SDU

University of
Southern Denmark

Material recognition using Hyperspectral Imaging

Project in Electronics

Author: Maria Teresa Maza Luengo
Supervisor: Ali Sahafi
Bachelor in Electronic and Automation Engineering
University of Southern Denmark
2 June 2025

Abstract

Asbestos is the name given to a group of six natural mineral fibers, widely used in construction during the 19th and 20th centuries for insulation. Nowadays, its use has been banned in many countries (more than 50 WHO Member States) after discovering that the inhalation exposure to the fibers that this material releases can cause cancer, asbestosis and other severe diseases that can ultimately lead to death. However, asbestos is still present in many buildings. Thus, real-time systems for the detection of this material have become of great importance. Previous research and projects have led to the development of different types of systems that detect this material, including real-time, portable asbestos detectors that use spectroscopy to detect different types of asbestos minerals. Based on this idea of real-time material detection, a Hyperspectral Imaging (HSI) system is used in this project to capture the spectrum of different construction materials. By feeding the collected data into a Support Vector Machine (SVM) machine learning algorithm, the pursued objective is to obtain an automated, fast and reliable method to differentiate several construction materials from one another. Although it was not possible to develop a fully autonomous system, a design that correctly and reliably identifies four different constructions materials was achieved. This shows the potential of this technology in fields like recycling, where automated classification of materials can be very useful, increasing efficiency and reducing the risk of human exposure to toxic substances.

Contents

List of Figures

List of Tables

1 Introduction

Asbestos is a group of 6 fibrous silicate minerals of natural origin: amosite (brown asbestos), crocidolite (blue asbestos), tremolite, anthophyllite, actinolite and chrysotile (white asbestos). These are usually divided into 2 groups according to the shape of their fibers. So, while the first 5 types are grouped under amphibole asbestos, defined by straight, pointed fibers, chrysotile belongs to the serpentine category, which has curly fibers and represents 95% of all the asbestos used in the world [2].

Asbestos is a strong, thermal and electrical insulating material, resistant to chemical and biological degradation. These properties made this material widely present in construction during the 19th and 20th century [2]. However, it is also one of the main causes of work-related deaths in the world, responsible for more than 200,000 deaths worldwide per year [3]. Since the 1970s and especially during the beginning of the 21st century, awareness about this harmful material has risen, after multiple studies, like the one carried out by the Pneumoconiosis Committee of the College of American Pathologists and the National Institute for Occupational Safety and Health in 1982 [4] or the one conducted by the U.S. Department of Health and Human Services in 2001 [5] among others, have proven all types of asbestos to be toxic and a serious health hazard. Banned for use in new constructions in all the EU since 2005, many old buildings still contain this toxic material, which is also still being mined and used in some other parts of the world. As the fibers of this material can spread very easily, it is essential to remove asbestos and asbestos-containing materials (ACM) to avoid contamination. All of this has made the interest in developing asbestos detectors increase significantly in the past recent years. Big research and projects were carried out and subsidized by several institutions. One of the most relevant works was the ALERT project [6]. Funded by the EU and carried out by a team of 11 organizations from across Europe, this 3 years study accomplished its goal of developing a real-time, portable asbestos detector.

Nowadays, there are multiple ways to detect asbestos such as Polarized Light Microscopy (PLM), which takes advantage of the particular optical characteristics of asbestos to distinguish it, and electron microscopy, which makes use of the interactions between an emitted beam of electrons and the material under study to obtain an image of its structure that allows identification. Two of the most widely used types of this last technique are Scanning Electron Microscopy (SEM) and Transmission Electron Microscopy (TEM) [7], which differ in some aspects like the resulting images and the way they are obtained: SEM uses backscattered electrons to obtain an image of the sample surface while TEM utilizes transmitted electrons to get an image of its internal structure [8]. Although all these detection processes are very precise and allow a good visualization of asbestos fibers, most of them require the recollection of samples that must be then analyzed in a lab, thus involving human exposure to the toxic material, requiring qualified personnel and certified equipment and being time-consuming. Another asbestos detection technique that can provide a more immediate result is light scattering [9], where a laser beam is used to detect particles in the air. However, this method cannot differentiate asbestos particles from non-hazardous ones. Infrared spectroscopy and X-ray fluorescence [10], where infrared light or X-rays are used respectively

to detect the material according to its distinctive interaction with the radiation, are other forms of on-site asbestos detection.

The use of Hyperspectral Imaging (HSI) represents a significant progress in asbestos detection, as the great amount of data this technology provides allows the use of machine learning to achieve an automated asbestos detection process. HySpex [11], a brand by the Norwegian company Norsk Elektro Optikk (NEO), develops HSI systems and has carried out experiments that prove how their cameras can be used to detect and classify construction materials, including asbestos.

This project is based on this HSI approach, focusing on the development of a method that allows an efficient, autonomous identification of construction materials, implementing machine learning algorithms to analyze the data obtained by an HSI system. By using this technique, it is possible to perform quick, on-site analysis of materials and to obtain results without the need of running lab exams on samples. Moreover, as this process does not require direct contact, it represents a cleaner and safer way of detection. On the downside, machine learning may consume a great amount of time and resources at the beginning to process initial data and obtain the first results. But this will eventually pay off because the more data it consumes, the better its ability to identify materials will be. This means that, after a first great effort to develop and train the algorithm, once the system is able to recognize materials, it can be improved by using it. However, although HSI technology can detect asbestos and its presence in other materials, it is important to observe that this method is not suitable for detecting airborne asbestos, as the microscopic fibers (between 2µm – 10µm long and only around 0.1µm wide [12]) scattered through the air would hardly be detected.

This project is developed at the University of Southern Denmark (SDU) as an academic project with the collaboration of the NanoSYD department that provides the lab and the HSI equipment and the Civil Engineering department that supplies the required construction materials.

In this project, the equipment used to obtain the materials spectral information is the NEWTEC HSI system Buteo [13]. The software Hyperspectral Analyzer, provided by Morten Sielnik Andersen, is used to work with the hyperspectral images and manage the information given by the HSI system. Then, for machine-learning, a Support Vector Machine (SVM) algorithm will be implemented in Python using Visual Studio Code. The reason why SVM is the method chosen among several types of classification algorithms like Random Forest or Naïve Bayes is because SVM performs well when given a significant amount of labeled data with a lot of features, which is the kind of data used in this project.

An important remark regarding the materials studied is that asbestos will not be used in this work due to the difficulties in obtaining the material, its toxicity and all the risks and problems it represents. Another construction materials (timber, steel, concrete and reinforced concrete) will be utilized instead because, once the differentiation among multiple construction materials is achieved, an analogous process can be followed to detect asbestos.

This report shows the process followed to try to achieve the desired goal, as well as

the results and conclusions obtained. First, the report presents HSI systems, briefly explaining how they work and how this technology was used in the project. Then, the process of working with the data to prepare them for machine learning is explained, the materials used to extract the data are presented and the most relevant aspects of their spectra are discussed. After that, the machine learning process is described and shown in detail, explaining how the algorithm was developed and trained to recognize the different materials. The results of this project are presented in the following section, followed by the final conclusions drawn from all the work and its outcomes. The Annex, divided into 2 sections, contains all the information that, being relevant to clarify some aspects of the project, was not included in the main document to avoid making it excessively long.

2 Hyperspectral Imaging

Hyperspectral Imaging (HSI) is a technique that uses spectroscopy and imaging to obtain spectral information from every pixel of an image [14]. Spectroscopy consists in studying the way electromagnetic radiation interacts with matter. This technique allows the acquisition of spectra (the amount of reflected, absorbed, emitted, transmitted or scattered radiation at different wavelengths of the electromagnetic spectrum). As every material has its own unique spectrum, HSI is a reliable, non-invasive technique to identify materials.

HSI systems include a hyperspectral camera or imaging spectrometer, which is a type of optical spectrometer. Optical spectrometers split the incoming light into its different wavelengths by focusing it through a slit at, usually, a diffraction grating, a film that breaks up light in all its constituents including UV and IR radiation. Then, they measure the intensity of light for every wavelength to obtain the spectrum. In the case of imaging spectrometers, the information obtained is not only about the spectrum of the studied material, but also about the spatial distribution of it.

Due to its flexibility (suitable for use in different lab and field applications) and its great performance (high temporal, spatial and spectral resolution [15]), this technology is becoming increasingly present in multiple sectors: from agriculture in the analysis of products quality or the mineral composition of the planting soil to criminology in the detection of explosives or blood [16]. Now, HSI is also starting to be used for material classification in the textile, plastic and construction industries.

For this project, focused on the classification of building materials, the NEWTEC HSI system Buteo [13] in the Sensor Teknologi Lab of the NanoSYD department was used. This system uses push broom, line scanning technique: spectral and spatial information is extracted from one line of pixels at a time, line by line, as the object moves in a direction perpendicular to the scanning line, and the resulting hyperspectral image is created by putting all these lines together. The equipment includes the hyperspectral smart camera Oculus, working with 900 channels in the visible-short-wave-infrared (Vis-SWIR) configuration, that means, in a spectral range from 430-1700 nm. This is a relevant factor, given that, if the characteristic features of one material's spectrum are out of this region, it would be very difficult to identify said material.

Other important aspects to take into account when working with HSI systems are the calibration of the system and the lighting. To calibrate the equipment, we use a modular calibration board provided by Morten Sielnik Andersen, PhD student at SDU. This board allows us to carry out intensity, spectral and spatial calibration in a simple, quick way. More details about the board and its operation can be found in the document *Calibration of Hyperspectral Cameras* by Morten Sielnik Andersen. An early version of the document was used in this work. However, the article is not included in the References because it had not been published by the end of this project. Regarding the lighting, the lab is equipped with six lamps, which are kept off during the scans. In addition, the room receives a relatively small amount of natural light, which can be almost fully suppressed by closing the curtains (although they are kept open for these

scans, as the amount of natural light is negligible compared to that provided by the light bulbs that the HSI equipment has). This way, in the scans, most of the light is provided by the 13 halogen bulbs incorporated in the system (6 bulbs on one side and 7 on the other side of the scanning band, all of them pointing at said band). The intensity of these bulbs can be adjusted to obtain the best resolution without saturating any pixels. Moreover, there are other parameters that can be adjusted, such as the speed of the conveyor belt and the image length. These two parameters are set to 40mm/s and 1000 pixels respectively to get the entire calibration board into the image, something essential for the calibration to work correctly.

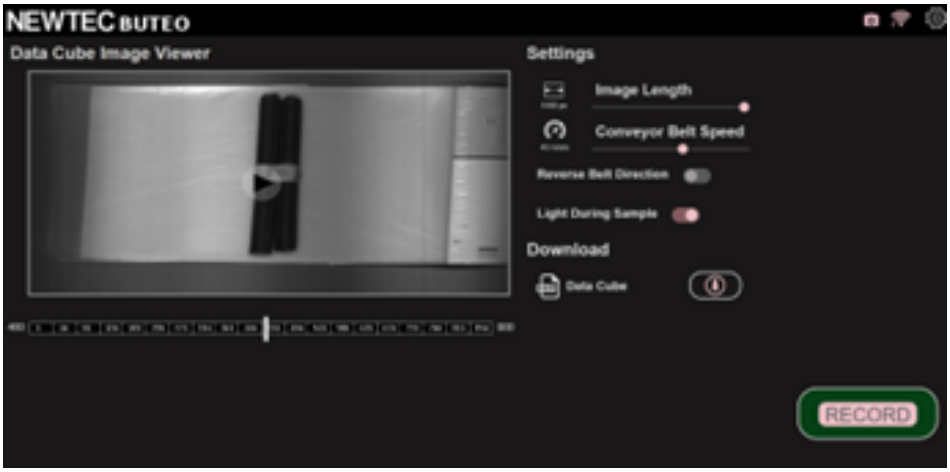


Figure 1: Main screen of the HSI system.

Then, when a scan is carried out, starting it by pressing the "record" button on the main screen (Fig. 1), the system turns the halogen bulbs on with the established intensity, runs the conveyor at the defined speed and measures, line by line, the light reflected by every pixel of the image for each of the 900 spectral bands or channels (each corresponding to a wavelength value) in the mentioned wavelength range. This way, each time a scenario is scanned with the system, a picture with detailed spectral information of each pixel is obtained. This is what is known as a "cube": a two-dimensional image where the third dimension is formed by channels/wavelengths.

The representation of the scanning process can be seen in Figure 2 where the yellow rays represent the light emitted by the halogen bulbs while the rainbow beam shows the light captured by the hyperspectral camera.

The hyperspectral cube resulting from the process provides the necessary data to obtain a reflectance spectrum, as the spectral information obtained is based on reflected light.

Although this technology provides a continuous representation of the materials spectra, unlike other techniques such as RGB or multispectral imaging, the extracted data is still discrete, with 900 values, each corresponding to a channel of the hyperspectral camera. However, these 900 values represent a much more detailed dataset than the one obtained through any of the other techniques, where the obtained data is more discrete (only three channels in the visible spectrum for RGB [14]). Therefore, HSI makes "analysis, identification, and separation of materials and substances more accurate" [14]).

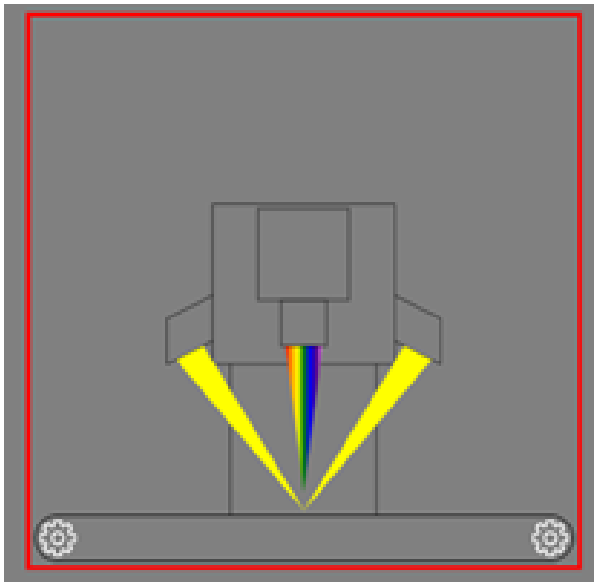


Figure 2: Graphical representation of the scanning process

The system can be connected to the computer and managed from it. To do this, it is necessary to plug in a LAN cable and configure the IP address as shown in Figure 3. Then, searching for the specific address (in this case, 10.100.10.100:8080) in the web browser, it is possible to access the camera from the computer. This also makes it possible to download the obtained hyperspectral cubes to the computer, facilitating their subsequent analysis.

Figure 3: Configuration of the IP address

3 Data

Thanks to the great amount of information and data that HSI provides, it is possible to obtain a detailed dataset that can later be used to run machine learning and train the system to detect the desired materials. However, before starting the algorithm training, it is very important to clean and structure the acquired data, as data and its quality are essential for good results. By preprocessing data, it is possible to obtain a clearer, more organized dataset, where data is easier to access and manage. This increases the efficiency of the training process and the precision and reliability of the results [17].

Using the Hyperspectral Analyzer software provided by Morten Sielnik Andersen, it is possible to select a region of the image to obtain the spectrum from it (Fig. 4) and get a graphical representation of said spectrum by plotting intensity versus channels or wavelength. For each point in the spectrum, the intensity value is a mean of the values obtained in each pixel of the selected area for that specific wavelength. As the Buteo system works with reflectance spectroscopy, intensity in these graphs represents the amount of reflected radiation. To record this intensity, an 8-bit analog to digital converter is used, meaning that the values (without units) will vary from 0 (when no light is reflected) to 255 (when the reflected intensity is maximum).

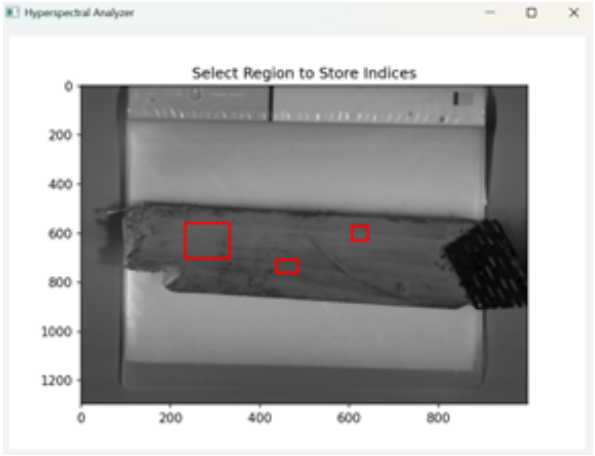


Figure 4: Selection of areas for obtaining spectra

Moreover, the software also offers the possibility of saving the acquired data in a csv file where wavelength and intensity values are stored. This way, the original unstructured data (the image that the HSI system provides) is transformed into semi-structured data (a csv file with wavelengths and their corresponding intensity).

This is a very convenient transformation because, although unstructured data is very flexible and can be easily and rapidly storage, its manipulation is complex, requiring expertise and particular tools. This data type, which represents about 80% of all enterprise-generated data [18], includes every type of data without a standardized format, from text files and sensor data to images and videos, and it is mainly used for generative AI. However, since the technology used in this project is ML, the aim is to obtain structured data, which is more appropriate for this approach as it is more easily manipulated by both humans and computers. In this context, semi-structured data

represents a "bridge" between these two types of data, having sections with a standardized format but without requiring "a predefined data model" [19] or a particular order.

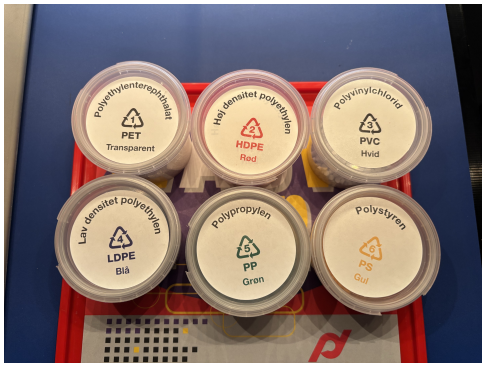
Therefore, once all the csv files are collected, each one of them is loaded into an Excel file that ends up containing all the collected spectra as tables with labels indicating the material corresponding to each spectrum.

This leads to a structured dataset, which is ideal for machine learning applications as structured data is easier to analyze, use and storage. Using this type of data with a standardized and predefined format, an organized form of storage consisting of rows and columns and clearly defined attributes [20] (in this case: wavelength, intensity and material) can be used, making it easier for the algorithm to process the data and quickly identify patterns [17]. Moreover, structured data is rather user-friendly: this type of data is more manageable and easier to understand also for humans and there is a great variety of tools to work with it [18]. Although there are also some drawbacks to structured data, these have to do mainly with the lack of flexibility (it has rigid schemes and can only be used for a specific goal) [18], thus not representing a problem for the project.

It is also important to observe that, as mentioned before, although both the Hyperspectral Analyzer software and Excel provide a continuous representation of the spectra when plotted in a graph, the extracted data is indeed discrete (900 values, each corresponding to a channel of the hyperspectral camera). This makes data more manageable and improves the model performance.

When data has been collected, it is essential to explore them [17]. By looking at the data, plotting graphs, etc. it is possible to get useful information, find missing data, study some behaviors and detect inaccuracies. This is why some graphs were plotted using the data in Excel. This exploration, facilitated by the use of a structured dataset, leads to some insights that help to get a better understanding of the dataset and finer prepare it for the machine learning process. The most relevant of these observations are presented next, while more detailed information can be found in Annex B.

Due to the need to get familiar with the HSI system, the first scans are done with plastics, from which there is already existing data collected by Patrick Vogelius and Hassibullah Noori, Physics and Technology students working on a project that involves the recollection of hyperspectral information of these materials. This way, the first data collection is done using the plastic samples shown in Figure 5, with photographs provided by Patrick Vogelius.



(a) Boxes indicating the type of plastic they contain

(b) Boxes open

Figure 5: Plastic types used in the project

By scanning these materials multiple times in different scenarios, it is possible to observe how the resulting spectrum may vary from one situation to another. To study this, the focus is mainly on two types of plastics: PVC and PET (Fig. 6). The tests show that, although the shape of the graph is always the same, intensity varies from one scan to another. For example, intensity when plastics are in the box is lower than when they are placed on the lid (Figure 7a). Moreover, scanning on different days leads to a noticeable difference in the values of intensity of the spectrum (Figure 7b). Although this could be attributed to different conditions in the lab from one day to another, scans carried out on the same day also led to different intensities even when doing a new calibration for each scan (Figure 7c), though the difference in this case is much lower.

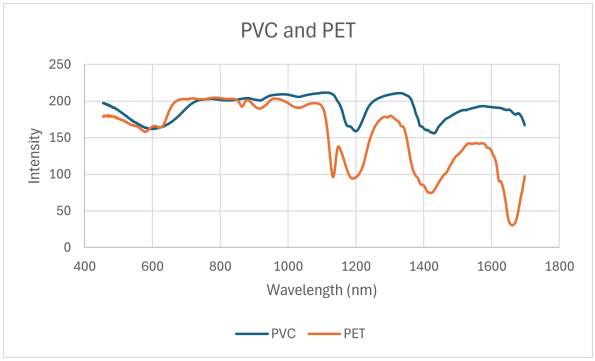
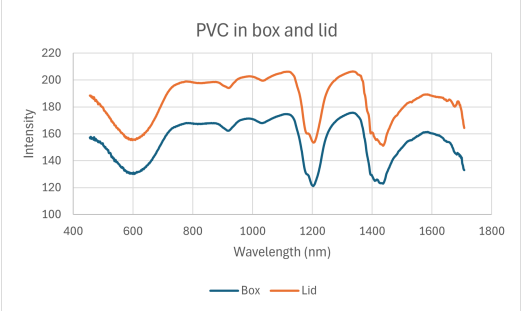
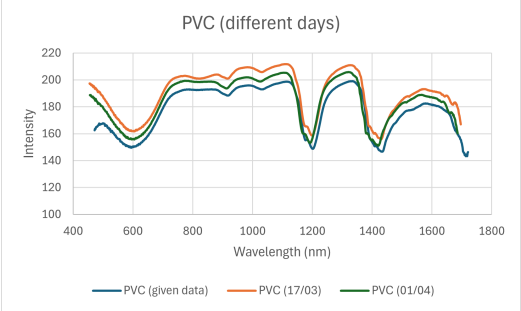


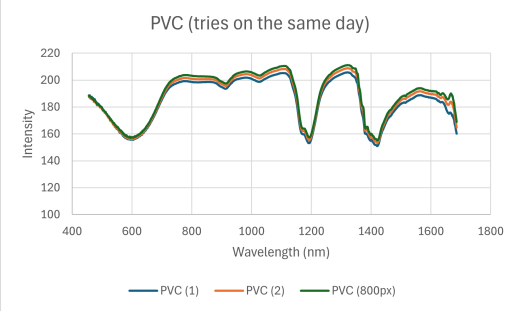
Figure 6: Spectra of PVC and PET



(a) Spectra when PVC is in the box or on the lid



(b) Spectra of PVC obtained on different days



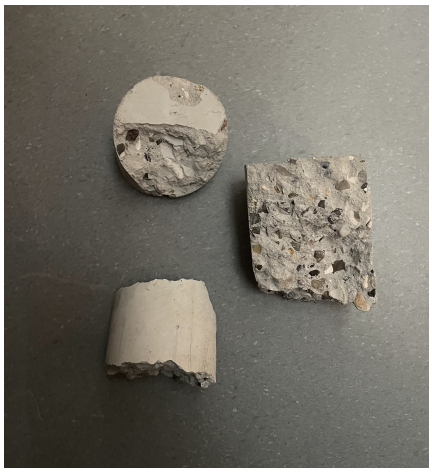
(c) Spectra of PVC obtained on the same day

Figure 7: Variations in plastic spectra

After working with plastics, several building materials provided by the Civil Engineering department will be used for the scanning and the machine learning process. The construction materials analyzed in the project are timber, steel, concrete and fiber reinforced concrete (Fig. 8).

(a) Timber

(b) Steel



(c) Concrete

(d) Reinforced concrete

Figure 8: Construction materials used in the project

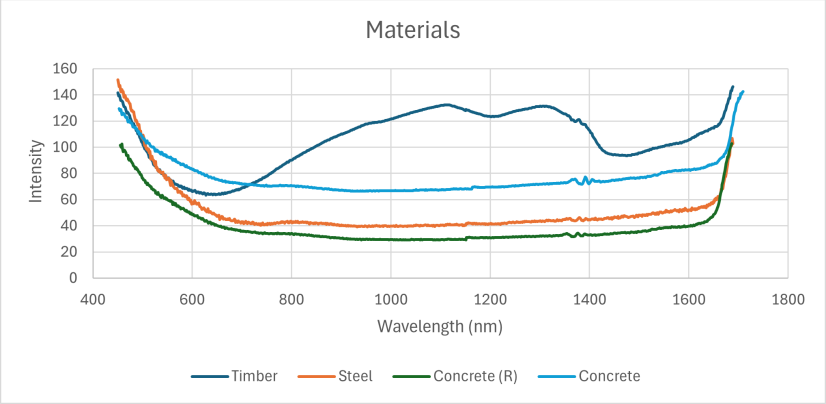
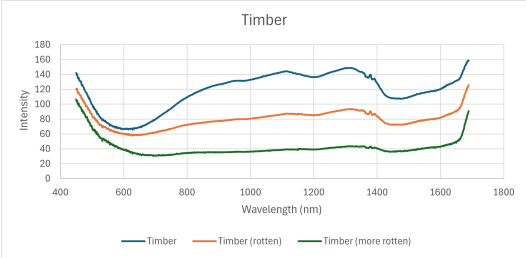


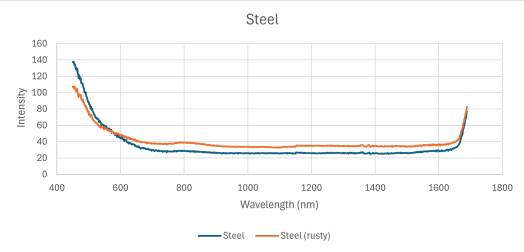
Figure 9: Spectra of the different construction materials

It can be observed in Figure 9 that, while timber has a spectrum with a very distinctive shape, the other 3 materials have more similar spectra. This suggests that, while timber will probably be easily identified by the ML algorithm, the other materials may present a challenge. Reinforced concrete is especially challenging because, having a spectrum with similar shape, its intensity is also close to that of steel. This makes sense, since this material is concrete with some added fibers which usually consist of steel, glass or polymers.

However, because spectra are defined by molecular structure, all the materials show differences in their spectrum depending on the state they are in. This can be easily seen in timber, which shows a lower reflected intensity when it is rotten (Figure 10a), and in steel, where the intensity in its spectrum also changes when the material is rusted (Figure 10b). However, as the effect of all these conditions on the intensity is mostly even across the spectrum, it only represents a vertical shift of the graph, whose shape remains essentially the same. More details about how different conditions affect the material spectra are studied in Annex B.



(a) Spectra of timber and rotten timber



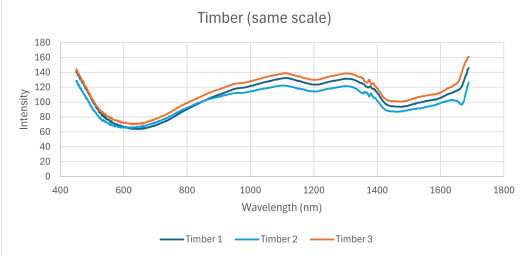
(b) Spectra of steel and rusty steel

Figure 10: Variations in spectra because of their conditions

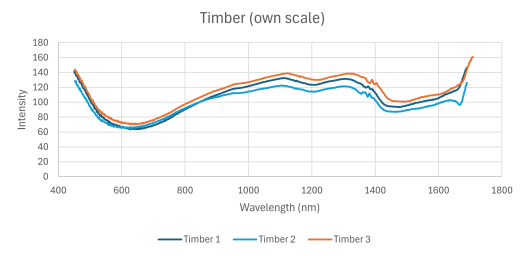
All of these are important details to take into account when taking data since, for the system to identify materials correctly regardless of their state, it is necessary to train it with samples of these materials in different conditions.

In this way, after several scans of the materials throughout several days, the data required for machine learning is obtained. This dataset contains 49 timber samples (some of them with rotten/written parts), 56 steel spectra (from steel in good condition, rusty steel and from screw-shaped parts), 46 samples for concrete and 49 for reinforced concrete.

Because wavelength values vary slightly from one day to another and the machine learning model needs only 1 wavelength column, the strategy adopted consists of selecting one of the wavelength columns and using it as the common wavelength values for all the spectra. After doing this, it is important to check that this modification does not have a significant effect on the spectra. As shown in Figure 11, this change does not affect the data drastically, making it safe to proceed.



(a) Spectra with only one wavelength scale



(b) Spectra with multiple wavelength scales

Figure 11: Variations in spectra when using common wavelength values

All the collected data is separated into training data, which will be used to train the model, and evaluation data, which will be used to check the accuracy of said model performance in new cases [21]. In this project, this division of the dataset is done in a way that 70% of all the data will be used for training while the remaining 30% will be used for evaluation.

Once again, it is important to explore and verify training data to ensure it includes multiple scenarios and forms an accurate and complete dataset, thus reducing the risk of bias.

4 Machine learning

Having collected the spectrum from a given material in multiple different occasions and having saved, organized and properly labeled the data, the next procedure is to train the machine learning model.

Machine learning is a subfield of AI that uses diverse techniques to train systems to recognize patterns, make decisions or predict outcomes based on given data. As stated by the AI pioneer Arthur Samuel in the 1950s, "it allows computers to learn without being explicitly programmed" [21]. By training a machine learning algorithm with a large amount of high-quality data, it is possible to obtain precise models [22] that can perform specific tasks with great accuracy and efficiency.

Nowadays, this technology is present in multiple industries and commercial activities, with 67% of companies already using it in 2020 [21]. With the increasing level of automation in the industrial sector, machine learning is becoming more and more important to optimize processes and make them more efficient and reliable. Thus, machine learning is already being used in this sector for inventory management, quality control and many other purposes [23].

In this case, machine learning is used for classification, trying to achieve a system that autonomously recognizes construction materials by analyzing data and recognizing patterns in their spectra. This is a type of hyperspectral image classification, a field where machine learning is already being applied and studied for new use cases. In this project, instead of analyzing the image pixel by pixel, the spectrum of a selected area is used to train the model to detect the material.

When applying machine learning, there are mainly 3 different paradigms that can be followed: supervised, unsupervised and reinforced machine learning. Supervised learning, which is the most widely used method nowadays, uses abundant annotated data to generate results. On the other hand, unsupervised learning does not require any annotation of data. The system is able to establish patterns and relations in data without human interaction. Finally, reinforced learning is based on "trial-and-error" [24] and feedback: the system tries to figure out the best output by itself, receiving positive responses when the outputs are good and hints to discard those that are not so accurate [24].

In this project, a supervised machine learning approach will be adopted, as it offers greater control over the model [24]. This method requires a good amount of annotated data to perform well, but this has already been obtained through the data preparation process (section 3).

Specifically, the supervised learning method used in this project is Support Vector Machine (SVM). This algorithm is widely used for classification, when the aim is to assign data to a specific category or discrete label. This is the purpose for which it is used in this project. However, SVM could be also used for regression applications, that is, for predicting continuous numerical values [25].

SVM models are characterized by the use of a hyperplane. In classification tasks, this hyperplane is used to define categories, and it is defined as the barrier that best separate classes. To achieve this, the algorithm uses different mathematical functions to find the maximum distance between the hyperplane and the data points closest to it. This distance is called "margin", which can be defined as hard (when none of the samples are allowed within the margins) or soft (when margins can contain some samples and even allow for the misclassification of some of them). Meanwhile, the points used to define the hyperplane are called "supporting vectors". Some of the advantages of this method are that it enables classification in multiple categories, it can manage nonlinear divisions effectively and it works well with data with a high number of features, even when the number of samples is not very big [26]. All of these aspects make SVM a good algorithm for the acquired dataset.

It is important to notice that even the same dataset can be approached in different ways depending on the desired results. Because the goal of this project is to distinguish between multiple construction materials, a supervised, multi-class classification method is applied: a great amount of labeled data is used as input to train a system that, when given new data after training, will provide a label as an output.

The algorithm is implemented in Python using the code editor Visual Studio Code. The final code can be found attached to this document (Annex A). Some observations about the most relevant aspects of it are included below.

The first remarkable point is that, to develop the code and make it work, it is necessary to use several libraries. The two main ones are "pandas" library, which is used to manage the data collected in the Excel file, and "scikit-learn", the library employ to implement machine learning in Python. Moreover, if the trained model needs to be saved for future use, the "joblib" library is required. Additionally, although this is not the case in this project, other libraries could be used, such as "matplotlib" to plot some results.

The first part of the code is dedicated to reorganizing the *Dataset* Excel file and giving it the adequate structure for machine learning (Figure 12).

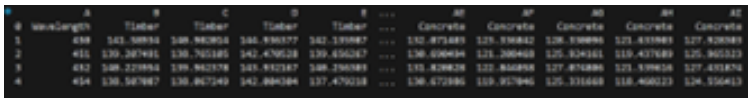
The "print(.head())" commands, when not commented out, provide an overview of the specified dataset. This is a way to check that the reorganization is done correctly.

Then, data is separated into X and y categories, with X containing the numerical intensity values of each spectrum and y including the names of each material. Afterwards, the data is scaled to adjust all the features to a specific scale and, after that, PCA can be applied to reduce the number of features. In this regard, it is important to note that all the new data given to the model in the future must undergo the same preprocessing to ensure that the results obtained are reliable.

Next, the preprocessed dataset is divided into training and test data. This is done with the "train_test_split" function, where the percentage of data used for evaluation is also specified by giving a value to "test_size". A value could also be assigned to "random_state" in this function, to obtain always the same dataset division. However, as the goal is to study how the model works with different data, no fixed value is

	A	B	C	D	...	AM	AN
0	Wavelength	Material 1	Material 1	Material 2	...	Material n	Material n
1	λ1	Sample1.Intensity1	Sample2.Intensity1	Sample3.Intensity1	...	Sample(n-1).Intensity1	Sample(n).Intensity1
2	λ2	Sample1.Intensity2	Sample2.Intensity2	Sample3.Intensity2	...	Sample(n-1).Intensity2	Sample(n).Intensity2
3	λ3	Sample1.Intensity3	Sample2.Intensity3	Sample3.Intensity3	...	Sample(n-1).Intensity3	Sample(n).Intensity3
...	...	...	...	...	...	...	...
899	λ899	Sample1.Intensity899	Sample2.Intensity899	Sample3.Intensity899	...	Sample(n-1).Intensity899	Sample(n).Intensity899
900	λ900	Sample1.Intensity900	Sample2.Intensity900	Sample3.Intensity900	...	Sample(n-1).Intensity900	Sample(n).Intensity900

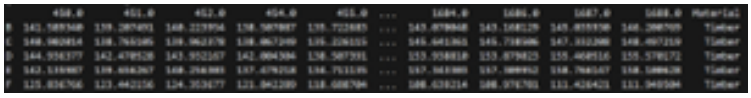
(a) Structure of the dataset in Excel



(b) Overview of the Excel dataset in Python

	λ1	λ2	λ3	...	λ899	λ900	Material
B	Sample1.Intensity1	Sample1.Intensity2	Sample1.Intensity3	...	Sample1.Intensity899	Sample1.Intensity900	Material1
C	Sample2.Intensity1	Sample2.Intensity2	Sample2.Intensity3	...	Sample2.Intensity899	Sample2.Intensity900	Material1
D	Sample3.Intensity1	Sample3.Intensity2	Sample3.Intensity3	...	Sample3.Intensity899	Sample3.Intensity900	Material2
...	...	...	...	...	...	...	...
AM	Sample(n-1).Intensity1	Sample(n-1).Intensity2	Sample(n-1).Intensity3	...	Sample(n-1).Intensity899	Sample(n-1).Intensity900	Material n
AN	Sample(n).Intensity1	Sample(n).Intensity2	Sample(n).Intensity3	...	Sample(n).Intensity899	Sample(n).Intensity900	Material n

(c) Structure of the dataset for machine learning



(d) Overview of the reorganized dataset in Python

Figure 12: Comparison of the original dataset and the one used for ML

established. In this way, it is ensured that each time the code is executed, different data will be used for training and testing.

The model is defined as SVC, which is the name of the class that uses SVM for classification [27]. This model requires some parameters to be specified. In this case, the values for the Kernel function, C and gamma are established. The Kernel function is the tool used by the algorithm to define classes. Setting the type of Kernel determines the way the model separates the different categories. In this case, a Radial Basis Function (RBF) Kernel is used, which is a nonlinear Kernel. For its part, C is a parameter that establishes the penalization of errors, therefore influencing the margin of the SVM model. With scikit-learn, C can be set to any positive value, defining a wide, soft margin when the number is low and a small, hard one when it is high. Finally, gamma is the parameter that determines how influential single training samples are, helping to define fit boundaries when gamma is high and simpler ones when it is low [26].

To find the best combination of parameters, it is also possible to use "GridSearchCV" [28]. This tool evaluates the model performance with all the possible combinations of multiple parameter values, previously defined, and obtains the best one using cross-validation. However, here, the parameters are adjusted manually, starting with the default values and analyzing the results after training the model until finding accurate ones.

After defining the parameters of the model, the method ".fit()" is used to train the model and then ".predict()" is applied to make predictions with the test data.

Finally, it is essential to evaluate the model to determine if it will produce good results when given new data. This is done through the commands on the "Evaluation of the model" code section, which provide a classification report, the confusion matrix and

clear visualization of the most important metrics. The classification report includes the main evaluation metrics for classification, which provide a lot of information about the model performance. These metrics include precision, recall and F1-score, for each class and for the whole model, plus accuracy. Precision (1) measures how many of the model predictions for each material were correct. Meanwhile, recall (2) indicates how well the model recognizes the material when it is present. In this way, F1-score (3) is defined as the harmonic mean of precision and recall [29], thus giving a better idea of how well the classification model works: how many of the assigned labels were right and how many samples were not missed. The higher the f1-score the better the model, with 1 being the maximum F1-score possible. Because this is a multi-class classification, precision, recall and F1-score are also calculated using the 'weighted' option. This makes an average of the metrics of the model for each class taking into account the number of samples per class, giving a more precise overview of the overall performance of the model (5). Lastly, accuracy (4) determines the percentage of correct results of the model [29]. Because the number of total predictions is needed to calculate this metric, accuracy can only be defined for the whole model and not for each class.

$$\text{Precision} = \frac{\text{Correct predictions of one class}}{\text{Total predictions of that class}} \tag{1}$$

$$\text{Recall} = \frac{\text{Correct predictions of one class}}{\text{Total real cases of that class}} \tag{2}$$

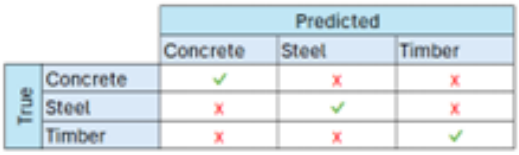
$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total samples}} \tag{4}$$

$$\text{Weighted average} = \sum(\text{Metric for class i x } \frac{\text{Number of test samples of class i}}{\text{Total number of test samples}}) \tag{5}$$

The confusion matrix shows how the samples were classified compared to their true label, making it possible to determine where the misclassifications occurred (Fig. 13).

By looking at the rows of the confusion matrix, it is possible to determine the recall of the model. On the other hand, precision can be calculated by inspecting the columns of this matrix. With the example in Figure 13b, it is possible to determine that precision is 0.67 for concrete (2 correct predictions divided by 3 samples predicted as concrete) and 1 for the other materials. In a similar way, recall can be calculated, being 1 for concrete and steel and 0.8 for timber (4 correct predictions divided by 5 actual timber samples, one misclassified as concrete). By introducing these values in the f1-score formula, it is easy to obtain the f1-score for each class. Finally, to obtain the desired weighted averages, all that is required is to sum the values of each class multiplied by



(a) Structure of the confusion matrix for 3 materials

```
Confusion matrix:
 [[2 0 0]
 [0 4 0]
 [1 0 4]]
```

(b) Example of the confusion matrix for 3 materials in Python

(c) Structure of the confusion matrix for 4 materials

```
Confusion matrix:
 [[ 7  0  2  0]
 [ 0 22  0  0]
 [ 0  2 13  0]
 [ 0  0  0 14]]
```

(d) Example of the confusion matrix for 4 materials in Python

Figure 13: Structure and examples of the confusion matrix for 3 and 4 materials

the number of samples per class (indicated in the "support" column) and divide the result by the number of categories. These are the calculations that the algorithm does to evaluate the model, and whose results are presented as in Figure 14, where the results for this specific example are shown.

```
              precision    recall  f1-score   support

    Concrete       0.67      1.00      0.80         2
       Steel       1.00      1.00      1.00         4
      Timber       1.00      0.80      0.89         5

    accuracy                           0.91        11
   macro avg       0.89      0.93      0.90        11
weighted avg       0.94      0.91      0.91        11

Confusion matrix:
 [[2 0 0]
 [0 4 0]
 [1 0 4]]
Precision:
 0.9393939393939393
Recall:
 0.9090909090909091
F1-score:
 0.9131313131313131
Accuracy: 0.9090909090909091
Real lasses: {'Steel', 'Timber', 'Concrete'}
Predicted classes: {'Steel', 'Timber', 'Concrete'}
```

Figure 14: Results of the explained three-material example

In Figure 14, it can be noticed that results match those expected. An important detail about these results is that weighted recall is equal to accuracy, as indicated in the sckit-learn documentation [30]. Moreover, it is also possible to observe that, from 11 test samples (30% of the total 34), 2 of them were concrete, 4 steel and 5 timber. Although the model will always take the specified percentage of the total number of samples (in this case, 30%), the number of samples of each material varies every time, as the split of the dataset is random.

Another classification metrics are logarithmic loss, area under the curve (AUC) and specificity. However, logarithmic loss is not used because, as it is based on probabilities [29], it is a metric better suited for probabilistic classifiers, those that provide the probability of a sample belonging to each class instead of just assigning a label to the sample as the model of this project does. AUC, which measures the discrimination ability of the model to differentiate between categories, was not applied either as it is mainly used in the evaluation of binary classifiers [29]. Specificity, which measures the capacity of the model to correctly detect negative cases, is also a metric mostly used in binary classification evaluation. In scikit-learn, it is defined as the recall of the negative class (pos_label = 0), and it can only be applied for binary classifications (average = 'binary') [31]. This is why specificity is not used in this project, where multi-class classification is applied. However, if desired, it would be possible to calculate the last two metrics using a "one-vs-all" method, where the one material chosen would be the "positive" class and all the others would be "negative". In this way, it would be possible to determine specificity and AUC for each category and then get the weighted average.

Finally, once the model is trained and shows a good performance, it can be saved. In this way, it can later be used to classify new samples without the need to go through the training process again.

5 Results

The first test consisted of training and evaluating a SVM model with a dataset of 34 samples, without reinforced concrete. In this first scenario, the data was scaled but PCA was not applied.

As a general rule, to obtain meaningful information about the model, the program is run 10 times, classification metrics are collected, and a mean of these 10 values is used to assess the model performance. The resulting mean values can be found in Tables 1 and 2. Another common factor in all the tests is that the parameters for the SVM model were kept the same, using an RBF kernel, a value of 1 for C, and gamma set to 'scale'.

This first model showed a great performance with an accuracy of 0.89, a weighted average of 0.94 for precision and 0.89 for recall, resulting in a 0.89 weighted f1-score. The confusion matrix showed that most of the misclassification involved timber, either because this material was classified as some different one or some of the other materials were wrongly identified as timber. However, these errors were still uncommon, and the model provided good results. Moreover, to study its performance on new data and prove that it could really recognize materials based on their spectra, the trained model was saved and then used to identify 4 new samples. These samples, taken from known materials (2 spectra extracted from timber, 1 spectrum from concrete and 1 from steel), were given unlabeled to the model to check if it was able to correctly classify them. This time, the algorithm showed great results, successfully distinguishing all the materials (Fig. 15).

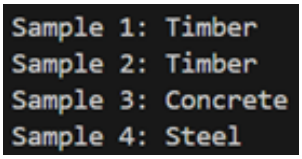


Figure 15: Identification of new materials with SVM, no PCA and 34 samples

In another case, PCA was implemented in this same model working with the same dataset (34 samples without reinforced concrete). By defining the PCA for it to maintain 95% of the original variance, the 900 original features were reduced to just 3 components. However, the model still showed great results, being able to assign the correct labels to the new samples. This was reflected in the classification metrics, which showed even higher values than for the case without PCA.

However, when the SVM model was used with a dataset that also contained reinforced concrete, some problems appeared. The main one had to do with the low number of samples to train and evaluate the model. Although it appeared as a warning, the small dataset resulted in some classes not being predicted during the evaluation process (Fig 16). This could happen because of 2 different reasons: either one material had no samples selected as test data, or the few samples taken were misclassified. In both cases, this made it impossible to correctly analyze the model performance, as the resulting metrics were not realistic. This happened regardless of whether PCA was applied or

not.

```
Warning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to cont
rol this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
Precision:
 0.5375
Real lasses: {'Steel', 'Timber', 'Reinforced', 'Concrete'}
Predicted classes: {'Steel', 'Timber', 'Concrete'}
```

Figure 16: Python warning when a class was not predicted during the test

To solve this problem, the volume of the dataset was increased. Multiple more samples were collected and added to the 39 original ones, forming the final dataset described in Section 3.

Then, the model was trained again with the new dataset, without applying PCA. This time the model presented great results, with higher accuracy, recall and F1-score than in any of the previous cases, and better precision than that shown by the same model with 34 samples. Although some steel samples (and, more rarely, timber ones) were sometimes incorrectly detected as reinforced concrete, these represented isolated cases. To make sure it worked well, the model was provided with unlabeled data taken from known materials: the previous 4 timber, steel, and concrete samples, plus 2 new ones obtained from reinforced concrete. This algorithm was able to correctly identify all of them (Fig. 17).

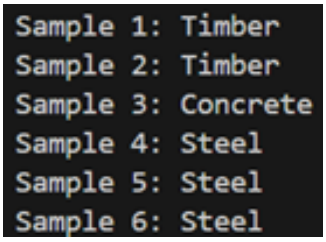
```
Sample 1: Timber
Sample 2: Timber
Sample 3: Concrete
Sample 4: Steel
Sample 5: Reinforced
Sample 6: Reinforced
```

Figure 17: Identification of new materials with SVM, no PCA and 200 samples

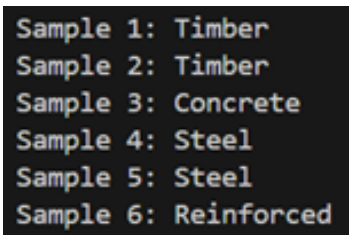
When PCA was applied to the scaled, full, big dataset, it was again set to preserve 95% of the variance and, in this way, it only took 2 components. The same SVM model was trained with this preprocessed data, and, this time, all the classification metrics decreased. Moreover, when using the trained model to predict new materials, it mistook reinforced concrete for steel most of the time (Fig. 18). This was coherent to the results obtained from the confusion matrix, where most of the misclassifications were reinforced concrete classified as steel or vice versa. Some concrete samples were also misclassified as steel and timber was sometimes wrongly detected as reinforced concrete. However, these two errors occurred less frequently, and they did not prevent the model from correctly labeling the new samples of these materials.

However, just by changing the percentage of preserved variance from 95% to 98%, which resulted in the algorithm using 3 components instead of 2, the performance of the model improved significantly. Although the metric values were still slightly lower than those of the model without PCA, this time the model was able to correctly identify all the new samples.

Additionally, a Random Forest (RF) algorithm, another supervised machine learning model for classification, was implemented. To make results comparable with those of



(a) Misclassification of the two reinforced concrete samples as steel



(b) Misclassification of one reinforced concrete sample as steel

Figure 18: Identification of new materials with SVM, PCA (95% variance) and 200 samples

the SVM model, the same classification metrics were used, and same percentages were established for the division of the dataset (30% for test and 70% for training). However, because of the way Random Forests work, neither scaling nor PCA was used. When applying RF with the 34-sample dataset, the model showed a good performance in most cases, although the metrics were slightly lower than those obtained when using the SVM model with the same dataset. Some misclassifications occurred during the evaluation process, mainly regarding timber. Nevertheless, these were rare, and, in this case, RF was also able to correctly classify the unlabeled samples. But using such a small dataset led to problems with this model too. Even though it only happened once in more than 20 runs, thus allowing the model to be trained, used and evaluated according to the initially established criteria, the situation where not all classes were predicted during the evaluation appeared again. Figure 19 shows a case where, because only 1 steel sample was randomly picked for the test dataset and it was misclassified, no steel labels were predicted, leading to unrealistic metrics.

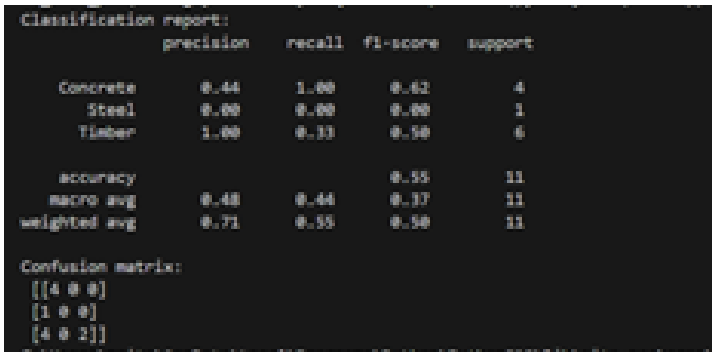


Figure 19: Results for a case in which no steel samples were predicted

However, although this complication already occurred using the dataset without reinforced concrete, the RF algorithm worked better than the SVM one with the 39-sample dataset including reinforced concrete. The non-predicted labels error kept occurring, but, while this happened to be a critical error for the SVM model, happening so often that it did not allow for a correct evaluation, the RF algorithm only presented it in 2 of more than 15 runs, making it possible to still assess its performance. Said performance was similar to that shown with the 34-sample dataset. The different metrics presented relatively high values (around 0.8 and 0.9) and, although steel was occasionally mis-

classified as reinforced concrete, the model still worked well.

Nevertheless, previous tests proved that working with more data led to better, more reliable results. This is why the RF model was also tested using the final, more complete dataset with 200 samples. Additionally, testing the algorithm with the same dataset used for the SVM model, allowed for a better comparison. In this case, the RM model presented very similar results to those of the SVM without PCA, with very high values for all the metrics and misclassifications rarely happening. This model was also able to recognize all the 6 types of material correctly when given new data.

Tables 1 and 2 present the metrics of the different models for the 34-sample dataset and the 200-sample one respectively.

Metrics	**SVM (No PCA)**	**SVM (PCA)**	**RF**
Precision	0.9377	0.9477	0.911
Recall	0.8908	0.9181	0.8454
F1-score	0.8932	0.916	0.84
Accuracy	0.8908	0.9181	0.8454

Table 1: Comparison of classification metrics for 34 samples

Metrics	**SVM (No PCA)**	**SVM (PCA 0.95)**	**SVM (PCA 0.98)**	**RF**
Precision	0.9392	0.8281	0.9256	0.935
Recall	0.925	0.795	0.9016	0.9283
F1-score	0.9256	0.799	0.905	0.9277
Accuracy	0.925	0.795	0.9016	0.9283

Table 2: Comparison of classification metrics for 200 samples

As a final observation, the time taken to train and evaluate the model was very similar in all the different cases, with all models working considerably fast.

6 Conclusions

The combined use of HSI and ML has proven to be an effective way of differentiating materials. Although in this project only 4 building materials were used, this technique allows the training of the system to detect a great number of different materials. In fact, this technology is already being used in the analysis and classification of textile fibers [32], plastics, etc. and even for other different purposes such as food quality control or the analysis of paintings [13].

The outcomes of this work show that the combined use of these two technologies can be applied in the sorting of construction materials to, for example, achieve a more efficient and autonomous recycling process. Moreover, even though it could not be explicitly proven in this project, other research, such as the one carried out by HySpex, confirms that it is possible to train a system that safely, quickly and autonomously detects asbestos [33], showing another application field for this technique. Additionally, while medical analysis and fertilized floor detection are some other examples of present use cases, research and technological advances suggest an even wider use of this technology in the future.

Some of the main constraints of this method have to do with the analyzed spectra. Two examples are the need to find the correct wavelength range for the analyzed materials, as different materials show distinctive spectral characteristics at different wavelengths [14], and the difficulties that can be encountered when trying to differentiate materials with a very similar spectrum (as occurred in this case with reinforced concrete and steel).

Moreover, there are also some challenges regarding the use of machine learning and some aspects that must be taken into account. One important point is that, to develop a good machine learning model, there is a need for a large number of diverse, good-quality samples that form a complete and representative dataset. Moreover, in most cases, data requires preprocessing before training the algorithm. Thus, appropriate methods have to be found, time and effort must be invested in preparing the data and, although preprocessing methods can usually be automated, every new sample given to the model will have to go through the same preprocessing for it to work well. Additionally, finding the best parameters for different methods can require time and effort. Not only for the machine learning model, where in the case of SVM there are some functions that help with the tuning, but also in the preprocessing methods, as experienced with the number of components for PCA. Time investment is especially notable for supervised methods, where all data must be labeled. However, a good model will make predictions quickly and efficiently, making all the initial effort worth. In this project, it took a great amount of time to acquire and label the data and to implement and train the model. Yet, once the model is developed, multiple spectra can be classified in real time.

Another important consideration regarding ML is that, because evaluating data takes time and resources, for some applications it might be better to reduce the data dimensionality with techniques like Principal Component Analysis (PCA), focusing just on the most significant information [32]. However, the SVM model obtained in this project

has shown an accurate and fast performance when given the whole collected spectrum of the materials, making it unnecessary to use these methods. Overall, the use of SVM has proven to be an efficient technique for the classification of materials according to their spectrum.

Regarding the final results, the system developed in this project, although successfully classifying construction materials, is not completely autonomous, as it requires scanned samples to be selected manually and converted into Excel files. However, the use of a structured file and the "pandas" library to work with it makes the process more automatic. The next step in creating an autonomous detection system will be finding a way for the HSI system to extract the spectra of each pixel and feed them into the model, which will classify them appropriately.

In conclusion, HSI is a powerful tool that, used together with ML, is gaining new fields of application. This promising combination of technologies, which is already being used in multiple sectors, is now showing a lot of potential, with a lot of research being done to improve its performance and applying it to new domains.

References

[1] OpenAI, "ChatGPT," *OpenAI*, 2025. https://chatgpt.com/(June 26, 2025).

[2] M. Whitmer, "Asbestos," *Asbestos.com*, 2025. https://www.asbestos.com/asbestos/(June 26, 2025).

[3] World Health Organization, "Asbestos," *World Health Organization*, 2024. https://www.who.int/news-room/fact-sheets/detail/asbestos(June 26, 2025).

[4] J. Craighead, J. Abraham, A. Churg, F. Green, J. Kleinerman, P. Pratt, T. Seemayer, V. Vallyathan, and H. Weill, "The pathology of asbestos-associated diseases of the lungs and pleural cavities: diagnostic criteria and proposed grading schema.," *Report of the Pneumoconiosis Committee of the College of American Pathologists and the National Institute for Occupational Safety and Health. Arch Pathol Lab Med.*, 1982 Oct 8;106(11):544-96. PMID: 6897166.

[5] U.S. Department of Health and Human Services: Agency for Toxic Substances and Disease Registry, "Toxicological profile for asbestos," *U.S. Department of Health and Human Services*, 2001. https://www.atsdr.cdc.gov/ToxProfiles/tp61-p.pdf(June 26, 2025).

[6] European Commission, "Final report summary - alert (portable real time detection of airborne asbestos fibres for tradespersons)," *CORDIS - European Commission*, 2014. https://cordis.europa.eu/project/id/243496/reporting(June 26, 2025).

[7] ALS, "Analytical methods - asbestos testing," *ALS*, 2025. https://www.alsglobal.se/en/environment/asbestos/analytical-methods#:~:text=The%20asbestos%20minerals%20can%20be,microscopy%20(TEM)%20are%20used(June 26, 2025).

[8] N. Gleichmann, "Sem vs tem," *Technology Networks*, 2025. https://www.technologynetworks.com/analysis/articles/sem-vs-tem-331262(June 26, 2025).

[9] airQ, "Detecting asbestos: What the air-q can do," *airQ*, 2025. https://en.air-q.com/blog/asbest-erkennen#:~:text=If%20the%20micrometer%2Dsized%20fibers,matter%20sensor%20provides%20initial%20indications(June 26, 2025).

[10] K. Wirth and A. Barth, "X-ray fluorescence (xrf)," *Geochemical Instrumentation and Analysis*, 2025. https://serc.carleton.edu/research_education/geochemsheets/techniques/XRF.html(June 26, 2025).

[11] HySpex, "ndustry leading hyperspectral solutions," *HySpex*, 2025. https://www.hyspex.com/(June 26, 2025).

[12] U. H. S. Agency, "Asbestos: toxicological overview," *GOV.UK*, 2025. https://www.gov.uk/government/publications/

asbestos-properties-incident-management-and-toxicology/asbestos-toxicological-overview(June 26, 2025).

[13] Newtec, "Pushbroom hyperspectral imaging system," *Newtec*, 2025. https://www.newtec.com/maskiner/kvalitets-kontrol/buteo(June 26, 2025).

[14] Specim, "What is hyperspectral imaging: A comprehensive guide," *Specim*, 2025. https://www.specim.com/technology/what-is-hyperspectral-imaging/#:~:text=Hyperspectral%20imaging%20is%20a%20technique,analyzing%20their%20unique%20spectral%20signatures(June 26, 2025).

[15] A. Bhargava, A. Sachdeva, K. Sharma, M. H. Alsharif, P. Uthansakul, and M. Uthansakul, "Hyperspectral imaging and its applications: A review," *Heliyon*, 2024. https://www.sciencedirect.com/science/article/pii/S2405844024092399(June 26, 2025).

[16] J. Huang, H. He, R. Lv, G. Zhang, Z. Zhou, and X. Wang, "Non-destructive detection and classification of textile fibres based on hyperspectral imaging and 1d-cnn," *Analytica Chimica Acta*, 2022. https://www.sciencedirect.com/science/article/abs/pii/S0003267022008091(June 26, 2025).

[17] F. Khan, "¿qué es el preprocesamiento de datos? definición, conceptos, importancia, herramientas," *Astera*, 2025. https://www.astera.com/es/type/blog/data-preprocessing/(June 26, 2025).

[18] A. Corbo, "What are data structures?," *built in*, 2025. https://builtin.com/data-science/data-structures#:~:text=Types%20of%20Data%20Structures,structures%20and%20graph%20data%20structures(June 26, 2025).

[19] A. Jonker and A. Gomstyn, "Structured vs. unstructured data: What's the difference?," *IBM*, 2025. https://www.ibm.com/think/topics/structured-vs-unstructured-data(June 26, 2025).

[20] Amazon Web Services (AWS), "¿qué son los datos estructurados?," *Amazon Web Services (AWS)*, 2025. https://aws.amazon.com/what-is/structured-data/(June 26, 2025).

[21] S. Brown, "Machine learning, explained," *MIT Sloan School of Management*, 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained(June 26, 2025).

[22] Google, "¿qué es el aprendizaje automático (aa)?," *Google Cloud*, 2025. https://cloud.google.com/learn/what-is-machine-learning?hl=es-419(June 26, 2025).

[23] Hewlett Packard Enterprise, "¿qué es el aprendizaje automático?," *Hewlett Packard Enterprise*, 2025. https://www.hpe.com/lamerica/es/what-is/machine-learning.html#:~:text=Los%20cuatro%20modelos%20principales%20de,y%20el%20aprendizaje%20de%20refuerzo.(June 26, 2025).

[24] Sigma AI, "Training data for machine learning: here's how it works," *Sigma AI*, 2025. https://sigma.ai/understanding-data-side-of-machine-learning/(June 26, 2025).

[25] GeeksforGeeks, "Machine learning tutorial," *GeeksforGeeks*, 2025. https://www.geeksforgeeks.org/machine-learning/(June 26, 2025).

[26] GeeksforGeeks, "Support vector machine (svm) algorithm," *GeeksforGeeks*, 2025. https://www.geeksforgeeks.org/support-vector-machine-algorithm/(June 26, 2025).

[27] scikit-learn, "Support vector machines," *scikit-learn*, 2025. https://scikit-learn.org/stable/modules/svm.html(June 26, 2025).

[28] scikit-learn, "Gridsearchcv," *scikit-learn*, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html(June 26, 2025).

[29] GeeksforGeeks, "Evaluation metrics in machine learning," *GeeksforGeeks*, 2025. https://www.geeksforgeeks.org/metrics-for-machine-learning-model/(June 26, 2025).

[30] scikit-learn, "recall_score," *scikit-learn*, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html(June 26, 2025).

[31] scikit-learn, "classification_report," *scikit-learn*, 2025. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html(June 26, 2025).

[32] specim, "Hyperspectral imaging reducing textile waste," *specim*, 2020. https://www.specim.com/hyperspectral-imaging-reducing-textile-waste/#:~:text=%E2%80%9CHyperspectral%20NIR%20image%20processing%20systems,synthetic%20fibers%2C%E2%80%9D%20explains%20Herrala. (June 26, 2025).

[33] HySpex, "Detection of asbestos: Classifying mixed demolition and renovation waste materials," *HySpex*, 2025. https://www.hyspex.com/use-cases-application-notes/asbestos/(June 26, 2025).

A Annex A - Python code

Code for SVM:

The commented-out (#) instructions are the ones related to PCA.

```
import sklearn
import pandas as pd
import joblib

from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix,
precision_score, recall_score, f1_score, accuracy_score
from sklearn.preprocessing import StandardScaler
# from sklearn.decomposition import PCA

# Load and prepare data
df = pd.read_excel("Dataset.xlsx", sheet_name="Data")
print(df.head())

# Extract wavelengths
wv = df["A"].iloc[1:].astype(float).values
spectranmat = df.drop(columns=["A"])
print(spectranmat.head())

# Extract material class
spectra = spectranmat.drop(0).astype(float)

# Transpose: intensity values in rows instead of columns (ML format)
spectra = spectra.transpose()
spectra.columns = wv

#Get material labels
labels = spectranmat.iloc[0].to_list()
spectra["Material"] = labels
print(spectra.head())

# Separation X and y
X = spectra.drop("Material", axis=1).values
y = spectra["Material"].values

# Scale data (and apply PCA if wanted)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# pca = PCA(n_components=0.98)
```

```
# X_pca = pca.fit_transform(X_scaled)
# print("Used componentes:", pca.n_components_)


# Division train/test data (Change X_Scaled for X_pca if PCA is used)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
test_size=0.3)

# Training SVM
model = SVC(kernel='rbf', C=1.0, gamma='scale')
model.fit(X_train, y_train)

# Evaluation of the model
y_pred = model.predict(X_test)
print("Classification report:\n", classification_report(y_test, y_pred))
print("Confusion matrix:\n", confusion_matrix(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred, average="weighted"))
print("Recall:", recall_score(y_test, y_pred, average="weighted"))
print("F1-score:", f1_score(y_test, y_pred, average="weighted"))
print("Accuracy:", accuracy_score(y_test, y_pred))

print("Real classes:", set(y_test))
print("Predicted classes:", set(y_pred))

#Save scaler and model (and PCA if used)
joblib.dump(model, "svm_model.pkl")
joblib.dump(scaler, "scaler.pkl")
# joblib.dump(pca, "pca.pkl")

print("Saved")
```

Code for RF:

```
import sklearn
import pandas as pd
import joblib

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score, precision_score, recall_score, f1_score

# Load and prepare data
df = pd.read_excel("Dataset.xlsx", sheet_name="Data")

print(df.head())

# Extract wavelengths
wv = df["A"].iloc[1:].astype(float).values
spectranmat = df.drop(columns=["A"])

print(spectranmat.head())

# Extract material class
spectra = spectranmat.drop(0).astype(float)

# Transpose: intensity values in rows instead of columns (ML format)
spectra = spectra.transpose()
spectra.columns = wv

#Get material labels
labels = spectranmat.iloc[0].to_list()
spectra["Material"] = labels

print(spectra.head())

# Separation X and y
X = spectra.drop("Material", axis=1).values
y = spectra["Material"].values

# Division in training and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

# Train model
model = RandomForestClassifier()
model.fit(X_train, y_train)

# Evaluate model
```

```
y_pred = model.predict(X_test)
print("Classification report:\n", classification_report(y_test, y_pred))
print("Confusion matrix:\n", confusion_matrix(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred, average="weighted"))
print("Recall:", recall_score(y_test, y_pred, average="weighted"))
print("F1-score:", f1_score(y_test, y_pred, average="weighted"))
print("Accuracy:", accuracy_score(y_test, y_pred))

print("Real lasses:", set(y_test))
print("Predicted classes:", set(y_pred))

joblib.dump(model, "svm_model.pkl")

print("Saved")
```

Code for making predictions on new data:

The commented-out (#) instructions are the ones related to PCA.

```
import joblib
import pandas as pd

model = joblib.load("svm_model.pkl")
scaler = joblib.load("scaler.pkl")
# pca = joblib.load("pca.pkl")

# Load data
data = pd.read_excel("Dataset.xlsx", sheet_name="NewPredictions(R)")

# Prepare it for machie learning
wv = data["Wavelength"]
new_spectra = data.drop(columns=["Wavelength"]).astype(float)
spectra = new_spectra.transpose()
print(spectra.head())
X_scaled = scaler.transform(spectra.values)
# X_pca = pca.transform(X_scaled) # => use X_pca for the model

# Let the model predict and show label
material = model.predict(X_scaled)
for nombre_col, pred in zip(spectra.index, material):
    print(f"{nombre_col}: {pred}")
```

B Annex B - Analysis of material spectra

4 construction materials were analyzed in this project. This annex provides a more in-depth study of these materials and their spectra.

In addition to the variations in rotten timber spectrum, mentioned in Section 3, this material also presents different intensities when it is peeled or has something written on it. Moreover, two different pieces of timber were analyzed to see if they presented significant disparities. As shown in Figure 20, the difference between the spectra of the two pieces is small, and the increase in intensity of peeled timber is also relatively low. The main variation in the spectra is found when the studied piece has something written on it. Then, the intensity of the spectrum decreases significantly as a result of the black ink added to the timber. However, the shape of the spectrum is still similar to those of the other cases, suggesting that it is still recognizable as timber.

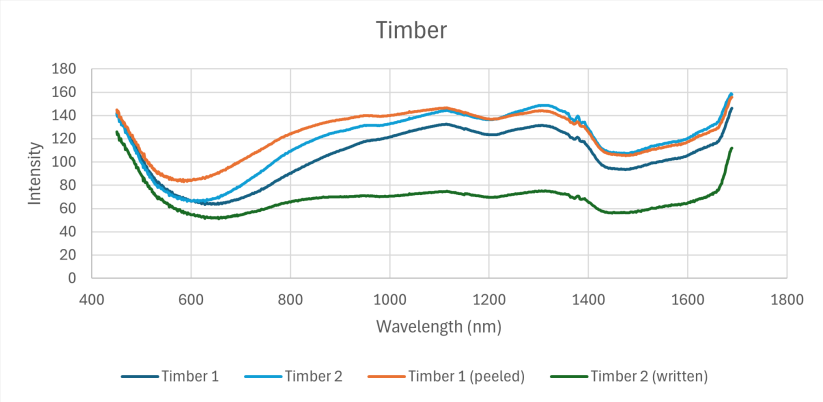
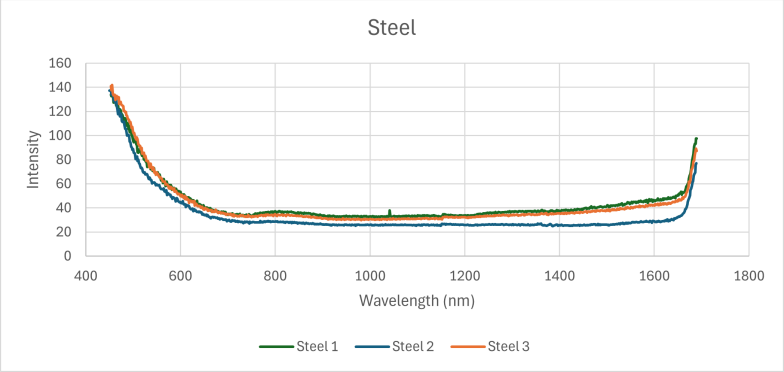


Figure 20: Spectra of timber in different conditions

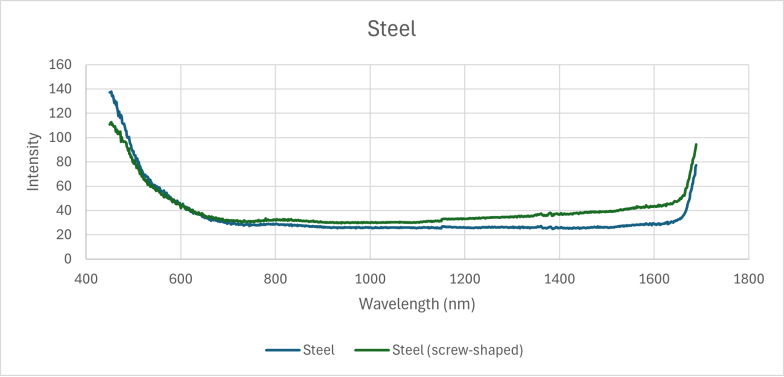
The changes in the spectrum of steel due to rust were also mentioned in Section 3, where it was shown that rust increased the amount of reflected light. However, other factors such as shape can also influence the spectrum. Again, several bars were analyzed to study the possible differences between them. Figure 21a illustrates that there is a minor difference between the spectra of the bars. This variation is especially small between bars 1 and 3, which are the ones taped together. Meanwhile, Figure 21b shows the effects of shape on the spectrum by focusing on one of the bars that has a section indented with a screw-shaped pattern. It can be observed that the spectrum from this section is similar to that of the non-indented part, only less intense. This means that shape does not represent a problem when identifying the material.

There are multiple types of concrete materials, which can make it difficult to detect and differentiate all of them. In this case, only concrete and fiber reinforced concrete were analyzed.

Concrete was provided in pieces. This allowed the study of the material from the outside, where the material was worked and even; the inside, where the aggregates could be clearly seen; and in sections where the even layer was broken, showing both internal and external parts.



(a) Spectra of different steel bars



(b) Spectra of steel bars with different shape

Figure 21: Variations in steel spectra because of its conditions

As can be seen in Figure 22, the spectra are essentially the same in these 3 different conditions. This makes sense because, in all cases, the material is the same. The slight differences in the intensity of the spectra, which do not follow a specific pattern depending on the scanned area, may be due to the equipment and the small variations it shows from one scan to another.

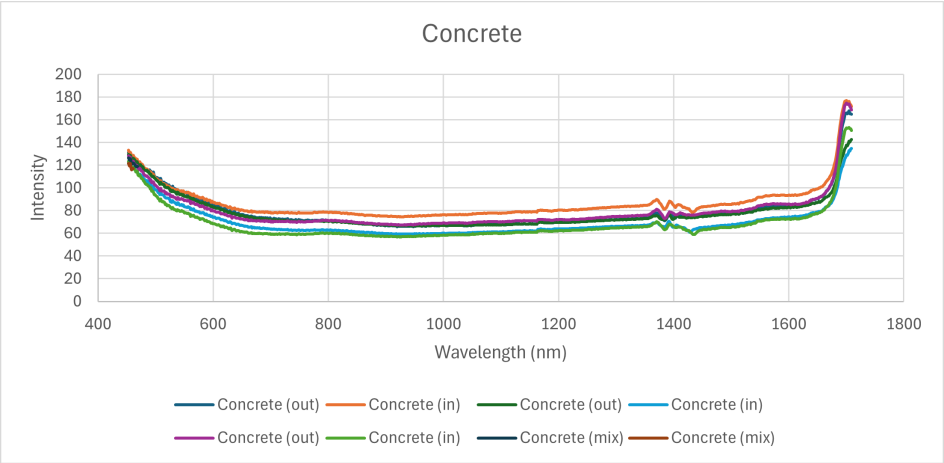


Figure 22: Spectra of different pieces of concrete from different sides

Meanwhile, reinforced concrete, which was provided as one solid block, could only be

examined from the outside. In this case, whether the selected area is a full region or a zone with small holes, there is almost no difference between spectra (Fig. 23).

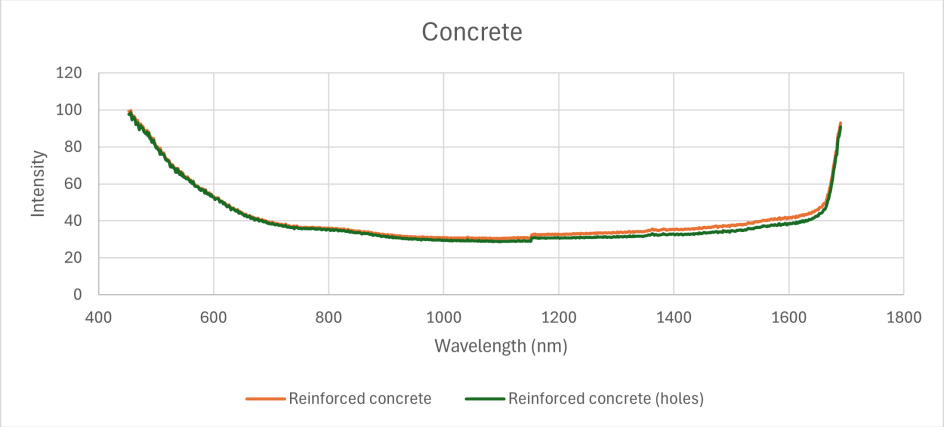


Figure 23: Reinforced concrete spectra

When comparing the two types of concrete, it can be observed that, although both of their spectra have a very similar shape, the intensity of concrete reflectance is considerably higher (Fig. 24). This means that these two types of concrete can be differentiated without much trouble.

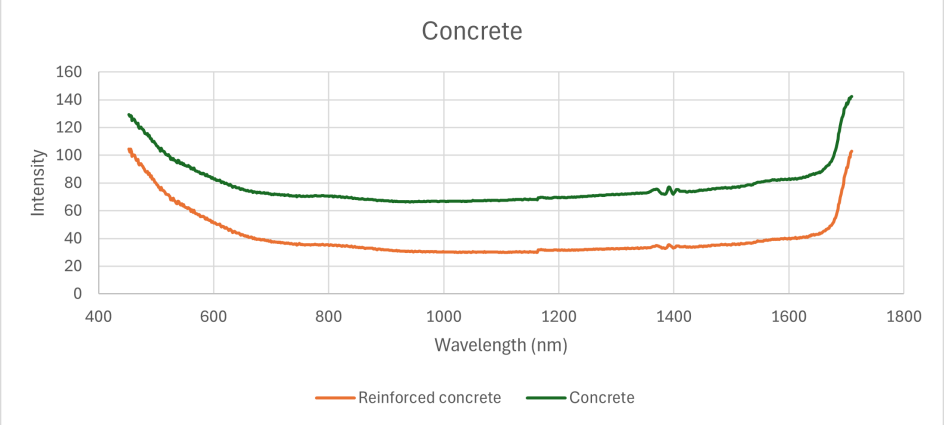


Figure 24: Concrete and reinforced concrete spectra

As mentioned in Section 3, the main problem can come from differentiating reinforced concrete and steel. This is because, as shown in Figure 25, their spectra are quite similar in both shape and intensity.

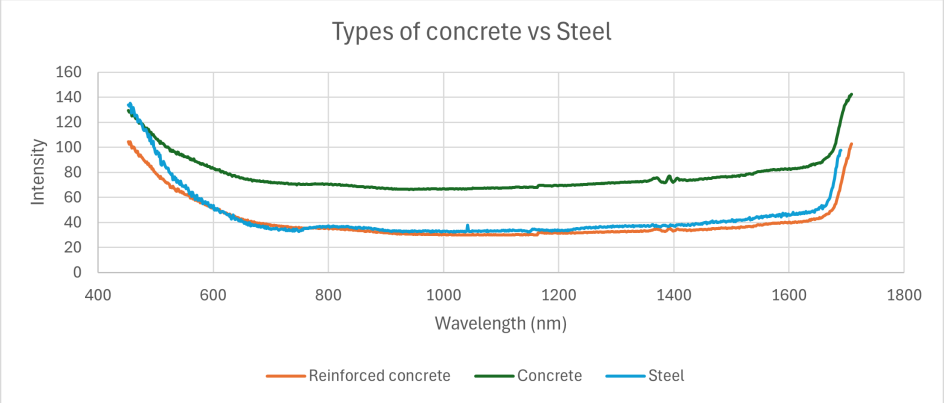


Figure 25: Spectra of different types of concrete vs steel spectra