “Spoilage and shelf life of sardines (Sardina pilchardus) packed in modified atmosphere,” *Eur. Food Res. Technol.*, vol. 222, no. 5–6, pp. 667–673, 2006, doi: 10.1007/s00217-005-0194-8.

[117] V. R. Kyrana and V. P. Lougovois, “Sensory, chemical and microbiological assessment of farm-raised European sea bass (Dicentrarchus labrax) stored in melting ice,” *International Journal of Food Science and Technology*, vol. 37, no. 3. pp. 319–328, 2002, doi: 10.1046/j.1365-2621.2002.00572.x.

[118] G. Olafsdottir, R. Jonsdottir, H. L. Lauzon, J. Luten, and K. Kristbergsson, “Characterization of volatile compounds in chilled cod (Gadus morhua) fillets by gas chromatography and detection of quality indicators by an electronic nose,” *J. Agric. Food Chem.*, vol. 53, no. 26, pp. 10140–10147, 2005, doi: 10.1021/jf0517804.

[119] C. Alasalvar, K. D. A. Taylor, A. Öksüz, F. Shahidi, and M. Alexis, “Comparison of freshness quality of cultured and wild sea bass (Dicentrarchus labrax),” *J. Food Sci.*, vol. 67, no. 9, pp. 3220–3226, 2002, doi: 10.1111/j.1365-2621.2002.tb09569.x.

[120] I. N. A. Ashie, J. P. Smith, and B. K. Simpson, *Spoilage and Shelf-Life Extension of Fresh Fish and Shellfish*, vol. 36, no. 1–2. 1996.

[121] G. Olafsdottir *et al.*, “Multisensor for fish quality determination,” *Trends Food Sci. Technol.*, vol. 15, no. 2, pp. 86–93, 2004, doi: 10.1016/j.tifs.2003.08.006.

[122] A. Pedrosa-Menabrito and J. M. Regenstein, “Shelf-life extension of fresh fish - A review part 1 - Spoilage of fish,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 1988.

[123] CEE, “COUNCIL REGULATION (EEC) No 103/76 of 19 January 1976 laying down common marketing standards for certain fresh or chilled fish,” *Off. J. Eur. Communities*, no. 42, pp. 29–34, 1976.

[124] S. Pons-Sánchez-Cascado, M. C. Vidal-Carou, M. L. Nunes, and M. T. Veciana-Nogués, “Sensory analysis to assess the freshness of Mediterranean anchovies (Engraulis encrasicholus) stored in ice,” *Food Control*, vol. 17, no. 7, pp. 564–569, 2006, doi: 10.1016/j.foodcont.2005.02.016.

[125] A. E. Massa, E. Manca, and M. I. Yeannes, “Development of quality index method for anchovy (Engraulis anchoita) stored in ice: Assessment of its shelf-life by chemical and sensory methods,” *Food Sci. Technol. Int.*, vol. 18, no. 4, pp. 339–351, 2012, doi: 10.1177/1082013211428014.

[126] A. Bensid, Y. Ucar, B. Bendeddouche, and F. Özogul, “Effect of the icing with thyme, oregano and clove extracts on quality parameters of gutted and beheaded anchovy (Engraulis encrasicholus) during chilled storage,” *Food Chem.*, vol. 145, pp. 681–686, 2014, doi: 10.1016/j.foodchem.2013.08.106.

[127] R. Marrone, G. Smaldone, G. Palma, R. Romano, D. Bortone, and A. Aniello, “Determination of cholesterol oxides in anchovies (Engraulis encrasicolus) treated with a commercial mixture of citric

acid, trisodium acid and hydrogen peroxide," *Ital. J. Food Saf.*, vol. 1, no. 6, pp. 42–45, 2012.

[128] F. Özogul, E. Tugce Aksun, R. Öztekin, and J. M. Lorenzo, "Effect of lavender and lemon balm extracts on fatty acid profile, chemical quality parameters and sensory quality of vacuum packaged anchovy (Engraulis encrasicolus) fillets under refrigerated condition," *Lwt*, vol. 84. pp. 529–535, 2017, doi: 10.1016/j.lwt.2017.06.024.

[129] R. Draisci, G. Volpe, L. Lucentini, A. Cecilia, R. Federico, and G. Palleschi, "Determination of biogenic amines with an electrochemical biosensor and its application to salted anchovies," *Food Chem.*, vol. 62, no. 2, pp. 225–232, 1998, doi: 10.1016/S0308-8146(97)00167-2.

[130] M. O. Varrà, S. Ghidini, A. Ianieri, and E. Zanardi, "Near infrared spectral fingerprinting: A tool against origin-related fraud in the sector of processed anchovies," *Food Control*, vol. 123, no. October 2020, 2021, doi: 10.1016/j.foodcont.2020.107778.

[131] H. M. Velioꞵlu, H. T. Temiz, and I. H. Boyaci, "Differentiation of fresh and frozen-thawed fish samples using Raman spectroscopy coupled with chemometric analysis," *Food Chem.*, vol. 172, pp. 283–290, 2015, doi: 10.1016/j.foodchem.2014.09.073.

[132] Y. Y. Pu, Y. Z. Feng, and D. W. Sun, "Recent progress of hyperspectral imaging on quality and safety inspection of fruits and vegetables: A review," *Compr. Rev. Food Sci. Food Saf.*, vol. 14, no. 2, pp. 176–188, 2015, doi: 10.1111/1541-4337.12123.

[133] T. Wu, N. Zhong, and L. Yang, "Application of VIS/NIR Spectroscopy and SDAE-NN Algorithm for Predicting the Cold Storage Time of Salmon," *J. Spectrosc.*, vol. 2018, 2018, doi: 10.1155/2018/7450695.

[134] A. A. Agyekum *et al.*, "Rapid and Nondestructive Quantification of Trimethylamine by FT-NIR Coupled with Chemometric Techniques," *Food Anal. Methods*, vol. 12, no. 9, pp. 2035–2044, 2019, doi: 10.1007/s12161-019-01537-0.

[135] C. Alamprese and E. Casiraghi, "Application of FT-NIR and FT-IR spectroscopy to fish fillet authentication," *LWT - Food Sci. Technol.*, vol. 63, no. 1, pp. 720–725, 2015, doi: 10.1016/j.lwt.2015.03.021.

[136] E. Ivorra, A. J. Sánchez, S. Verdú, J. M. Barat, and R. Grau, "Shelf life prediction of expired vacuum-packed chilled smoked salmon based on a KNN tissue segmentation method using hyperspectral images," *J. Food Eng.*, vol. 178, pp. 110–116, 2016, doi: 10.1016/j.jfoodeng.2016.01.008.

[137] J. H. Cheng, D. W. Sun, and Q. Wei, "Enhancing Visible and Near-Infrared Hyperspectral Imaging Prediction of TVB-N Level for Fish Fillet Freshness Evaluation by Filtering Optimal Variables," *Food Anal. Methods*, vol. 10, no. 6, pp. 1888–1898, 2017, doi: 10.1007/s12161-016-0742-9.

[138] S. Khoshnoudi-Nia and M. Moosavi-Nasab, "Prediction of various freshness indicators in fish fillets by one multispectral imaging system," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019, doi: 10.1038/s41598-019-51264-z.

[139] M. Bachrun Alim, A. Suhaeli Fahmi, L. Purnamayati, and T. W. Agustini, "Non-destructive freshness assessment of Cyprinus carpio based on image analysis," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 530, no. 1, 2020, doi: 10.1088/1755-1315/530/1/012014.

[140] S. Negi, N. Yadav, R. Rawat, and R. Singh, "An effective technique for determining fish freshness using image processing," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special Issue, pp. 460–464, 2019, doi: 10.35940/ijitee.I1073.0789S19.

[141] H. Mohammadi Lalabadi, M. Sadeghi, and S. A. Mireei, "Fish freshness categorization from eyes and gills color features using multi-class artificial neural network and support vector machines," *Aquac. Eng.*, vol. 90, no. October 2019, p. 102076, 2020, doi: 10.1016/j.aquaeng.2020.102076.

[142] A. Yapar and H. Yetim, “Determination of anchovy freshness by refractive index of eye fluid,” *Food Res. Int.*, vol. 31, no. 10, pp. 693–695, 1998, doi: 10.1016/S0963-9969(99)00047-2.

[143] A. Wendel, J. Underwood, and K. Walsh, “Maturity estimation of mangoes using hyperspectral imaging from a ground based mobile platform,” *Comput. Electron. Agric.*, vol. 155, no. June, pp. 298–313, 2018, doi: 10.1016/j.compag.2018.10.021.

[144] J. Gasteiger and J. Zupan, “Neural Networks in Chemistry,” *Angew. Chemie Int. Ed. English*, vol. 32, no. 4, pp. 503–527, Apr. 1993, doi: 10.1002/anie.199305031.

[145] C. M. Bishop, “Neural networks for pattern recognition,” 1995.

[146] M. F. Abbod, J. W. F. Catto, D. A. Linkens, and F. C. Hamdy, “Application of Artificial Intelligence to the Management of Urological Cancer,” *J. Urol.*, vol. 178, no. 4, pp. 1150–1156, Oct. 2007, doi: 10.1016/j.juro.2007.05.122.

[147] C. W. DAWSON and R. WILBY, “An artificial neural network approach to rainfall-runoff modelling,” *Hydrol. Sci. J.*, vol. 43, no. 1, pp. 47–66, Feb. 1998, doi: 10.1080/02626669809492102.

[148] A. Zell *et al.*, “SNNS (Stuttgart Neural Network Simulator),” 1994, pp. 165–186.

[149] M. Miljanovic, “Comparative analysis of Recurrent and Finite Impulse Response Neural Networks in Time Series Prediction,” *Indian J. Comput. Sci. Eng.*, vol. 3, no. 1, pp. 180–191, 2012.

[150] J. Kelleher, B. Mac Namee, and A. D’Arcy, *Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, vol. 53, no. 9. 2020.

[151] D. Marcu, M. Danubianu, B. Adina, and C. Simionescu, “Algorithms for Classifying the Results at the Baccalaureate Exam - Comparative Analysis of Performances,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 8, 2021.

[152] Z. Shao *et al.*, “Memory-Efficient CNN Accelerator Based on Interlayer Feature Map Compression,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 69, no. 2, pp. 668–681, 2022, doi: 10.1109/TCSI.2021.3120312.

[153] “Raspberry Pi Foundation - About us.” https://www.raspberrypi.org/about/ (accessed Jan. 28, 2023).

[154] C. L. Beltran, J. V. Cortez, A. Z. López, and S. Cordero, “Implementación de sistema de visión inteligente para reconocimiento de rostros en robot de código abierto BOB,” *Res. Comput. Sci.*, vol. 148, no. 8, pp. 777–790, 2021.

[155] “IP5305, Injoinic Technology.” https://aitendo3.sakura.ne.jp/aitendo_data/product_img/ic/charger/IP5305/IP5305-Injoinic.pdf.

[156] “SHT40, Sensirion.” https://sensirion.com/products/catalog/SHT40/.

[157] “TPL5110, Texas Instrument.”https://www.ti.com/lit/ds/symlink/tpl5110.pdf?ts=1674910543075&ref_url=https%253A%252F%252Fwww.google.com%252F.

[158] “MAX5161, Maxim Integrated.” https://www.stg-maximintegrated.com/en/products/analog/data-converters/digital-potentiometers/MAX5161.html.

12. Acknowledgments

The thesis presented is the result of 3 years of work, developed mainly in the context of the Electrical, Electronic, and Information Engineering Department (DEI) of the University of Bologna (Cesena Campus). During my course of study, the activity was carried out under the supervision of prof. Tartagni, and made possible thanks to the help of the other DEI professors; of prof. Berardinelli, from the Centre Agriculture Food Environment (C3A) of the University of Trento; and thanks to the collaboration with other Unibo Departments, like the Department of Agricultural and Food Science (DISTAL) and the Civil, Chemical, Environmental and Materials Engineering Department (DICAM), as well as with the Nanotech Analysis startup.

In particular, I would like to gladly thank the following for the collaboration and help given to me during the PhD:

- My advisor Prof. Marco Tartagni, from DEI; and my co-advisor Prof. Annachiara Berardinelli, from C3A
- Prof. Marco Crescentini and Prof. Aldo Romani, from DEI
- Prof. Luigi Ragni, Junior Assistant Prof. Eleonora Iaccheri and Senior Assistant Prof. Chiara Cevoli, from DISTAL
- Prof. Claudio Mazzotti and Prof. Elisa Franzoni from DICAM
- Raffaele Correale and Carla Ciricugno from Nanotech Analysis (NTA)