THE HAGUE UNIVERSITY OF APPLIED SCIENCES

Photonics Research Group

Plastic Detection

Optimizing Plastic Sorting with Handheld NIR Devices through Machine Learning

Master's Thesis

Fahimeh Pourmohammadi
(23100060)

Next Level of Engineering

Faculty of Technology, Innovation and Society

Supervisor: Hedde Van Hoorn
Coach: Caroline Mok Kai Rine

June, 2024

Abstract

Plastic pollution presents a significant global challenge, especially in regions lacking adequate waste management infrastructure. Accurate plastic sorting is crucial for effective recycling efforts. This research project, is a part of a bigger project defined with the photonic research group, which aims to contribute to the development of handheld, affordable, and accurate devices for detecting plastic types. We evaluated three devices: the Plastic Scanner, SpectraPod, and NIR Spectrometer, exploring various machine learning and preprocessing methods to compare their accuracy.

Our findings revealed that the NIR Spectrometer achieved an impressive accuracy of 0.90, whereas the SpectraPod and Plastic Scanner reached 0.74 and 0.56, respectively. By consolidating HDPE and LDPE into a single category and introducing an "Unknown" category for flat spectra, accuracies improved to 1.0 for NIR, 0.85 for SpectraPod, and 0.58 for Plastic Scanner. However, despite employing various data science techniques, we couldn't elevate the accuracy of the Plastic Scanner and SpectraPod to the desired 0.95 threshold, underscoring the importance of understanding the limitations of these devices.

To investigate these limitations, we applied feature selection methods to spectra collected by the NIR Spectrometer, which served as an accurate reference. Our analysis identified significant wavelengths around 1700 nm, where both the Plastic Scanner and SpectraPod exhibit no responsiveness. Furthermore, we discovered that the ability to detect narrow bands of wavelengths significantly contributes to achieving higher accuracy, whereas the Plastic Scanner provides output averaged over broad wavelengths. Although the SpectraPod offers higher resolution, its output still relies on averaging in certain parts of the reflected spectra.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Plastic pollution is a global challenge that affects ecosystems, wildlife, and human well-being [1]. Plastic production has increased twenty-fold in the last 50 years, and globally, about 9,200 million tons (Mt) of plastic have been produced, and more than 6,900 tons have ended up to landfills or contributed to environmental pollution [2, 3]. In 2019 alone, global plastic production reached 368 million tonnes (mt), which is expected to double in the next 20 years [1]. Due to its slow degradation and unsustainable production, use and disposal, plastic pollution has emerged as a significant cross-border threat to natural ecosystems, human health and sustainability [4, 5]. Increasing evidence indicates potential risks to human health through the ingestion of plastics present in agricultural soils and aquatic life consumed as food [6, 7]. Recent studies have clearly shown that the entire life cycle of plastic contributes to climate change and biodiversity loss [8, 9]. As a result, the importance of recycling plastic waste is increasing.

However, Only about 9% of plastic waste has ever been recycled globally, 12% has been incinerated and a remarkable 79% has accumulated in natural ecosystems [2]. Alarming predictions by Borrelle et al. [10] suggest that plastic waste entering aquatic ecosystems, estimated at 19 to 23 million tons globally in 2016, is expected to increase to 53 million tons per year by 2030. Meanwhile, inadequate plastic management, especially in low- and middle-income countries (LMICs), worsens the problem. Effective sorting of plastics is essential for recycling, but the lack of proper sorting methods, especially in these countries, contributes significantly to persistent plastic pollution [11]. The Figure 1.1 shows the percentage of the recycled plastic in 2016 per countries.

To produce valuable, high-quality products from recycled plastic, sorting based on plastic type is crucial, as different plastics have varying melting temperatures, and failure to sort them results in unknown, mixed material properties, and partially burned or degraded plastic.

Various methods have been suggested for sorting plastic for recycling. Lim and Chan [12] propose thermal-adhesion sorting, controlling surface temperature to exploit softening temperature differences among plastics. Bauer et al. [13] suggest float-sink sorting, determining plastic buoyancy in a liquid solution. However, it struggles with small specific gravity differences. Howell [14] introduces dry zig-zag sorting, using air flow to separate plastics by specific gravity. Tilmatine et al. [15] propose electrostatic sorting, charging plastic flakes for separation. While these methods effectively preprocess plastic waste, they have limitations, such as specific gravity constraints. These traditional methods rely on physical properties like density and conductivity, lacking a feedback mechanism for constant quality monitoring, hindering plastic traceability for recycling [16].

In contrast, chemometric methods utilize chemical data from spectroscopy methods for automatically sorting plastic waste. Chemometrics, extensively employed in quality control within the food [17] and phar-

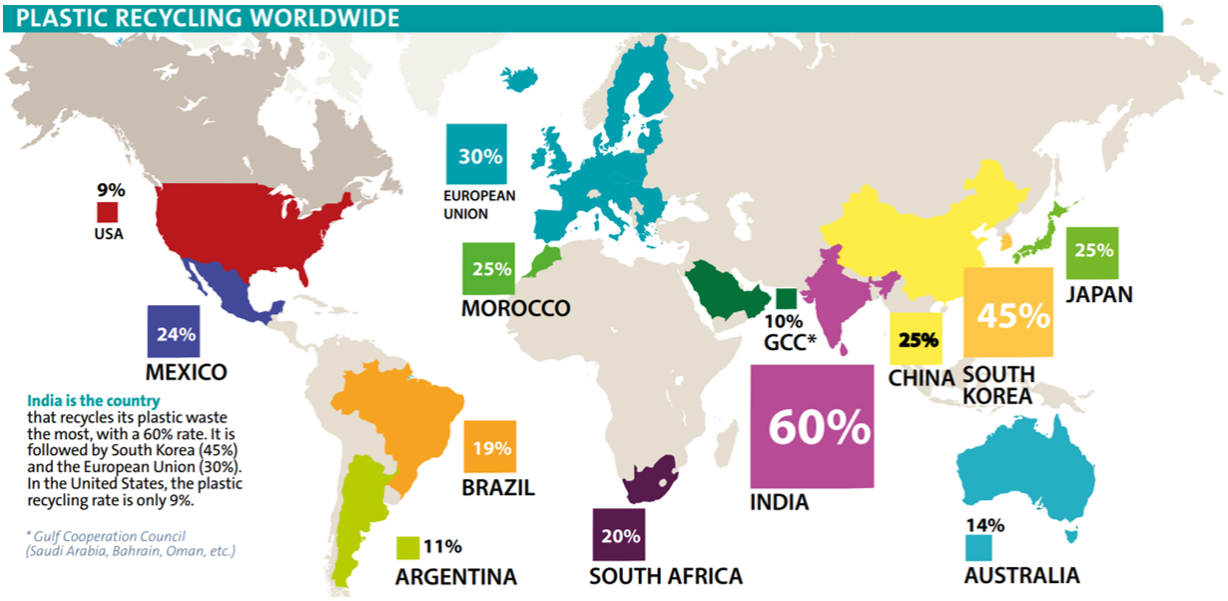


Figure 1.1: percentage of recycled plastic in each country in 2016. Plastic packaging rates in Europe – which ranked 3rd – vary from 26% to 52%.

maceutical industries [18], environmental modeling [19], and forensics [20], has recently gained popularity in addressing plastic waste challenges. Chemometric techniques utilize spectroscopy analysis to examine how electromagnetic radiation interacts with the molecules of a substance. While traditional methods depend on physical characteristics like density and conductivity, chemometric methods offer enhanced precision and reliability through the detailed chemical analysis of plastics. This higher accuracy is essential for producing high-quality recycled plastic products. Additionally, chemometric techniques often include feedback mechanisms for continuous quality monitoring, which is crucial for maintaining plastic traceability throughout the recycling process. By ensuring consistent and accurate sorting, chemometric methods not only improve the efficiency of plastic recycling but also contribute to the production of superior recycled materials, addressing the limitations of traditional methods.

Spectroscopy is a scientific method that investigates how materials interact with electromagnetic radiation across different wavelengths. Among spectroscopy techniques, Near-Infrared Spectroscopy (NIRS) is a well-established method that analyzes material properties by their interaction with electromagnetic radiation in the 700–2500 nm wavelength range[21]. It has found applications in various fields, from monitoring industrial processes to evaluating the chemical composition of products. NIRS is advantageous due to its speed, non-destructive nature, minimal sample preparation requirements, and ability to provide information on multiple components simultaneously. However, traditional benchtop NIR spectrometers are characterized by their large size, high cost, complexity, and sensitivity to vibrations due to moving parts. Thus, there is a growing interest in miniaturized, robust, and low-cost NIR sensors[22]. In the case of plastic sorting, it is essential to expand their application beyond dedicated stations in industrial settings and analytical labs into the hands of non-specialists working on-site, making them affordable and usable in any location, including low-income countries.

Technological advances in photonics and fabrication have enabled cost and size reduction [23]. The design of portable NIR spectrometers is mostly inspired by conventional benchtop instruments [11, 22]. This research project is a part of the project defined by the Photonics Research Group focusing on improving plastic type detection accuracy while also addressing the challenges related to the size, cost, and complexity

of the devices. In this thesis, we use the term ”plastic type detection” instead of ”plastic sorting” to emphasize the focus on accurately identifying specific types of plastics. Currently, two plastic type detection devices, the ”Plastic Scanner” and the ”SpectraPod,” which are acceptable in terms of size and portability, are available in the photonics laboratory. However, their accuracy and limitations require thorough evaluation. This thesis focuses on utilizing machine learning and data science techniques to assess and enhance the performance accuracy of these devices.

To achieve this, machine learning models have been trained for classifying plastic types. By comparing and selecting the best models, the accuracy of the devices has been estimated. Additionally, this accuracy has been compared to that of a benchtop NIR spectrometer, which is also available in the photonics lab and serves as a high-accuracy reference device. Further data science techniques have been employed to identify critical wavelengths and properties that contribute to higher accuracy. This investigation helps identify the limitations and challenges that lead to lower accuracy in the two portable devices. The findings can contribute to the enhancement of these devices or the design of new, higher-accuracy instruments.

Figure 1.2: a) NIR Spectrometer. b) SpectraPod. c) Plastic Scanner

Based on all of these considerations, the objective of this research is to systematically compare and enhance the accuracy of three plastic detection devices: Plastic Scanner, SpectraPod, and NIR Spectrometer, utilizing data science approaches. This includes steps such as sampling, preprocessing, feature selection, model training, evaluation, and validation.

The research question and sub-questions that arise are as follows:

Research Question:
Can data science and machine learning methods be applied to achieve accuracy exceeding 95% for Plastic Scanner, SpectraPod, and NIR Spectrometer in classifying plastic types, while comparing their performance and identifying challenges and limitations in the plastic detection process?

Sub-Questions:

1. What are the best preprocessing methods and ML models to classify plastic type for each of the three devices: Plastic Scanner, SpectraPod, and NIR Spectrometer?
2. What are the accuracy levels of plastic type classification for each of the three methods?
3. What are the important wavelengths crucial for improved plastic type classification?
4. What challenges and limitations exist in achieving highly accurate plastic type classification with handheld devices?

Modifying the software and hardware of the devices falls outside the scope of this project. Data were trained and tested on a GPU server rather than directly on the devices' boards. Implementing the trained models onto the devices' boards is not within the scope of this project. Additionally, it should be noted that, in addition to this report, a documented Python notebook will be delivered to the Photonic group, our project's client

1.1 Thesis Structure

Chapter 2, the Theory chapter, delves into the fundamental theories and concepts of photonics and spectroscopy. Additionally, it discusses data science methods that can be employed for the analysis and classification of spectral data. This chapter serves to provide a comprehensive background on the subject matter. Furthermore, it presents a detailed overview of the three devices employed in this study.

Chapter 3 focuses on the methodology employed in this research. It elaborates on the data collection process and the methodologies utilized in this study. Additionally, it provides insights into the approaches used for data analysis and classification.

Chapter 4, the Results chapter, presents the findings obtained from the experiments and analysis. This chapter aims to provide a thorough analysis and interpretation of the results obtained from the data analysis techniques discussed in Chapter 3, while also suggesting potential enhancements or modifications to Plastic Scanner and SpectraPod based on the identified limitations and areas for improvement.

Lastly, Chapter 5, the Conclusion chapter, summarizes the key findings of the study.

Chapter 2

Theory

When sorting plastic, the plastic recycling codes can be checked located on the plastic material. Plastic recycling codes, also known as resin identification codes, are symbols on plastic products that identify the type of plastic resin used, aiding in sorting for recycling. These codes, ranging from 1 to 7 and usually enclosed in a triangle of arrows, include the six common types; PET (Polyethylene Terephthalate), HDPE (High-Density Polyethylene), PVC (Polyvinyl Chloride), LDPE (Low-Density Polyethylene), PP (Polypropylene), PS (Polystyrene); and an additional category, 7, which is not commonly recycled due to mixed plastics [24]. Figure 1 illustrates these codes and their common product uses.

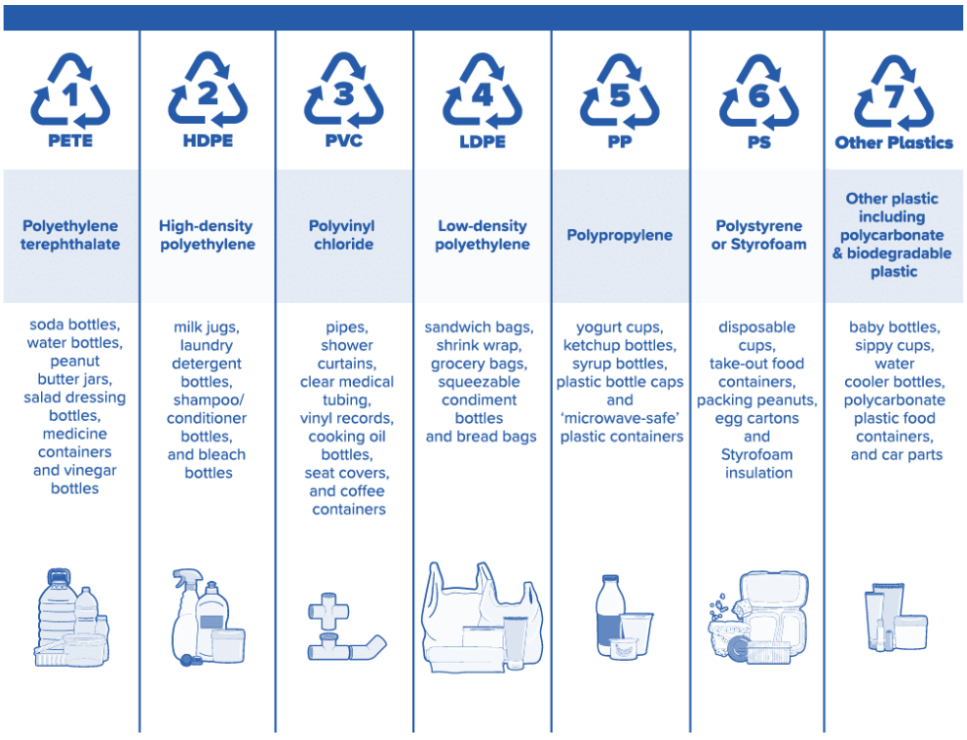


Figure 2.1: Plastic recycling codes and their common product uses [24]

2.1 Molecular structure of plastics

Plastics are made of polymers, which consist of multiple repeating units called monomers bonded together. The molecular structures of these polymers can experience internal vibrations. Depending on the bonding between the atoms in the molecule, the molecule can vibrate in different directions. Each vibration creates absorption bands that form an absorption spectrum. Figure 2.2 shows the monomer structure of the five type of plastic: PET, HDPE, PVC, PP, PS.

The C-H, O-H, N-H, and C-O bonds are particularly important for characterizing specific types of plastic by their absorbance in the NIR spectrum [25]. The absorption bands in Figure 2.3 represent the fundamental normal modes.

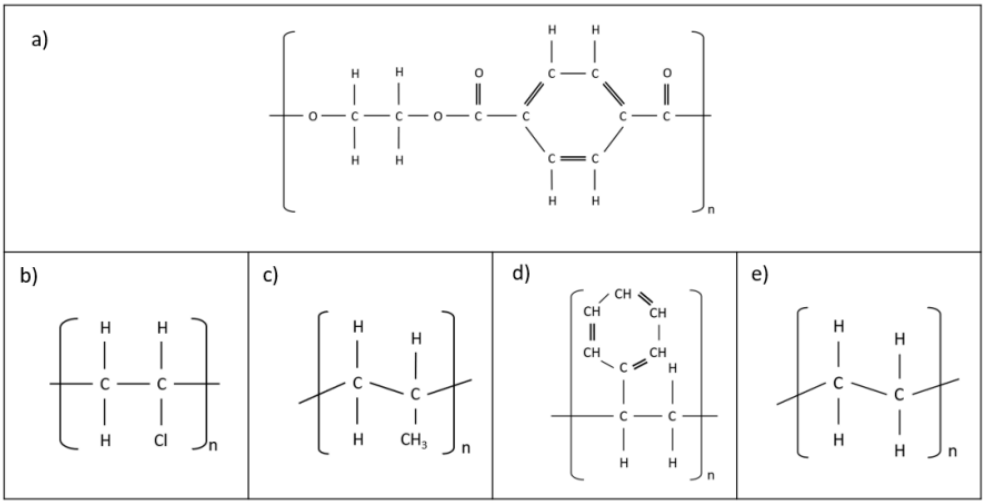


Figure 2.2: The monomer structures of the five most common types of plastic. a) PET, b) PVC, c) PP, d) PS and e) HDPE.[25]

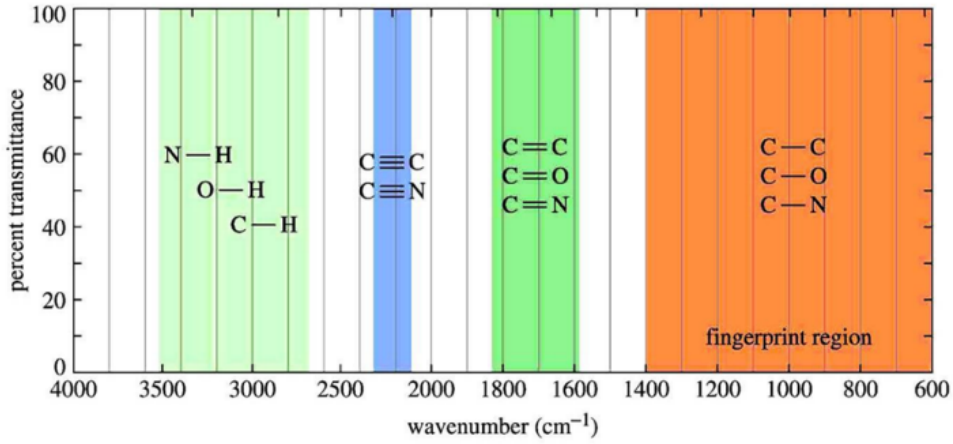


Figure 2.3: The absorption bands of specific C-bonds at a wavenumber range in the IR spectrum [25]

When monomers are joined together to form polymers, the vibrations within the molecules weaken. This weakening of vibrations results in a shift of the absorption bands associated with the molecular vibrations to the NIR range. The NIR spectrum ranges from 700 nm to 2.5 µm. These wavelengths are relatively high,

so they are often expressed in wavenumbers. The wavenumber (σ) is a measure of radiation in waves per centimeter:

$$\sigma = \frac{1}{\lambda} \tag{2.1}$$

In equation 2.1, λ is wavelength in cm. The convenience of using a wavenumber scale lies in its linearity with respect to energy (Figure 2.4).

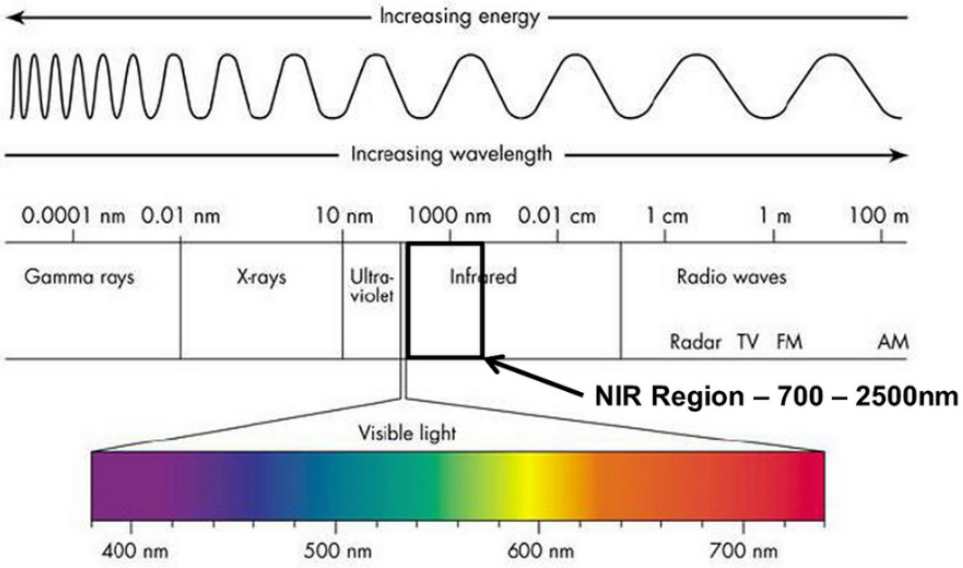


Figure 2.4: The NIR Region

2.2 Spectrum and Reflectance

Each type of plastic has multiple C-H, C-O, and other bonds which form an absorption spectrum. The different spectra produced by the molecules of the plastics can be used for identification through spectroscopy. When electromagnetic radiation is directed at a material, it can be absorbed, reflected, transmitted, or scattered [26] (Figure 2.5 illustrates these phenomena) depending on the frequency of its normal modes, which are specific patterns of vibrational motion within a molecule. The measurement of electromagnetic radiation absorbed or emitted by atoms, or molecules as they undergo state changes is called spectrometry. Figure 2.6 shows how the reflectance spectra of different types of plastic vary, with different peaks and deeps appearing at different wavelengths. The spectra are shown with an offset to clearly distinguish between the types of plastic [27].

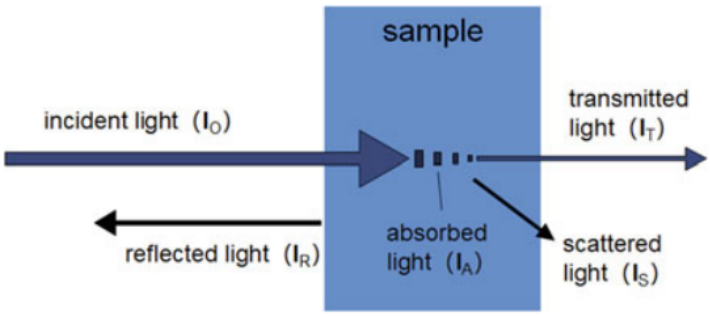


Figure 2.5: Schematic diagram of the interaction of electromagnetic radiation with matter [26]

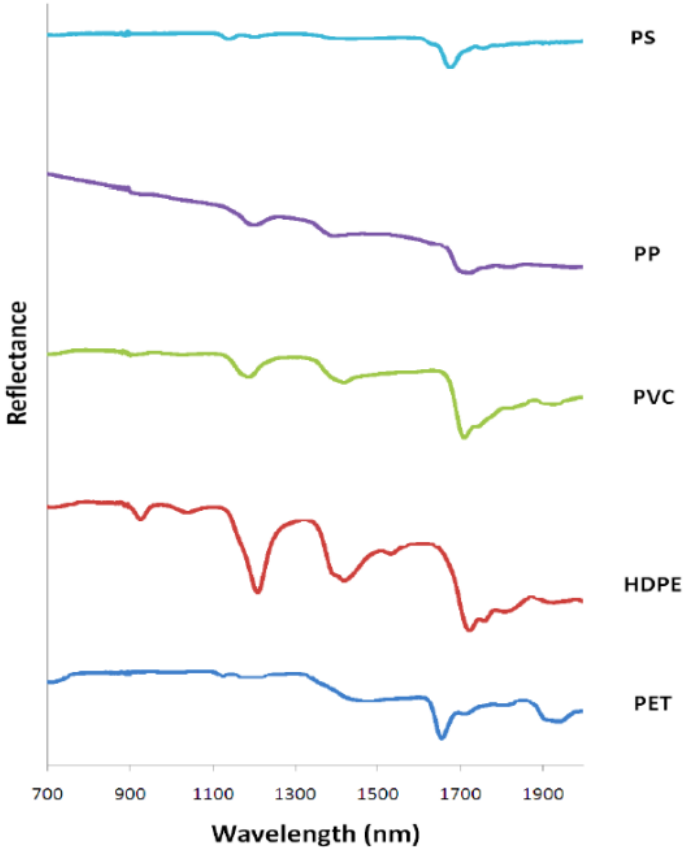


Figure 2.6: The reflectance of five of the most common types of plastic. The spectra are plotted with an offset to avoid interference and enhance clarity. The signature dips of the types of plastic are most present between 1100 nm to 1500 nm and around 1700 nm wavelength. [27]

The reflectance of a sample is determined by dividing the reference reflection of a reference tile, mainly made of PTFE, with the reflectance of the sample [28]:

$$R = \frac{I_r}{I_0} \tag{2.2}$$

2.2.1 Removing Dark Background

In this step, the aim is to remove the dark background. This involves subtracting the dark current or background signal from the measurements of the reference tile and the sample before calculating reflectance (equation 2.2) [28]. The dark current is obtained by scanning the reference and the samples separately while the light source of the device is off, relying only on the natural light of the environment.
The reflectance while removing dark background is calculated as follows:

$$R_{\text{sample}}(\lambda) = \frac{I_{\text{sample}}(\lambda) - I_{\text{dark\_sample}}(\lambda)}{I_{\text{reference}}(\lambda) - I_{\text{dark\_reference}}(\lambda)} \tag{2.3}$$

where:

- $I_{\text{sample}}(\lambda)$ is the intensity of light reflected from the sample.
- $I_{\text{reference}}(\lambda)$ is the intensity of light reflected from the reference standard.
- $I_{\text{dark\_sample}}(\lambda)$ is the intensity measured from the sample in the absence of the light source (dark current).
- $I_{\text{dark\_reference}}(\lambda)$ is the intensity measured from the reference standard in the absence of the light source (dark current).

By subtracting the dark current measurements ($I_{\text{dark\_sample}}(\lambda)$ and $I_{\text{dark\_reference}}(\lambda)$), we can correct for any background signals and accurately determine the reflectance of the sample.

2.2.2 Beer-Lambert law

When a sample is measured under ideal conditions, i.e., in transmission, at low concentration of the analyte(s) of interest and without light scattering, the absorbance of a substance can be accurately determined using the Beer-Lambert law:

$$A_0(\lambda) = -\log\left(\frac{I(\lambda)}{I_0(\lambda)}\right) = \epsilon(\lambda)LC \tag{2.4}$$

where $A_0(\lambda)$ is the absorbance at wavelength λ, $I(\lambda)$ is the intensity of the transmitted light at the same wavelength, $I_0(\lambda)$ is the intensity of the incident light at the same wavelength, $\epsilon(\lambda)$ is the molar absorptivity (a measure of how strongly the substance absorbs light at that wavelength), C is the concentration of the substance, and L is the path length of the light through the sample [29].

However, under current measurement conditions, a number of phenomena are added to the molecular absorption, causing Beer–Lambert's law to no longer apply. The interaction of radiation with particles and changes in the optical index affect the path of photons, resulting in light scattering. This scattering has two consequences, as illustrated by Figure 2.7. The first is a lengthening of the optical path, which introduces a multiplicative term. The second is a loss of photons, which are falsely counted as absorption, thus introducing an additive term.

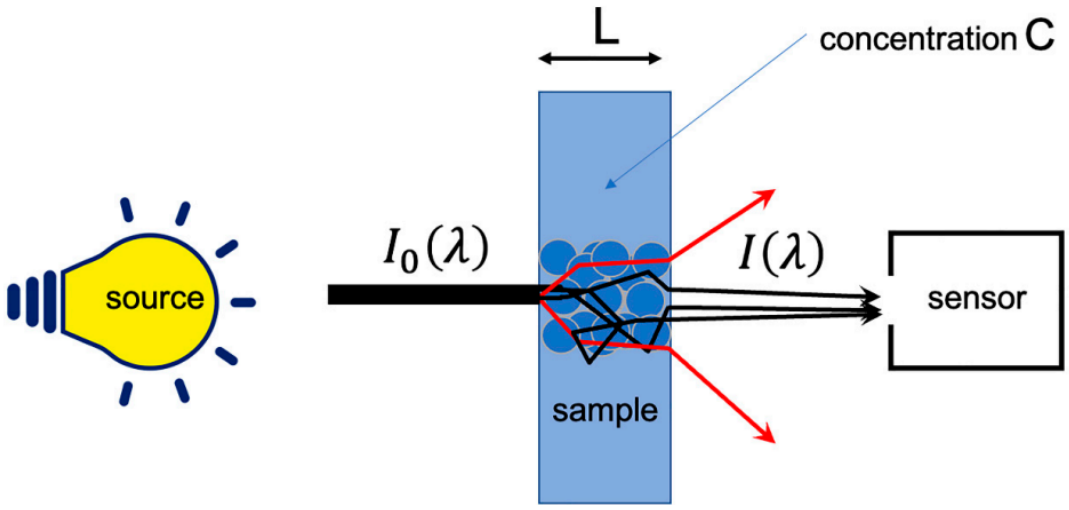


Figure 2.7: Effects of light scattering on molecular absorption measurements [29].

2.3 Spectroscopy

NIR spectroscopy relies on three key aspects: fundamentals, instruments, and data analysis.

The fundamentals include four measurement modes: transmittance, transflectance, diffuse reflectance, and interactance (Figure 2.8). The sample material determines the appropriate measurement mode. Transmittance is used for gases, liquids, or semi-solids in cuvettes, with NIR applied on one side and measured on the other. Transflectance is for semi-solids without cuvettes, where NIR penetrates the sample and reflects off a metal surface, resulting in a light path twice as long as in transmittance mode. Diffuse reflectance is for solids, measuring NIR scattering and absorption from one side. Interactance is also for solids, measuring absorption at a greater distance from the NIR source and potentially affected by ambient NIR signals.For plastic materials, diffuse reflectance is typically the appropriate measurement mode, as it is best suited for solid samples [29].

In the case of instruments, there are different types of NIR instruments, and the choice depends on application requirements, cost, signal-to-noise ratio, and measurement speed. The least expensive instruments use LEDs, with each LED producing a distinct NIR wavelength [21]. In the next section, the details and technology behind the three devices—Plastic Scanner, SpectraPod, and NIR Spectrometer are explained in detail.

Data analysis is essential for mapping NIR absorption or transmittance values to desired sample properties, as it is illustrated in Figure 2.9. Machine learning (ML) algorithms are used for this, involving training and testing phases. During training, the algorithms learn from light absorption values and desired outcomes. In testing, they predict outcomes based on new absorption values.

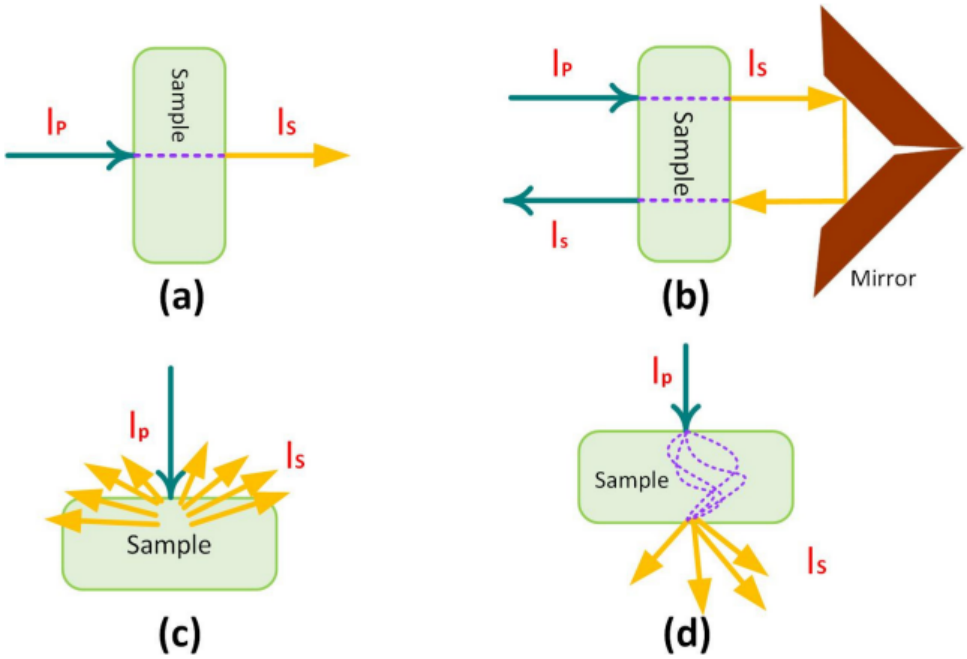


Figure 2.8: (a) Transmittance measurement mode, which is used with gases and semi-solids placed in a cuvette; (b) transflectance measurement mode, which is used with semi-solids without a cuvette;(c) diffuse reflectance measurement mode, which is used with solids where the measurement is taken from the NIR incidence; (d) transmittance through a scattering medium. [29]

NIR spectroscopy uses various multivariate analysis techniques based on machine learning, which can be divided into traditional methods and deep network architectures. Traditional methods, such as partial least squares (PLS) and Random Forest (RF), have few or no hidden layers and require expert feature engineering. Deep network architectures, like AlexNet and GoogLeNet, have multiple hidden layers and use raw features, often outperforming traditional methods with large datasets but facing overfitting and high computational costs with limited data [29].

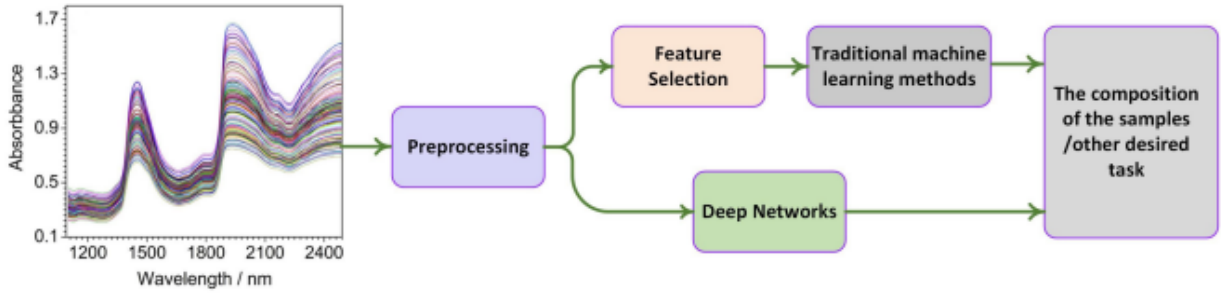


Figure 2.9: Architecture of machine learning for NIR spectroscopy

Deep architectures train end-to-end, learning local and temporal patterns, this can lead to overfitting with limited data. Traditional ML methods often form a pipeline architecture that includes preprocessing, feature engineering, and modeling. In Section 2.2 to 2.5, the theoretical concepts of the data analysis phase, including preprocessing and some popular ML models, are explained [29].

2.4 Spectroscopy Devices

A typical spectroscopy device, called a spectrometer, consists of three main components: a light source, a mechanism for isolating specific wavelengths of radiation, and a detector. In spectroscopy, a spectrum of light is directed onto a sample, and the resulting transmission or reflection spectrum is captured by the detector. The main difference between the devices are correspond to their different light source, detector and the mechanism that they isolate the wavelengths. In following subsections technology and the main components of the three devices are explained.

2.4.1 NIR Spectrometer

The NIR spectrometer includes a halogen lamp for illuminating NIR wavelengths and AVASPEC-NIR256-2.0TEC as the detector. AVASPEC-NIR256-2.0TEC is a grating-based device capable of distinguishing different NIR wavelengths with high resolution (see Appendices for more details).

Grating Based Devices

Grating-based spectrometers are devices that operate on the principle of diffraction by using a grating to disperse light into its component wavelengths (Figure 2.10). The light source in a grating-based spectrometer is typically a broad-spectrum lamp, such as a halogen or xenon lamp, which emits light across a wide range of wavelengths. This broad-spectrum light source ensures that the full spectral characteristics of the materials can be analyzed [28].

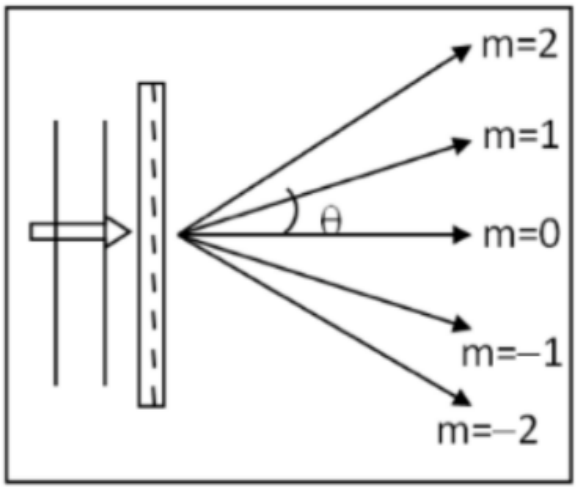


Figure 2.10: Dispersing light into a spectrum using a diffraction grating

The diffraction grating acts as the mechanism for isolating specific wavelengths of light. A diffraction grating is an optical component with a regular pattern of closely spaced lines or grooves. When light strikes these lines, it is diffracted, or bent, at different angles depending on its wavelength. This property allows

the grating to disperse a beam of light into a spectrum. The behavior of light interacting with a diffraction grating is governed by the grating equation:

$$d \sin \theta = m\lambda \tag{2.5}$$

where d is the spacing between adjacent grating lines (known as the grating constant), θ is the angle at which a particular wavelength λ is diffracted, and m is the order of the diffraction (an integer that can be 0, ± 1, ± 2, etc.). This equation shows that different wavelengths of light will be diffracted at different angles, allowing the grating to separate them distinctly.

While increased dispersion spreads out the wavelengths, it is the sharpness of the peaks that determines how distinctly neighboring wavelengths can be separated. This sharpness is the resolving power of a diffraction grating, which means the capacity to produce distinct peaks for wavelengths that are very close together in a specific order. The resolving power R is given by:

$$R = \frac{\lambda}{\Delta\lambda} = mN \tag{2.6}$$

where λ is the wavelength of light, $\Delta\lambda$ is the smallest difference in wavelengths that can be distinguished, m is the diffraction order, and N is the total number of lines utilized on the grating. Higher resolving power means the grating can separate wavelengths that are very close together.

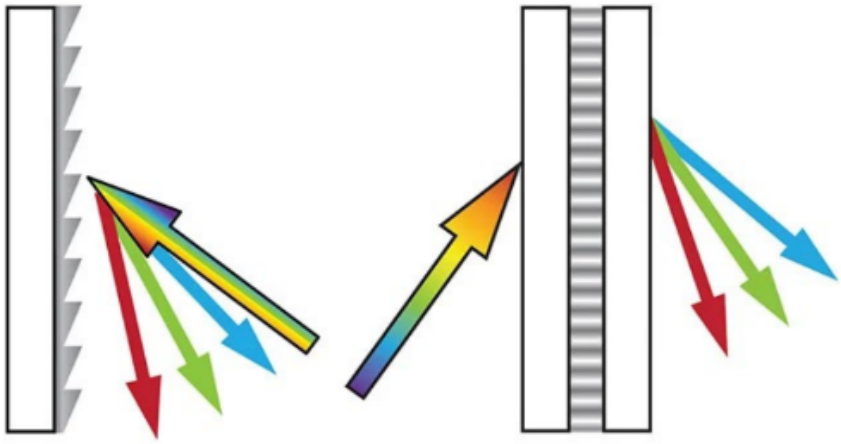


Figure 2.11: Two types of gratings: Reflective Grating (left) and Transmissive Grating (right)

In the domain of spectral analysis, numerous device configurations leverage gratings as spectral dispersing elements. One such configuration, used in the AvsSpec_NIR256_2.0 TEC, is the Czerny-Turner spectrometer (Figure 2.12). This architecture comprises three primary components: an entrance slit, two mirrors, and a diffraction grating.

The entrance slit serves as the initial point where light enters the spectrometer. Its function is to ensure that a narrow, well-defined beam of light is directed to the optical components, which is essential for maintaining high spectral resolution. Following the entrance slit, the Czerny-Turner design employs two mirrors: a collimating mirror and a focusing mirror. The collimating mirror, positioned after the entrance slit, transforms the diverging light from the slit into a parallel beam. This parallel beam is necessary for precise diffraction by the grating. The focusing mirror, placed after the diffraction grating, collects the dispersed light and focuses it onto the detector or exit slit. The configuration and orientation of the grating determine the range and resolution of the wavelengths that the spectrometer can analyze. The detector records the

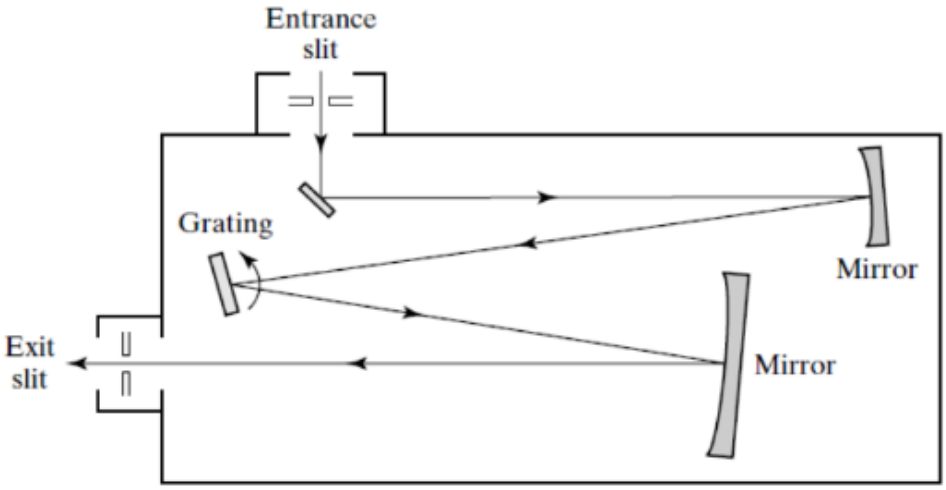


Figure 2.12: Czerny-Turner spectrometer architecture [28]

intensity of light at each wavelength, and this data is processed to generate the spectrum of the incoming light [28].

2.4.2 Handheld Plastic Scanner

The Plastic Scanner, developed by De Vos [30]. It employs a few LEDs with different wavelengths in the NIR spectrum as the light source. Unlike grating-based methods, wavelength selection occurs at the light source, enhancing cost efficiency. The scanner illuminates the plastic sample using eight LEDs sequentially, each with a different NIR wavelength (850, 950, 1050, 1200, 1300, 1450, 1550, and 1650 nm), and then measures the spectrum for each illumination using an InGaAs detector. The internal design of the Plastic Scanner is shown in Figure 2.13. An Arduino board is programmed to control the components, light the LEDs, and read the detected reflectance. A software called PsPlot is developed to use the Plastic Scanner in connection with a computer. More information about the device can be found in the appendix.

The InGaAs detector has specific responsivity, meaning it can detect a particular range of wavelengths. Figure 5 shows the plotted responsivity data of the detector, extracted using an online program named Web-PlotDigitizer by De Rijke [25].De Rijke conducted research to determine if the correct LEDs and InGaAs detector were chosen for the Plastic Scanner. Plastic samples' spectra were measured in three ways: using a halogen lamp and NIR spectrometer as a reference, using LEDs with the NIR spectrometer, and using the Plastic Scanner with LEDs and an InGaAs detector. These investigation are valuable to characterize and investigate source lights behaviour as well as detectors. Figure 2.14 shows the resulted spectra measured by these three ways.

Results suggested replacing the 1460 nm LED with a 1400 nm LED to enhance intensity around the 1400 nm identification dip and replacing the existing 1650 nm LED with a higher power version. The 1720 nm LED was deemed unnecessary due to spectral overlap with the 1650 nm LED. The study also explored the influence of plastic color, revealing that darker-colored plastics had less defined spectra, making it challenging to distinguish plastic types.

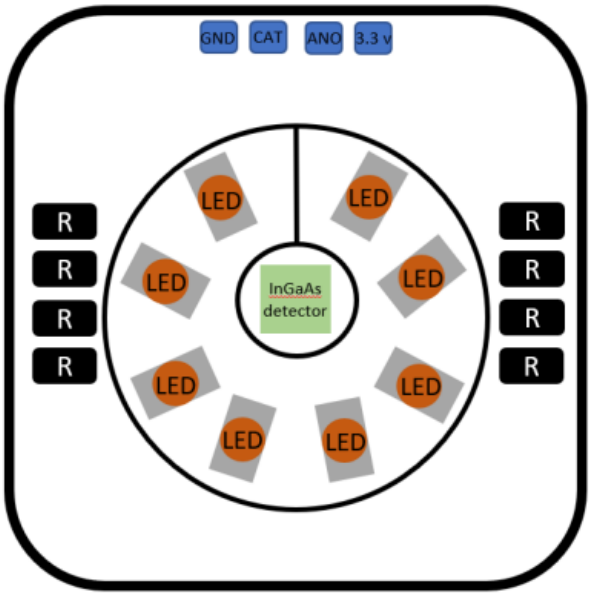


Figure 2.13: The internal design of the spectroscopic side of the plastic scanner. The InGaAs detector is surrounded by 8 LEDs, each emitting light in an different wavelength. The resistors for the LEDs are placed on the sides of the board (R). The Arduino, to manage the LEDs and analyse the reflected spectra collected from the InGaAs detector, is connected via the four connectors at the top of the board. The GND and 3.3v deliver the voltage to the board and the CAT and ANO are the connectors for the detector. The circle around the LEDs is the spacer between the plastic samples and the LEDs, usually 10 mm high. The circle around the InGaAs detector is used to block any direct light from the LEDs into the detector, usually 4 mm high.[25]

In addition, De Rijke plotted the responsivity of the InGaAs detector in the same graph as the spectra of the NIR LEDs for further investigation. The results showed that the spectra of the 940 nm, 1650 nm, and 1720 nm LEDs are outside the responsivity range of the InGaAs detector. The detected spectra areas are shown in grey for the 940 nm LED, yellow for the 1650 nm LED, and deep pink for the 1720 nm LED. This suggests that either the 1650 nm or the 1720 nm LED can be removed from the design. Due to the larger detectable range of the 1650 nm LED, it is recommended to remove the 1720 nm LED. However, De Rijke suggests keeping the 940 nm LED in the design as HDPE shows an identification dip around 950 nm.

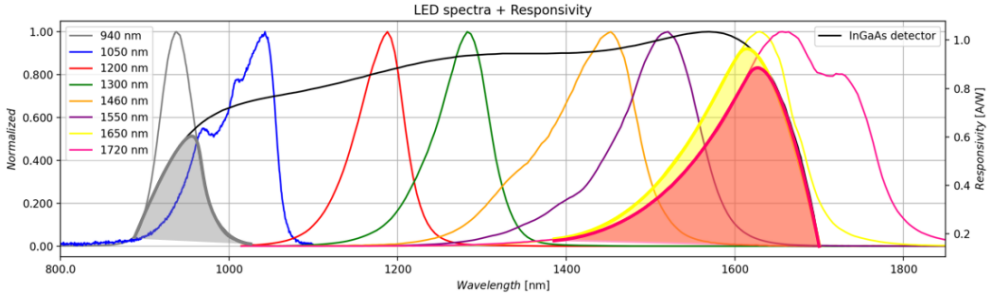
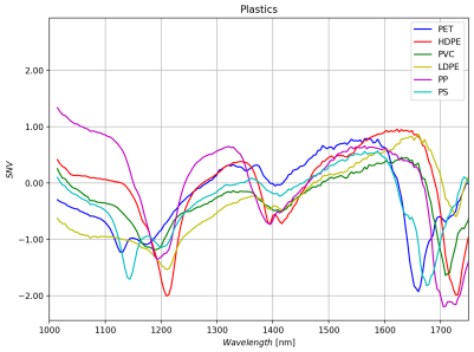
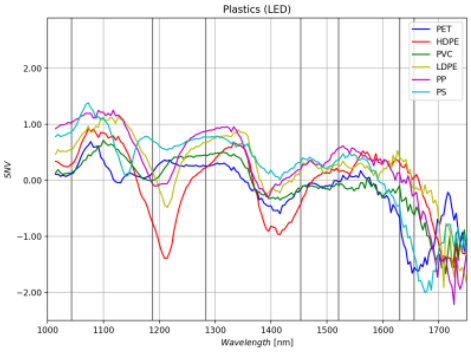


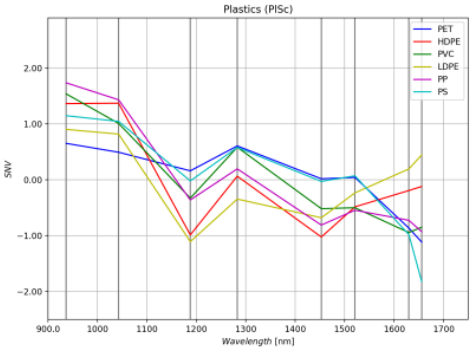
Figure 2.15: The spectra of the 8 different LEDs and responsivity of the InGaAs detector are plotted against the wavelength. The left y-axis The coloured areas show how much the spectra of the LEDs are detected by the detector.[25]



(a) Measured with a halogen lamp as a light source and the NIR spectrometer as the detector.



(b) Measured with the LEDs as a light source and the NIR spectrometer as the detector.



(c) Measured with the prototype of the plastic scanner.

Figure 2.14: The spectra of the different types of plastic are plotted against the wavelength. (a) are the spectra of the types of plastic measured with a halogen lamp, (b) are the spectra measured with the NIR LEDs and (c) are the spectra measured using the plastic scanner prototype.[25]

2.4.3 SpectraPod

The SpectraPod uses a halogen light as source light and a special type of detector called a resonant-cavity-enhanced (RCE) detector, which is designed for near-infrared (NIR) spectral sensing. The RCE detector works by using an array of photodetectors, each integrated with its own filter, to detect specific wavelengths of light within the 850–1700 nm range.
Each pixel in the detector array has a thin absorbing layer positioned inside a Fabry-Perot (FP) cavity (2.16. This cavity creates a strong spectral dependence in the quantum efficiency of the detector, meaning each pixel is sensitive to different wavelengths of light. The wavelength each pixel detects can be tuned by adjusting the thickness of a tuning element inside the cavity, allowing the detector to cover a broad range of wavelengths without the need for mechanical adjustments [31].

The array contains 16 pixels, and each pixel can be tuned differently. This tuning changes the length of the FP cavity, shifting the wavelengths each pixel detects. The optical response of each pixel is simulated and shows distinct peaks at different wavelengths, which can be adjusted by modifying the cavity's structure. The result is a robust, integrated detector that eliminates alignment errors and mechanical tuning, providing high efficiency and low dark current.

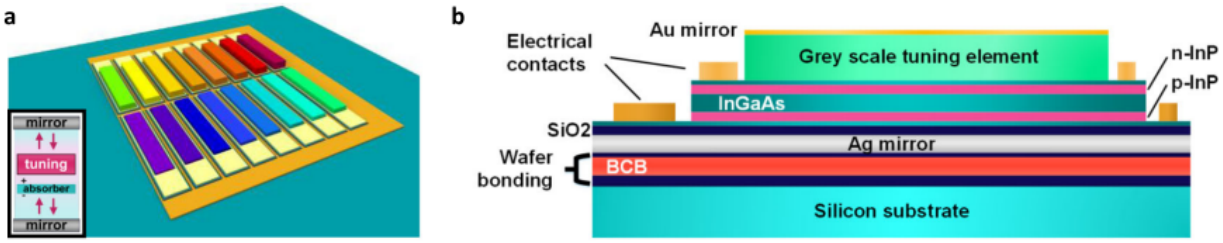


Figure 2.16: Mechanism of a resonant-cavity-enhanced (RCE) multi-pixel array.: a) Top view sketch of a multi-pixel array where each pixel (indicated by the different colors) has a different wavelength response. Inset: Sketch of an RCE detector, where both the absorber and the tuning element are positioned within the vertical-cavity structure. b) Cross section of a single RCE detector (not to scale). [31]

During operation, the SpectraPod's detector array is illuminated with light from an unknown spectrum. Each pixel generates a photocurrent depending on its responsivity to the incident light's wavelengths. This photocurrent data is then used to determine the spectrum of the incident light, which can be used for various sensing applications including plastic type detection. Equation 2.7 represents the calculation of photocurrent in the SpectraPod detector array:

$$I_i = \int_{\lambda_1}^{\lambda_2} S(\lambda) R_i(\lambda)\, d\lambda, \quad (i = 1, 2, \ldots, N) \tag{2.7}$$

I_i represents the photocurrent produced by the $i - thpixel$ in the SpectraPod detector array. $S(\lambda)$ signifies the incident light spectrum, showing the intensity distribution across different wavelengths. $R_i(\lambda)$ indicates the responsivity of the $i-th$ pixel to light at a specific wavelength λ. By integrating the product of (λ) and $R_i(\lambda)$ over the wavelength range (λ_1 to λ_2), the equation calculates the total contribution of incident light to I_i [31].

2.4.4 devices comparison

The size of features provided by the output data of each device is detailed in Table 1. The output of the plastic scanner and the NIR spectrometer consists of the intensity of the reflection for different wavelengths, while the output of SpectraPod is photocurrent for different channels.

Table 2.1: Summary of the devices comparison

	Light source	Detector	Place of isolating λ	Output size
Plastic Scanner	LEDs	InGaAs	Light Source	8
SpectraPod	Halogen Lamp	Array of RCEs	Detector	16
NIR Spectrometer	Halogen Lamp	Grating	Detector	237

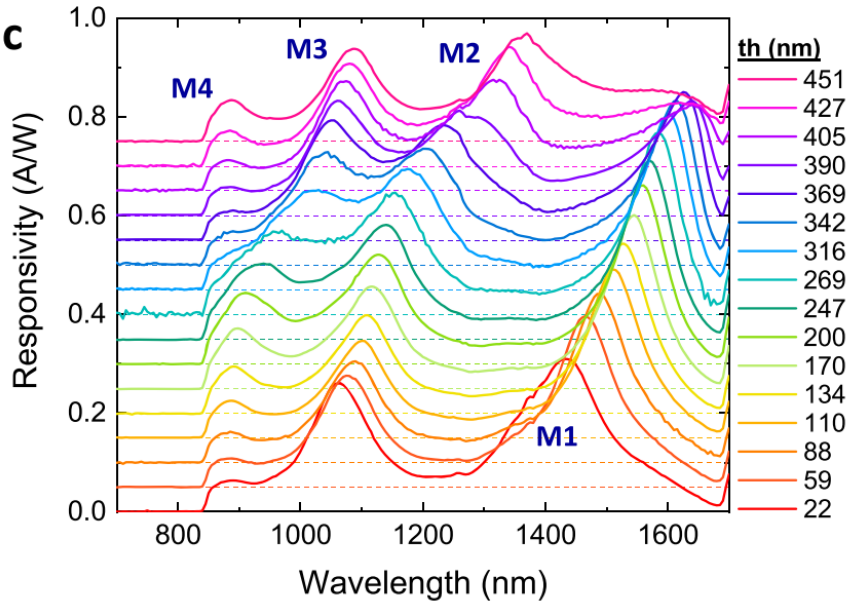


Figure 2.17: Measured responsivity for 16 pixels of the same array with measured Ma-N thickness increasing from 22 to 451 nm. [31]

2.5 Data Analysis

Using the output of devices to classify plastic types is typically a classification problem that can use machine learning (ML). This project focuses mainly on supervised learning methods. A supervised classifier is trained on a training set in which each data point has some features and a label. Each label denotes a class. A feature is a characteristic or property of the data that can represent some aspect of the data, and be used for analysis [32]. For example, in spectral data, features can be:

- A set of wavelengths measured directly by the device.
- A set of combinations or functions of these wavelengths, such as averages, differences, or other transformations that capture important patterns or relationships in the data.

To keep the notation and explanation in this chapter clear, each data is represented as a separate row in a feature matrix X, where each feature is stored as a separate column. For example, a dataset consisting of 150 samples and four features can be written as a 150×4 matrix:

$$X \in \mathbb{R}^{150\times 4} :$$

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{150,1} & x_{150,2} & x_{150,3} & x_{150,4} \end{bmatrix} \tag{2.8}$$

Here, each x_{ij} represents the value of the j-th feature for the i-th sample.

Similarly, labels are noted as a 150-dimensional column vector:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{150} \end{bmatrix} \in \mathbb{R}^{150} \tag{2.9}$$

Figure 2.18 illustrates a roadmap for building a supervised machine learning system. As shown in Figure 2.18, a typical workflow for using a supervised machine learning system has four dominant stages: Preprocessing, Learning, Evaluation, and Prediction. In this section, the theories of methods in these four stages are explained.

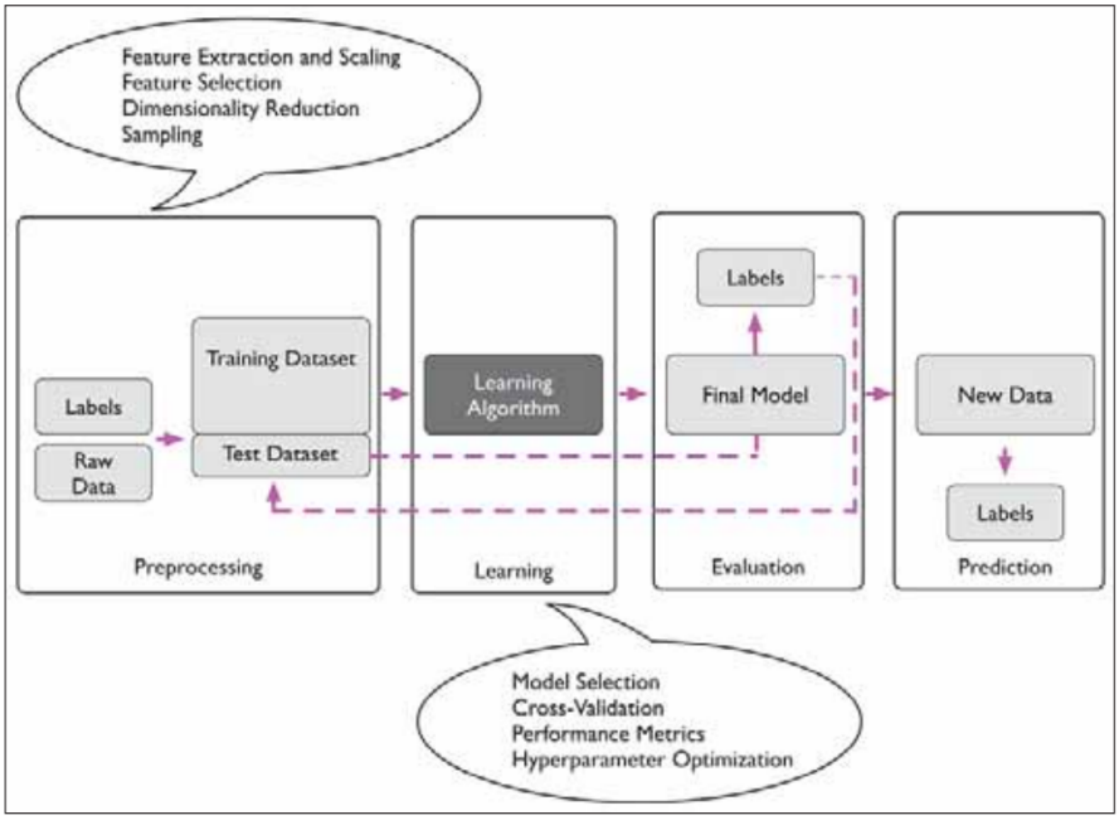


Figure 2.18: A roadmap for building a supervised machine learning system. Preprocessing: This stage involves applying transformations to the data, such as scaling, which can improve the accuracy of analysis and classification. Additionally, methods for reducing the dimensionality of the data can be employed.Learning: In this stage, the appropriate machine learning model is selected and its hyperparameters are tuned using techniques like cross-validation. The tuned model is then trained on the training dataset. Evaluation: The trained model is tested on a separate test dataset, and its performance is evaluated using evaluation metrics such as accuracy. Prediction: Once the model is evaluated and deemed satisfactory, it can be deployed for real-world prediction tasks. [32]

2.5.1 Preprocessing

Preprocessing, or data cleaning, is an important step in analyzing spectroscopy data. It removes unwanted parts of the data that can hide useful information needed for building accurate models. Spectral data often have variations that do not help the model, and some may even be harmful, like baseline shifts in spectroscopy [26].

Baseline shifts are a common problem in spectroscopy data and can affect the model's performance. To fix baseline shifts, methods like calculating the first and second derivatives of the spectra are often used. These derivatives remove vertical offsets and sloping baselines. Techniques like Savitzky-Golay (SG) smoothing are then used to smooth the data further. Noise is another issue that can lower the quality of spectral data. SG smoothing also reduces noise levels while keeping important spectral features. In addition, normalization and scaling are key preprocessing steps to ensure consistency and comparability of spectral data and minimizing unwanted variations. Common techniques include standard normalization and standard normal variate (SNV). Furthermore, dimensionality reduction techniques like Principal Component Analysis (PCA) reduce the complexity of spectral data by transforming it into a simpler form. PCA finds the main components that capture most of the data's variance [26].

In the subsequent subsections, each of these methods will be elaborated upon in detail to provide a deeper understanding of their implementation and effectiveness in spectroscopy preprocessing.

Normalization and Scaling Methods

- Standard Normalization (SN):

  Standard normalization for spectroscopy data involves adjusting each wavelength's intensities to have a mean of zero and a standard deviation of one across all spectra [33]. The normalization equation is expressed as:

  $$R_{\text{normalized}}(\lambda) = \frac{R(\lambda) - \mu_\lambda}{\sigma_\lambda} \tag{2.10}$$

  Here, $R(\lambda)$ represents the reflectance at each wavelength λ, μ_λ denotes the mean reflectance across all spectra at wavelength λ, and σ_λ signifies the standard deviation of reflectance across all spectra at wavelength λ.

- SNV:

  In SNV, each spectrum's reflectance are scaled to achieve a mean of zero and a standard deviation of one [33]. This is calculated for each spectrum R as:

  $$R_{\text{SNV}}(\lambda) = \frac{R(\lambda) - \mu_R}{\sigma_R} \tag{2.11}$$

  Here, $R(\lambda)$ represents the reflectance at each wavelength λ, μ_R denotes the mean reflectance across all wavelengths for the given spectrum R, and σ_R signifies the standard deviation of reflectance across all wavelengths for the given spectrum R.

Savitzcy-Golay

The spectral signals obtained by the spectrometer contain both useful information and random noise. Signal smoothing is a common de-noising method, primarily used to reduce noise and improve the signal-to-noise ratio when the noise is zero-mean random white noise. One effective method for signal smoothing is the SG smoothing [26].

The SG smoothing method involves fitting a polynomial to a subset of data points within a moving window and then replacing the central data point in this window with the value of the fitted polynomial. This approach smooths the signal while preserving the important features like peak height and width. The window of smoothing has a certain width $(2w + 1)$, with an odd number of wavelength points in each window. The smoothed value for the k-th wavelength point is calculated as:

$$x_{k,\text{smooth}} = \frac{1}{H} \sum_{i=-w}^{w} x_{k+i} h_i \tag{2.12}$$

In Equation 2.12, h_i and H are the smoothing factor and the normalization factor, respectively, where $H = \sum_{i=-w}^{w} h_i$. The purpose of multiplying each measurement by the smoothing factor h_i is to reduce the effect of smoothing on useful information. The h_i values can be obtained by using polynomial fit based on the principle of least squares. The way that h_i is obtained is explained in detail in [26].

The effectiveness of SG smoothing depends on the width of the smoothing window. A too-small window width may not sufficiently reduce noise, while a too-large window width may smooth out important spectral features, resulting in signal distortion (as shown in Figer 2.19).

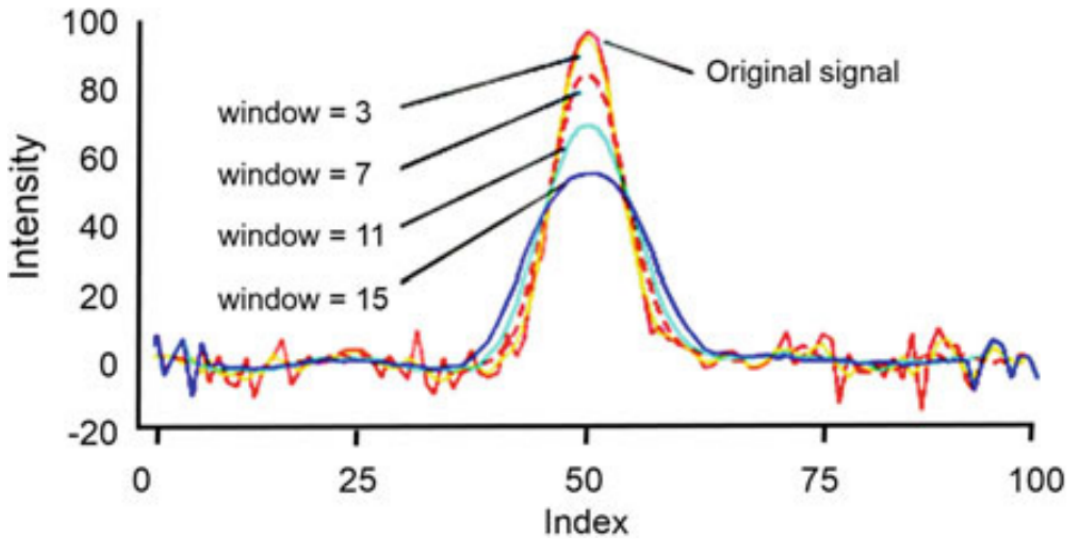


Figure 2.19: Schematic diagram of window moving smoothing method [26]

In addition to smoothing, the Savitzky-Golay method can also compute derivatives, which are commonly used for baseline correction and resolution enhancement in spectral analysis. Derivative spectra can effectively eliminate baseline interference and other background noise, distinguish overlapping peaks, and improve resolution and sensitivity. However, taking derivatives can also introduce noise.

The first derivative and second derivative are particularly useful. The general equation for the SG derivative is:

$$y_k^{(m)} = \sum_{i=-w}^{w} C_i^{(m)} x_{k+i} \tag{2.13}$$

where $y_k^{(m)}$ is the m-th derivative of the signal at point k, and $C_i^{(m)}$ are the Savitzky-Golay coefficients for the m-th derivative.

Principal component analysis

PCA is used to reduce the dimensionality of a high-dimensional dataset by reducing its large feature set to a smaller one while retaining most of the information in the large set. The new features are called principal components, each of which is a linear combination of the original features [34]. These components are uncorrelated, and the first components contain most of the information in the original variables.

For a dataset with p features, each data point is a point in a p-dimensional space. However, not all dimensions have the same effect on the observations' pattern. PCA finds the dimensions(orginal features) that have the most effect and stores them in principal components. Principal components are linear combinations of features created as follows:

For a set of features $X_1, X_2, \ldots, X_p$, the first principal component is:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p \tag{2.14}$$

where

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{2.15}$$

and the elements $\phi_{11}, \ldots, \phi_{p1}$ are the loadings of the first principal component. The loading vector ϕ_1 (a vector of $\phi_{11}, \ldots, \phi_{p1}$) is a direction in feature space along which the data vary the most. The importance of features is related to their loading values.

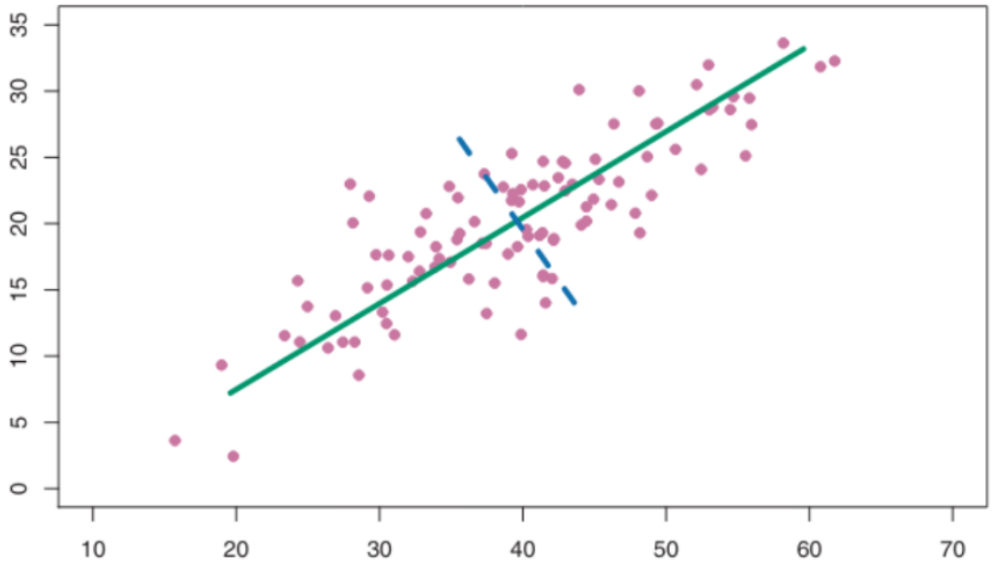


Figure 2.20: The green line is the first principal component along which the data vary the most.The axis are represented to two original features [34]

In Figure 2.20, the green line represents the first principal component. Projecting n data points onto the green line creates n scores for each data point referred to as $z_{11}, \ldots, z_{p1}$, which are calculated by:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip} \tag{2.16}$$

As seen in Figure 4.1, the variety of scores along the green line is greater than any variety along other lines in the feature space.

The optimization problem (Equation 10.3) is used to find the desirable loading vector for the first principal component:

$$\max_{\phi_{11},\ldots,\phi_{p1}} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \tag{2.17}$$

subject to

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{2.18}$$

Here, $\sum_{j=1}^{p} \phi_{j1} x_{ij}$ is z_{i1}, which projects the i-th observation on the direction defined by ϕ_1. Squaring it gives the squared distance of the score from the original data point. The second summation gives the total variance along the direction. The constraint is used for normalization to prevent arbitrarily large variance.

The second principal component Z_2 is a linear combination of features that has the maximum variance among all linear combinations of features that are uncorrelated with the first principal component. The form of the second principal component's i-th score is:

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip} \tag{2.19}$$

The same process is applied to determine other principal components.

2.5.2 Classification Machine Learning Models

PLS_DA

Partial Least Squares Discriminant Analysis (PLS-DA) is a variant of PLS that is specifically designed for classification tasks [34]. PCA identifies directions that best represent the predictors $X_1, \ldots, X_p$ without using labels, making it an unsupervised method. Partial least squares (PLS), on the other hand, is a supervised alternative. Like PCA, PLS reduces dimensionality by finding new features $Z_1, \ldots, Z_M$ that are linear combinations of the original features. However, PLS incorporates the labels Y when identifying these new features. This means that the new features not only represent the original features in a reduced dimensional space but also maintain a relationship with the labels.

The first direction Z_1 in PLS is computed by:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p \tag{2.20}$$

Here, the loading ϕ_{j1} for each feature X_j is proportional to how strongly X_j is correlated with Y. This relationship is determined using simple linear regression:

$$Y = \beta_0 + \beta_j X_j \tag{2.21}$$

The coefficient β_j is proportional to the correlation between X_j and Y. A higher β_j means a higher correlation. PLS sets ϕ_{j1} to be proportional to β_j, ensuring ϕ_{j1} reflects the strength of the relationship between X_j and Y.

Next, for each feature X_j, the residual between X_j and Z_1 is calculated. The residual is the difference between the actual value and the value predicted by Z_1. These residuals, which are parts of X_j not captured by Z_1, form a new feature set called $X_j^{(1)}$. This new set is then used to calculate Z_2 in the same way Z_1 was calculated, but using $X_j^{(1)}$ instead of X_j.

The process is repeated to calculate $Z_3, \ldots, Z_M$.

Additionally, PLS is not only used for dimensionality reduction but also serves as a powerful prediction tool in machine learning models. By incorporating the response variable Y, PLS can predict outcomes and make informed decisions based on the learned relationships between features and labels.

The above explanation is used when the Y labels are continuous values. However, PLS can be used for classification by treating the class labels as numerical values.

PLS-DA is a method specifically designed for classification tasks. It combines the concepts of PLS with discriminant analysis to handle categorical outcome variables (class labels). PLS-DA reduces the dimensionality of the predictor space while preserving the ability to discriminate between different classes. This is achieved by finding new components that are linear combinations of the original features, tailored to maximize the separation between classes.

In PLS-DA, the outcome variable Y represents class labels, which are transformed into a numerical format suitable for analysis, often using one-hot encoding. The method identifies components $Z_1, Z_2, \ldots, Z_M$ that are strongly correlated with these class labels.

Once the PLS-DA model is trained, it can be used to classify new observations. For a new observation, calculate the scores for each component $Z_1, Z_2, \ldots, Z_M$:

$$Z_{\text{new}} = \begin{pmatrix} \phi_{11}X_{1,\text{new}} + \phi_{21}X_{2,\text{new}} + \ldots + \phi_{p1}X_{p,\text{new}} \\ \phi_{12}X^{(1)}_{1,\text{new}} + \phi_{22}X^{(1)}_{2,\text{new}} + \ldots + \phi_{p2}X^{(1)}_{p,\text{new}} \\ \vdots \\ \phi_{1M}X^{(M-1)}_{1,\text{new}} + \phi_{2M}X^{(M-1)}_{2,\text{new}} + \ldots + \phi_{pM}X^{(M-1)}_{p,\text{new}} \end{pmatrix}$$

The new observation is classified based on the scores calculated. The class label is assigned by finding the class whose scores best match the observed scores.

Support Vector Machine

In classification, SVM is particularly effective for binary classification problems, where the goal is to separate data points into two classes. The main idea behind SVM is to find a hyperplane that best separates the data points of different classes in the feature space [34].

The hyperplane serves as the decision boundary between the classes. It is defined by the equation:

$$B_0 + \sum_{j=1}^{p} \beta_j x_{ij} > 0 \quad \text{if } y_i = 1$$

$$B_0 + \sum_{j=1}^{p} \beta_j x_{ij} < 0 \quad \text{if } y_i = -1$$

Equivalently:

$$y_i(B_0 + \sum_{j=1}^{p} \beta_j x_{ij}) > 0 \quad \text{for all } i = 1, \ldots, n$$

This equation could represent the black solid line in Figure 2.21 where $p = 2$.

In other words, for each observation i, if y_i equals 1, the point lies on one side of the hyperplane, and if y_i equals -1, it lies on the other side. The objective is to maximize the margin M.

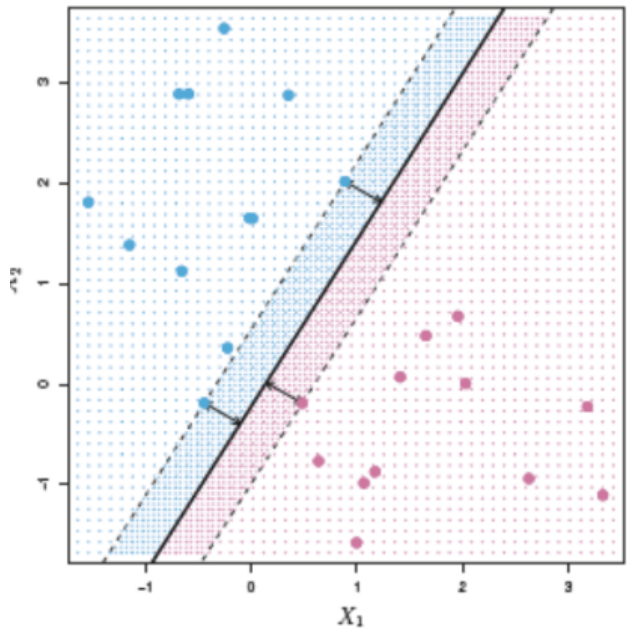


Figure 2.21: Example of SVM hyperplane. The dataset consists of two classes The blue one and the purple one, the solid black line is the hyperplane with the maximum margin. [34]

The minimal distance from the observations to the hyperplane is the margin. There can be multiple separating hyperplanes, but the maximal margin hyperplane is chosen because it minimizes classification errors by having approximately equal distances to each side. Thus, the goal of the maximal margin classifier is to maximize the margin:

$$\text{maximize } M \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(B_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \geq M \quad \text{for all } i = 1, \ldots, n$$

In cases where the data is not linearly separable, SVM employs the kernel trick to map the input features into a higher-dimensional space where they can be linearly separated. This allows SVM to find nonlinear decision boundaries by transforming the input features using kernel functions like the polynomial kernel:

$$K(x_i, x_j) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{ij}\right)^d$$

where d is the degree of the polynomial kernel.

By using the kernel function, SVM can capture complex relationships between features and improve classification performance.

For multiclass classification, SVM can be extended using two approaches: One-Versus-One and One-Versus-All. One-Versus-One creates classifiers for each pair of classes, while One-Versus-All constructs one classifier per class against all others. The final classification is determined based on the results of these classifiers.

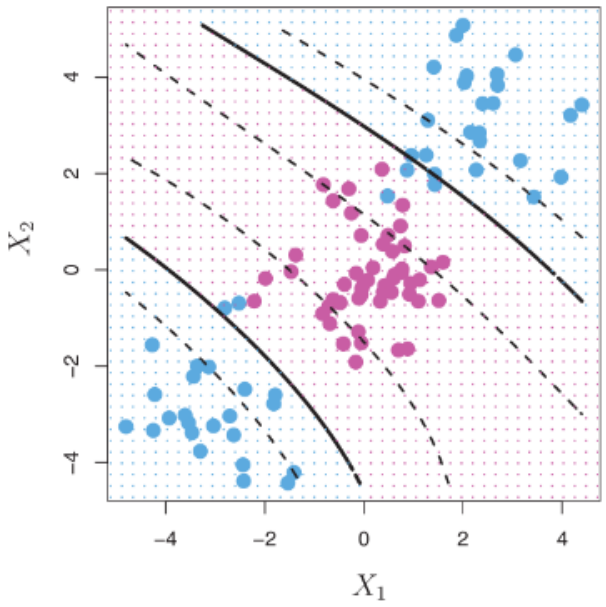


Figure 2.22: Example of non linear kernel function in SVM to have a more flexible decision boundaries. [34]

Some important hyperparameters in SVM include the choice of kernel function type (e.g., linear, polynomial), and the degree of the polynomial kernel function. These hyperparameters can significantly impact the performance of the SVM model and should be carefully tuned.

Artificial Neural Network

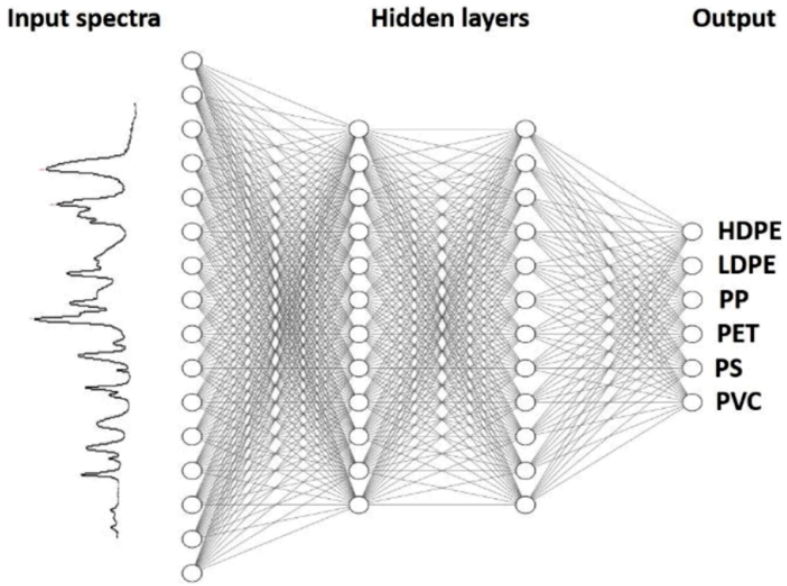


Figure 2.23: Sample ANN architecture for plastic classification.

A neural network is a computational model inspired by the human brain's information processing. It's composed of several key components [34]:

- **Input Layer**: Neurons in this layer represent features of the input data.
- **Hidden Layers**: These layers process information from the input layer and pass the results to the next layer. Each hidden layer contains various neurons that perform computations on the input data.
- **Output Layer**: This layer processes the final output of the network. The number of neurons in this layer corresponds to the desired number of outputs.
- **Weights and Bias**: Weights determine the strength of connections between neurons. Bias terms are added to the weighted sum of inputs to introduce flexibility and enable the network to learn complex relationships in the data.
- **Activation Function**: Neurons apply an activation function to the weighted sum of inputs to introduce non-linearity into the network.

Forward Propagation:
In each hidden layer neuron, a linear combination of inputs is computed using weights:

$$a_j^{(1)} = \sum_{i=1}^{M} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{2.22}$$

Here, $a_j^{(1)}$ represents the activation of neuron j in the first hidden layer, $w_{ji}^{(1)}$ is the weight associated with the connection from input neuron i to hidden neuron j, and $w_{j0}^{(1)}$ is the bias term for neuron j.

Then, an activation function $h(.)$ is applied to introduce non-linearity:

$$z_j^{(1)} = h(a_j^{(1)}) \tag{2.23}$$

This process is repeated for subsequent layers, with the output of each layer serving as the input to the next layer.

A loss function measures the difference between predictions and actual labels. For binary classification, binary cross-entropy loss is used, while categorical cross-entropy loss is used for multiclass classification.

Hyperparameters such as the number of layers, neurons per layer, and activation function are critical design choices that affect the performance of the network.

Random Forest

Random Forests is a Classification method based on decision trees. A decision tree has some nodes, in each node, an observation is asked if a certain feature value is greater than a threshold or not. So each node splits into two nodes. For example in Figure 2.25, the first node is used to check the value of $P1$ of the given observation, if it is less than 100, the observation will be passed to the left node, otherwise, to the right one. The same process will be done for each node unless the observation reaches a leaf, a node that doesn't split anymore. Each node represents one of the classes.

Each node must decide whether to split or not, the Gini index is a technique that is used to find if a node is pure or not. A pure node is a leaf. Gini Index is used for measuring variance across the classes in a node. The small value of the Gini index indicates that the node is leaf and no more splitting is needed.

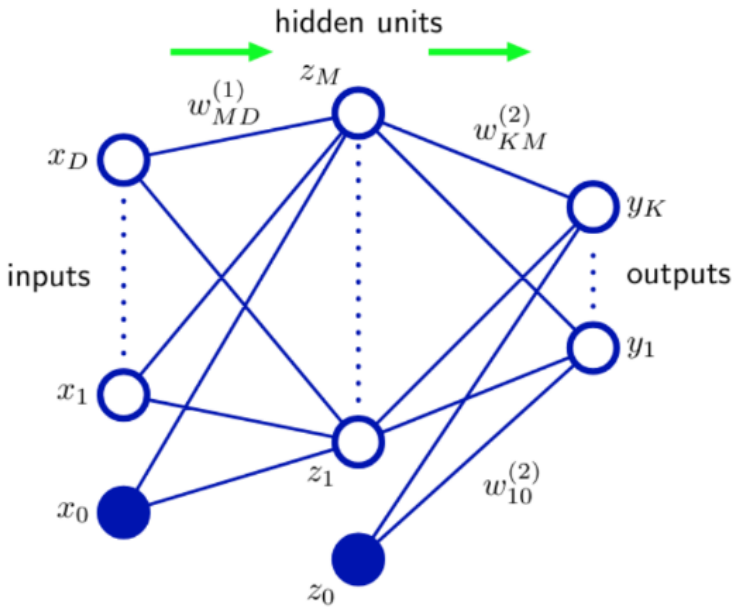


Figure 2.24: ANN architecture. The input layer contains neurons representing input data features. Hidden layers process information and pass results to subsequent layers. The output layer generates the final output. Weights and biases determine connection strengths and introduce flexibility.

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{2.24}$$

p_{mk} is the proportion of training observations in the mth node that are from the kth class.

One method to determine the threshold for splitting a node is to select the threshold that results in the minimum Gini index in the two nodes created after the split.

Bagging, short for Bootstrap Aggregation, is a technique employed to enhance classification performance using multiple trees. Each tree evaluates the given observation and predicts a class. The final class of the observation is determined by the class that receives the majority of votes from all the trees. The trees are different because each one is trained on a different subset of the original dataset. Each subset is called a bootstrap. A bootstrap is created by randomly drawing observations, with replacement, from the original dataset (Figure 2.26).

Random Forest process is like Bagging except that in each split, only a random subset of features is considered for splitting. If there are p features, in each splitting we consider m features that $m \leq p$. Adding randomization in feature selection reduces variance and improves accuracy.

Some hyperparameters in Random Forests:

- Number of Trees.
- Maximum Depth of the Tree.
- Minimum sample per Leaf.
- Minimum samples Split.

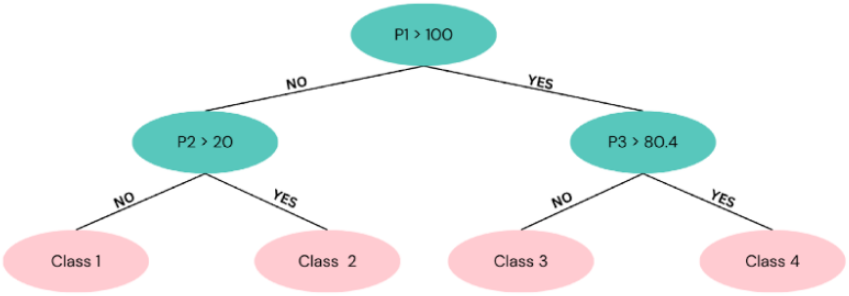


Figure 2.25: A decision tree with 4 classes, each leaf represents one class.

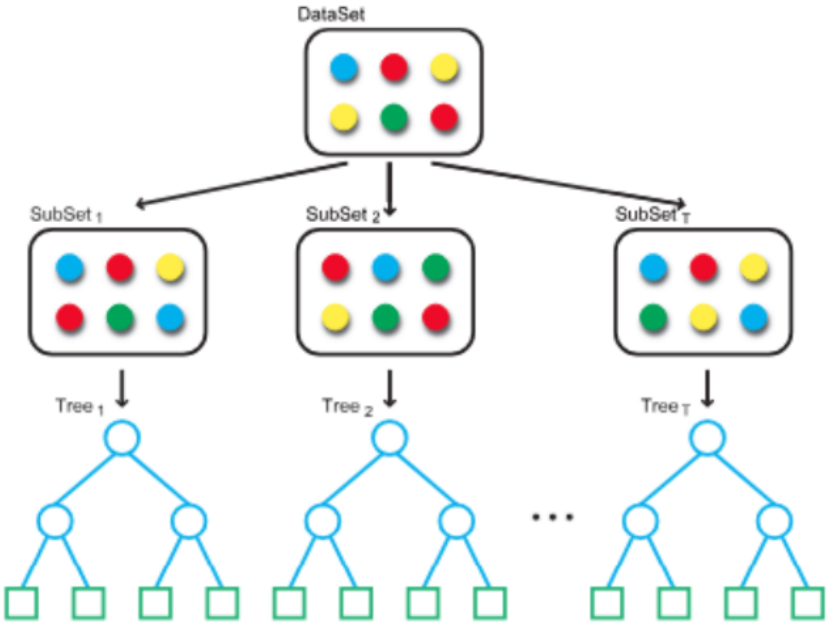


Figure 2.26: Bagging architecture. Multiple trees are used to improve performance. Each tree is trained in a different bootstrap.

2.5.3 AdaBoosting

In Boosting, the model is an ensemble of weak learners. They are called weak because their predictive power is only slightly better than random guessing. An example of a weak learner is a tree stump, which is a decision tree with a single split that makes its decision based on a single feature.

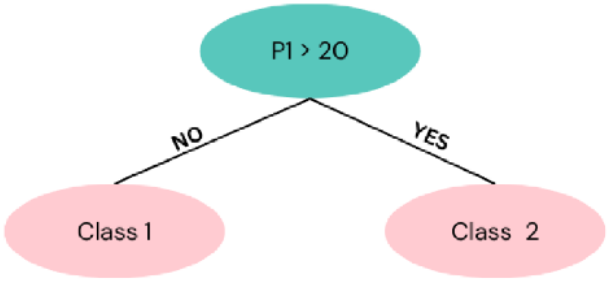


Figure 2.27: A stump tree with a single split.

In AdaBoosting, a sequence of weak learners is trained, with each one focusing on the misclassified data from the previous learner. This process aims to reduce misclassifications progressively. Each data point has a weight that influences the learner's decision, and the sum of all weights equals 1. For the next classification, the weights of the misclassified data points are increased. By doing this, the next learner emphasizes more on the previously misclassified data points. Each learner also has a weight related to the number of observations it classifies correctly. In the final step, the results of all learners are considered, but learners with higher weights contribute more to the final decision. The AdaBoosting algorithm can be explained as follows [32]:

1. For the first weak learner, all data points have equal weight, $w_i = \frac{1}{n}$.
2. For j in m boosting rounds, do the following:
   (a) Make the j-th input set, where its size is equal to the size of the original dataset, but the inputs with higher weight are repeated multiple times.
   (b) Train a weak learner C_j using the dataset created in the last step.
   (c) Apply C_j to the original training dataset.
   (d) Measure the performance of the weak classifier by computing the weighted error rate ϵ_j by summing the weights of the misclassified points:
   $$\epsilon_j = \sum_{i=1}^{n} w_i \cdot I(y_i \neq \hat{y}_i) \tag{2.25}$$
   where I is 1 if the prediction is incorrect and 0 otherwise.
   (e) Calculate the coefficient α_j for the weak learner, which measures the learner's influence in the final model:
   $$\alpha_j = 0.5 \cdot \log\left(\frac{1-\epsilon_j}{\epsilon_j}\right) \tag{2.26}$$
   (f) Update the weights to emphasize the misclassified points. This is done by:
   $$w_i = w_i \cdot \exp(-\alpha_j \cdot y_i \cdot \hat{y}_i) \tag{2.27}$$
   where y_i is the true label and $\hat{y}_i$ is the predicted label.
   (g) Normalize the updated weights so that they sum to 1.
3. Compute the final prediction $\hat{y}$ by considering the weighted prediction of all weak learners:
   $$\hat{y} = \text{sign}\left(\sum_{j=1}^{m} \alpha_j \cdot C_j(X)\right) \tag{2.28}$$

AdaBoost process illustrated. In Subfigure 1, the training set for binary classification is represented where all training samples are assigned equal weights. A decision stump is trained based on this training set (shown as a dashed line). In Subfigure 2, a higher weight is assigned to the two previously misclassified samples and the weights of correctly assigned samples are lowered. The second training stump will now focus on the samples with the highest weights. The weak learner shown in Subfigure 2 misclassifies three different samples, which will have higher weights, as shown in Subfigure 3. This AdaBoosting has 3 rounds of boosting, then the combination of these three weak learners by a weighted majority vote results in Subfigure 4.

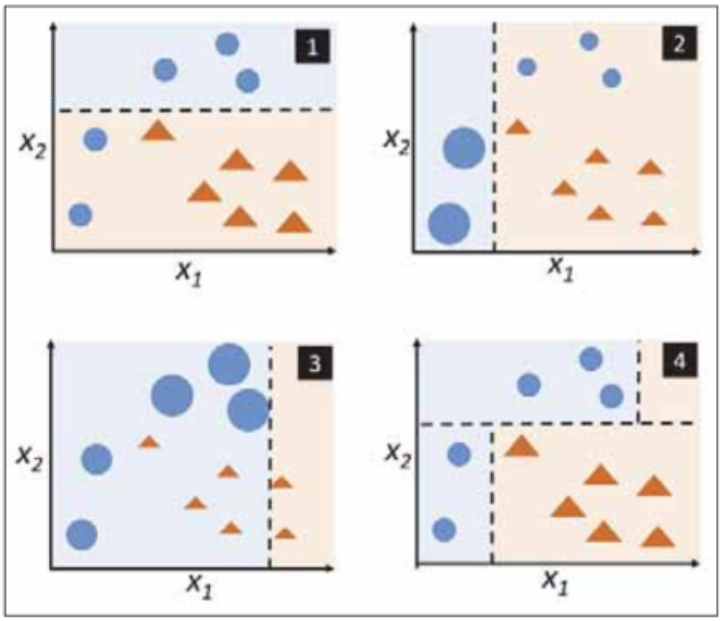


Figure 2.28: (1) The training set for binary classification is represented where all training samples are assigned equal weights. A decision stump is trained based on this training set (shown as a dashed line). (2) A higher weight is assigned to the two previously misclassified samples and the weights of correctly assigned samples are lowered. The second training stump will now focus on the samples with the highest weights. The weak learner shown in (2) misclassifies three different samples, which will have higher weights, as shown in (3). This AdaBoosting has 3 rounds of boosting, then the combination of these three weak learners by a weighted majority vote results in (4). [32]

2.5.4 Splitting data and Cross Validation

Cross-Validation is a technique used to prevent overfitting and to assess how the model generalizes to an independent dataset. One common method is k-Fold Cross-Validation, where the dataset is randomly divided into k groups (folds) of approximately equal size. One of the folds is used as the validation fold while the others are used as training folds. This process is repeated k times, each time with a different fold as the validation set. The error rate of k-Fold Cross-Validation is calculated as:

$$\text{k-Fold CV error} = \frac{1}{k}\sum_{i=1}^{k}\text{Error}_i \tag{2.29}$$

where Error_i is the error on the i-th validation fold.

Hyperparameter tuning is a crucial step in developing machine learning models. It involves selecting the best set of hyperparameters that control the learning process. In developing ML models, the dataset is first divided into a training set and a test set. The test set is set aside and used only for the final evaluation of the model to ensure unbiased performance metrics. Cross-validation is used in step hyperparameter tuning to ensure that the hyperparameters chosen provide the best performance on unseen data. By evaluating different sets of hyperparameters using cross-validation, it is possible to find the combination that optimizes the model's performance (Figure 2.29).

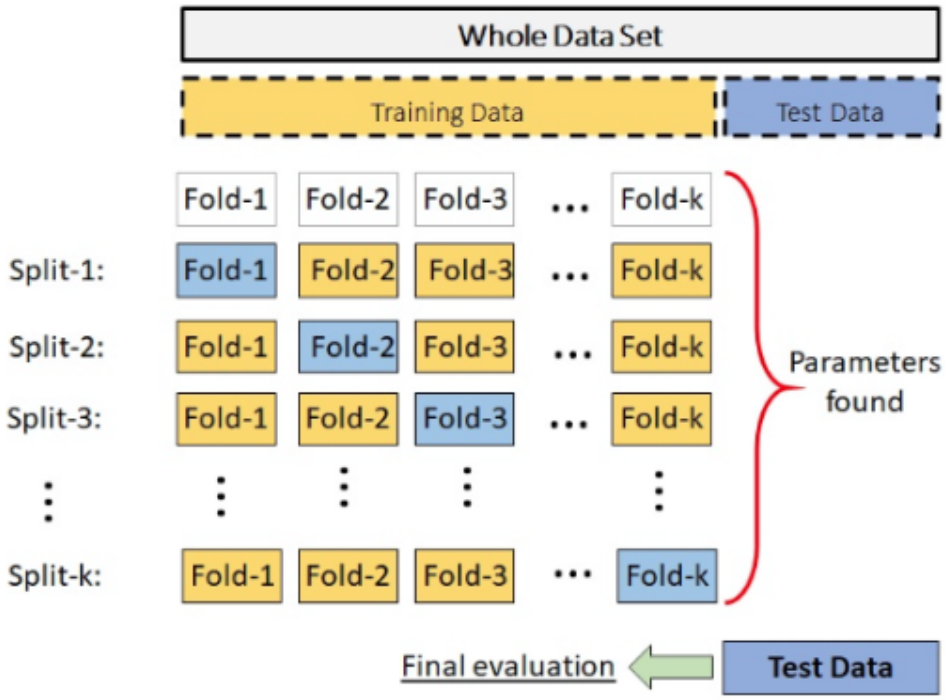


Figure 2.29: Cross-Validation architecture

2.5.5 Evaluation Metrics

Validation is important to evaluate how well a machine learning model performs. It involves testing the model on a separate dataset that was not used during training. This helps in understanding how the model will perform on new, unseen data. One common way to measure this performance is by using accuracy.

Accuracy is a simple metric that shows how often the model's predictions are correct. It is calculated by dividing the number of correct predictions by the total number of predictions. This gives a straightforward percentage of how many predictions were right.

The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2.30}$$

High accuracy means the model correctly predicts a large number of instances. This indicates that the model is performing well. Low accuracy suggests that the model is often incorrect, indicating poor performance.

Accuracy is easy to understand and use and most chemometric classification studies only report the accuracy of plastic sorting, which is useful as a general indication of the model performance.

2.5.6 Tackling Over fitting

Overfitting is a common challenge in supervised learning, where a model performs exceptionally well on the training data but struggles with unseen data. This phenomenon arises due to various factors, notably the limitations of training data and algorithmic complexity. The constraints imposed by the training dataset, including its limited size and the presence of noise, pose significant hurdles in achieving optimal generalization. Additionally, the complexity of algorithms, often characterized by numerous parameters, increases the risk of overfitting because models might memorize random details instead of learning the important patterns. A range of strategies has been proposed to address the multifaceted nature of overfitting. Early- stopping methods help stop the training process before the model focuses on random details, finding a balance between underfitting and overfitting. Network-reduction strategies, such as pruning, selectively eliminate less meaningful or irrelevant data, streamlining the model and enhancing interpretability. Furthermore, expanding the training data is an important strategy, as is adding more examples to the dataset to help fine-tune the hyperparameter in complex models effectively.

2.5.7 Wavelength Selection

When the number of features an algorithm must consider increases, it requires more computation and time to classify, while some features might be unimportant and their absence does not affect the classification result.

RFE

Recursive Feature Elimination (RFE) is a method to identify and remove these unimportant features.RFE is a backward selection method, which starts with the full model containing all potential features and then iteratively removes the least important predictor to enhance the model. The process of determining the importance of features, also known as variable ranking, typically involves the following steps:

1. **Train the model on the training set using all P predictors.**
2. **Calculate model performance.** This could involve metrics such as accuracy, or any other relevant performance measure depending on the problem.
3. **Calculate variable importance or rankings.** Variable importance is often determined by the model itself. For example:
   - **In linear models**, the importance of a feature can be determined by the magnitude of its coefficient.
   - **In tree-based models**, like decision trees or random forests, the importance of a feature can be measured by how much the feature decreases the measure of disorder (e.g., Gini index) across all the trees in the forest.
   - **In SVMs**, the importance can be inferred from the weight coefficients of the support vectors.
4. **For each subset size S_i, $i = 1, \ldots, S$, do:**
   (a) Keep the S_i most important variables. Based on the rankings, select the top S_i features.

Optional Pre-process the data. This could include normalization, scaling, or any other necessary data transformations.

(b) Train the model on the training set using S_i predictors. Re-train the model using only the selected subset of features.

(c) Calculate model performance. Evaluate the performance of the model with the reduced feature set.

Optional Recalculate the rankings for each predictor. If the rankings might change with the reduced set, recalculate the importance of the remaining features.

5. **End loop.**

6. **Calculate the performance profile over the S_i.** Analyze the performance of models with different numbers of features.

7. **Determine the appropriate number of predictors** (i.e., the S_i associated with the best performance).

8. **Fit the final model based on the optimal S_i.** Train the final model using the optimal subset of features.

Chapter 3

Experiments and Methods

The research design primarily relies on quantitative methodology, using data science techniques and machine learning methods to evaluate and improve the accuracy of three photonic devices for plastic type detection. Although the study is primarily quantitative, the research method has been used for steps such as identifying the best potential machine learning options, conducting detailed investigations of the results, and comparing them with previous work. Figure 3.1 presents a general graph of the methodology.

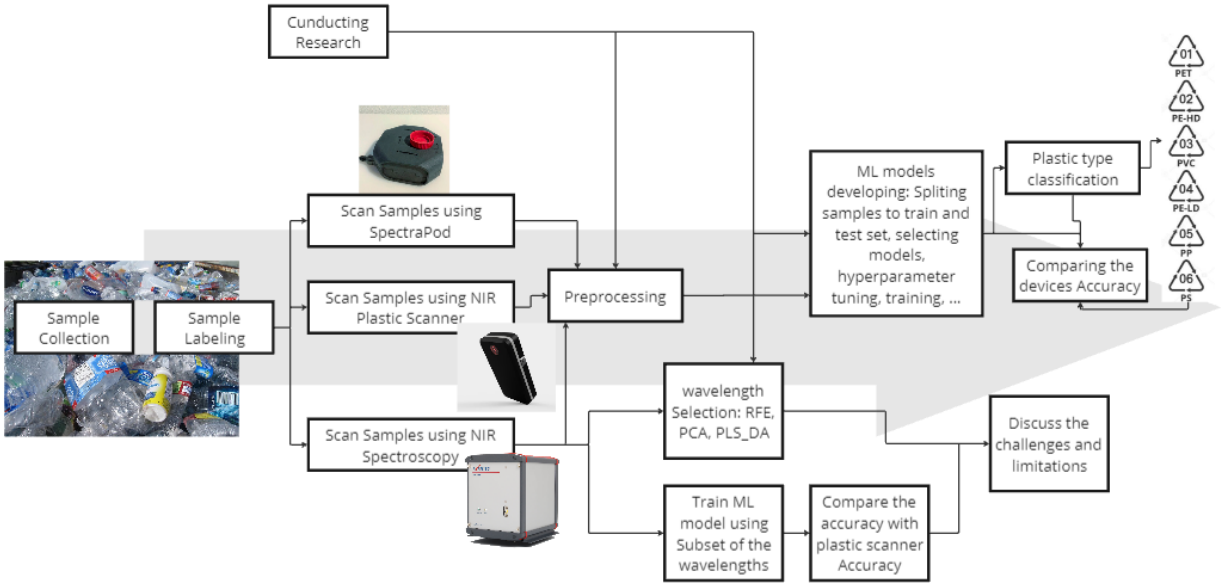


Figure 3.1: Methodology Overview

The study begins with primary data collection by gathering an appropriate number of samples for each of the six target plastic categories and directly acquiring spectral data from the plastic scanner, SpectraPod, and NIR spectrometer. Instead of relying solely on descriptive methods, device accuracy comparisons are made after identifying the best machine learning models for each device. Additionally, to determine the critical wavelengths, the study uses spectrometer data. The detected wavelengths are then compared to see if the plastic scanner and the SpectraPod cover these necessary spectral ranges. This approach aims to provide a detailed understanding of the devices' performance and wavelength coverage, offering valuable insights into plastic detection technology. In the next step, machine learning models are trained using a subset of wavelengths from the NIR spectrometer that are close to those covered by the plastic scanner,

and their accuracies are compared. This comparison helps determine whether the limiting factor is solely the coverage or lack of coverage of wavelengths, or if other issues, such as the accuracy of wavelength separation, are also significant.

In the following sections, these steps and the methods for carrying them out are explained in more detail.

3.1 Sample Preparing

For this research, the initial reference sample, Box A, provided by the Plastic Scanner team, is utilized. However, to thoroughly assess the accuracy of the plastic detection devices, additional samples are gathered. Increasing both the size and variety of the sample set is crucial, especially when applying machine learning methods. A larger and more diverse dataset improves the model's ability to generalize, leading to more reliable accuracy in real-world scenarios. Conversely, a limited sample set can result in overfitting or bias towards certain features , such as color, if most samples in one class share the same color.

To gather more samples, assistance has been sought from various sources, including students and lecturers from the master's program in NLE at The Hague University of Applied Sciences, as well as the Plastic Scanner team.

Regarding the machine learning approach, reliable labels for samples are crucial. One factor in collecting samples is ensuring they have the standard plastic code, as illustrated in Figure 2.1, which can be effectively used as labels for training the machine learning models.

After obtaining new samples, each one is assigned a specific code for the possibility of tracking their behavior, especially in terms of spectral data and training results. For instance, some samples might have flat spectra or pose challenges for the classification model in predicting their type. Using the code allows for tracking these samples and investigating their properties such as color, transparency, etc. This approach provides insights into the challenges and limitations faced. Table 3.1 summarizes the number of samples collected for each type of plastic.

Table 3.1: The number of collected plastic sample for each type of the plastic

	PET	HDPE	PVC	LDPE	PP	PS
Number of Samples	44	39	22	13	56	24

3.2 Data Collection

In the next phase, each device is utilized to capture spectra from every plastic sample. Each device operates with different software for controlling the device and collecting data. Specifically, the Plastic Scanner utilizes Psplot software, SpectraPod employs SpectraByte software, and the NIR spectrometer uses AvaSoft8. The size of features provided by the output data of each device is detailed in Table 2.1. The output of the plastic scanner and the NIR spectrometer consists of the intensity of the reflection for different wavelengths, while the output of SpectraPod is photocurrent for different channels.

During the data collection phase, each sample undergoes multiple scans, typically ranging from 4 to 10 times, using each device. These scans are conducted from various sides, orientations, and locations to ensure comprehensive data collection. Table 3.2 summarizes the number of the collected spectra, categorized by each plastic type.

Transparent samples, by nature, allow most of the light to pass through and reflect less during scanning. To address this, during scanning, the opposite side of transparent samples is consistently covered by the

Table 3.2: The number of scanned spectra from all samples of each type of plastic

	PET	HDPE	PVC	LDPE	PP	PS
NIR Spectrometer	171	161	128	56	239	128
SpectraPod	198	185	91	71	259	141
Plastic Scanner	168	182	83	65	203	111

reference tile. This method ensures maximum reflectance by providing a standardized surface against which the light can bounce back, thus optimizing the scanning process for transparent materials.

3.3 Data splitting

To ensure the validity and reliability of the model's performance, it is essential to properly split the data into training and test sets. This process must be done in a manner that avoids any overlap of information between the two sets, thus enabling an accurate assessment of the model's ability to generalize to unseen data. This involves several considerations:

- The data is divided based on unique sample codes. Each sample, along with all its associated spectra, is assigned to either the training set or the test set.
- Preprocessing steps, such as standard normalization along the wavelengths, should be fitted only on the training set. The parameters derived from this fitting (e.g., mean and variance) must then be applied to transform the test data to ensure consistent scaling.
- Data augmentation should only be done on the training set

In this project, the data split is performed with an 80:20 ratio, where 80% of the data is used for training and 20% for testing. To meet the first criterion of maintaining sample integrity, the splitting is done based on sample codes rather than directly on the spectra data. Here's the step-by-step process:

- Sample-based Splitting: The data is divided based on unique sample codes.
- Assignment of Spectra: Once the samples are divided, the spectra associated with these samples are used to form the training and test sets.
- Maintaining Class Ratios: The ratio of the number of samples for each class in the training and test sets is approximately aligned with the ratio in the total dataset. For instance, by comparing Tables 3.3 and 3.1, it can be seen that the class "LDPE" has a lower number of samples in the test set, reflecting its proportion in the overall data.

Figure 3.2 visually depicts the samples included in the test set.

Table 3.3: The number of plastic samples in the test set for each type of plastic

	PET	HDPE	PVC	LDPE	PP	PS
Number of Samples	7	7	5	3	10	6

Figure 3.2: Photo of test samples, including all six types of plastic (PET, HDPE, PVC, LDPE, PP, PS) in various colors such as green and gray. 3.1

3.4 Data Relabeling

As being mentioned, the aim is to detect six different types of plastics: PET, HDPE, PVC, LDPE, PP, and PS, which serve as the main categories for classification. However, certain challenges arise, particularly with samples that exhibit flat spectra due to factors like dark color, resulting in insufficient information to classify them accurately. Additionally, it is the project client's preference that if a sample has flat and noninformative spectra, the ML models can recognize them as plastic samples that cannot be detected instead of being detected without acceptable confidence.

To address this issue, a new class has been introduced to categorize such samples separately. This allows the machine learning model to differentiate samples with flat spectra as unrecognizable rather than attempting to assign them to one of the main six classes blindly, which could potentially reduce accuracy without adding any meaningful information. Table 3.4 provides an overview of the number of samples in each class after this new categorization. In this thesis, 'pu' type of category, which stands for "plus unknown"

Table 3.4: The number of samples in each category including: PET,HDPE, PVC, LDPE, PP, PS, Unknown

	PET	HDPE	PVC	LDPE	PP	PS	Unknown
Number of Samples	6	6	4	3	9	5	4

To identify samples with flat spectra, the spectra of each sample are plotted and compared with the mean spectra of samples from the same plastic type. This comparison aids in determining whether a sample should be categorized as unknown. Figures 3.3(b) and 3.3(c) illustrate examples of two samples with flat spectra. This approach ensures that samples lacking sufficient spectral information are appropriately handled, contributing to more accurate classification results.

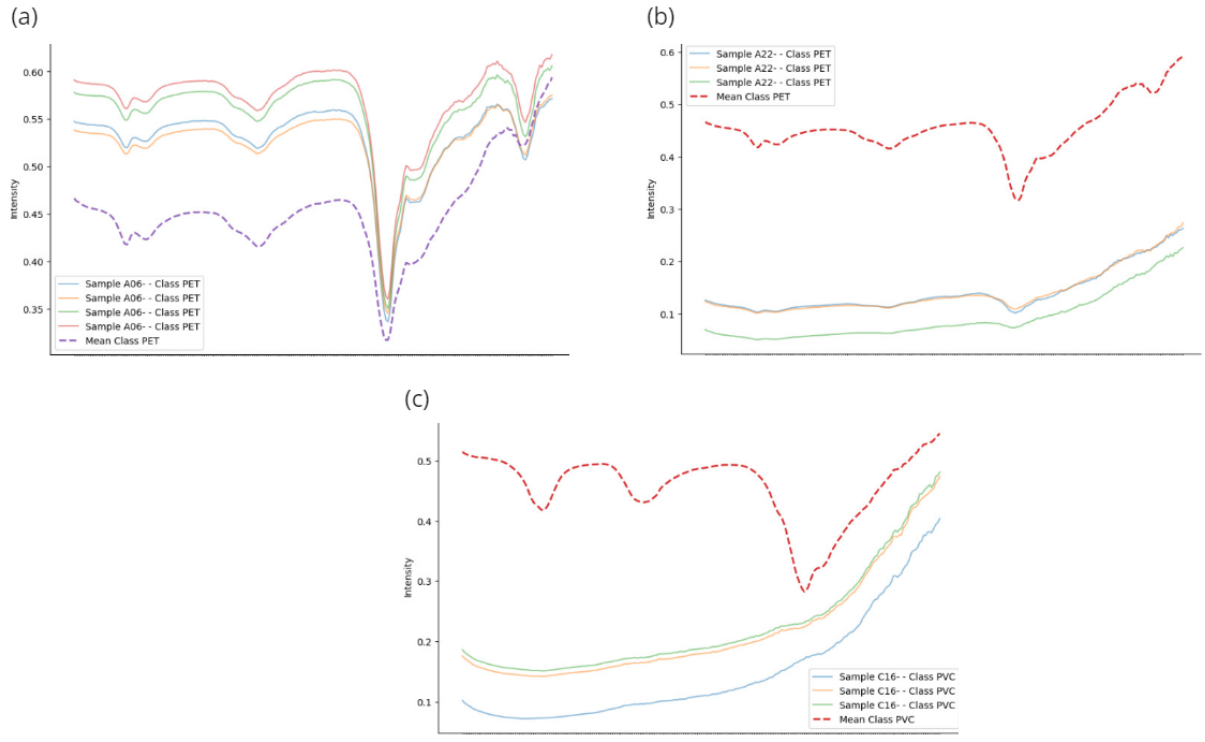


Figure 3.3: Example spectra of three samples, each plotted with the mean spectrum of all samples of the corresponding plastic type. This comparison is used to check if the spectra of the samples include meaningful peaks and deeps similar to the mean of the class. Samples without at least two similar peaks or deeps compared to the mean will be relabeled as "Unknown." (a) Spectra of a PET sample(A06, transparent) with strong peaks and deeps and the mean of spectra in PET category. (b) Spectra of a PET sample(A22, Black) which just includs a weak deep similar to the mean of spectra in PET category. (c) Spectra of a PVC sample(C18, Black) which is flat and the mean of spectra in PVC category.

IN addition, given the close chemical resemblance between LDPE and HDPE [35], distinguishing between them can be challenging. Moreover, obtaining an equal number of samples for LDPE proved difficult, resulting in imbalanced classes. However, since these plastics share a similar structure and can be recycled together [35], there's no significant advantage to classifying them separately, especially in recycling scenarios.

To address these challenges, the classes of LDPE and HDPE have been merged into a new category termed "PE". Table 3.5 provides the distribution of samples across the different classes for the newly obtained categories. In this thesis, 'cpu' type of category.

Table 3.5: The number of samples in each category including: PET,PE, PVC, PP, PS, Unknown

	PET	PE	PVC	PP	PS	Unknown
Number of Samples	6	9	4	9	5	4

3.5 Preprocessing

After preparing and scanning the samples, they should first be preprocessed. The initial step for the spectrum data is calculating Reflectance using equation 2.3, which includes removing the dark offset. While the NIR spectrometer and SpectraPod devices are designed to automatically remove the dark offset, this is not implemented by default in the Plastic Scanner. Therefore, some minor changes have been made in the Plastic Scanner's software source code and the Arduino board to scan the dark offset for the sample and reference tile.

In the next step, for removing the mean and rescaling the spectra, SNV and SN are separately applied, and the transformed data are used to train ML models (ANN and SVM). The final accuracy is compared to determine which method works better. Additionally, combining these two methods (SN and SNV) is also investigated to see if it can contribute to more accurate classification.

SG is another preprocessing method that is investigated. However, since this method smooths spectra data over several wavelengths (e.g., 15), it cannot be used for the Plastic Scanner and SpectraPod's output. The total size of their output is small, and smoothing over a large portion of the output can lead to loss of information.

Additionally, the SMOTE augmentation method is implemented to balance the size of classes (the number of spectra for each type of plastic + an "Unknown" category for the flat ones) and investigated by training models on them. PCA is also explored the same way

3.6 Training Classification Models

This research adopts an exploratory approach to determine the most effective algorithm for classifying plastic types. Several supervised models, including SVM, PLS_DA, ANN and RF are developed and tuned through hyperparameter optimization. Supervised models are preferred for classification tasks because they are trained using labeled data, allowing them to learn patterns and make accurate predictions based on the provided labels. Each algorithm is chosen for its demonstrated effectiveness in similar tasks.

The hyperparameter optimization process involves adjusting algorithmic parameters to achieve the highest accuracy in plastic type classification. Each model is tuned and trained separately for the Plastic Scanner, SpectraPod, and NIR Spectrometer. This strategy in model development and optimization seeks to identify the algorithm that excels in accurately categorizing plastics for each specific detection device. The models will be implemented using Python and Scikit-learn (Sklearn).

Scikit-learn also provides built-in functions like GridSearchCV for hyperparameter tuning. The GridSearchCV function is used to evaluate possible combinations of hyperparameters using cross-validation.

3.7 Evaluation

For comparing the models and devices performance, the accuracy is used. However as the number of approaches are a lot for being more readable the accuracy result will be presented by bar charts, and just the best one will be reported in tables.

3.8 Identification of Important Wavelengths

This section describes the methods used for selecting important wavelengths: PCA, PLS-DA, and RFE. Figure 3.4 provides an overview of the methodology.

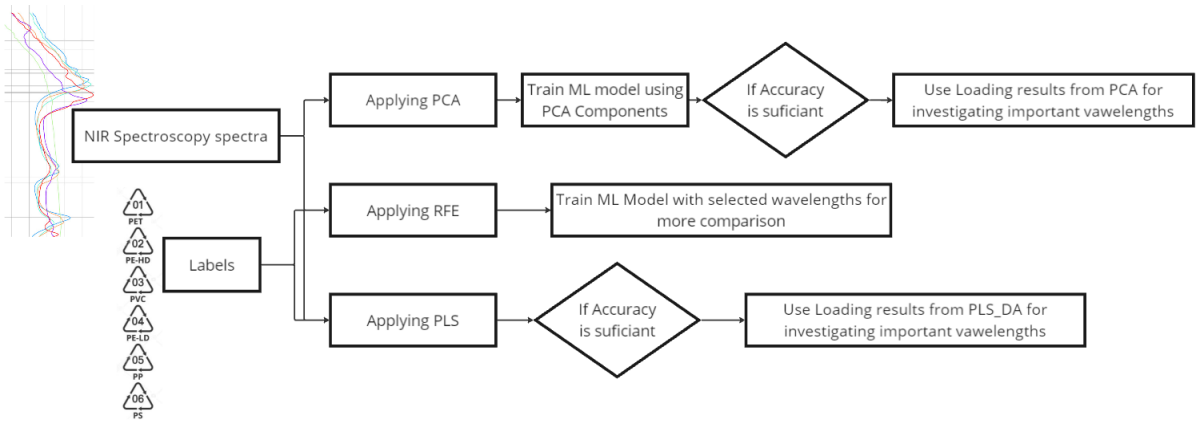


Figure 3.4: Methodology Of important wavelength selection by more details than Figure 3.1. To assess the adequacy of accuracy, two approaches are employed: comparing the resulting accuracy across the three methods and ensuring that the decrease in accuracy after wavelength selection does not exceed 0.15.

Principal Component Analysis

To identify significant wavelengths, PCA is applied to the training set, and the derived principal components are then applied to the test set. An ANN model is developed using these principal components, which are combinations of the original wavelengths. The criteria for evaluating the effectiveness of this method is that the accuracy should not decrease by more than 0.15 compared to the best-performing model. This threshold ensures that the model's performance remains acceptable while reducing the complexity of the wavelength data.

Partial Least Squares Discriminant Analysis

PLS-DA is developed and its accuracy is compared to other methods. PLS-DA demonstrates perfect accuracy (1.0) on the dataset, indicating its reliability in identifying relevant wavelengths. The loading data from PLS-DA are analyzed to determine which wavelength regions are significant for the model. PLS-DA is trained with a tuned number of components equal to 9, meaning there are 9 loading plots. Each plot shows which wavelengths are most significant for each component, providing insights into important wavelengths. However, this method does not offer the same level of granularity as RFE, which evaluates each wavelength individually.

Recursive Feature Elimination

RFE is also explored for wavelength selection. This method involves iteratively removing wavelengths and evaluating the model's performance using cross-validation. Even though RFE inherently assesses the impact of each wavelength by training models during the elimination process, it is essential to validate the selected wavelengths further. The reason is that RFE uses SNV filtering, which depends on the mean and variance of the spectra across all wavelengths. Hence, the information from a removed wavelength may still influence the remaining wavelengths through SNV filtering.

To ensure a thorough evaluation, after selecting the wavelengths, SNV is recalculated using only the selected wavelengths from the reflectance data and a new model is trained. This step guarantees that the wavelengths' information is accurately reflected in the new model.

For both the RFE process and the final evaluation step, Gradient Boosting is used as the model. This

choice is made because the Gradient Boosting implementation in Scikit-learn provides a feature importance attribute, which is crucial for implementing RFE and understanding the significance of each wavelength.

Comparison and Selection

After evaluating the accuracy of models using PCA, PLS-DA, and RFE, it is found that PLS-DA and RFE provide higher accuracy compared to PCA. Therefore, PLS-DA and RFE are chosen as the reference methods for wavelength selection in this study.

PLS-DA, while providing valuable insights into important wavelength regions through its loading plots, does not offer the same detailed analysis as RFE, which investigates each wavelength individually.

Chapter 4

Results and Discussion

This chapter presents and discusses the results to address the main research questions. The focus is on comparing the accuracy of different devices and identifying the most important wavelengths that influence their performance. This analysis helps in understanding the limitations of the Plastic Scanner and SpectraPod.

4.1 Data Insights

Looking at the mean spectra for each plastic class gives a basic understanding of the differences captured by each device. Plot 4.1 shows the mean spectra for each class using the Plastic Scanner, plot 4.2 does the same for the SpectraPod, and plot 4.3 shows the NIR Spectrometer data. Each plot uses different colors for the various plastic types, with error bars to show the standard deviation within each class. These plots help to see how consistent each device is and how much the readings vary.

The plots show that the Plastic Scanner has a relatively high standard deviation, meaning its readings for the same sample vary a lot. This variation is less in the SpectraPod, and even less in the NIR Spectrometer. The NIR Spectrometer spectra show clear, visible peaks and dips that help distinguish each plastic type. The most noticeable differences are seen between 1640 and 1740 nm, with additional useful information in the 1125 to 1235 nm range.

However, De Rijke's [25] research suggests that spectra using LEDs around 1700 nm and higher are noisier and less distinct (Figure 2.14). Conversely, in the 1125 to 1235 nm range, spectra illuminated by LEDs display more noticeable features. Comparing the mean plots of the SpectraPod output with the other devices is challenging due to the different output types: photcurrent for the SpectraPod and reflectance for the other two devices.

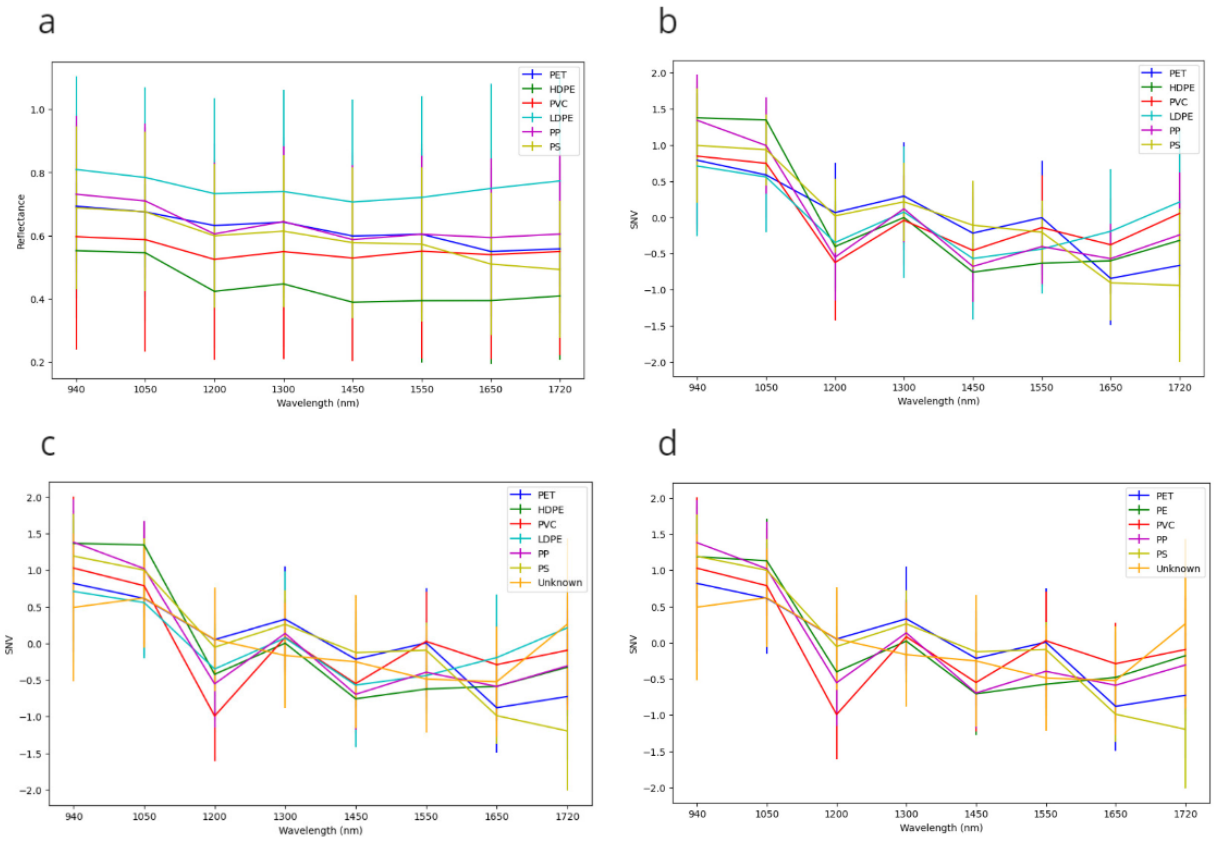


Figure 4.1: (a) mean reflectance of Plastic Scanner output per each plastic type. (b) The same, after applying SNV. As it is significant the deeps and peaks are enhanced. (c) The mean of SNV for relabeled spectra by separating flat spectra as a new class named ”Unknown”. (d) The same, after combining HDPE and LDPE as a same class named ”PE”.

Figures 4.1(b), 4.2(b) and 4.3 illustrate the effect of SNV on the spectra of the three devices. Three main changes are observed: 1. The offset between the spectra is removed. 2. The variance decreases. 3. The peaks and deeps are enhanced. The first two changes make the spectra more comparable for training models. Additionally, since the offset in spectra can be caused by factors like color, texture, transparency, or scattering, removing this offset means eliminating some irrelevant information for the project’s aim. The third change, by enhancing the peaks and deeps, improves the spectral features, aiding in better discrimination and classification.

Plots (c) and (d) in Figures 4.1, 4.2, and 4.3 respectively illustrate how the mean of each class changes by relabeling the classes of plastic types. Plot (c) demonstrates the effect of adding an ”Unknown” class, while plot (d) shows the impact of combining LDPE and HDPE into a single class labeled as PE.

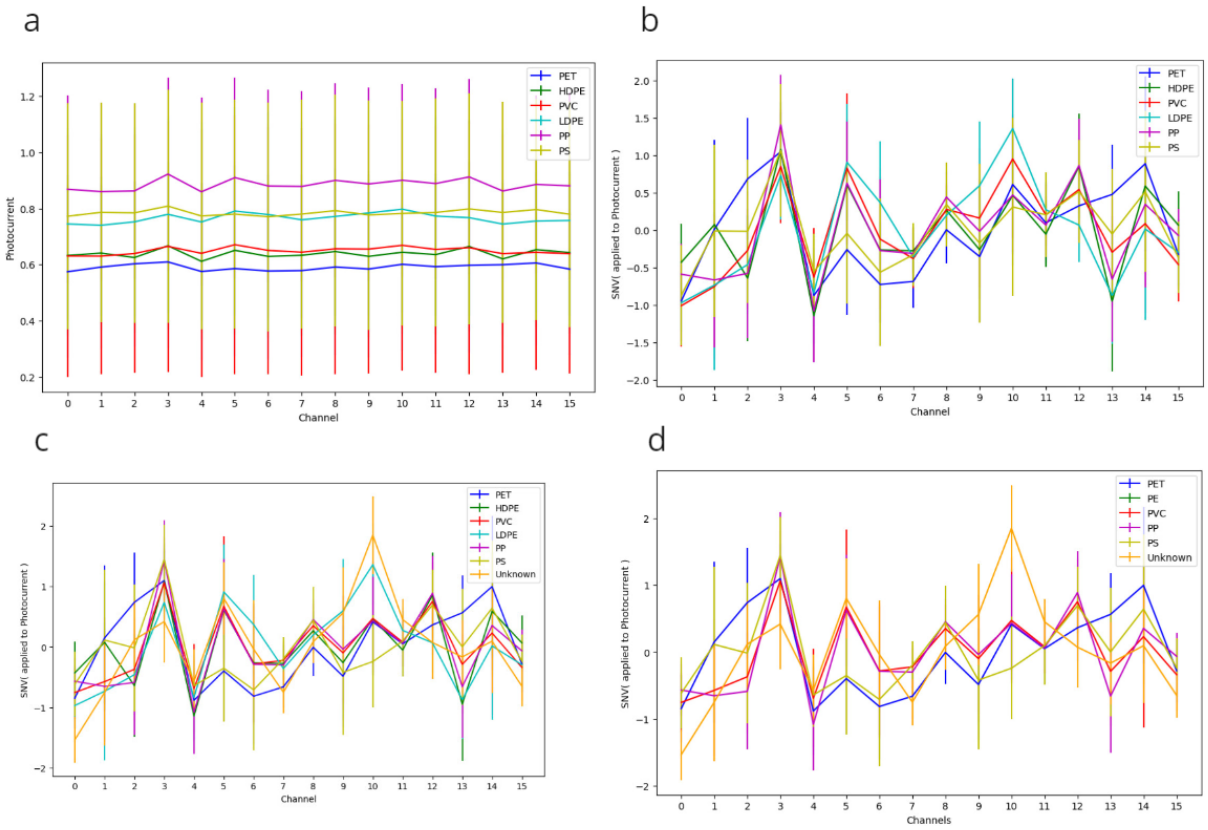


Figure 4.2: (a) mean reflectance of SpectraPod output per each plastic type. (b) The same, after applying SNV. As it is significant the deeps and peaks are enhanced. (c) The mean of SNV for relabeled spectra by separating flat spectra as a new class named ”Unknown”. (d) The same, after combining HDPE and LDPE as a same class named ”PE”.

It’s evident, for instance in Figure 4.1(c), that the deep in the PVS spectrum at 1200 nm is enhanced after separating flat spectra into a distinct class. Similarly, in Figure 4.3(c) for the NIR Spectrometer spectra data, particularly for PVC and PS, enhancements are noticeable, especially in the wavelength range of 1620 to 1730 nm. Though it’s not straightforward to interpret from spectra plots alone specially for SpectraPod data, the impact of relabeled classes can be examined by comparing the accuracy of models trained on spectra data with different labeling approaches.

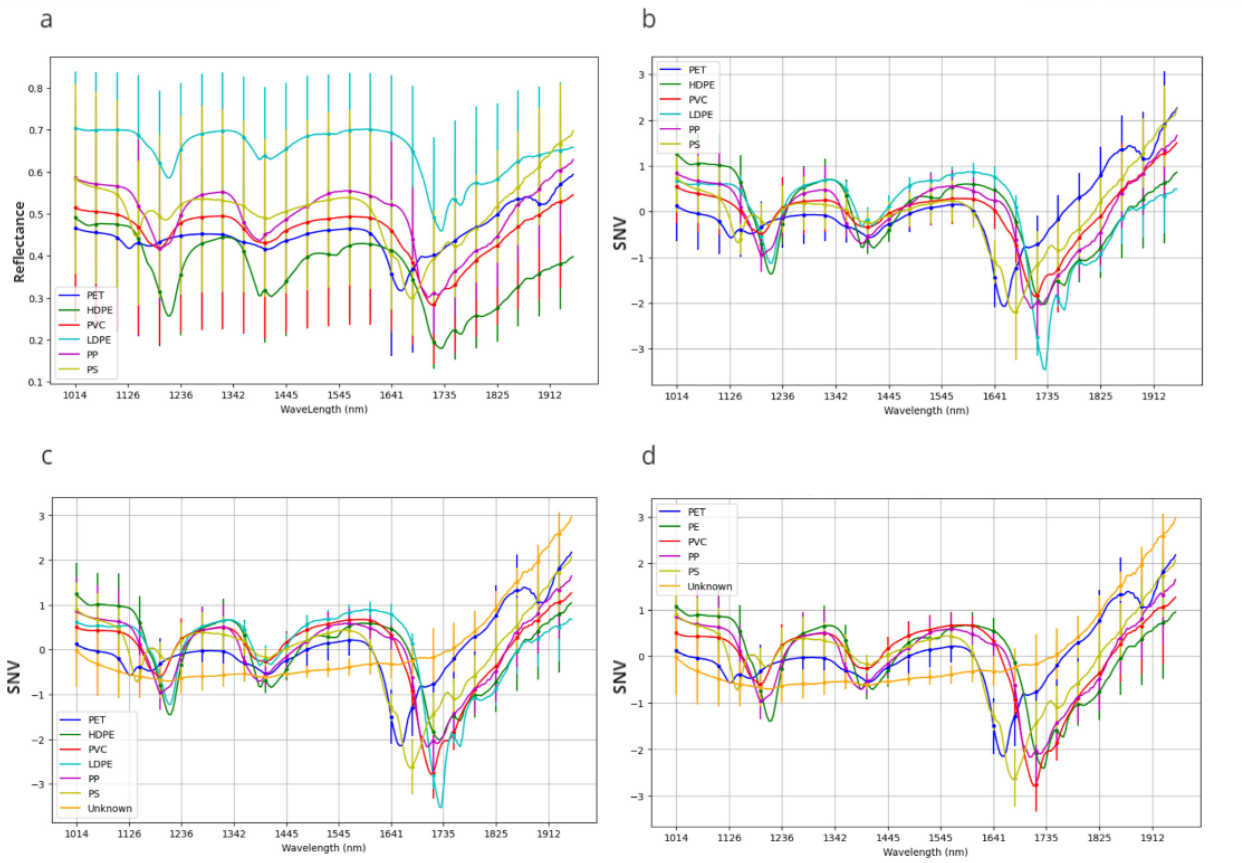


Figure 4.3: (a) mean reflectance of NIR Spectrometer output per each plastic type. (b) The same, after applying SNV. As it is significant the deeps and peaks are enhanced. (c) The mean of SNV for relabeled spectra by separating flat spectra as a new class named "Unknown". (d) The same, after combining HDPE and LDPE as a same class named "PE".

4.2 Classification

Different machine learning models have been developed, and their accuracies have been computed. Various preprocessing methods and class labeling strategies, as mentioned in the previous chapter, have been tried. To make it easy to compare and understand the results, the achieved accuracies are presented in Figure 4.4 (as a bar plot). Additionally, the highest accuracy achieved for each device and the model that resulted in that accuracy are shown in Table 4.1 for the regular six plastic type categories, and in Table 4.2 for the modified categories that include the Unknown class and the combined HDPE and LDPE class.

The results show that the best normalization method for the data from all three devices is SNV . The accuracy for the NIR spectrometer is higher than 0.9 for most of the models. However, the accuracy decreases when using SN and PCA. PLS-DA combined with SNV and CPU labeling results in 100% accuracy. Additionally, an accuracy of 0.98 is achieved with the ANN model when SNV and SG preprocessing are applied to NIR spectrometer data. However, this high accuracy does not mean the NIR spectrometer detects the type of all samples in test set; as plastics in the "Unknown" category are not assigned to the six plastic types. Instead, it indicates that this device can identify samples lacking enough information in the reflectance spectra and correctly classify the others.

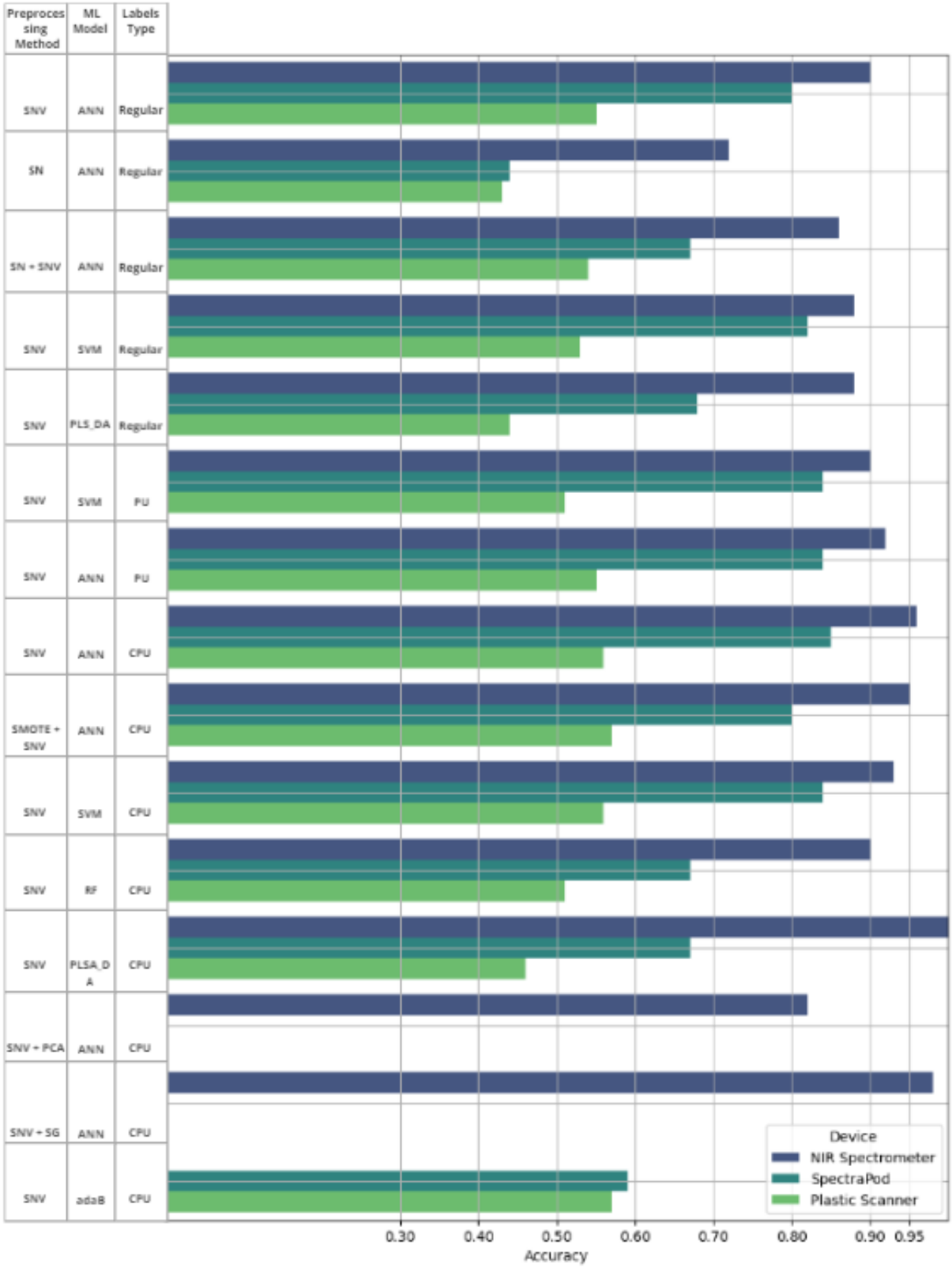


Figure 4.4: Overview of the implemented preprocessing and classification method performance (accuracy). The definitions of abbreviations can be found in the Abbreviations section.

Table 4.1: The Best for each device and its accuracy considering category including: PET,HDPE, PVC,LDPE, PP, PS

	Preprocessing Method	Model	Accuracy
NIR Spectrometer	SNV + SG	ANN	0.90
SpectraPod	SNV	ANN or SVM	0.74
Plastic Scanner	SNV	ANN	0.57

Table 4.2: The Best for each device and its accuracy considering category including: PET,PE, PVC, PP, PS, Unknown

	Preprocessing Method	Model	Best Accuracy
NIR Spectrometer	SNV	PLS_DA	1.0
SpectraPod	SNV	ANN or SVM	0.85
Plastic Scanner	SNV + smote	AB	0.58

Combining HDPE and LDPE into a single category and adding a new category for flat spectra increases the accuracy of the models. For NIR spectrometer data, PLS_DA yields the highest accuracy with CPU categories, but for regular classification of six common types of plastic, ANN achieves higher accuracy compared to other models. The Savitzky-Golay (SG) filter works well with NIR spectrometer data but is unsuitable for the other two devices due to their smaller output size and potential loss of meaningful information from smoothing.

For the SpectraPod, the highest accuracies are achieved by SVM and ANN. Overall, the accuracy of the Plastic Scanner is much lower compared to the other two devices, with its highest accuracy at 0.58 using ANN or AdaBoosting, indicating predictions are less than 60% correct. This highlights the need to identify its limitations and areas for improvement. The SpectraPod, with an accuracy of 0.85, shows promise for practical applications but still its accuracy is less than 0.95 which is the goal in research question. Identifying important wavelengths can further refine the SpectraPod's performance and suggest areas for improvement.

4.3 Wavelengths Selection

In this section, the aim is to identify important wavelengths that convey useful information from the spectra. We primarily focus on the NIR spectrometer data as a reference for applying wavelength selection techniques. The results obtained will be used to find the limitations of both the Plastic Scanner and SpectraPod.

The first approach involves the RFE technique, which systematically removes irrelevant wavelengths from consideration. This method helps in identifying the wavelengths that contribute most significantly to distinguishing between different plastic types. The RFE technique was employed three times in this study, progressively reducing the number of wavelengths from 237 to 135, then to 45, and finally to 13. During the phase of removing features, Gradient Boosting was utilized. This choice was made due to its feature importance score, a crucial factor for employing the Boosting function. By iteratively eliminating less relevant features, RFE aimed to identify the most significant wavelengths essential for accurate classification. Following the feature selection process, the trained model was evaluated using Gradient Boosting, ensuring consistency in methodology and facilitating direct comparison of the selected subsets of wavelengths. The results, presented in Figure 4.5, provide insights into the impact of wavelength reduction on model accuracy. By comparing the accuracy of models trained on the full set of wavelengths with those trained on the reduced subsets, we gain valuable understanding of the importance of individual wavelengths in classification.

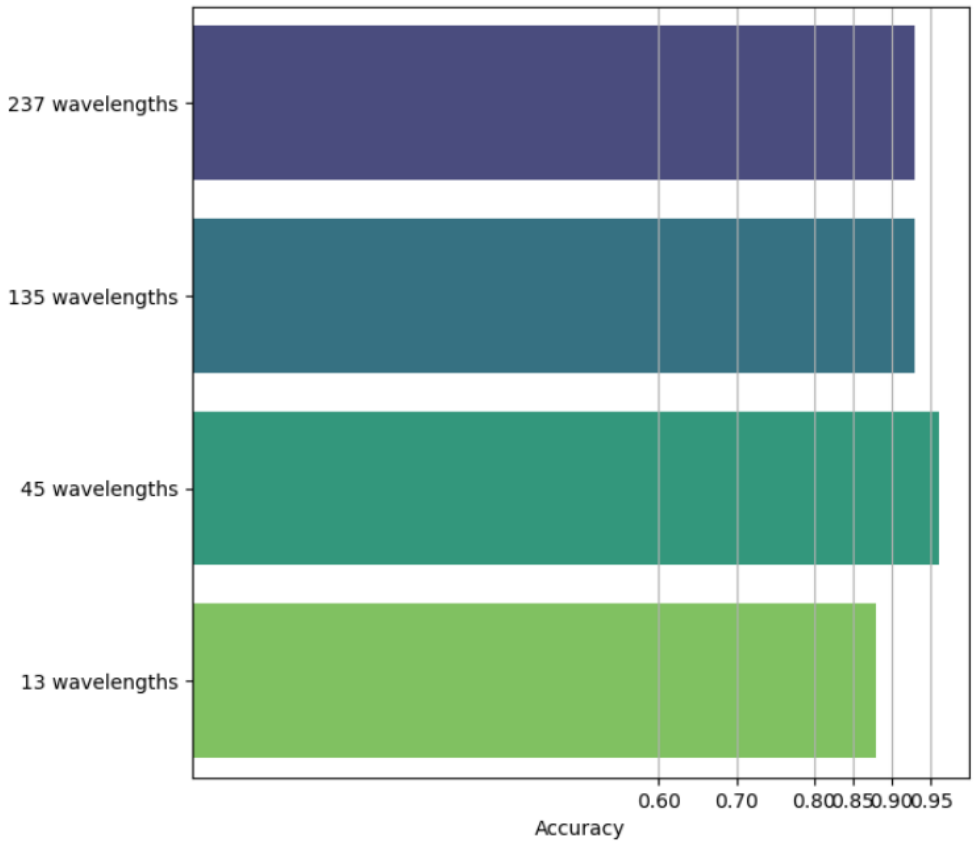


Figure 4.5: Comparing the accuracy of the model after applying RFE for different size of selected wavelengths set.

By comparing the accuracy of models trained on the full set of wavelengths with those trained on the reduced subsets, it becomes evident that eliminating some features not only fails to decrease the accuracy of

the model but, in some cases, even leads to an increase. For instance, when reducing the feature set to just 45 selected features, the accuracy of the model showed improvement.

The observed phenomenon where the accuracy of the model either remains unaffected or even increases after eliminating some features can be attributed to several factors. Firstly, the removal of irrelevant or noisy features during the feature selection process can lead to a reduction in overfitting. By focusing on the most informative features, the model becomes more generalized and better equipped to handle unseen data, ultimately improving accuracy.

Additionally, feature reduction can enhance the model's interpretability by simplifying the decision-making process. With fewer features to consider, the model may identify clearer patterns in the data, resulting in more accurate predictions.

Moreover, it's possible that some of the features removed during the RFE process were redundant or highly correlated with other features. Eliminating such redundant features can streamline the model and reduce computational complexity, potentially leading to improved performance.

Figure 4.6 illustrates the 45 selected wavelengths, showcasing the subset of features that contributed to the improved accuracy of the model.

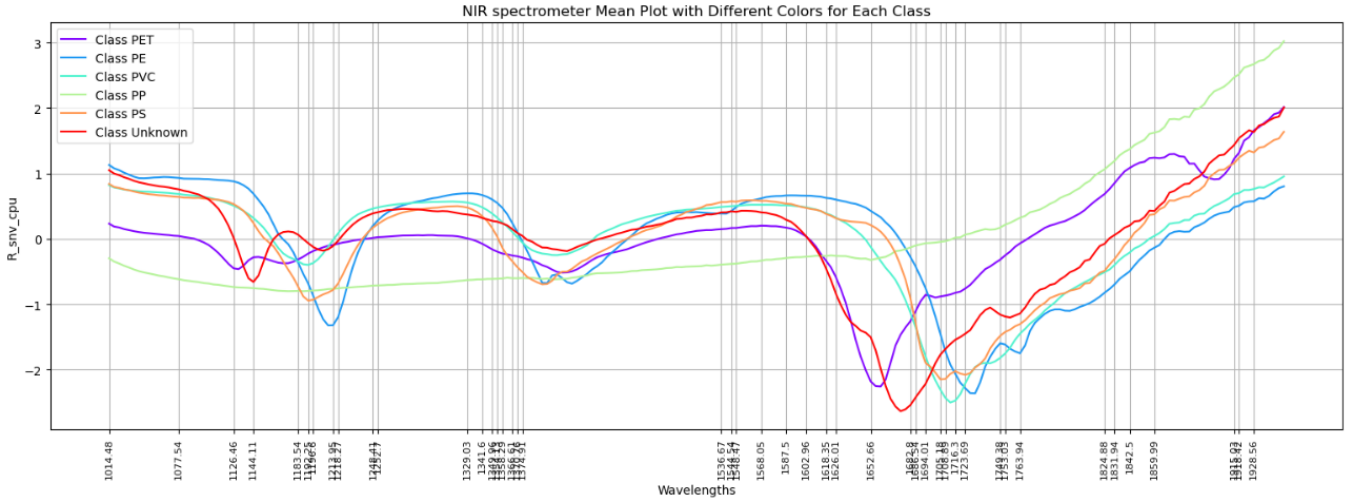


Figure 4.6: investigating the 45 selected wavelengths with the avrage mean spectra of each class

Furthermore, even when reducing the feature set to just 13 wavelengths, the accuracy declines only to 0.88. This decrease, although noticeable, still maintains a high level of accuracy, indicating that these 13 wavelengths carry a substantial portion of the useful information. This finding is particularly valuable for the Plastic Scanner, which covers only 8 wavelengths. The selection of these 13 wavelengths provides insight into which wavelengths are crucial to be covered, even when reducing the number of features to a very small subset. Figure 4.7 displays these 13 selected wavelengths, highlighting their significance in classification tasks.

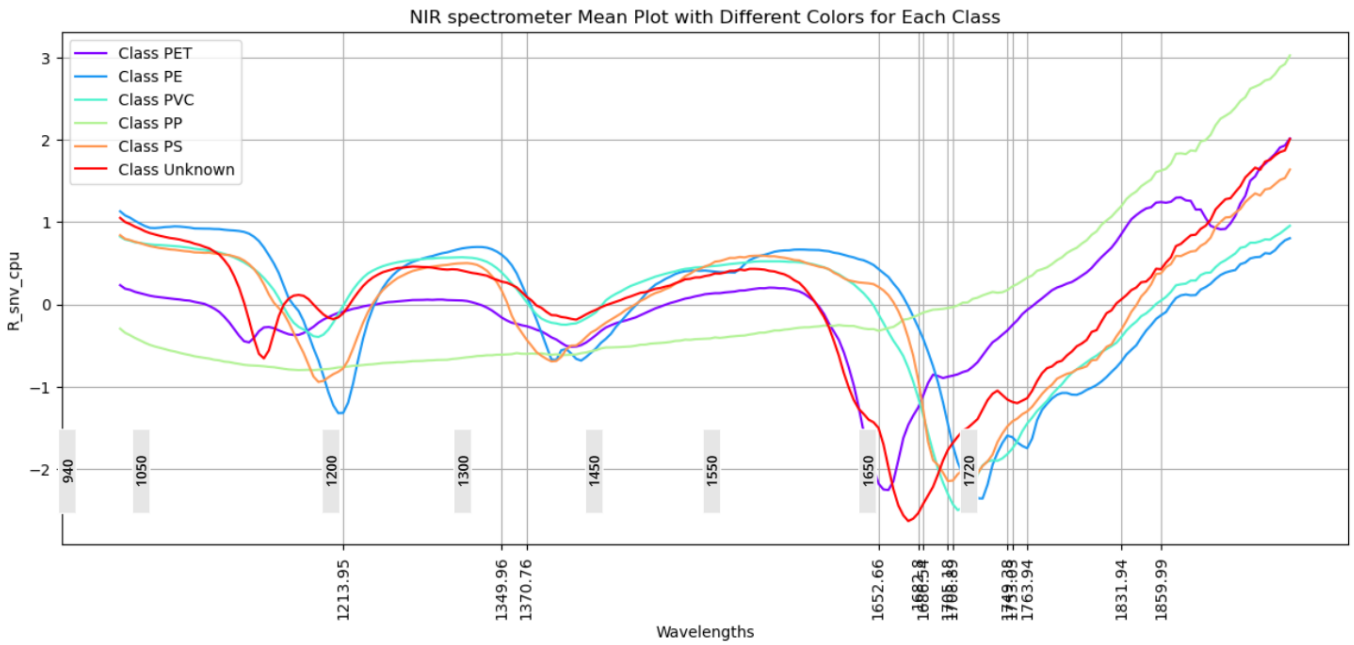


Figure 4.7: investigating the 13 selected wavelengths with the avrage mean spectra of each class

The results from Figure 4.6 indicate that certain wavelength ranges can be considered redundant and easily removed from the feature set. Specifically, the ranges from 1375 to 1537 and 1255 to 1329 are identified as such.

Moreover, the findings suggest that the importance of wavelengths extends beyond those with distinct peaks or deeps. Instead, there appears to be a need for a set of wavelengths in close proximity to these areas, with small differences between them (high resolution). This observation suggests that understanding the behavior of the spectra in these areas, rather than solely focusing on the magnitude of reflectance for a single wavelength, is crucial for accurate classification. Figure 4.7 confirms this as well. The selected wavelengths are mostly in the same areas, showing a tendency to pick wavelengths close together in short ranges rather than spreading them out across all wavelengths.

4.3.1 Limitations Analysis

For further comparison, the wavelengths covered by the Plastic Scanner are depicted in Figure 4.7 using gray rectangles. Interestingly, three of the LEDs have wavelengths very close to some of the selected wavelengths: 1200, 1650, and 1720. However, upon revisiting the InGaAs responsivity and examining the actual spectra of the wavelengths emitted by each LED, additional issues become apparent.

Firstly, InGaAs has no responsivity for wavelengths higher than 1700 and decreases sharply for wavelengths higher than 1650. Secondly, LEDs emit light across a broad distribution of wavelengths rather than illuminating a narrow range. Consequently, it becomes challenging to isolate the reflectance of closely spaced wavelengths to track the spectra's behavior in certain ranges, such as the area highlighted by the large yellow rectangle in Figure 4.9. Furthermore, the situation is compounded by the significant overlap between LED 1650 and 1720, making it even more difficult to distinguish between these wavelengths. Additionally, the small rectangle in the figure highlights an important gap in wavelengths.

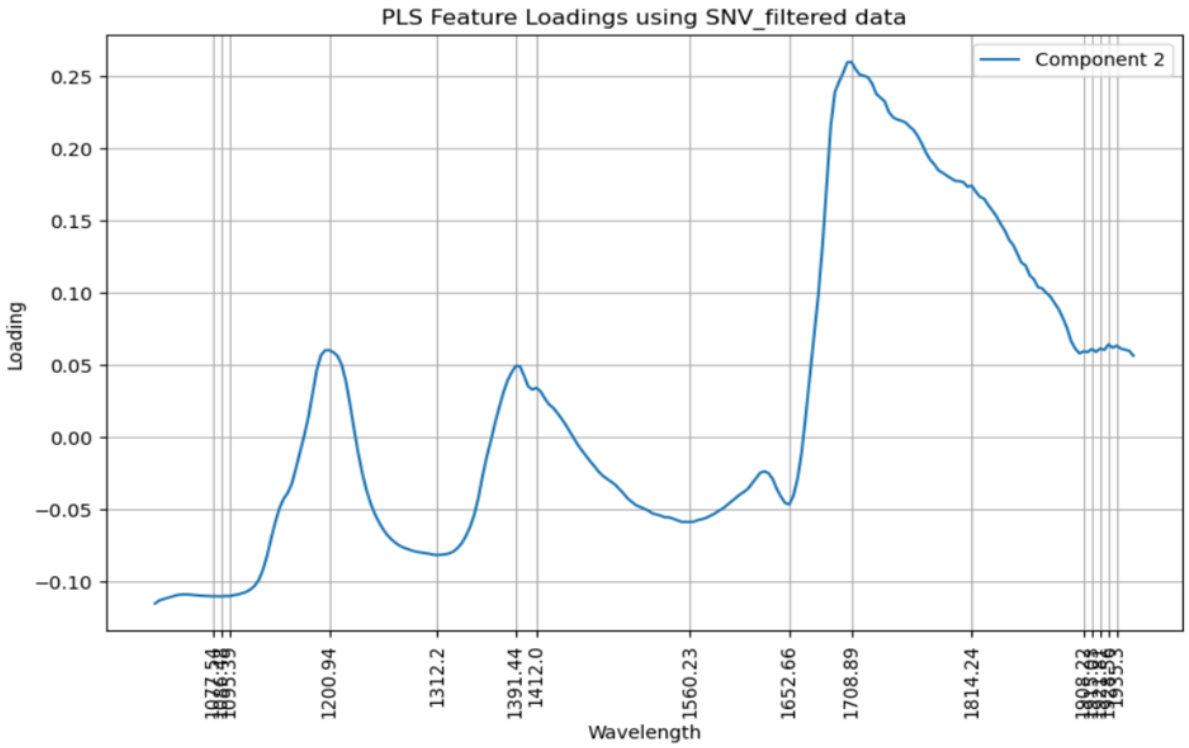


Figure 4.8: Plot of the loading results using the PLS approach, with the PLS model trained for 18 components. Here is the result for component 2. A strong peak around 1720 indicates that component 2 scores wavelengths in this range highly and recognizes it as an important range of wavelengths.

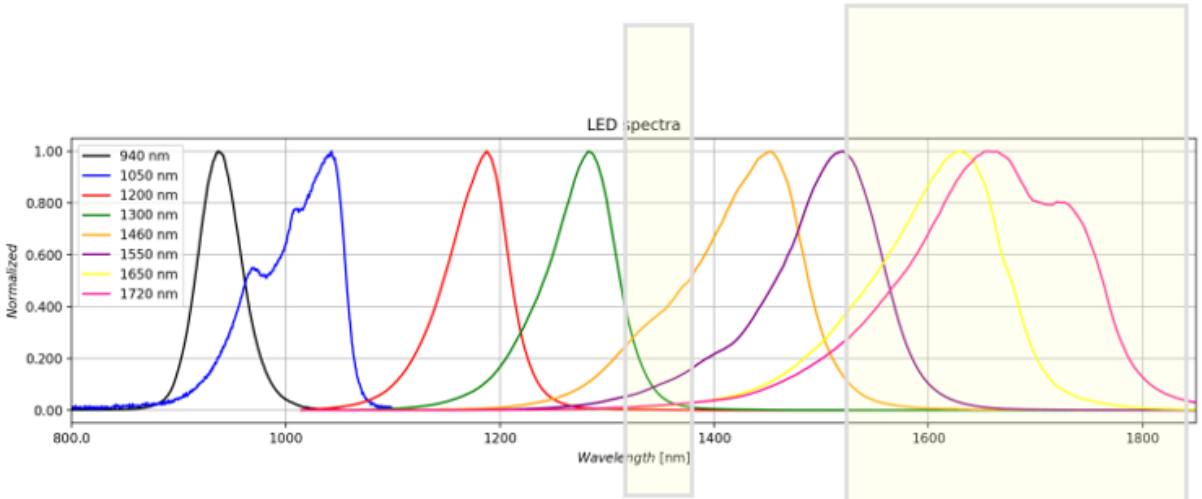
To further investigate the importance of a narrow range of wavelengths, another approach was taken: selecting the reflectance of wavelengths close to those used by the Plastic Scanner. An SVM model (one of the more accurate models) was trained on these selected wavelengths and used to predict test samples. The chosen wavelengths were 1050.62, 1200.94, 1299.53, 1448.69, 1548.47, 1648.86, and 1720. Since the NIR spectrometer wavelengths start from 1014, no wavelength close to 950 could be selected. Thus, the model was trained with 7 wavelengths instead of 8. Interestingly, the accuracy of this model was 0.776, significantly higher than the highest accuracy for the Plastic Scanner, which was 0.57.

This comparison shows that not only the specific wavelengths but also the precision of the device in detecting a narrow range of wavelengths is important. Additionally, it demonstrates that even with low resolution, as long as the wavelengths are narrow enough, some meaningful information can still be transformed for classifying plastic.

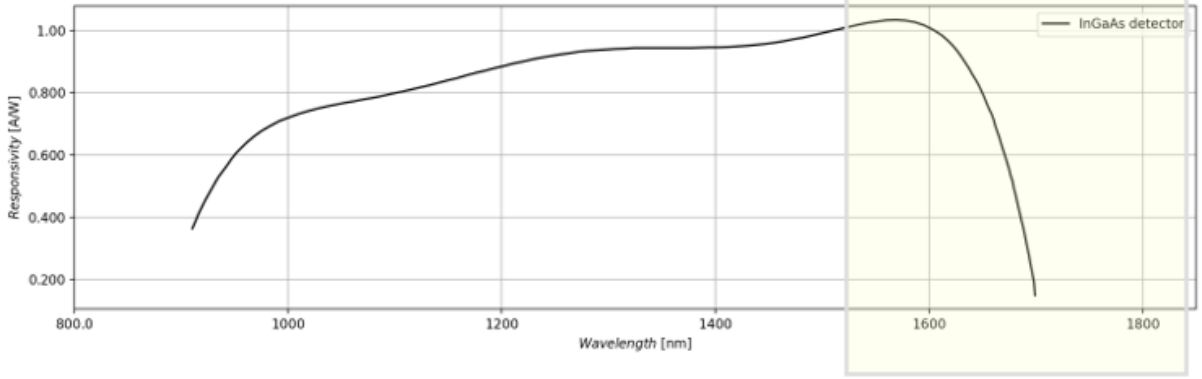
To summarize, the analysis shows that three factors contribute to the accuracy of the devices: covering important wavelengths (in both the light source and detector), isolating narrow wavelength ranges, and covering high resolution in some parts of the spectra (e.g., in the range 1650 to 1750).

To investigate SpectraPod's advantages and limitations in these areas, we can examine the responsivity plot of its 16 pixels for comparison (Figure 4.10). It is evident that SpectraPod, like the Plastic Scanner, has limited sensitivity for wavelengths higher than 1650. This means none of the pixels detect wavelengths around 1700 and higher, which is problematic since several important wavelengths fall in this range.

As mentioned in Chapter 3, the wavelength selection mechanism for SpectraPod is in the detector. Therefore, to investigate whether SpectraPod detects narrow wavelength ranges, we need to examine the responsivity of each pixel. According to the relevant equation, each channel is calculated by integrating



(a) The measured LED spectra plotted in a single graph. The reflectance has been normalised.



(b) The responsitivity of the InGaAs detector plotted with the extracted data from the data sheet.

Figure 4.9: Investigating plastic scanner limitations based on the LEDs spectra and InGaAs responsitivity [25]

the amount of corresponding pixel response across the light source's wavelength distribution. Figure 4.10 shows that each pixel has responsivity for a broad range of wavelengths. The existence of several peaks in the responsivity plots indicates that each channel detects wavelengths from multiple ranges rather than a continuous one, unlike the Plastic Scanner and NIR Spectrometer. For example, Pixel 1 has responsivity for two ranges: approximately 980 to 1160 and 1300 to 1580.

Comparing this to the Plastic Scanner, LED 1550 spectrum has significant intensity for wavelengths in the range 1400 to 1570 (Figure 4.9), while LED 1720 covers approximately 1450 to 1800. Although both the Plastic Scanner and SpectraPod detect narrow ranges similarly, the SpectraPod's pixels produce different (shifted) peaks of responsivity, revealing hidden patterns that can be leveraged by machine learning models to extract more detailed features.

The patterns produced by these peaks can also increase the resolution of the results, as they are shifted in each channel within a relatively small range of wavelengths. This shift allows for a more detailed analysis and enhances the ability to distinguish between different types of plastic, even when using broader wavelength ranges.

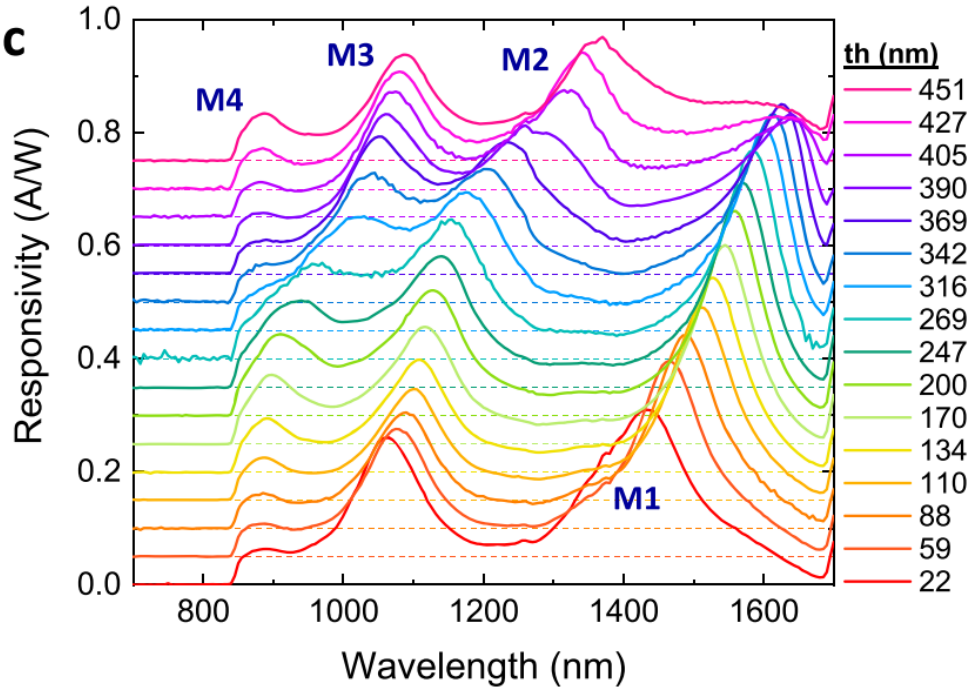


Figure 4.10: Note:The same approach can be taken to analyze the SpectraPod limitation [31]

4.4 Recommendations

To improve the accuracy of handheld devices, two main approaches can be taken: enhancing machine learning methods and upgrading the hardware of the devices.

4.4.1 Data Analysis

Although various machine learning and preprocessing methods have been investigated in this project, more modeling approaches can still be explored. One approach is to develop models(such as ANN), that can take multiple spectra from a single sample as input rather than analyzing each spectrum separately. For example, for the Plastic Scanner, this means classification will be done by analyzing several spectra of a sample simultaneously, and the prediction will be made after scanning the sample multiple times.

While this approach could increase the overall accuracy, considering the hardware limitations of the Plastic Scanner, the improvement might not be significant.

4.4.2 Hardware Upgrade

Adding LEDs

The first option for improving Plastic Scanner is to increase the number of LEDs or change the type of LEDs to ones with different wavelengths to cover gaps in the current spectrum. This is helpful not only because it can cover more wavelengths, but if the number of wavelengths is much larger and their spectra do not overlap around their peaks, then, similar to a spectrometer, the result can reveal hidden patterns. These

patterns can be trained by machine learning models to compensate for resolution and broad wavelength ranges.

Changing the combination of source light and detector

The second approach involves combining the sensor used in the SpectraPod with LEDs as the light source.

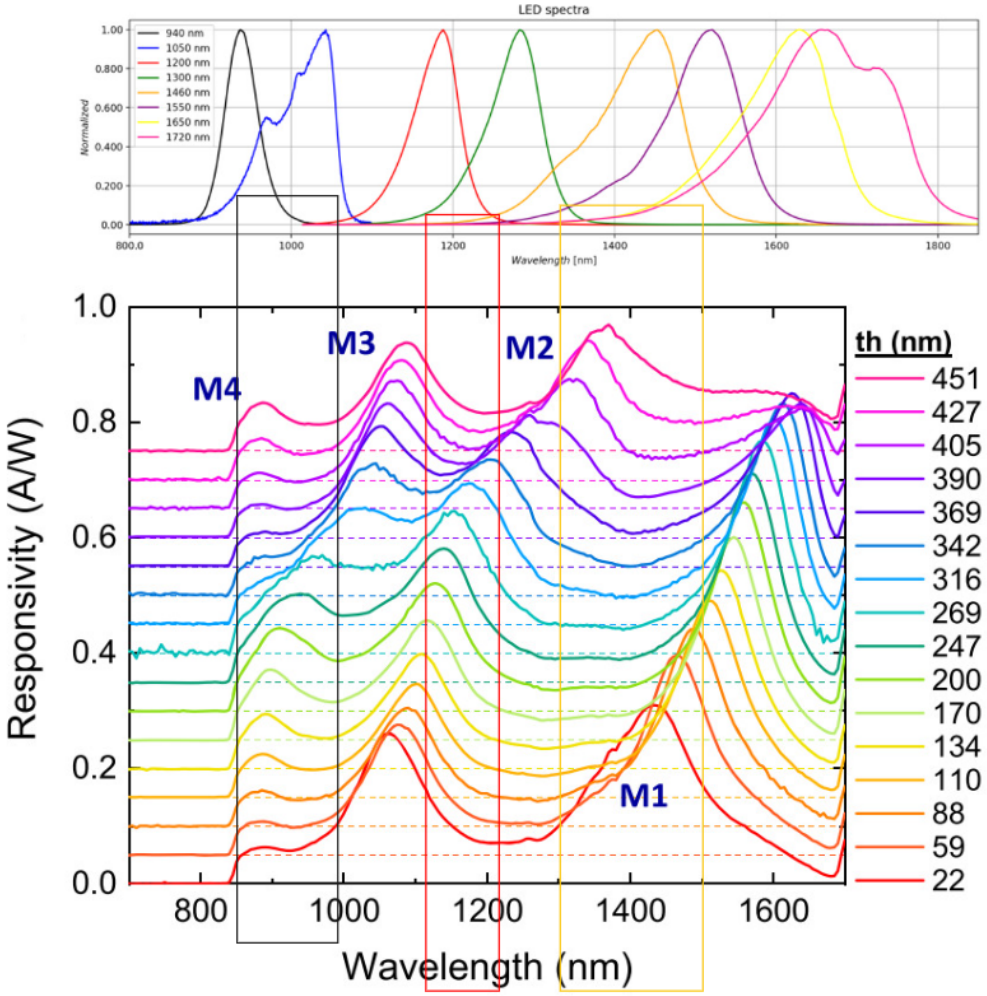


Figure 4.11: Explanation should be added

As mentioned in chapter 2, the amount of each channel is calculated by the equation 2.7. In the current SpectraPod, $R_i(\lambda)$ is the ith channel's corresponding pixel responsivity and $S(\lambda)$ is the distribution of the halogen lamp. Since both cover a wide range of wavelengths, the equation averages out a lot of information, causing the loss of some helpful details for more accurate plastic detection. In contrast, combining this sensor with LEDs instead of halogen lamps will decrease the range of wavelengths in which $S(\lambda)R_i(\lambda)$ is not zero, thus averaging out over smaller wavelength ranges. For example, as seen in Figure 4.11, channels using LED 1450 will average $S(\lambda)R_i(\lambda)$ only from 1300 to 1500 nm, as $S(\lambda)$ is almost zero for other wavelengths. The main idea is that in this way isolating the wavelengths is done in either source light and detector. It is expected that the accury will be higher than both of Plastic Scanner and SpectraPod. The main idea is that to isolate the wavelengths simultaneously in both the source light and the detector. This approach also increases the number of features. Instead of producing only one value for each LED, it shows

16 different values (16 channels) per LED, providing more information about the intensity for wavelengths in the range of that LED's spectrum. The total output will be 6 LEDs $\times$ 16 channels = 96 features, conveying more resolution and information about different types of plastic spectra. It is expected that the accuracy will be higher than that of both the Plastic Scanner and SpectraPod.

Chapter 5

Conclusion

This study aimed to find and compare the accuracy of three devices: Plastic Scanner, SpectraPod, and NIR Spectrometer. In the first step, we explored various ML learning and preprocessing techniques to enhance the accuracy of plastic classification using these devices. The models were trained on a dataset comprising 198 samples, including plastic samples with a variety of colors from white to black and even transparent ones. The results show that the NIR Spectrometer has an accuracy of about 0.90 in classifying six common types of plastic: PET, HDPE, PVC, LDPE, PP, and PS. SpectraPod and Plastic Scanner, with respective accuracies of 0.74 and 0.56, follow.

As flat spectra containing not enough information for detecting the type of plastics, models were also developed by relabeling samples into categories including: ”PET”, ”PE”, ”PVC”, ”PP”, ”PS”, and ”Unknown”. In these new classes, HDPE and LDPE are also combined together as class PE , as they have similar molecular structure. The result was an increase in accuracy: 0.1 for NIR, 0.85 for SpectraPod, and 0.58 for Plastic Scanner. However, still Plastic scanner and SpectraPod were less than 0.95. This results significant that Data science approaches can not increase the accuracy of these devices to more than 0.95, and there is a need for hardware changes.

As the accuracy of the NIR Spectrometer is high, its output can be used as a reference to investigate limitations for the other two devices by employing data analysis techniques like important feature selection to find more contributing NIR wavelengths for detecting plastic types. By employing RFE, we systematically reduced the number of wavelengths from 237 to 45 and 13, aiming to identify the most important wavelengths for accurate classification.

Our results reveales that reducing the wavelengths set did not necessarily lead to a decrease in accuracy. In fact, in some cases, the accuracy of the model even improved after eliminating certain features. This phenomenon can be attributed to the removal of irrelevant or noisy features, leading to a reduction in overfitting and improved generalization.

The selected wavelengths introduce some important wavelengths around and higher than 1700 nm, where the detectors of the two other devices have no responsivity and fail to detect in that range. Furthermore, our analysis underscored the importance of wavelength resolution in some wavelength ranges, while narrow wavelength ranges with small differences between them were crucial for accurate classification. By selecting these key wavelengths, even when reducing the feature set to a very small subset, we were able to maintain a high level of accuracy.

Achieving these results from analysis and investigating the responsivity of detectors and spectra of LEDs provides insight into the limitations of the other two devices. It appears that the Plastic Scanner cannot detect narrow ranges of wavelength reflectance, nor can it detect close wavelength reflectance separately (lacking high resolution). Even though SpectraPod improved this issue significantly by adding different pixels with varying responsivity to different wavelengths, some wavelengths' reflectance is still averaged out over broad

wavelength ranges in each channel.

It is recommended to overcome the limitations of handheld spectrometers and increase their accuracy; there is a need for hardware upgrades, such as increasing the number of LEDs or combining sensors with LEDs as the light source. These upgrades not only expand wavelength coverage but also improve resolution, leading to more accurate classification results.

In conclusion, our study demonstrates the significance of wavelength selection, resolution, and coverage in handheld spectrometers for plastic classification. By optimizing these factors, we can enhance the accuracy and effectiveness of plastic detection methods, ultimately contributing to environmental conservation efforts.

Bibliography

[1] Tony R Walker and Lexi Fequet. Current trends of unsustainable plastic production and micro (nano) plastic pollution. *TrAC Trends in Analytical Chemistry*, 160:116984, 2023.

[2] Roland Geyer, Jenna R Jambeck, and Kara Lavender Law. Production, use, and fate of all plastics ever made. *Science advances*, 3(7):e1700782, 2017.

[3] Amy L Brooks, Shunli Wang, and Jenna R Jambeck. The chinese import ban and its impact on global plastic waste trade. *Science advances*, 4(6):eaat0131, 2018.

[4] Joana C Prata, Ana L Patrício Silva, João P Da Costa, Catherine Mouneyrac, Tony R Walker, Armando C Duarte, and Teresa Rocha-Santos. Solutions and integrated strategies for the control and mitigation of plastic and microplastic pollution. *International journal of environmental research and public health*, 16(13):2411, 2019.

[5] Ana L Patrício Silva, Joana C Prata, Tony R Walker, Armando C Duarte, Wei Ouyang, Damià Barcelò, and Teresa Rocha-Santos. Increased plastic pollution due to covid-19 pandemic: Challenges and recommendations. *Chemical engineering journal*, 405:126683, 2021.

[6] Samaneh Karbalaei, Parichehr Hanachi, Tony R Walker, and Matthew Cole. Occurrence, sources, human health impacts and mitigation of microplastic pollution. *Environmental science and pollution research*, 25:36046–36063, 2018.

[7] Tony R Walker, Lei Wang, Alice Horton, and Elvis Genbo Xu. Micro (nano) plastic toxicity and health effects: Special issue guest editorial. *Environment International*, 2022.

[8] Xia Zhu. The plastic cycle–an unknown branch of the carbon cycle. *Frontiers in Marine Science*, 7:609243, 2021.

[9] Linn Persson, Bethanie M Carney Almroth, Christopher D Collins, Sarah Cornell, Cynthia A De Wit, Miriam L Diamond, Peter Fantke, Martin Hassellov, Matthew MacLeod, Morten W Ryberg, et al. Outside the safe operating space of the planetary boundary for novel entities. *Environmental science & technology*, 56(3):1510–1521, 2022.

[10] Stephanie B Borrelle, Jeremy Ringma, Kara Lavender Law, Cole C Monnahan, Laurent Lebreton, Alexis McGivern, Erin Murphy, Jenna Jambeck, George H Leonard, Michelle A Hilleary, et al. Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. *Science*, 369(6510):1515–1518, 2020.

[11] Jerry de Vos. Micro (nano) plastic toxicity and health effects: Special issue guest editorial. *https://repository.tudelft.nl/islandora/object/uuid*

[12] L Ki and C Chan. Separation of wasted plastics by thermal adhsion. J. Korean Inst. Resour. Recycl*, 4:44–50, 2003.*

[13] Markus Bauer, Markus Lehner, Daniel Schwabl, Helmut Flachberger, Lukas Kranzinger, Roland Pomberger, and Wolfgang Hofer. Sink–float density separation of post-consumer plastics for feedstock recycling. Journal of material cycles and waste management*, 20:1781–1791, 2018.*

[14] S Garry Howell. A ten year review of plastics recycling. Journal of hazardous materials*, 29(2):143–164, 1992.*

[15] Amar Tilmatine, Karim Medles, Salah-Eddine Bendimerad, Fodil Boukholda, and Lucien Dascalescu. Electrostatic separators of particles: Application to plastic/metal, metal/metal and plastic/plastic mixtures. Waste management*, 29(1):228–232, 2009.*

[16] Edward Ren Kai Neo, Zhiquan Yeo, Jonathan Sze Choong Low, Vannessa Goodship, and Kurt Debattista. A review on chemometric techniques with infrared, raman and laser-induced breakdown spectroscopy for sorting plastic waste in the recycling industry. Resources, Conservation and Recycling*, 180:106217, 2022.*

[17] Ning Liang, Sashuang Sun, Chu Zhang, Yong He, and Zhengjun Qiu. Advances in infrared spectroscopy combined with artificial neural network for the authentication and traceability of food. Critical Reviews in Food Science and Nutrition*, 62(11):2963–2984, 2022.*

[18] Alessandra Biancolillo and Federico Marini. Chemometric methods for spectroscopy-based pharmaceutical analysis. Frontiers in chemistry*, 6:576, 2018.*

[19] James Chapman, Vi Khanh Truong, Aaron Elbourne, Sheeana Gangadoo, Samuel Cheeseman, Piumie Rajapaksha, Kay Latham, Russell J Crawford, and Daniel Cozzolino. Combining chemometrics and sensors: Toward new applications in monitoring and environmental analysis. Chemical Reviews*, 120(13):6048–6069, 2020.*

[20] Georgina Sauzier, Wilhelm van Bronswijk, and Simon W Lewis. Chemometrics in forensic science: approaches and applications. Analyst*, 146(8):2415–2448, 2021.*

[21] Wenwen Zhang, Liyanaarachchi Chamara Kasun, Qi Jie Wang, Yuanjin Zheng, and Zhiping Lin. A review of machine learning for near-infrared spectroscopy. Sensors*, 22(24):9764, 2022.*

[22] Fang Ou, Anne van Klinken, Petar Ševo, Maurangelo Petruzzella, Chenhui Li, Don MJ van Elst, Kaylee D Hakkel, Francesco Pagliano, Rene PJ van Veldhoven, and Andrea Fiore. Handheld nir spectral sensor module based on a fully-integrated detector array. Sensors*, 22(18):7027, 2022.*

[23] Zongyin Yang, Tom Albrow-Owen, Weiwei Cai, and Tawfique Hasan. Miniaturization of optical spectrometers. Science*, 371(6528):eabe0722, 2021.*

[24] The puzzle of plastics recycling, by the numbers. https://www.cohenusa.com/blog/the-puzzle-of-plastics-recycling-by-the-numbers/.

[25] Queena L.D. Rijke. Characterisation of the LEDs and InGaAs detector of the handheld plastic scanner. *PhD thesis, The HAGUE University of applied science, Delft, 2023.*

[26] Xiaoli Chu, Yue Huang, Yong-Huan Yun, and Xihui Bian. Chemometric methods in analytical spectroscopy technology. *Springer, 2022.*

[27] Hamed Masoumi, Seyed Mohsen Safavi, and Zahra Khani. Identification and classification of plastic resins using near infrared reflectance. Int. J. Mech. Ind. Eng*, 6:213–220, 2012.*

[28] Frank L Pedrotti, Leno M Pedrotti, and Leno S Pedrotti. Introduction to optics*. Cambridge University Press, 2017.*

[29] Wenwen Zhang, Liyanaarachchi Chamara Kasun, Qi Jie Wang, Yuanjin Zheng, and Zhiping Lin. A review of machine learning for near-infrared spectroscopy. Sensors*, 22(24):9764, 2022.*

[30] Jerry de Vos. Plastic Identification Anywhere*. PhD thesis, Delft University of Technology, Delft, 2021.*

[31] Kaylee D Hakkel, Maurangelo Petruzzella, Fang Ou, Anne van Klinken, Francesco Pagliano, Tianran Liu, Rene PJ Van Veldhoven, and Andrea Fiore. Integrated near-infrared spectral sensor based on near-infrared detector arrays. In CLEO: Science and Innovations*, pages STu1A–4. Optica Publishing Group, 2021.*

[32] Sebastian Raschka. Pdf python machine learning: Machine learning and deep learning with python, scikit-learn, and tensorflow, by. 2017.

[33] Yiping Jiao, Zhichao Li, Xisong Chen, and Shumin Fei. Preprocessing methods for near-infrared spectrum calibration. Journal of Chemometrics*, 34(11):e3306, 2020.*

[34] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. An introduction to statistical learning*, volume 112. Springer, 2013.*

[35] DS Achilias, Ch Roupakias, P Megalokonomos, AA Lappas, and EV Antonakou. Chemical recycling of plastic wastes made from polyethylene (ldpe and hdpe) and polypropylene (pp). Journal of hazardous materials*, 149(3):536–542, 2007.*