UNIVERSITÀ DEGLI STUDI DI PAVIA
DIPARTIMENTO DI FISICA
CORSO DI LAUREA MAGISTRALE IN SCIENZE FISICHE

Understanding and Classifying Domestic Microplastics through Raman Spectroscopy and Machine Learning

Relatore:
Prof. Galinetto Pietro
Co-Relatore:
Prof. Cusano Claudio

Tesi di laurea di:
Canevari Simone

Anno Accademico 2024/2025

Contents

1 Abstract **4**

2 Introduction: The Challenge of Microplastics Pollution and Analytical Advancements **6**

3 The Micro/Nano-plastic water Pollution issue: social, environmental, health and economic aspects **8**
3.1 Plastics Type and Sources 8
3.2 Life Cycle of MPs 9
3.3 Effects of MPs on the Environment 9
3.4 Impact on the Terrestrial Ecosystem 10
3.5 Impact on the Aquatic Ecosystem 10

4 Human Exposure to Microplastics **11**
4.1 Routes of Exposure: Ingestion, Inhalation, and Dermal Contact 11
4.2 Health Effects of Microplastics 11

5 State of the Art in microplastics Detection and the Role of SERS **13**

6 Comparison of Raman and FTIR Spectroscopy for microplastics Detection **16**
6.1 Fundamentals and Comparison of Raman and FTIR Spectroscopy 16
6.2 SERS as an Enhancement of Raman Spectroscopy 18
6.3 Advantages of SERS over Conventional Raman Spectroscopy 18
6.4 Limitations and Challenges of SERS 18
6.5 Future Directions for SERS Development 19

7 The Physics of Raman Spectroscopy **20**
7.1 The Raman Effect 20
7.2 Mathematical Framework of Raman Scattering 21
7.3 Selection Rules and Symmetry 21
7.4 Advantages and Limitations of Raman Spectroscopy 22
7.5 Applications of Raman Spectroscopy 22

8 The Physics of Surface-Enhanced Raman Scattering (SERS) **24**
8.1 Fundamental Principles of SERS 24
8.2 Nanostructures in SERS 24
8.3 Theoretical Framework of SERS 25
8.4 Applications of SERS 25
8.5 Challenges and Future Directions 26

9 Physical and Chemical Properties of the Plastics Used **27**
9.1 Polyethylene Terephthalate (PET) 27
9.2 High-Density Polyethylene (HDPE) 28
9.3 Low-Density Polyethylene (LDPE) 29
9.4 Polypropylene (PP) 30
9.5 Polystyrene (PS) 31
9.6 Polyvinyl Chloride (PVC) 32

9.7 Other Plastics 33

10 The Machine Learning Approach **34**

11 Introduction to Machine Learning **35**
11.1 Classification in Machine Learning 36

12 Motivation and Application: ML for Microplastics Classification **37**
12.1 Why Use Machine Learning for Microplastics Classification 37
12.2 Advantages over Traditional Analytical Methods 37
12.3 Related Work and Previous Studies 38

13 Dataset Description **39**
13.1 Structure and Composition of the Dataset 39
13.2 Spectral Format and Feature Representation 40
13.3 Relevance for Supervised Classification 41

14 Preprocessing and Feature Engineering **42**
14.1 Methodological Choices and Rationale 42

15 Model Architecture and Training **43**
15.1 Chosen Model Type (SVM) 43
15.1.1 Handling Non-Separable Data 44
15.1.2 Non-Linear Separation and Kernels 44
15.1.3 Application to Raman Spectroscopy 45
15.2 Model Configuration and Hyperparameter Optimization 45
15.3 Training Strategy and Frameworks Used 45

16 Evaluation and Results **46**
16.1 Evaluation Metrics 46
16.2 Performance on Test Set 46
16.3 Cross-Validation and Generalization 46
16.4 Intra-Class Variability 47
16.5 Interpretation 47

17 Discussion of the Results **49**
17.1 Analysis of Model Performance 49
17.2 Visualization of Class Separability 49
17.2.1 Principal Component Analysis (PCA) 49
17.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE) 56
17.3 Spectral Contribution Analysis 57

18 Correlation Between Spectral Peaks and Model Weights **63**

19 Real-World Case Study: Classification of Unknown Mixed Samples **68**
19.1 Case 1: Ground vs Plastics 68
19.2 Case 2: Plastic Mixtures (Two or More Classes) 71
19.3 Case 3: TiO_2 Matrix vs Plastics 74

20 Conclusion **77**

1 Abstract

The pervasive diffusion of micro- and nanoplastics (MPs and NPs) across environmental matrices—including freshwater, marine ecosystems, soil, and atmospheric particles—has raised increasing concerns regarding their long-term ecological and human health impact. MPs are ubiquitous pollutants, that persist in the environment for decades, fragmenting into sub-micron particles, and entering biological systems through ingestion, inhalation, or dermal absorption. The complexity of their chemical composition, surface morphology, and environmental interactions makes their detection and classification particularly challenging. This thesis addresses these challenges by integrating physical analysis of microplastics with advanced spectroscopic and machine learning (ML) methodologies.

The first part of the work offers an interdisciplinary review of the mechanisms of plastic degradation, the dynamics of MP transport, and their interactions with biotic and abiotic components. It also discusses the societal and regulatory aspects of microplastic contamination, including the role of consumer habits, industrial practices, and policy frameworks. Following this contextual overview, a comparative assessment is presented between Raman and Fourier-transform infrared (FTIR) spectroscopy, with a focus on their respective advantages in microplastic detection. Surface-Enhanced Raman Spectroscopy (SERS) is also explored as a potential solution for overcoming sensitivity limitations of traditional Raman techniques.

The core experimental and computational work involved the development of a supervised classification pipeline based on Raman spectra collected from real-world plastic waste. More than 11,700 spectra were acquired from seven major polymer categories (PET, HDPE, LDPE, PVC, PP, PS, and Other), including both transparent and colored fragments. Each spectrum was standardized and treated as a high-dimensional feature vector, and a Support Vector Machine (SVM) classifier was trained using a linear kernel. Hyperparameter tuning was conducted via grid search on the training set, and the model achieved a classification accuracy of 99.49% on a held-out test set, with cross-validation accuracy exceeding 99.7%. These results confirm the discriminative power of Raman spectra, even in the presence of variability due to pigmentation and heterogeneous sample quality.

Beyond performance evaluation, interpretability and robustness were central to the model analysis. Dimensionality reduction techniques (PCA, t-SNE) were applied to visualize spectral separability between classes. The learned weight vectors from the SVM were examined to identify the most influential wavenumbers, providing physical insight into which spectral regions contribute most to each classification decision. A comparative analysis between average spectra and SVM weights was also conducted, revealing that key Raman peaks often coincide with high-weight regions in the model. This strengthens the link between spectral features and polymer-specific chemical signatures, demonstrating that the classifier does not rely solely on statistical artifacts but captures chemically meaningful patterns.

To test the model's applicability in field-like scenarios, additional unlabeled samples were collected and analyzed. These included mixtures of plastics and ground material, binary plastic blends, and plastic fragments embedded in TiO_2 matrices. By introducing a dedicated "ground" class and analyzing the predictions through confidence thresholds and t-SNE visualization, the model was shown to retain strong discriminative capacity even under ambiguous or noisy conditions. This exploratory application illustrates the system's potential for rapid screening of environmental or industrial samples, supporting

further automation in microplastic monitoring.

In conclusion, this thesis demonstrates the feasibility and relevance of combining Raman spectroscopy with interpretable machine learning for the classification of microplastics. The approach offers an efficient and scalable analytical tool that can support laboratory workflows and environmental monitoring efforts, and it highlights the value of data-driven methods in advancing material identification in complex ecological contexts.

2 Introduction: The Challenge of Microplastics Pollution and Analytical Advancements

The widespread presence of microplastics (MPs) and nanoplastics (NPs) in natural environments has emerged as a critical global concern, impacting ecosystems, public health, food economics and food safety. Global plastic production exceeds 350 million tons annually, with a significant portion accumulating in the environment due to inadequate waste management and accidental release [1]. This persistent contamination results in at least 14 million tons of plastic entering oceans yearly, representing 80% of marine debris from surface waters to deep-sea sediments [1]. MPs and NPs, defined by their minute sizes, exacerbate this crisis through their high bioavailability and propensity to adsorb toxic contaminants, forming biofilms and releasing harmful substances into organisms [1].

Human exposure to MPs occurs via ingestion, inhalation, and dermal absorption, leading to their detection in human tissues (see ch. *Human Exposure to Microplastics*) including the bloodstream [1]. Such exposure raises severe concerns regarding long-term health implications, including the accumulation of plastic-derived toxins and their potential interactions with biological systems [2]. Furthermore, MPs may act as vectors for persistent organic pollutants, heavy metals, and pathogens, complicating their ecological and health impacts [1].

Recent studies have shown that MPs not only persist in aquatic environments but are also ingested by marine organisms, leading to bioaccumulation along the food chain [2]. This phenomenon highlights the urgent need for comprehensive and scalable detection methods that can operate effectively across diverse environmental matrices. Regulatory frameworks addressing MPs remain nascent, particularly for secondary microplastics resulting from the fragmentation of larger plastic debris [1]. Despite growing awareness, existing legislation—such as the EU Directive 2020/2184—highlights the lack of structured protocols for sampling, detecting, and quantifying MPs and NPs in water systems and food chains [1]. The directive emphasizes the urgency of standardizing methods for emerging pollutants like MPs by 2024 to safeguard human health and ecosystems [1].

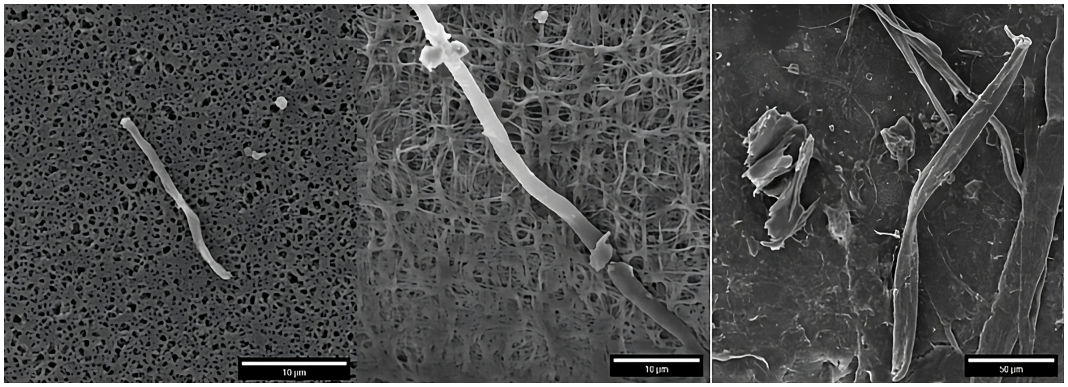

Figure 1: TESCAN scanning electron microscopy (SEM) image of examples of micro or nanoplastics in drinking water. Scales: 10 μm, 10 μm , 50 μm [1]

[1] https://www.labcompare.com/10-Featured-Articles/587247-Microplastics-in-Drinking-Water-SEM-Investigation/

This thesis investigates the application of machine learning techniques for the classification of Raman spectroscopy data from plastic samples belonging to the seven standard recycling codes (01 to 07). The experimental work focused on the collection and spectroscopic analysis of both transparent/white and colored plastic fragments, aiming to build a comprehensive dataset representative of real-world variability.

The core of the research lies in the development, training, and interpretation of supervised machine learning models—specifically Support Vector Machines (SVMs)—to accurately identify polymer types based on their vibrational spectra. This includes all stages of the data analysis pipeline, from preprocessing and feature engineering to model evaluation and interpretation of spectral relevance.

While the spectroscopic techniques employed (Raman and Surface-Enhanced Raman Scattering, SERS) are well-established, the innovative aspect of this work consists in the integration of ML approaches to automate and enhance the classification of microplastic particles.

The thesis also presents background chapters on the physical principles of Raman and SERS, the environmental and health impact of microplastic contamination, and the chemical properties of the most common polymers encountered in consumer waste.

Ultimately, the goal is to provide actionable insights and tools to support regulatory bodies and environmental stakeholders. This framework aims to ensure sustainable management of plastic pollution while fostering methodological innovation in environmental diagnostics.

3 The Micro/Nano-plastic water Pollution issue: social, environmental, health and economic aspects

Microplastics (MPs) are pervasive pollutants derived from the degradation of plastic materials or directly introduced into the environment as small particles. Their persistence, widespread distribution, and ability to accumulate in organisms make them a global environmental challenge. MPs interact with physical, chemical, and biological systems, leading to complex environmental and health implications.

3.1 Plastics Type and Sources

Plastics are classified based on their size and origin into two main categories:

- **Primary MPs**: These are intentionally manufactured small particles, such as plastic pellets used in industrial applications, microbeads found in cosmetics and personal care products, and microfibers released from synthetic textiles during washing. Their direct introduction into ecosystems often bypasses filtration systems, making them a significant contributor to environmental contamination [3].
- **Secondary MPs**: These are formed by the fragmentation of larger plastic debris due to physical, chemical, or biological processes. Photodegradation under UV light, oxidative weathering, and mechanical abrasion are key contributors to this process, particularly in marine and terrestrial environments [3].

The most common polymers found in MPs include polyethylene (PE), polypropylene (PP), polystyrene (PS), polyvinyl chloride (PVC), high-density polyethylene (HDPE) and low-density polyethylene (LDPE). These materials dominate applications in packaging, construction, and consumer goods. MPs are introduced into ecosystems through industrial discharges, urban runoff, mismanaged waste, and atmospheric deposition.

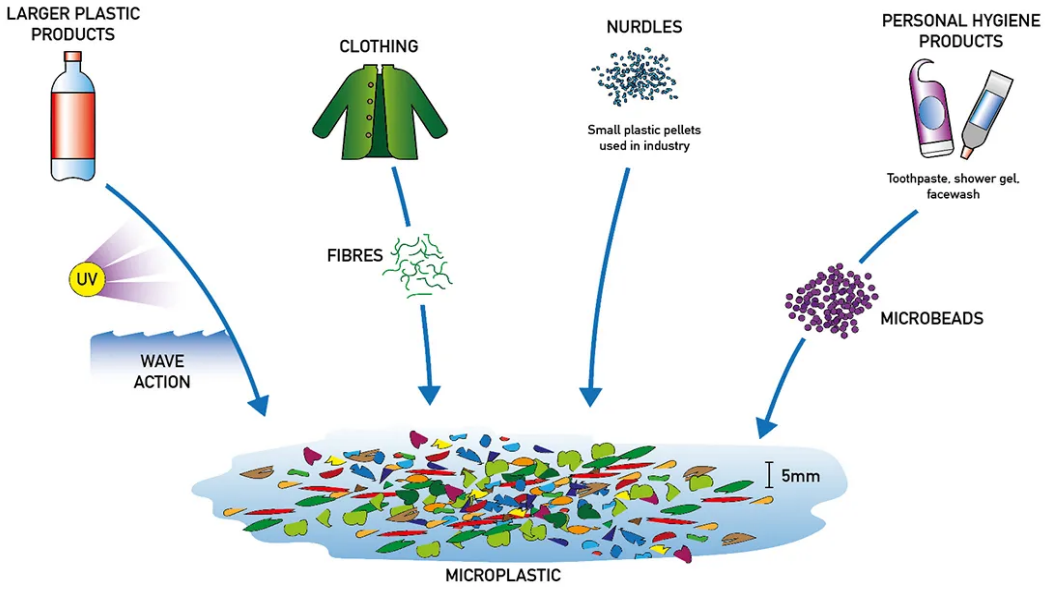

Figure 2: Main MP sources[2]

3.2 Life Cycle of MPs

The life cycle of MPs begins with their production and use in industrial and consumer applications. Improper waste management and accidental releases contribute to their entry into natural environments. Over time, environmental forces fragment larger plastics into MPs, which persist due to their resistance to degradation. Key stages in their life cycle include:

- **Production and Use**: MPs are utilized in a variety of products, from industrial abrasives to cosmetic exfoliants. Synthetic textiles are a major source, shedding microfibers during washing [3].
- **Environmental Transport**: MPs are carried through wind, water, and biological vectors, reaching remote regions such as polar ice caps and deep-sea sediments. They are detected in terrestrial soils, surface waters, and even the atmosphere [3].
- **Accumulation and Impact**: MPs accumulate in ecosystems, where they interact with organisms and environmental processes, leading to bioaccumulation and toxicity.

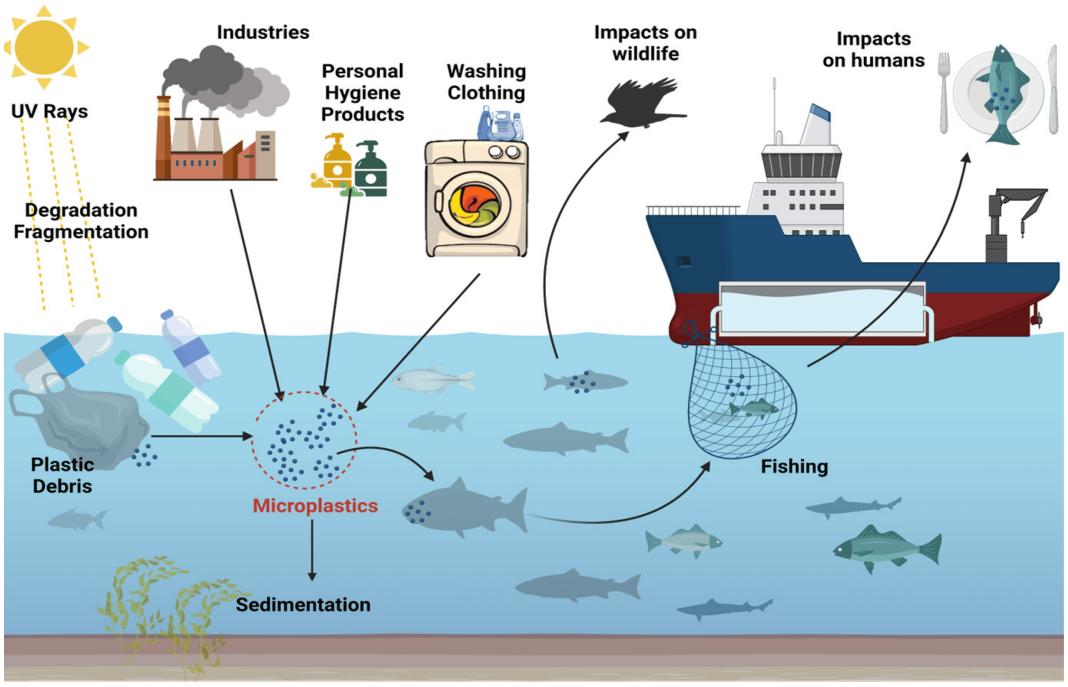

Figure 3: Life cycle of MPs (from origin to disposal) [3]

3.3 Effects of MPs on the Environment

MPs have far-reaching effects on environmental systems, influencing physical habitats, chemical processes, and ecological dynamics. Their small size and large surface area allow them to adsorb and transport hazardous chemicals, amplifying their ecological impact.

[2] https://encounteredu.com/multimedia/images/sources-of-microplastics
[3] https://www.mdpi.com/2072-6643/15/3/617

- **Physical Effects**: The accumulation of MPs in aquatic and terrestrial environments can alter habitats by creating physical barriers and reducing the availability of resources. Ingestion by wildlife leads to gastrointestinal blockages, starvation, and impaired reproduction [3].
- **Chemical Effects**: MPs act as carriers for persistent organic pollutants (POPs), heavy metals, and other toxic substances. These chemicals leach into surrounding environments, disrupting hormonal systems in organisms and contaminating food webs [3].
- **Microbial Interactions**: MPs create surfaces for microbial colonization, forming biofilms that alter microbial community structures and enhance the mobility of pathogens [3].

3.4 Impact on the Terrestrial Ecosystem

In terrestrial ecosystems, MPs are most prevalent in agricultural soils, where they are introduced through fertilizers, irrigation with contaminated water, and atmospheric deposition. MPs alter soil structure, reducing porosity and water retention capacity, which affects plant growth and soil microbial activity. Studies show that earthworms exposed to MPs exhibit reduced mobility, growth, and reproduction, leading to cascading effects on soil health and nutrient cycling [3].

MPs also interact with pesticides and other agrochemicals, influencing their distribution and persistence in soils. These interactions may exacerbate soil contamination and further disrupt ecosystem functions.

3.5 Impact on the Aquatic Ecosystem

Aquatic ecosystems are among the most affected by MP contamination. In marine environments, MPs are distributed throughout the water column, with an estimated 70% settling in sediments, 15% near coastal areas, and the rest floating in open waters [3]. MPs are ingested by a wide range of organisms, from zooplankton to large marine mammals, causing severe biological and ecological impacts:

- **Toxicological Effects**: MPs leach hazardous chemicals, such as bisphenol A (BPA) and phthalates, which disrupt endocrine functions and accumulate in tissues [3].
- **Feeding and Growth Impairments**: Ingestion of MPs reduces feeding efficiency and growth rates in marine species, including fish, mollusks, and crustaceans. This affects population dynamics and ecosystem stability [3].
- **Transport of Pollutants**: MPs serve as vectors for hydrophobic contaminants, introducing additional risks to marine food webs and human health through bioaccumulation [3].

Freshwater ecosystems are also significantly impacted, with rivers and lakes acting as conduits for MPs to enter marine environments. Studies indicate that MPs accumulate in sediments and are ingested by benthic organisms, affecting their survival and reproduction.

4 Human Exposure to Microplastics

Humans are exposed to microplastics (MPs) and nanoplastics (NPs) primarily through ingestion, inhalation, and dermal contact. These particles, ranging from 1 nm to 5 mm in size, have been detected in human tissues, including lungs, placenta, and bloodstream [4, 5].

4.1 Routes of Exposure: Ingestion, Inhalation, and Dermal Contact

Ingestion is the most significant pathway for MP exposure. Food items such as seafood, bottled water, salt, and even vegetables irrigated with contaminated water serve as major sources. MPs have been found in 81% of global tap water samples, with concentrations reaching up to 10,000 particles per liter in bottled water [5]. On average, humans may consume up to 5 g of MPs per week, an amount equivalent to the weight of a credit card [5]. MPs also accumulate in marine organisms, particularly in those consumed without gastrointestinal removal, such as mussels and oysters [4].

Airborne MPs and NPs, primarily from synthetic textiles, tire wear, and urban dust, pose significant risks via inhalation. Indoor air concentrations are often higher than outdoor levels due to sources like household textiles and furniture. Once inhaled, MPs can deposit in the respiratory tract and alveoli, potentially entering systemic circulation. Occupational exposure, particularly in industries handling synthetic materials, has been associated with obstructive bronchiolitis and hypersensitivity pneumonitis [4, 5].

Although less studied, dermal exposure to MPs and NPs is possible through cosmetics, personal care products, and textiles. Experimental evidence suggests that particles smaller than 40 nm can penetrate the skin barrier, potentially interacting with immune cells and triggering localized inflammation [5].

4.2 Health Effects of Microplastics

The health risks posed by MPs and NPs are mediated through physical irritation, chemical toxicity, and biological interference. These effects span multiple organ systems:

- **Gastrointestinal system:** MPs can irritate the gastrointestinal tract, leading to inflammation, oxidative stress, and dysbiosis of the gut microbiota. Elevated levels of fecal MPs have been correlated with inflammatory bowel diseases (IBD), while animal studies show increased intestinal permeability and metabolic disruption upon chronic exposure [4, 5].

- **Respiratory system:** Chronic inhalation of MPs, particularly those derived from polystyrene and polyvinyl chloride, induces oxidative stress and inflammation in the lungs. Studies highlight an association between MP exposure and conditions such as asthma and fibrosis. Workers exposed to synthetic fibers report higher incidences of respiratory ailments [4, 5].

- **Endocrine and reproductive systems:** MPs release endocrine-disrupting chemicals, including bisphenol A and phthalates, which interfere with hormonal regulation and reproductive health. These particles can cross the placental barrier, exposing fetuses to potential developmental risks. Human studies have detected MPs in placentas, raising significant concerns about prenatal exposure [5].

- **Neurological effects:** Animal studies suggest that MPs can cross the blood-brain barrier, causing neuroinflammation, oxidative stress, and cognitive impairments. Chronic exposure to styrene-based MPs has been linked to fatigue, dizziness, and reduced cognitive function in industrial workers [4, 5].
- **Immune system and carcinogenic risks:** MP exposure activates innate immune responses, triggering chronic inflammation and genotoxic effects. Occupational studies indicate increased risks of DNA damage and carcinogenesis associated with prolonged exposure to certain MP types. MPs also serve as carriers for heavy metals and persistent organic pollutants, compounding their toxicological impacts [5].

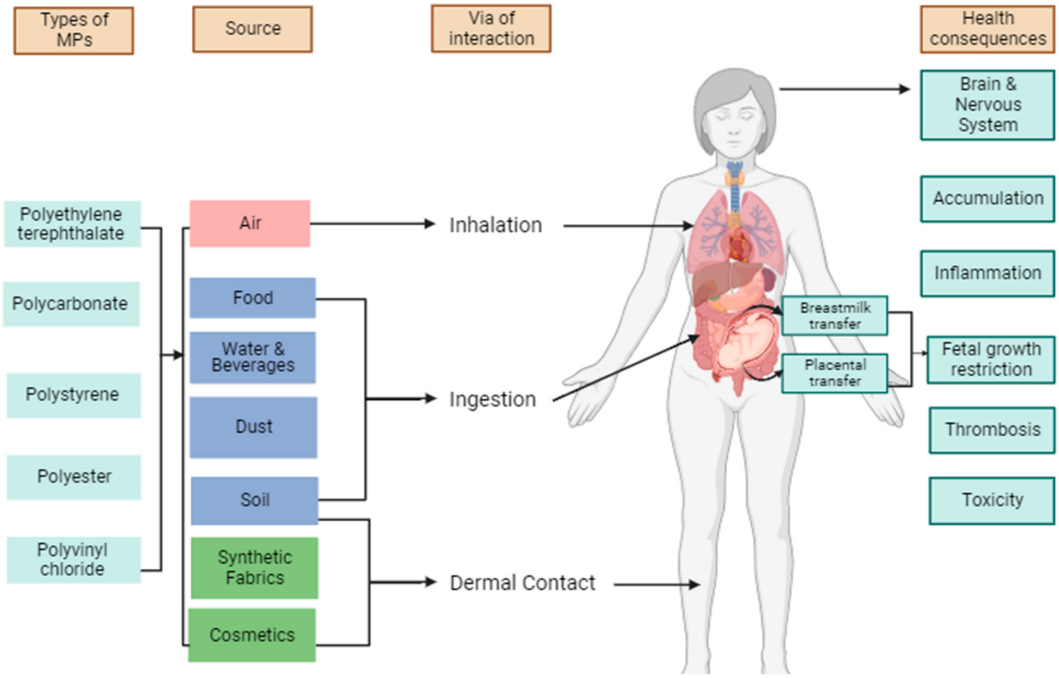

Figure 4: Pathways of human exposure to microplastics and their types, modes of interactions, and possible associated health consequences.[4]

[4] https://www.mdpi.com/2673-8929/3/1/4

5 State of the Art in microplastics Detection and the Role of SERS

Conventional methods for MP detection often rely on Fourier-transform infrared spectroscopy (FTIR), Raman spectroscopy, and fluorescence microscopy [6]. These techniques are capable of identifying particles down to a few microns [1]. FTIR spectroscopy is widely used due to its ability to identify polymer types based on their characteristic absorption bands. However, its detection limit is constrained to particles larger than 30 microns [6]. On the other hand, Raman spectroscopy offers greater sensitivity, detecting particles as small as a few microns, but emits a weak signal output and is limited by fluorescence interference, particularly when analyzing complex environmental samples [2].

Fluorescence microscopy, often used in conjunction with staining agents like Nile Red, allows visualization and quantification of MPs in biological and environmental matrices [1]. Despite its utility, this method suffers from false positives due to non-specific staining, necessitating complementary techniques for validation [2].

Recent advancements highlight the potential of combining multiple techniques to overcome individual limitations. For instance, the synergistic use of FTIR and Raman spectroscopy provides comprehensive chemical and morphological characterization of MPs [1]. However, the integration of these techniques requires significant instrumentation, setup and expertise, limiting their application to laboratory settings [1].

Surface-Enhanced Raman Scattering (SERS) represents a new advancement in MP detection. By utilizing plasmonic nanostructures, SERS amplifies Raman signals by factors as high as 10^{10} [2]. This enhancement enables the detection of MPs at nanometer scales and low concentrations, addressing critical gaps in sensitivity [2]. Notably, SERS substrates, such as gold and silver nanoparticles, can be functionalized to improve selectivity for specific polymer types [1].

Biopolymeric probes, such as rhodamine-functionalized hyaluronic acid (HA-RB), have been developed to enhance the specificity of SERS-based detection [1]. These probes exhibit distinct fluorescence and Raman signatures upon interaction with MPs, facilitating their identification and characterization [1, 2]. Additionally, fluorescence lifetime imaging microscopy (FLIM) has been integrated with SERS to enable real-time, high-resolution imaging of MPs in a combined environment [1].

Furthermore, machine learning algorithms for image recognition are increasingly being incorporated into SERS-based detection frameworks [2]. These algorithms analyze spectral data to classify polymer types, quantify concentrations, and identify potential interferents in complex samples [2]. By combining deep learning with multimodal imaging, researchers aim to achieve automated, high-throughput analysis of MPs [1]. This approach holds promise for scaling environmental monitoring efforts while ensuring reproducibility and accuracy [1].

Emerging portable devices leveraging consumer-grade components are bridging the gap between laboratory and field applications [1]. These devices integrate SERS with optical sensing technologies, enabling remote and real-time monitoring of MPs in water and food matrices [2]. For instance, IoT-enabled platforms have been proposed to transmit data to centralized facilities for comprehensive analysis [1]. Such innovations align with global efforts to establish standardized protocols for MP detection and quantification [1].

The integration of pre-treatment methods, such as oxidative degradation and filtration, further enhances the effectiveness of SERS-based detection in environmental samples [1]. These methods reduce the impact of organic and inorganic interferents, improv-

ing the reliability of analytical results [2]. Moreover, advancements in SERS substrates, including anisotropic nanoparticles and nanostructured surfaces, are expanding the capabilities of this technique [2].

Recent studies have also explored the potential of combining SERS with microfluidic devices for real-time, in situ analysis of MPs [2]. Microfluidic-SERS platforms offer advantages such as reduced sample preparation, enhanced detection sensitivity, and the ability to analyze small volumes of complex environmental samples [1]. These integrated systems are particularly promising for field applications, where rapid and accurate identification of MPs is required [2].

Another emerging approach involves the functionalization of SERS substrates with molecularly imprinted polymers (MIPs), which provide highly selective recognition sites for specific MP types [1]. MIP-SERS sensors combine the high sensitivity of SERS with the specificity of molecular imprinting, allowing precise identification of polymer compositions in environmental samples [2]. This technique holds promise for addressing challenges related to MP heterogeneity and environmental matrix complexity [1].

Additionally, novel deep-learning models are being developed to analyze SERS spectra, enhancing the accuracy of MP classification [2]. These models employ convolutional neural networks (CNNs) to differentiate between polymer types with high precision, even in the presence of spectral noise [1]. The integration of AI-driven spectral analysis with SERS detection frameworks is expected to significantly improve the robustness and scalability of MP monitoring systems [2].

Recent advancements also focus on expanding SERS-based detection beyond water matrices to include air and soil contamination. Airborne MPs, originating from synthetic textiles, industrial emissions, and urban dust, pose challenges for detection due to their small size and dispersive nature [1]. The adaptation of SERS to aerosol-based sampling systems enables the capture and identification of airborne MPs in real time [2]. Similarly, novel soil extraction techniques combined with SERS allow the identification of MPs in agricultural and urban soils, providing a more comprehensive assessment of environmental contamination [1].

In recognition of their potential health risks, the European Directive 2020/2184 has mandated the development of harmonized analytical methods for microplastics in drinking water by January 2024, and requires a formal evaluation of their health impact by 2029 [7].

Future research is also exploring the potential for multi-modal spectroscopy approaches, where SERS is integrated with advanced imaging techniques such as hyperspectral and terahertz spectroscopy [2]. These hybrid systems offer new capabilities for detecting MPs in highly complex environmental matrices, reducing false positives and improving overall detection accuracy [1].

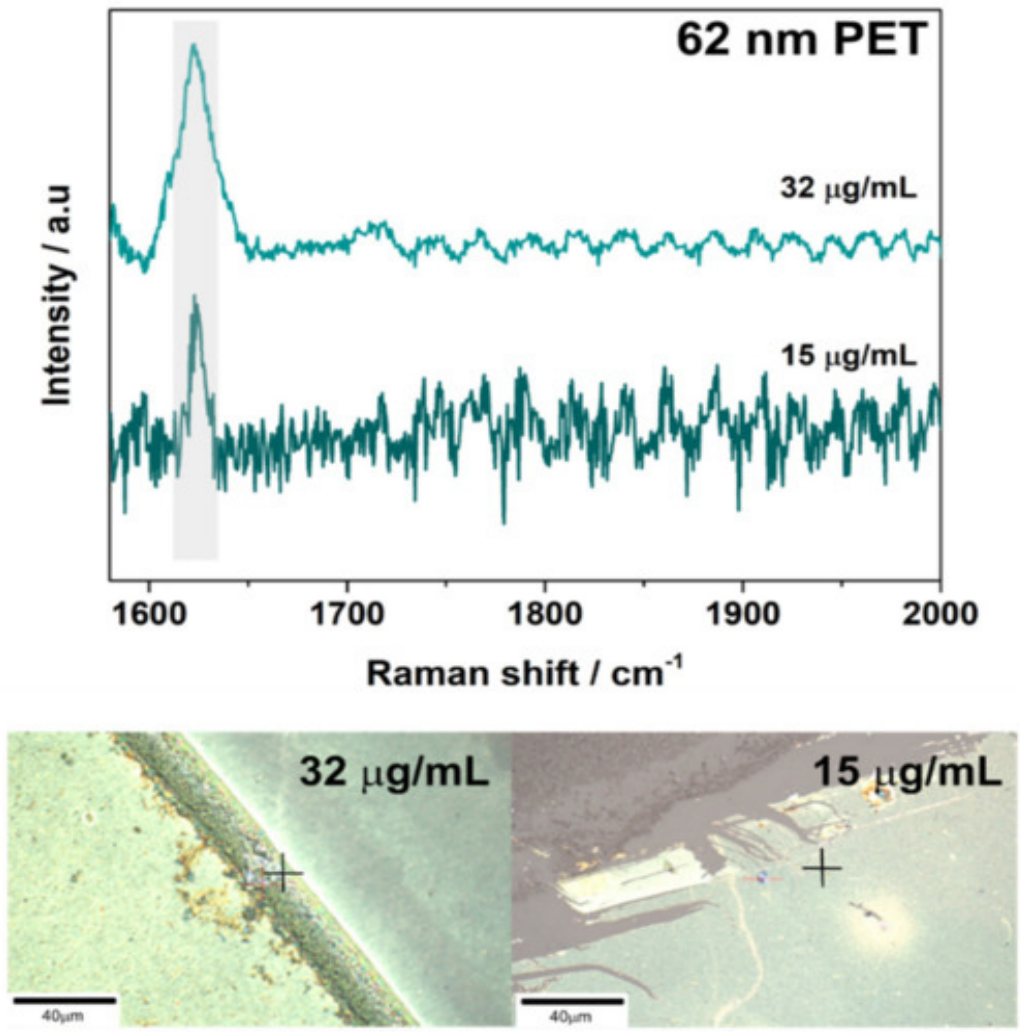

Figure 5: A representative SERS spectrum obtained for 62 nm PET nanoparticles on 46 nm AuNP substrates.[5]

[5] https://www.mdpi.com/2079-4991/11/5/1149

6 Comparison of Raman and FTIR Spectroscopy for microplastics Detection

The detection and characterization of microplastics (MPs) rely on various spectroscopic techniques, with Raman microscopy and Fourier-transform infrared (FTIR) microscopy being among the most commonly employed. Each method has distinct advantages and limitations, influencing its applicability depending on sample type, size, and analytical constraints. This section presents a detailed comparison of Raman and FTIR microscopy in MP detection, drawing insights from their principles, performance, and practical implementation.

6.1 Fundamentals and Comparison of Raman and FTIR Spectroscopy

Raman microscopy exploits the inelastic scattering of light to probe molecular vibrations, providing a unique spectral fingerprint of polymers. When monochromatic laser light interacts with a sample, it undergoes scattering, with a small fraction of the scattered light experiencing a frequency shift corresponding to molecular vibrational modes [8]. This technique is particularly effective for MP identification due to its ability to analyze particles as small as a few microns without extensive sample preparation.

One of the key strengths of Raman microscopy is its capability to analyze MPs in aqueous environments, as water has a weak Raman signal. This makes it advantageous for detecting MPs in marine and freshwater samples [8]. However, Raman spectroscopy is limited by fluorescence interference, which can obscure Raman signals, especially when analyzing complex environmental matrices. Fluorescence suppression techniques, such as longer excitation wavelengths and time-gated detection, have been explored to mitigate these effects [8].

FTIR microscopy is based on the absorption of infrared radiation by molecular bonds, providing chemical composition information through characteristic absorption bands [8]. This technique is widely used for MP detection due to its ability to analyze a broad range of polymer types and its compatibility with automated spectral libraries for material identification.

One of the primary advantages of FTIR microscopy is its efficiency in detecting larger MPs (typically >20 µm), making it well-suited for environmental samples with high MP concentrations. Additionally, FTIR operates in transmission and attenuated total reflectance (ATR) modes, allowing flexibility in analyzing different sample types [8]. However, FTIR spectroscopy is less effective for detecting smaller MPs due to diffraction limits and requires extensive sample purification to remove interfering organic matter [8].

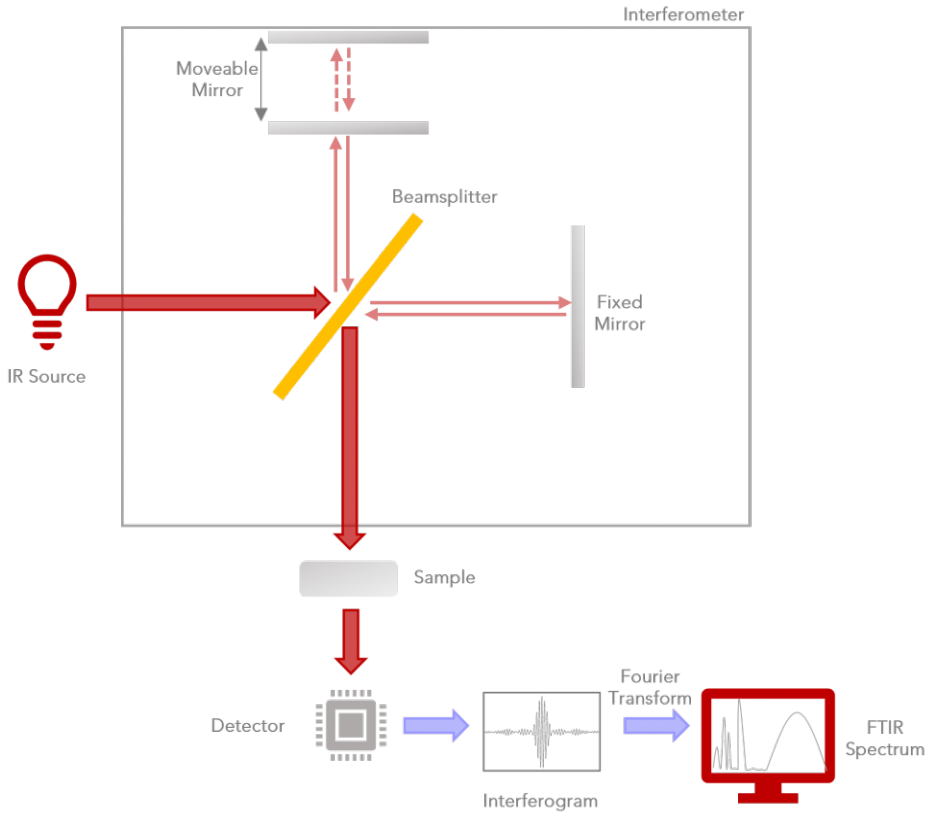

Figure 6: Interferometer schematic in an FTIR spectrometer.[6]

A direct comparison between FTIR and Raman techniques highlights key differences that impact their effectiveness in MP analysis. Table 1 summarizes the strengths and limitations of both methods.

Feature	FTIR Spectroscopy	Raman Spectroscopy
Detection Limit	>20 µm	<1 µm
Sensitivity to Water	High interference	Minimal interference
Sample Preparation	Requires clean samples	Minimal preparation required
Fluorescence Sensitivity	Low	High
Speed	Fast for larger MPs	Slower but highly detailed

Table 1: Comparison of FTIR and Raman spectroscopy for MP detection [8].

The choice between FTIR and Raman spectroscopy largely depends on the sample characteristics and the specific analytical requirements. FTIR is preferred for larger MPs in relatively clean matrices, whereas Raman spectroscopy excels in high-resolution analysis of smaller MPs and particles embedded in complex environmental samples [8].

Both Raman and FTIR spectroscopy offer valuable tools for MP detection, with their applicability depending on factors such as particle size, sample complexity, and fluorescence interference. While FTIR remains the dominant technique for large-scale MP screening due to its speed and ease of use, Raman spectroscopy provides superior resolution for small particle detection. Recent advancements in instrumental design, including

[6] `https://www.edinst.com/resource/what-is-ftir-spectroscopy/`

the integration of machine learning algorithms and multi-modal analysis, continue to enhance the accuracy and efficiency of both techniques [8]. Future research should focus on hybrid approaches that combine the strengths of both methods, offering comprehensive solutions for MP detection in environmental and biological matrices.

6.2 SERS as an Enhancement of Raman Spectroscopy

Surface-Enhanced Raman Spectroscopy (SERS) is a powerful advancement over conventional Raman spectroscopy, designed to overcome its intrinsic limitations, particularly in terms of sensitivity and detection limits. By exploiting the enhancement of the Raman signal through interaction with nanostructured metallic surfaces, SERS significantly improves the performance of Raman-based analytical techniques [9]. An early application example is the detection of copper phthalocyanine molecules on GaP nanoparticles, which demonstrated a Raman enhancement factor of nearly 700. This case illustrated how SERS could be effectively used for the analysis of industrial pigments and organometallic compounds, and not just biological or noble-metal systems. [10]

6.3 Advantages of SERS over Conventional Raman Spectroscopy

SERS provides several key advantages over traditional Raman spectroscopy, making it a superior tool for trace-level detection and complex sample analysis:

- **Enhanced Sensitivity**: The signal amplification due to localized surface plasmon resonance (LSPR) can increase Raman scattering by factors of 10^6 to 10^8, allowing for the detection of extremely low analyte concentrations, down to the single-molecule level [9].
- **Improved Signal-to-Noise Ratio**: The enhancement effect leads to clearer, more defined spectra, reducing background noise and increasing reliability in complex sample matrices [9].
- **Non-Destructive Nature**: Similar to standard Raman spectroscopy, SERS is a non-destructive technique, preserving the integrity of the analyzed samples [9].
- **Selective Detection**: Functionalized SERS substrates can be tailored to target specific analytes, increasing selectivity and reducing interference from unwanted components [9].

6.4 Limitations and Challenges of SERS

Despite its advantages, SERS also presents several challenges that must be addressed to ensure reproducibility and applicability in real-world scenarios:

- **Reproducibility Issues**: The enhancement effect is highly dependent on the uniformity and stability of nanostructured metallic surfaces. Small variations in the fabrication of SERS substrates can lead to significant discrepancies in signal intensity [9].
- **Substrate Degradation**: Over time, SERS-active substrates may degrade or become contaminated, reducing their efficiency and reliability [9].

- **Interference from Complex Matrices**: In environmental and biological samples, interfering substances can adsorb onto the SERS substrate, altering enhancement effects and complicating spectral interpretation [9].
- **Cost and Fabrication Challenges**: High-quality SERS substrates require precise nanofabrication techniques, which can be costly and limit large-scale implementation [9].

6.5 Future Directions for SERS Development

Ongoing research is addressing the limitations of SERS through:

- **Advanced Nanofabrication**: Development of more uniform and stable nanostructures to improve reproducibility [9].
- **Machine Learning Integration**: AI-driven spectral analysis to enhance data interpretation and automate classification [9].
- **Portable SERS Devices**: Miniaturization of SERS instruments for real-time, in-field applications [9].

While SERS presents clear advantages over traditional Raman spectroscopy, its implementation in routine analysis requires overcoming challenges related to reproducibility, cost, and substrate stability. With continuous advancements in nanotechnology and data processing, SERS is expected to become an increasingly viable tool for environmental monitoring, biomedical diagnostics, and material science applications [9].

7 The Physics of Raman Spectroscopy

Raman spectroscopy is a vibrational spectroscopic technique that provides molecular-level information by analyzing the inelastic scattering of light. Rooted in quantum mechanical and electromagnetic principles, this technique is widely used for material characterization, environmental monitoring, and biomedical applications due to its ability to provide unique molecular fingerprints.

7.1 The Raman Effect

The Raman effect occurs when incident light interacts with the vibrational, rotational, or electronic energy levels of a molecule, resulting in inelastic scattering. Unlike Rayleigh scattering, where the scattered photons retain the same energy as the incident photons, Raman scattering involves an exchange of energy between the photon and the molecule. This exchange produces a shift in the energy of the scattered photons, corresponding to the vibrational modes of the molecule [11].

There are two primary types of Raman scattering:

- **Stokes Scattering**: The scattered photon loses energy, as it transfers a portion of its energy to excite a molecular vibrational mode. The frequency of the scattered light is therefore lower than that of the incident light.
- **Anti-Stokes Scattering**: The scattered photon gains energy, as it interacts with a molecule already in an excited vibrational state. The frequency of the scattered light is higher than that of the incident light.

Stokes scattering is generally more intense than Anti-Stokes scattering, as the majority of molecules at room temperature are in their ground vibrational states. The shifts in energy, expressed in wavenumbers (cm^{-1}), are specific to the molecular bonds and structures, making Raman spectroscopy a highly selective analytical tool [11].

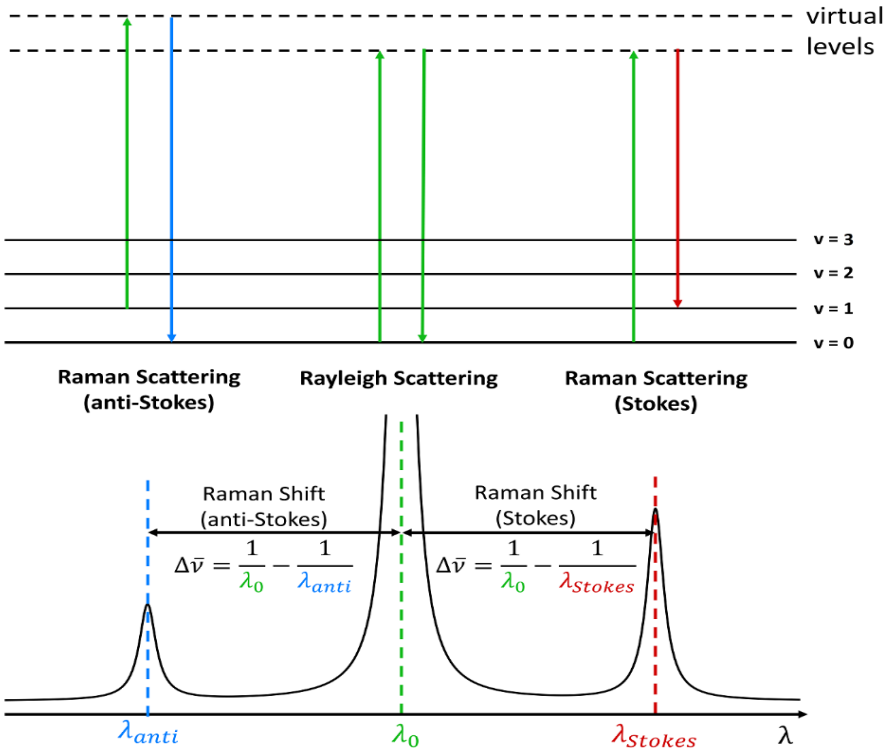

Figure 7: Schematic of the Raman scattering illustrating Rayleigh, Stokes, and Anti-Stokes scattering paths.[7]

7.2 Mathematical Framework of Raman Scattering

The Raman effect is fundamentally governed by the interaction of the molecule's polarizability tensor, α, with the electric field of the incident light. When an external electric field E interacts with a molecule, it induces a dipole moment, p, expressed as:

$$p = \alpha \cdot E,$$

where p is the induced dipole moment, and α is a second-rank tensor that represents the molecule's polarizability.

For a vibrational mode to be Raman-active, the polarizability tensor must change during the vibration. This requirement distinguishes Raman-active modes from IR-active modes, which depend on a change in the dipole moment [11]. The intensity of Raman scattering is directly proportional to the square of the rate of change of the polarizability tensor:

$$I_{\text{Raman}} \propto \left(\frac{\partial \alpha}{\partial q}\right)^2,$$

where q is the vibrational coordinate.

7.3 Selection Rules and Symmetry

The selection rules for Raman scattering are dictated by the symmetry properties of the molecule and its vibrational modes. Vibrational modes are classified based on their

[7] https://www.edinst.com/blog/what-is-the-stokes-shift/

symmetry with respect to the molecule's point group. Only modes that induce a change in the molecule's polarizability are Raman-active. For instance, symmetric stretching vibrations are often strongly Raman-active due to significant changes in polarizability during vibration [11].

7.4 Advantages and Limitations of Raman Spectroscopy

Raman spectroscopy offers several advantages over other vibrational techniques such as IR spectroscopy:

- **Non-destructive Analysis**: Raman spectroscopy does not require extensive sample preparation and can be applied to a wide range of materials, including solids, liquids, and gases.
- **Water Compatibility**: Unlike IR spectroscopy, Raman is less affected by water, making it suitable for aqueous samples.
- **High Spatial Resolution**: Coupled with confocal microscopy, Raman spectroscopy achieves sub-micron spatial resolution, enabling the analysis of microstructures and heterogeneous samples [11].

However, the Raman effect is inherently weak, with only about 1 in 10^7 photons undergoing Raman scattering. This low signal intensity is further challenged by fluorescence interference, which can obscure Raman signals, particularly in complex samples. Advanced techniques such as resonance Raman spectroscopy and SERS are often employed to enhance sensitivity and overcome these limitations [11].

7.5 Applications of Raman Spectroscopy

Raman spectroscopy is a highly flexible methodology very useful across various scientific fields. As a non destructive technique it's used for material characterization and quality control both on reserach areas and industrial frames.

In the following a not exhaustive list of applications is proposed:

- **Materials Science**
  - **Material characterization**: Identification of molecular structures and defects in polymers, nanomaterials, and semiconductors.
  - **Polymorph analysis**: Differentiation of crystalline forms, important for pharmaceuticals and advanced materials.
  - **Quality control**: Monitoring of chemical consistency and impurity levels in production lines.
- **Pharmaceuticals**
  - **Raw material verification**: Authentication of ingredients during drug manufacturing.
  - **Formulation analysis**: Evaluation of active ingredients and excipient composition.
  - **Counterfeit detection**: Rapid screening for falsified drugs.

- **Forensic and Environmental Analysis**
  - **Trace evidence**: Investigation of fibers, explosives, and unknown residues at crime scenes.
  - **Microplastic detection**: Identification of plastic particles in soil and water samples.
  - **Pollutant monitoring**: Analysis of chemical contaminants in environmental matrices.
- **Biomedical Applications**
  - **Disease diagnostics**: Detection of cancer biomarkers and pathological changes in tissues.
  - **Biomolecule identification**: Label-free detection of proteins and nucleic acids.
  - **In vivo analysis**: Non-invasive examination of tissues with minimal preparation.
- **Cultural Heritage and Industry**
  - **Artwork analysis**: Non-destructive characterization of pigments and historical materials.
  - **Process control**: Real-time monitoring of chemical reactions in industrial synthesis.
  - **Food safety**: Detection of contaminants and adulterants in food products.

8 The Physics of Surface-Enhanced Raman Scattering (SERS)

Surface-Enhanced Raman Scattering (SERS) is an advanced extension of Raman spectroscopy that dramatically enhances the Raman signal through the use of nanostructured metallic surfaces. This enhancement, often in the range of 10^6 to 10^{12}, enables the detection of single molecules, making SERS an invaluable tool for ultrasensitive chemical and biological analysis.

8.1 Fundamental Principles of SERS

The enhancement in SERS arises from two primary mechanisms:

- **Electromagnetic Enhancement (EE)**: This is the dominant mechanism, caused by the localized surface plasmon resonances (LSPRs) of metallic nanostructures. LSPRs occur when incident light interacts with the conduction electrons on the metallic surface, inducing collective oscillations. These oscillations amplify the local electromagnetic field, which, in turn, enhances the Raman scattering intensity. The enhancement is proportional to the fourth power of the local field intensity:

$$EF_{\text{EM}} \propto |E_{\text{local}}|^4 ,$$

  where E_{local} is the electric field near the metallic surface [11].

- **Chemical Enhancement (CE)**: This secondary mechanism involves charge transfer between the analyte molecules and the metal surface. The interaction modifies the polarizability of the molecule, enhancing its Raman cross-section. CE is particularly significant for molecules adsorbed directly onto the metallic surface [11].

8.2 Nanostructures in SERS

The efficiency of SERS depends critically on the morphology, composition, and arrangement of the metallic nanostructures used as substrates. Gold and silver are the most commonly employed metals due to their strong plasmonic properties in the visible and near-infrared spectral regions. Various nanostructures, such as nanospheres, nanorods, and nanostars, are engineered to create "hotspots"—regions where the local electromagnetic field is highly concentrated.

These hotspots, typically formed at the junctions between nanoparticles or at sharp edges, are the primary contributors to the SERS effect. The enhancement factor at a hotspot can exceed 10^{10}, enabling the detection of single molecules under optimized conditions [11].

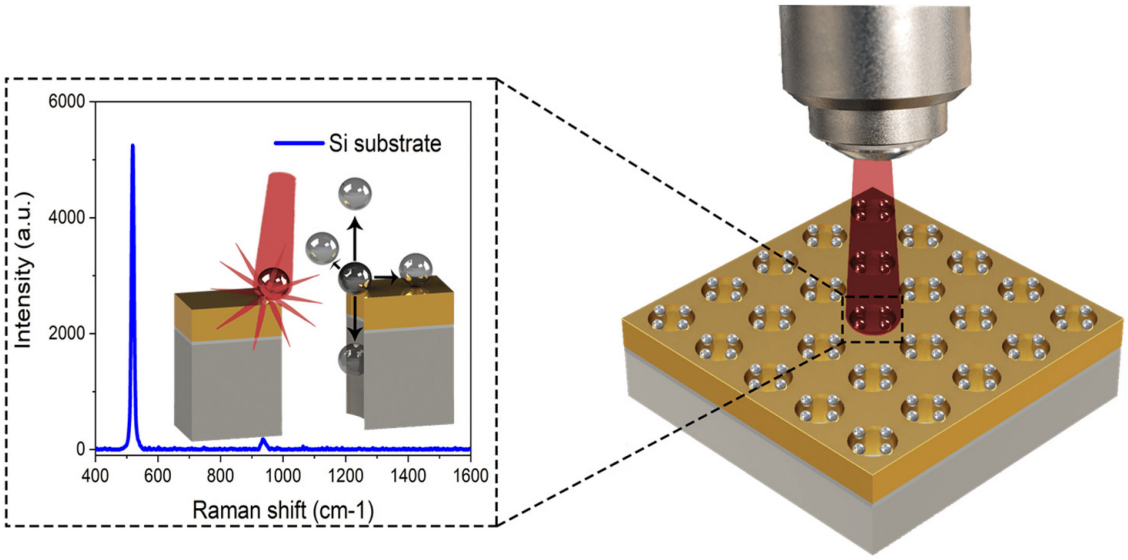

Figure 8: Sers interaction with nanostructured metallic substrate and hotspots.[8]

8.3 Theoretical Framework of SERS

The theoretical understanding of SERS is built upon classical electromagnetic theory and quantum mechanical models.

- **Electromagnetic Theory**: The enhancement occurs when the frequency of the incident light matches the plasmon resonance frequency of the metallic nanostructure. This resonance condition maximizes the local field intensity, which amplifies both the incident and scattered light [11].
- **Unified Theory of SERS**: Lombardi and Birke proposed a unified framework combining electromagnetic enhancement, charge transfer, and molecular resonance effects. This comprehensive model provides a robust explanation for the variability of SERS signals across different molecular systems and substrates [11].

8.4 Applications of SERS

SERS has been widely adopted in diverse fields due to its unparalleled sensitivity and molecular specificity. Key applications include:

- **Environmental Monitoring**: Detection of trace pollutants, such as pesticides, heavy metals, and microplastics in water and soil.
- **Biomedical Diagnostics**: Identification of biomolecules and disease markers, including cancer diagnostics through single-cell analysis.
- **Material Science**: Characterization of nanostructured materials and semiconductors, with insights into their chemical composition and physical properties [11].

[8] `https://www.mdpi.com/2079-6374/12/2/128`

8.5 Challenges and Future Directions

Despite its advantages, SERS faces several challenges:

- **Reproducibility**: Variability in the fabrication of nanostructured substrates often leads to inconsistent enhancement factors.
- **Matrix Effects**: Complex sample matrices can suppress the SERS signal or introduce spectral noise.
- **Scalability**: The production of high-quality, cost-effective SERS substrates for large-scale applications remains a technical bottleneck [11].

Ongoing research aims to address these limitations by developing advanced nanofabrication techniques, integrating machine learning algorithms for spectral analysis, and creating portable SERS devices for real-time, in-field applications.

9 Physical and Chemical Properties of the Plastics Used

Plastics are among the most widespread materials in modern society, with applications ranging from packaging to electronics. However, their environmental persistence and tendency to fragment into microplastics raise serious concerns about long-term ecological impact. Characterizing the physical and chemical properties of polymers is crucial for identifying their origin, tracking their environmental fate, and improving both analytical and recycling strategies.

Raman spectroscopy offers a non-destructive and highly specific technique for identifying polymer types based on their molecular vibrations. Each polymer exhibits a unique "spectral fingerprint" that reflects its chemical structure. In this section, we summarize the structural and spectroscopic properties of the plastic types analyzed in this work, based on the characterization reported by PhysicsOpenLab (2022) [12].

9.1 Polyethylene Terephthalate (PET)

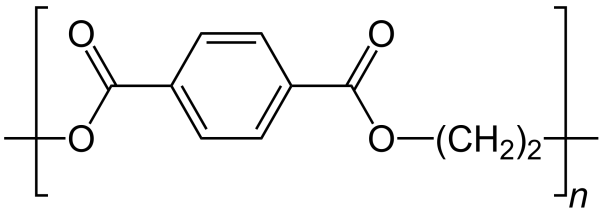

Figure 9: Caption

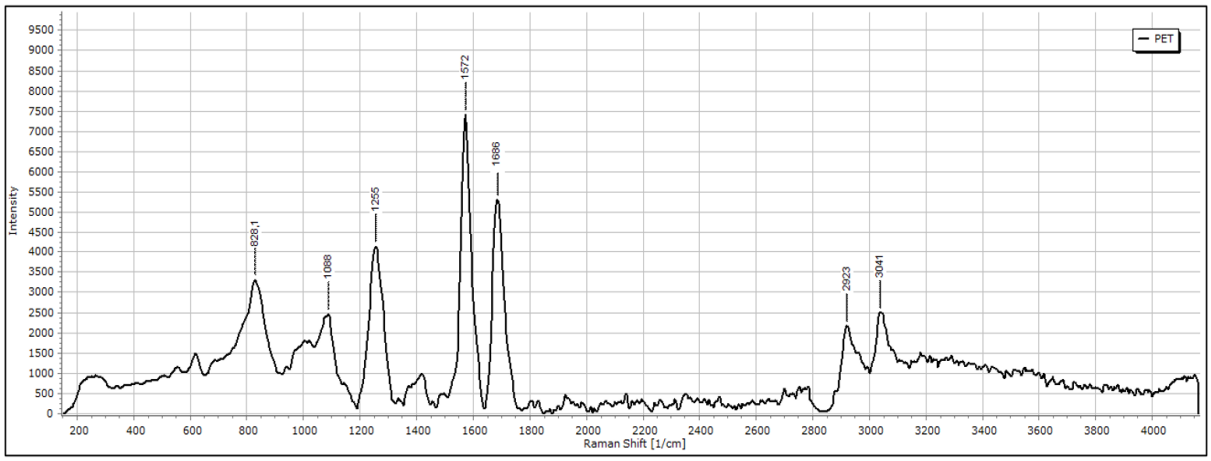

Figure 10: Spectrum of PET

PET is a polyester composed of terephthalic acid and ethylene glycol units. It exists in amorphous or semi-crystalline form depending on processing conditions. It is commonly used in beverage bottles, food packaging, and textiles.

The Raman spectrum of PET includes:

- Strong C=O stretching near 1700 cm^{-1}
- Aromatic C=C stretching around 1600 cm^{-1}
- CH stretching above 2900 cm^{-1}

9.2 High-Density Polyethylene (HDPE)

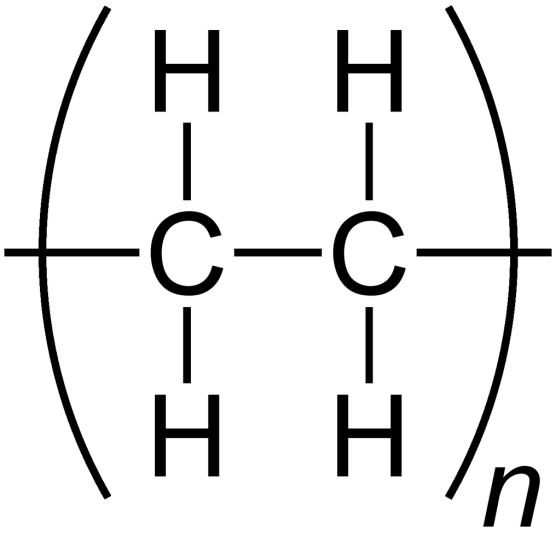

Figure 11: Polyethylene (monomer of HDPE)

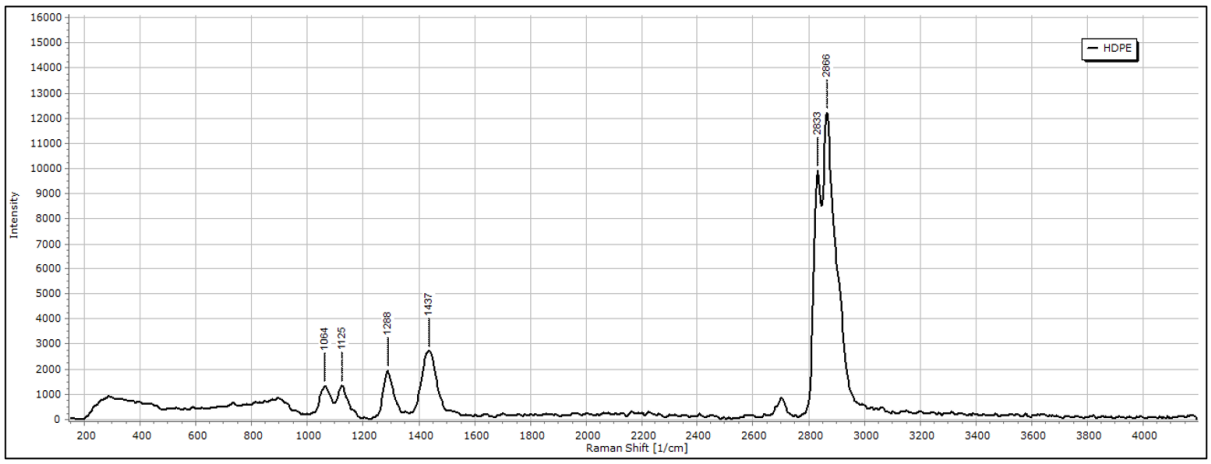

Figure 12: Spectrum of HDPE

HDPE is a linear, semi-crystalline polymer composed of repeating ethylene units. It is known for its chemical resistance and is widely used in detergent bottles, pipes, and crates.

Its Raman spectrum closely resembles that of LDPE, with:

- C-H stretching near 3000 cm^{-1}
- CH_2 bending modes around 1400 cm^{-1}
- C-C stretching between 1000–1200 cm^{-1}

9.3 Low-Density Polyethylene (LDPE)

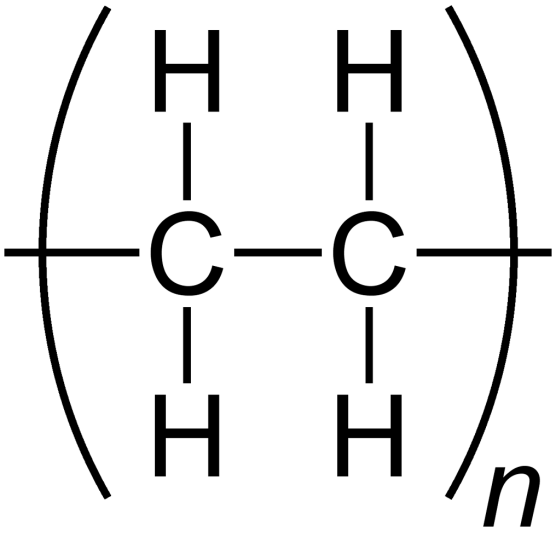

Figure 13: Polyethylene (monomer of LDPE)

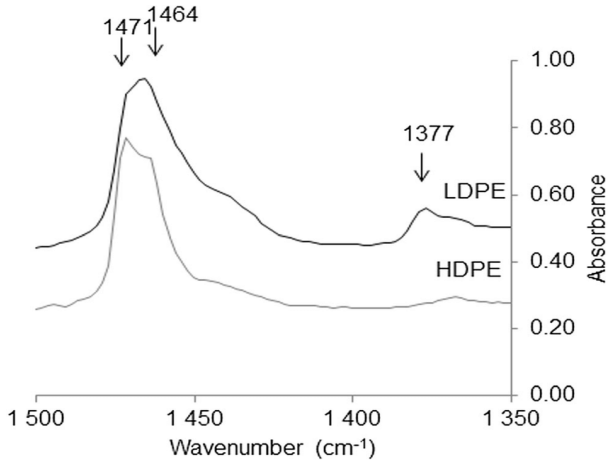

Figure 14: Local zoomed difference example between LDPE and HDPE spectra

Since the differences in spectra between LDPE and HDPE are small, a zoom of a local difference is shown instead of showing the same spectrum as in the HDPE section.

LDPE differs from HDPE in having a more branched molecular structure, leading to lower density and flexibility. It is used in plastic bags, films, and squeeze bottles.

Its Raman spectral features are similar to HDPE but typically show broader bands due to lower crystallinity.

9.4 Polypropylene (PP)

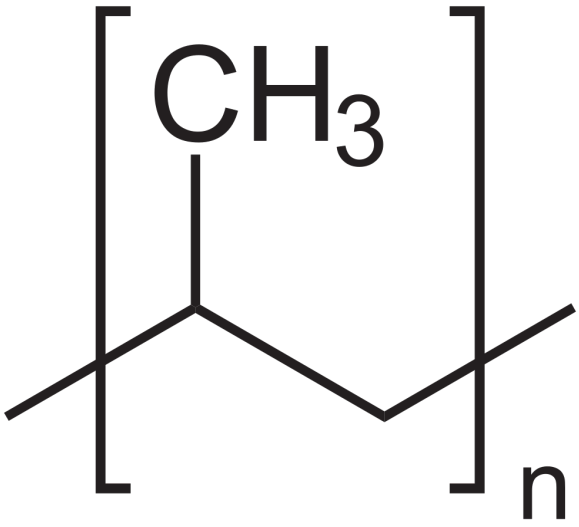

Figure 15: Caption

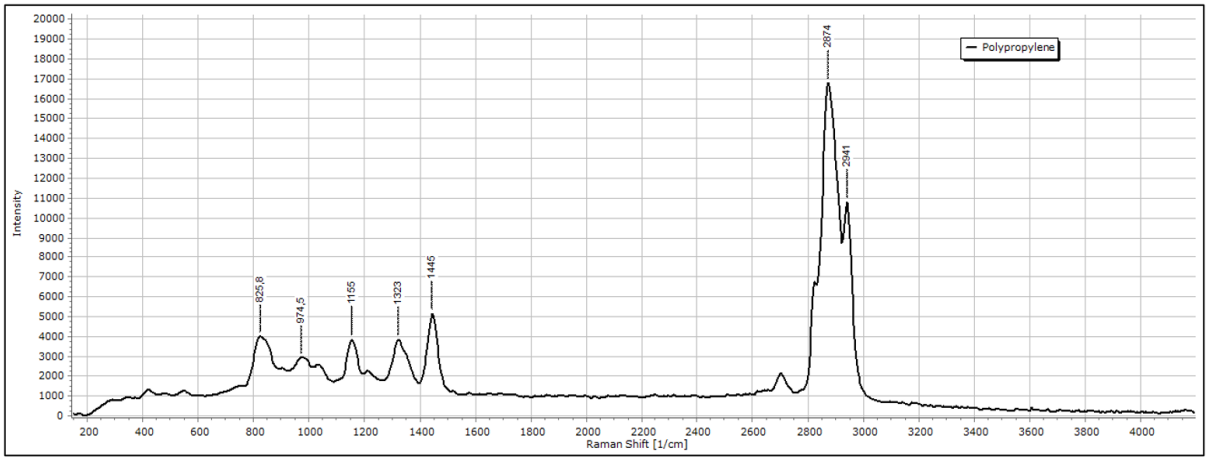

Figure 16: Spectrum of HDPE

PP is a semi-crystalline thermoplastic derived from propylene monomers. It exists in different tacticities (mainly isotactic in commercial products), which influence its mechanical properties.

Raman peaks include:

- CH_3 symmetric and asymmetric stretching
- C-C skeletal vibrations

Common PP items include bottle caps, medical syringes, and containers.

9.5 Polystyrene (PS)

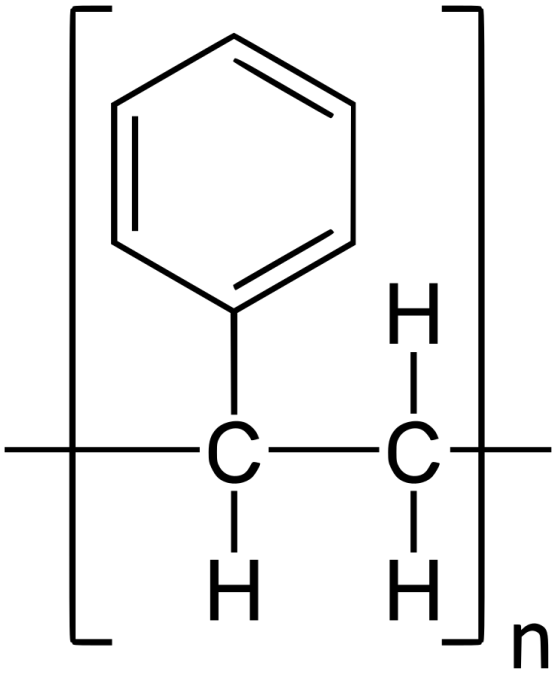

Figure 17: Polystyrene monomer

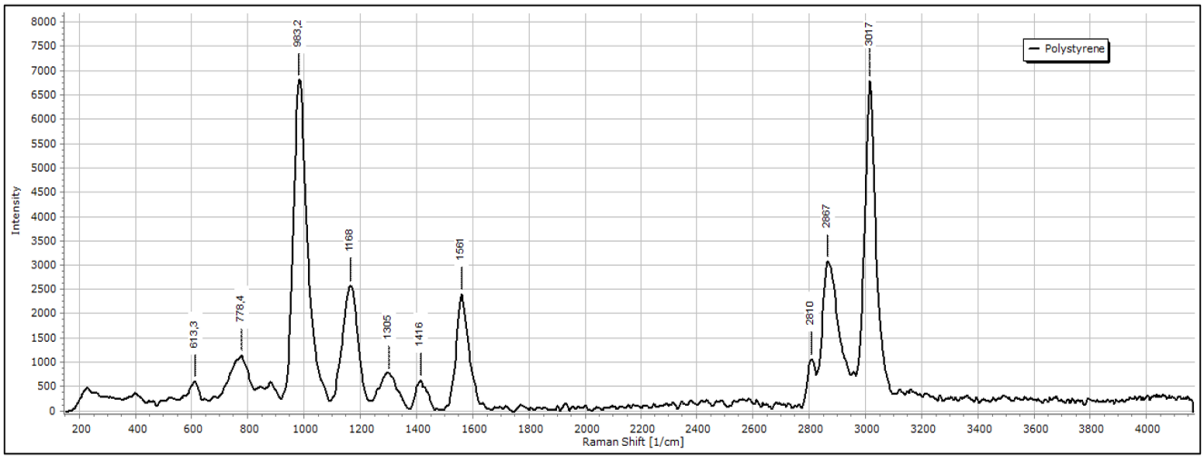

Figure 18: Spectrum of Polystyrene

PS is an aromatic thermoplastic formed from styrene monomers. It is rigid, transparent, and widely used in packaging and disposable products.

Characteristic Raman bands:

- Aromatic C=C stretching around 1600 cm^{-1}
- C-H vibrations near 2900 and above 3000 cm^{-1}
- A sharp peak near 1000 cm^{-1} from benzene ring modes

9.6 Polyvinyl Chloride (PVC)

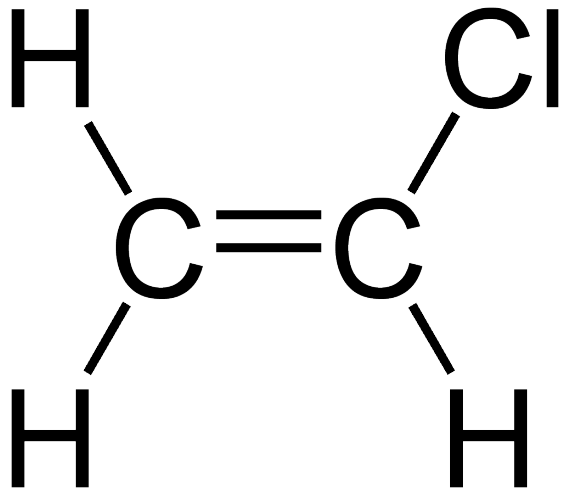

Figure 19: Polyvinyl chloride monomer

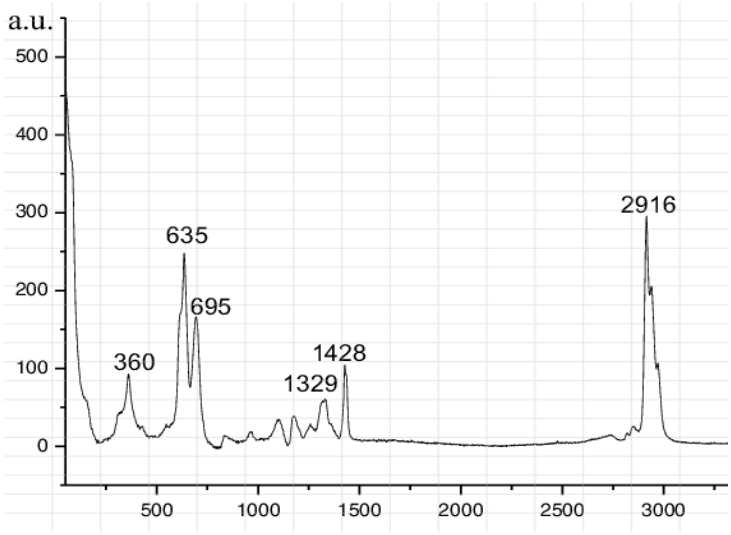

Figure 20: Spectrum of Polyvinyl chloride

PVC is a versatile polymer formed by polymerization of vinyl chloride monomers. It is used in pipes, flooring, and medical devices.

Although not deeply detailed in the source, PVC generally exhibits:

- C-Cl stretching modes
- CH bending and stretching vibrations

9.7 Other Plastics

The class labeled "Other" includes various materials not covered by standard recycling codes, such as:

- Polycarbonate (PC)
- Polymethylmethacrylate (PMMA)
- Acrylonitrile butadiene styrene (ABS)
- Nylon (PA)

These materials are chemically diverse and exhibit complex or overlapping Raman spectra. Their inclusion in a single class introduces heterogeneity in both chemical composition and spectral features, which complicates classification.

10 The Machine Learning Approach

The first part of this thesis was dedicated to a theoretical and contextual investigation of the microplastics issue, with particular attention to the physical mechanisms governing microplastics formation and transport, their interaction with materials and the environment, and their detection through surface-enhanced Raman spectroscopy (SERS). This framework provided the necessary scientific background to understand the complexity of microplastics contamination, both in terms of its origins and its analytical challenges.

In order to address these challenges from a complementary perspective, the second part of this work focuses on the development and application of machine learning techniques for automated microplastics identification and classification. Unlike traditional spectroscopic analysis, which is often time-consuming and reliant on expert interpretation, machine learning offers a scalable and reproducible approach to data interpretation. It is particularly suited to spectral data, where subtle variations in signal intensity and peak position encode critical information about the material composition.

The Raman spectra analyzed in this study were collected from real-world plastic fragments retrieved from domestic waste sources, providing a realistic and heterogeneous dataset. The primary goals of the analysis are twofold: to determine whether a given spectrum corresponds to a plastic material and to classify the spectrum into one of the main recyclable polymer types (PET, PVC, PP, etc.)

This section of the thesis therefore presents a practical and data-driven extension of the previous theoretical work, with the long-term vision of enabling rapid, in situ or laboratory-based classification of microplastics. Such an approach could support monitoring strategies and source attribution efforts by linking polymer type to potential contamination sources.

11 Introduction to Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence concerned with the design of algorithms that allow computers to learn from data and improve their performance without being explicitly programmed [13]. ML models are typically used to solve problems that are too complex to be addressed through conventional algorithmic approaches, such as image recognition, natural language understanding, and anomaly detection.

At its core, a machine learning system consists of three main components: a model, a training algorithm, and a dataset. The *model* (also referred to as a hypothesis) defines the input-output relationship, often represented as a parametric function $h_\theta(x) \to y$, where θ are the parameters to be learned. The *learning algorithm* aims to optimize the parameters by minimizing a *loss function*, which quantifies the discrepancy between the model's predictions and the true outcomes.

A central concept in ML is **generalization**, the ability of a trained model to perform well on unseen data. Without generalization, ML would merely be a form of sophisticated memorization. Generalization is typically assessed using validation strategies such as hold-out validation or cross-validation, which simulate the model's performance on independent test data.

Three primary learning paradigms are distinguished in machine learning:

- **Supervised Learning**: The model is trained on labeled data, where each input example is paired with a desired output. Tasks include classification (e.g., spam detection) and regression (e.g., predicting house prices). This is the most commonly used approach, as it provides a direct mapping between inputs and outputs.

- **Unsupervised Learning**: The model is trained on unlabeled data and attempts to identify inherent structures or patterns, such as clusters or anomalies. Common tasks include clustering, anomaly detection, and dimensionality reduction. A key advantage is that unlabeled data are often easier and cheaper to collect.

- **Reinforcement Learning**: The model interacts with an environment and learns by receiving feedback in the form of rewards or penalties. This approach is particularly suited for sequential decision-making tasks such as robotics or game playing.

Each paradigm has its own strengths and limitations. Supervised learning benefits from clear objectives but requires expensive labeled data. Unsupervised learning offers flexibility but lacks direct evaluation metrics. Reinforcement learning enables adaptive strategies but is often computationally intensive and sensitive to reward formulation.

ML systems are known to behave as "black boxes", which means their internal reasoning is difficult to interpret. This lack of transparency can be problematic in sensitive applications, such as medical diagnostics or autonomous driving. Another limitation is the dependency on high-quality data: biased, incomplete, or noisy datasets can lead to inaccurate or even harmful predictions.

It is also important to recognize that machine learning models are fundamentally statistical, not causal. As such, while they may detect correlations and make accurate predictions, they do not uncover causal relationships. In real-world applications, this distinction must be clearly understood to avoid misinterpretations of model outputs [13].

Despite these challenges, the versatility and performance of machine learning make it an essential tool in modern data analysis, especially in domains like environmental

monitoring and material classification, where complex, high-dimensional data such as spectroscopy signals must be interpreted efficiently.

11.1 Classification in Machine Learning

Classification is one of the most common and fundamental tasks in machine learning. It consists in finding a rule that assigns each input instance to one category among a finite set of predefined, mutually exclusive, and exhaustive classes [13]. This task is pervasive across real-world applications: for example, in medical diagnosis, patients are categorized as healthy or sick based on clinical data; in email filtering, messages are marked as spam or legitimate; and in fraud detection, transactions are labeled as either fraudulent or trustworthy.

From a formal point of view, a classifier is a function $h : X \to Y$, where X is the space of input features and Y is the set of possible class labels, typically represented as integer values $\{0, 1, \ldots, k-1\}$. While the labels are often encoded numerically for computational convenience, they do not possess any intrinsic order or metric relationship.

The input data X can take various forms, although it is commonly represented as vectors of real-valued features. Each training sample is a pair (x_i, y_i), with $x_i \in X$ and $y_i \in Y$, forming the training set $\{(x_0, y_0), \ldots, (x_{m-1}, y_{m-1})\}$. Supervised learning methods attempt to learn a classifier by optimizing a function within a parametric family $\{h_\theta : \theta \in \Theta\}$ that best replicates the correspondence between inputs and labels.

Classification problems form the backbone of many scientific and industrial applications of machine learning, including the identification of microplastics based on spectroscopic features, where each material sample must be assigned to a known polymer class.

12 Motivation and Application: ML for Microplastics Classification

12.1 Why Use Machine Learning for Microplastics Classification

Microplastics (MPs) are highly heterogeneous particles found in various environmental matrices such as water, sediment, and biota. Their variability spans size (from micrometers to millimeters), shape (fragments, fibers, films), and polymer type (e.g., PE, PET, PVC, PS). This physical and chemical diversity is reflected in the spectroscopic data used for their identification—especially Raman spectra—making the classification task extremely complex.

Traditional methods struggle to define deterministic rules that capture this variability. Spectra often contain overlapping peaks, noise, baseline shifts, and fluorescence background, all of which complicate analysis. Furthermore, MPs in environmental samples are often weathered, leading to chemical degradation and further distortion of spectral features.

Machine learning (ML) is a natural fit for this scenario. It excels at learning non-linear relationships in high-dimensional data, and it can generalize well even in the presence of noise and partial information. Supervised ML models learn from labeled examples to distinguish patterns associated with different polymers, enabling the classification of new, unseen samples with minimal human intervention. ML also supports adaptability: the same algorithm can be retrained on new datasets as more microplastics types or sampling conditions are encountered.

12.2 Advantages over Traditional Analytical Methods

Conventional identification techniques such as manual spectral interpretation, reference library matching, or chemical staining are labor-intensive and error-prone. These approaches often require significant domain expertise and are poorly suited for large-scale studies. Even automated methods based on rule-based algorithms can struggle with the spectral variability inherent in MPs.

ML techniques, in contrast, offer:

- **Scalability**: Once trained, ML models can process large datasets rapidly.
- **Automation**: Feature extraction, dimensionality reduction, and classification can be performed end-to-end with minimal human input.
- **Robustness**: ML models are less sensitive to minor spectral distortions or shifts, which are common in real-world samples.
- **Adaptability**: Models can be re-trained or fine-tuned with new data, allowing rapid incorporation of new classes or environmental conditions.

Furthermore, deep learning models like Convolutional Neural Networks (CNNs) or Support Vector Machines (SVM) can work directly on raw or minimally processed Raman spectra, learning to extract relevant features automatically. This reduces preprocessing effort and leads to improved performance on difficult classification tasks.

12.3 Related Work and Previous Studies

A comprehensive review by Sunil et al. (2024) [14] consolidates recent advances in machine learning applications for the classification of microplastics using Raman spectroscopy. The study emphasizes the growing importance of ML as a core analytical tool in environmental monitoring of MPs.

The authors describe the integration of various ML algorithms—including classical models like k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), and more complex architectures such as Multi-Layer Perceptrons (MLPs), Autoencoders, and CNNs—for polymer identification from Raman spectra. These approaches consistently achieve high classification accuracies, often exceeding 95%, even under challenging spectral conditions such as fluorescence interference or low signal-to-noise ratios.

Several case studies are cited in which ML pipelines were successfully applied to environmental and commercial samples, demonstrating the models' ability to generalize beyond the laboratory setting. In particular, 1D-CNN models trained on raw Raman spectra have been shown to outperform traditional methods in both speed and accuracy. Autoencoders and PCA-based methods have also proven effective in extracting meaningful representations from high-dimensional spectral data, improving interpretability and downstream classification.

Importantly, Sunil et al. stress that model performance is heavily influenced by preprocessing steps (e.g., baseline correction, normalization, dimensionality reduction) and the structure of the input data (e.g., full spectrum vs. selected features). The paper recommends the development of standardized workflows combining spectroscopy, ML, and domain knowledge to ensure repeatability and comparability across studies.

In conclusion, this body of work demonstrates that ML, when properly integrated with Raman spectroscopy, is not only a viable but a superior alternative for microplastics classification. It enables scalable, accurate, and reproducible analysis, making it a valuable tool in environmental science and pollution monitoring.

13 Dataset Description

13.1 Structure and Composition of the Dataset

The dataset used in this study consists of Raman spectra collected from real-world plastic fragments sourced from domestic waste. Each sample corresponds to one of the seven standard recycling codes (01 to 07), which categorize the most common recyclable polymers: polyethylene terephthalate (PET, code 01), high-density polyethylene (HDPE, code 02), polyvinyl chloride (PVC, code 03), low-density polyethylene (LDPE, code 04), polypropylene (PP, code 05), polystyrene (PS, code 06), and a mixed category of other or multilayer plastics (code 07).

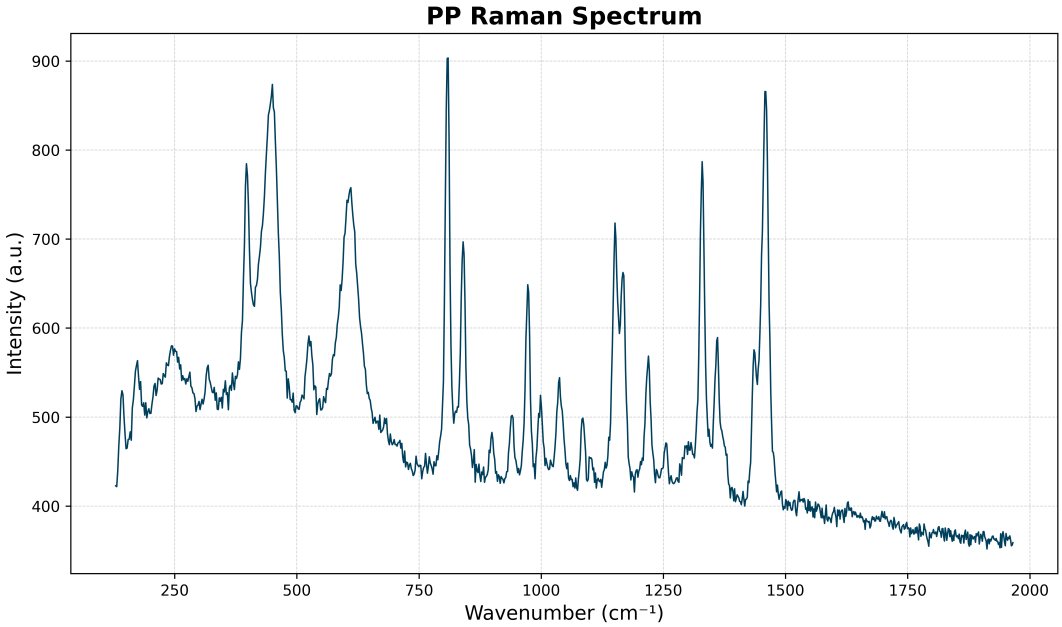

Figure 21: Example Raman spectrum from a transparent polypropylene (PP) sample.

For each plastic type, 900 spectra were collected from *transparent or white* samples, which generally provide clean and easy-to-interpret signals. In addition to these, the dataset also includes a total of 5413 spectra from *colored* samples, introducing greater variability and making the dataset more realistic. These colored samples introduce realistic complexity due to the presence of pigments, dyes, and surface treatments, which can alter the spectral features through absorption or fluorescence effects.

The breakdown by plastic type and sample type is summarized in Table 2. Notably, the quantity of colored spectra varies widely across classes, from 0 in PET to over 1700 in polypropylene. This distribution reflects both the availability of real-world material and the practical relevance of pigmented samples in environmental contexts.

Table 2: Distribution of Raman spectra per plastic type in the dataset

Plastic Type	**Code**	**Transparent Samples**	**Colored Samples**
Polyethylene Terephthalate (PET)	01	900	324
High-Density Polyethylene (HDPE)	02	900	1033
Polyvinyl Chloride (PVC)	03	900	324
Low-Density Polyethylene (LDPE)	04	900	1638
Polypropylene (PP)	05	900	1700
Polystyrene (PS)	06	900	354
Other Plastics (e.g., multilayer)	07	900	364
Total	–	**6300**	**5737**

The high representation of colored samples allows for more robust training of machine learning models, especially with respect to real-world variability. However, the uneven distribution of pigmented samples across classes may still introduce class imbalance issues and must be accounted for during model evaluation and training.

Additionally, class 07 (Other Plastics) includes multilayer and unidentified polymers, making it inherently heterogeneous. This introduces label ambiguity, which may lead to reduced classification performance and should be considered when interpreting the model's results.

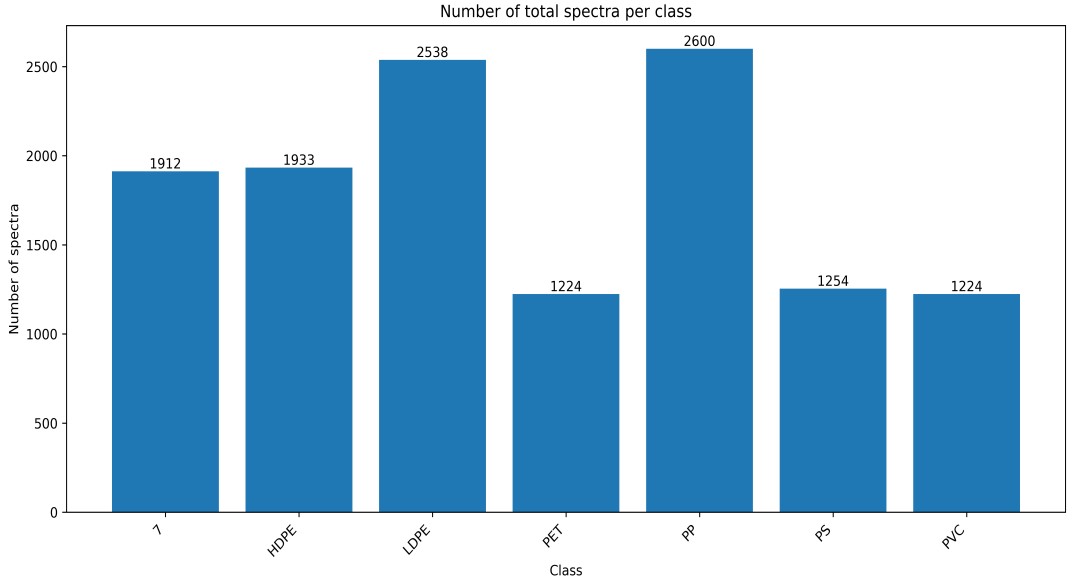

Figure 22: Histogram representing the database distribution.

13.2 Spectral Format and Feature Representation

Each Raman spectrum is stored as a plain text file containing two columns:

- The first column lists the **wavenumbers** (in cm^{-1}), which serve as the feature indices. Each spectrum includes 1005 discrete wavenumber values, consistently sampled across all files.
- The second column contains the corresponding **intensity values**, representing the Raman signal measured at each wavenumber. These values are the primary input features used for classification.

Consequently, each spectrum is represented as a 1005-dimensional real-valued feature vector. This fixed-length structure is ideal for integration into standard machine learning pipelines, facilitating training and evaluation of classifiers that expect tabular numerical inputs.

13.3 Relevance for Supervised Classification

The dataset used in this study is well-suited for supervised learning, as each Raman spectrum is explicitly labeled with its corresponding plastic type, from 01 (PET) to 07 (Other). This allows the model to learn associations between spectral features and polymer identity in a structured and traceable way.

A major strength of the dataset lies in its diversity. It includes both transparent/white and colored plastic samples, reflecting realistic conditions where visual appearance and chemical composition may vary due to additives, pigmentation, or environmental degradation. Such diversity enhances the model's ability to generalize and perform reliably on heterogeneous, real-world inputs.

However, this variability is not evenly distributed across classes. Some polymers—such as PP and LDPE—are heavily represented among colored samples (with more than 1600 spectra each), while others—such as PET—have no colored examples. This imbalance may introduce bias during training and result in uneven class-wise performance, particularly in cases where pigmentation affects spectral features.

Additionally, the class labeled as 07 (Other Plastics) includes multilayer, blended, or unknown polymer types. Its inherently heterogeneous nature results in label ambiguity and increased spectral overlap with other classes. These factors can limit the model's ability to learn reliable classification boundaries for this group and reduce metrics such as precision and recall.

Despite these challenges, the dataset represents a realistic and valuable foundation for developing machine learning models in the context of microplastics analysis. Its complexity mirrors the variability encountered in laboratory and environmental settings, making it a robust platform for evaluating classification strategies and exploring future improvements.

14 Preprocessing and Feature Engineering

14.1 Methodological Choices and Rationale

In this study, each Raman spectrum is represented as a 1005-dimensional numerical vector, where the i-th component corresponds to the Raman intensity at the i-th wavenumber. These vectors are extracted from the second column of each `.txt` file, assuming consistent formatting across the dataset.

Before training the classifier, the data undergoes preprocessing to improve model performance and generalization. Specifically, feature scaling is applied using z-score standardization (i.e., centering and scaling to unit variance) via `StandardScaler`. This transformation ensures that all features contribute equally to the SVM decision function, which is particularly important when using kernels such as the radial basis function (RBF).

No other transformations—such as smoothing, baseline correction, or dimensionality reduction—were applied. The analysis was performed directly on the raw standardized spectra to preserve the intrinsic structure of the Raman signal.

The class label for each spectrum (i.e., the plastic type) is inferred from the filename prefix, and used as the categorical target variable during training and evaluation.

No dimensionality reduction techniques, such as Principal Component Analysis (PCA), were applied during this phase of the project. The decision to retain all original 1005 spectral dimensions was made to preserve the full informational content of the spectra and to assess the baseline performance of the classification model without any information loss.

The decision to avoid preprocessing, normalization, or feature reduction was intentional. The primary goal was to test the ability of a simple machine learning pipeline to classify microplastics types using the original spectral intensities as-is, without any transformation or denoising.

This approach offers two advantages: (i) it reduces the complexity of the pipeline, making it more interpretable and reproducible, and (ii) it avoids introducing biases or artifacts that may arise from aggressive preprocessing. Furthermore, this choice aligns with the exploratory nature of this project, where the emphasis is on establishing a robust baseline before integrating more advanced preprocessing or feature extraction strategies.

To ensure proper evaluation, the dataset was split into training and test subsets using a stratified sampling approach. Specifically, 70% of the data for each plastic type (randomly chosen) was allocated to the training set, while the remaining 30% was used as a hold-out test set. This ensured that all the samples were represented proportionally in both sets, preserving the diversity and class balance throughout the analysis.

15 Model Architecture and Training

15.1 Chosen Model Type (SVM)

In this work, a Support Vector Machine (SVM) was selected as classification model. SVMs are powerful supervised learning algorithms that have proven particularly effective in high-dimensional spaces and under limited sample conditions—both of which are typical characteristics of Raman spectral datasets. Their foundations are rooted in statistical learning theory, specifically in the principles of margin maximization and structural risk minimization [15].

The fundamental idea behind SVMs is to find a decision boundary—formally, a hyperplane—that separates data points belonging to different classes while maximizing the margin, i.e., the distance between the hyperplane and the nearest data points from each class. This optimal separating hyperplane is defined by the equation $w \cdot x + b = 0$, where w is the weight vector orthogonal to the hyperplane and b is the bias term. The margin is given by $2/\|w\|$, and maximizing it reduces the model's complexity and improves generalization.

The data points that lie on the boundary of this margin are called *support vectors*. These are the most critical samples for the classifier, as they determine the position and orientation of the decision surface. Points farther away from the hyperplane do not influence the solution, making SVMs naturally resistant to overfitting in high-dimensional spaces.

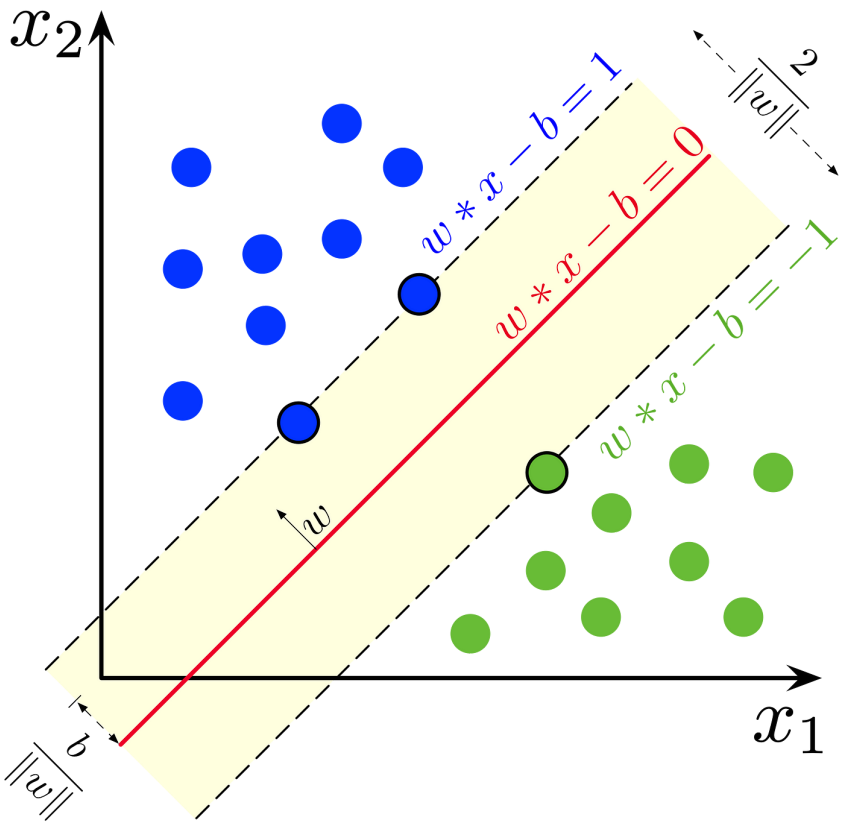

Figure 23: Mathematical representation of the basic concept of SVM classification.

15.1.1 Handling Non-Separable Data

In most real-world datasets, including Raman spectra of microplastics, perfect linear separability is rarely achievable. This can be due to spectral overlap, noise, or intrinsic variability within and between polymer classes. To handle such cases, SVMs introduce *slack variables* ξ_i that allow some misclassifications while still aiming to maximize the margin. The trade-off between maximizing the margin and minimizing classification error is controlled by the regularization parameter C:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

A small value of C allows a wider margin with more tolerance for misclassification, while a large C forces the model to prioritize accuracy over generalization.

15.1.2 Non-Linear Separation and Kernels

If the classes are not linearly separable even with slack variables, SVMs can apply the *kernel trick*, which implicitly maps the data into a higher-dimensional feature space where a linear separation may become possible. Popular kernel functions include:

- **Polynomial kernel:** suitable when interaction features are important.
- **Radial Basis Function (RBF):** ideal for capturing complex boundaries; it measures similarity in a local sense.
- **Sigmoid kernel:** similar in behavior to a neural network activation.

These kernel functions allow the algorithm to operate in an implicit feature space without explicitly computing the transformation, preserving computational efficiency.

Why a Linear Kernel Was Chosen

In this study, a linear kernel was selected for several reasons:

- **Interpretability:** Linear SVMs allow direct analysis of feature importance through the weight vector w. This is particularly valuable for Raman spectroscopy, where identifying influential wavenumbers aids physical interpretation.
- **Simplicity and robustness:** Initial exploratory analyses (e.g., PCA) showed partial linear separability in the dataset, justifying the use of a simpler linear model over more complex kernels.
- **Computational efficiency:** Linear SVMs train faster and are less prone to overfitting when the number of features is large and the sample size moderate, as in this case.

Furthermore, linear SVMs have been shown in the literature to perform very effectively on spectroscopic data, particularly when the input features (intensity values) are standardized—as was done here using z-normalization.

15.1.3 Application to Raman Spectroscopy

The suitability of SVMs for Raman spectral classification lies in their ability to handle high-dimensional, sparse, and partially overlapping data. Raman spectra often contain subtle yet class-defining differences that are distributed across many wavenumbers. SVMs are designed to capture these differences by focusing only on the most informative samples (support vectors), ignoring irrelevant variance. This makes them ideal for microplastics classification, where spectra may vary due to colorants, degradation, or instrumental factors.

Additionally, the linear SVM provides a practical benefit: the magnitude of the learned coefficients directly reflects the contribution of each wavenumber to the classification decision. This has been exploited in this work to generate feature importance plots and understand the physical basis of the model's predictions.

Overall, the SVM offers a principled, interpretable, and effective solution for multi-class classification of Raman spectra from microplastics fragments.

15.2 Model Configuration and Hyperparameter Optimization

The SVM model was implemented using the `scikit-learn` framework. To identify the optimal hyperparameters, a grid search with a 5-fold cross-validation was performed on the training set. The search space included multiple values for the regularization parameter C, kernel type, and the γ parameter (kernel coefficient). The best-performing configuration found was:

- **Kernel**: `linear`
- **Regularization parameter** C: `10`

This configuration was chosen because it yielded the highest cross-validation accuracy on the training set.

15.3 Training Strategy and Frameworks Used

The dataset was divided into training and test sets using a stratified split to preserve class distributions. Model training and hyperparameter search were conducted using the `GridSearchCV` utility from `scikit-learn`. Data loading, preprocessing, and formatting were handled with `pandas`, ensuring compatibility with the ML pipeline.

Batching was not required due to the relatively small dataset size and the nature of the SVM algorithm, which is typically trained on the full dataset at once. The `linear` kernel was selected for its computational efficiency and robustness in linearly separable or near-separable cases, which is often the case when Raman spectra are properly normalized and denoised.

16 Evaluation and Results

16.1 Evaluation Metrics

To assess the performance of the classification model, several standard metrics were employed: accuracy, precision, recall, F1-score, and confusion matrix. These metrics are defined as follows:

- **Accuracy**: the proportion of correctly predicted samples over the total number of predictions.
- **Precision**: the proportion of true positives among all samples predicted to belong to a given class.
- **Recall**: the proportion of true positives correctly identified out of all actual samples in a class.
- **F1-score**: the harmonic mean of precision and recall, providing a balance between the two.
- **Confusion Matrix**: a matrix that reports the number of correct and incorrect predictions for each class.

16.2 Performance on Test Set

The model achieved an overall accuracy of **99.57%** on the test set. Here's the output of the model:

```
Accuracy: 0.9957

Classification Report:
```

	precision	recall	f1-score	support
7	1.00	0.99	0.99	386
HDPE	1.00	0.99	1.00	379
LDPE	0.99	1.00	0.99	508
PET	1.00	1.00	1.00	248
PP	0.99	0.99	0.99	513
PS	1.00	1.00	1.00	262
PVC	1.00	1.00	1.00	241

16.3 Cross-Validation and Generalization

To evaluate the generalization capability of the model, a 5-fold cross-validation was conducted on the training set. The mean cross-validation accuracy was **99.65%**, indicating excellent consistency across folds:

- Fold 1: 99.66%

- Fold 2: 99.61%
- Fold 3: 99.61%
- Fold 4: 99.56%
- Fold 5: 99.80%

The small variation between folds demonstrates high robustness and low variance of the classifier, suggesting that the model is not overfitting to specific subsets of the data.

16.4 Intra-Class Variability

To investigate the internal consistency of each plastic class, the standard deviation of the intensity values across all spectra in each class was computed and averaged. This provides an estimate of intra-class spectral variability:

- 7: average std = 3232.18265
- HDPE: average std = 3043.94013
- LDPE: average std = 4617.74830
- PET: average std = 3358.13853
- PP: average std = 3584.45470
- PS: average std = 823.98465
- PVC: average std = 625.58262

These values highlight the significant variability present in classes such as LDPE and PP, which often include pigmented or environmentally altered samples. In contrast, PET and PS show lower variability, possibly due to the homogeneity of their transparent sample subset.

16.5 Interpretation

The overall results confirm that the model is highly accurate and stable across both clean and complex spectral classes.

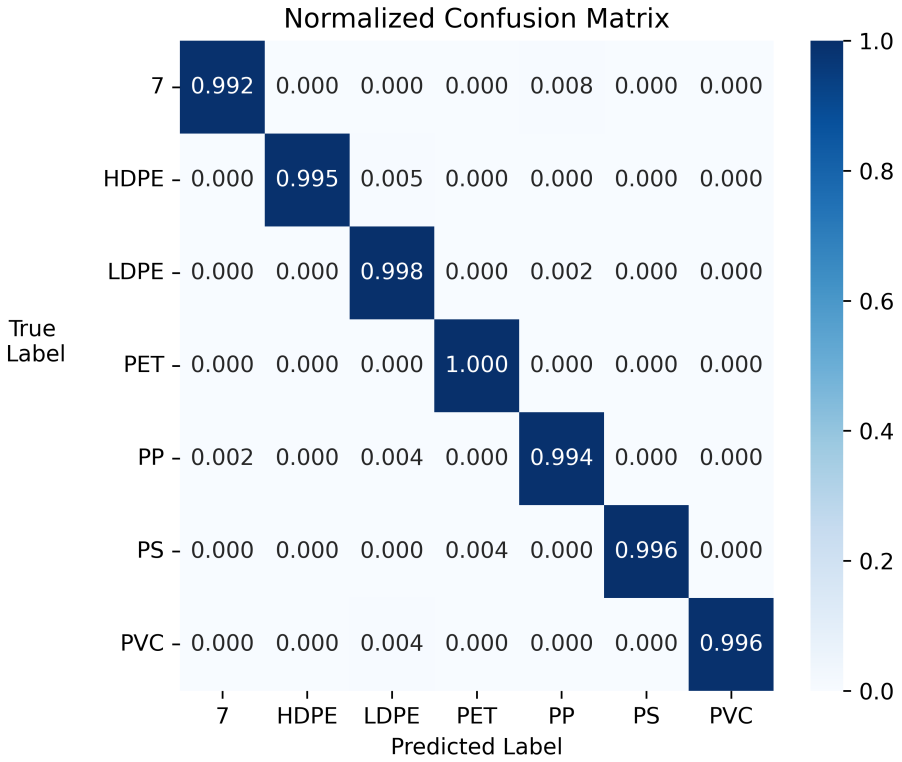

Figure 24: Confusion matrix on the test set showing classification performance per class.

17 Discussion of the Results

17.1 Analysis of Model Performance

The results obtained by the SVM classifier on the test set demonstrate excellent overall performance, with an accuracy of 99.49% and high precision and recall across all major classes. Misclassifications were rare, amounting to less than 1% of the test samples, indicating that the model has learned highly discriminative boundaries in the feature space.

Nonetheless, some variation in class-specific performance is observable. The lowest F1-scores were associated with classes `7` (mixed/other plastics), `LDPE`, and `PS`. These classes also exhibited higher intra-class variability, likely due to pigment content, manufacturing differences, or environmental degradation. As a result, the model may struggle to define consistent classification boundaries for these polymers.

17.2 Visualization of Class Separability

To better understand the structure of the dataset and the separability of the classes, dimensionality reduction techniques were applied to project the 1005-dimensional spectra into two-dimensional space.

17.2.1 Principal Component Analysis (PCA)

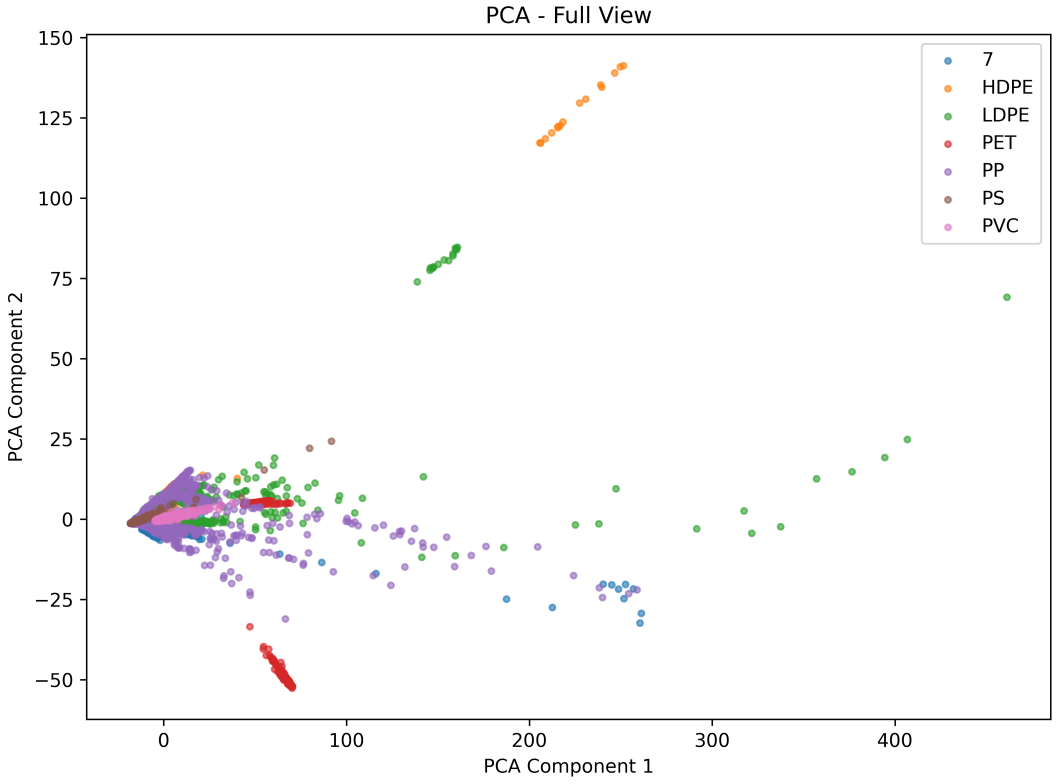

Figure 25: PCA projection of Raman spectra from the test set onto the first two principal components.

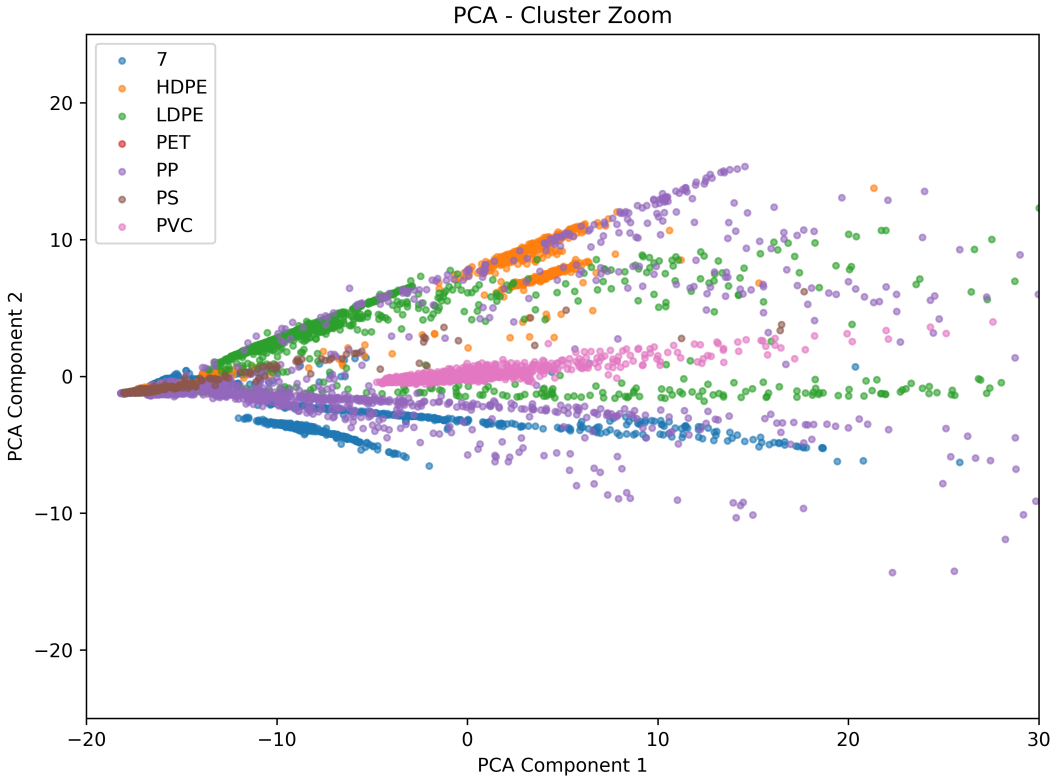

Figure 26: PCA projection of Raman spectra from the test set onto the first two principal components with a zoom on the cluster around the origin.

Principal Component Analysis (PCA) is a linear technique that projects high-dimensional data onto a lower-dimensional space by identifying the directions (principal components) that capture the most variance in the dataset. In the case of Raman spectra, PCA helps to identify whether certain plastic types tend to differ consistently across specific spectral features.

When interpreting a PCA plot:

- **Tightly clustered points** of the same class suggest high intra-class similarity — the spectra within that class are very similar, and the model is likely to classify them consistently.
- **Widely spread or elongated clusters** indicate high intra-class variability. This might correspond to physical variations such as pigment differences, sample degradation, or multilayer contamination.
- **Overlapping clusters from different classes** suggest that those polymers may share similar spectral features, making them harder to distinguish based on the selected components.
- **Outliers** (isolated points) may represent noisy spectra, unusual plastic fragments, or even mislabelled samples.

It is important to note that PCA captures only global linear relationships. Thus, some structures or similarities may not be visible in this projection, especially if they are non-linear in nature.

Plots of the first 7 main components are now shown below to improve the interpretability of the analysis by PCA.

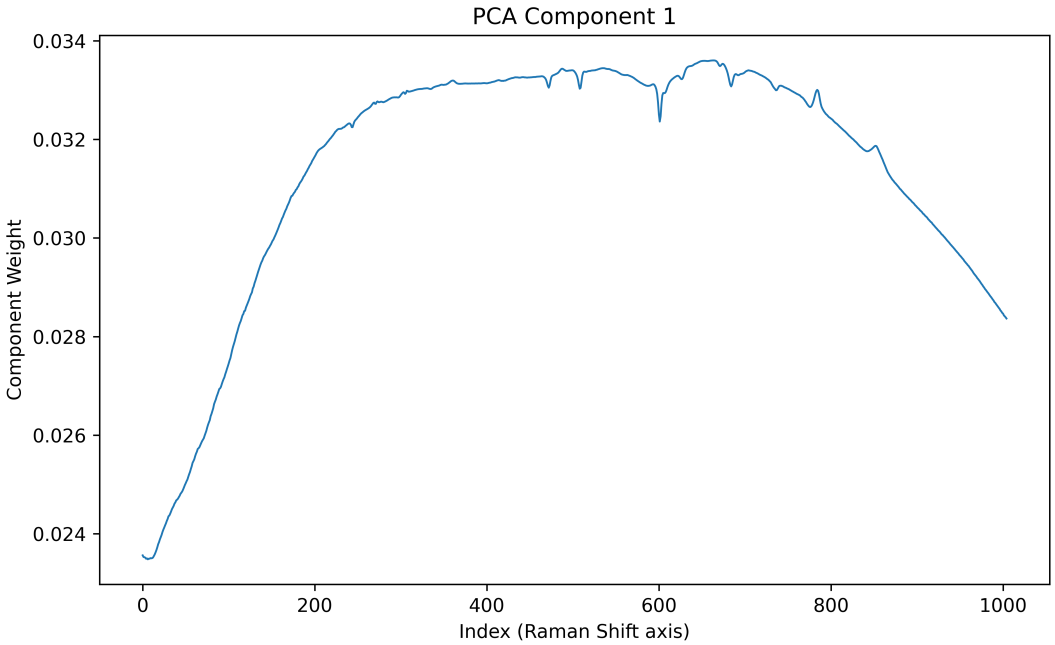

Figure 27: Plot of the first component of the PCA

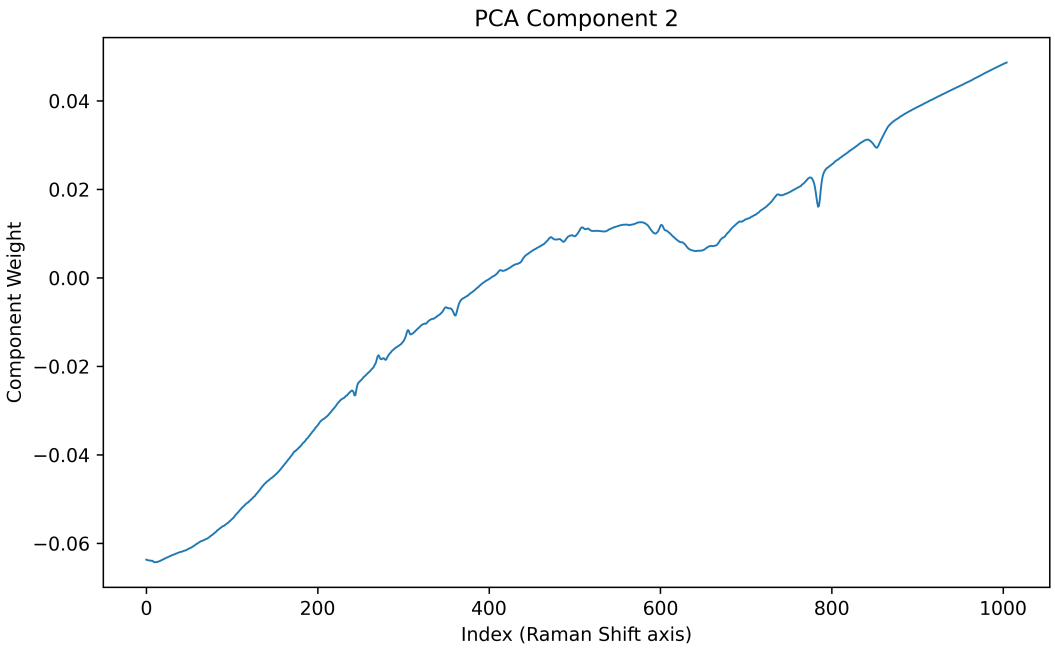

Figure 28: Plot of the second component of the PCA

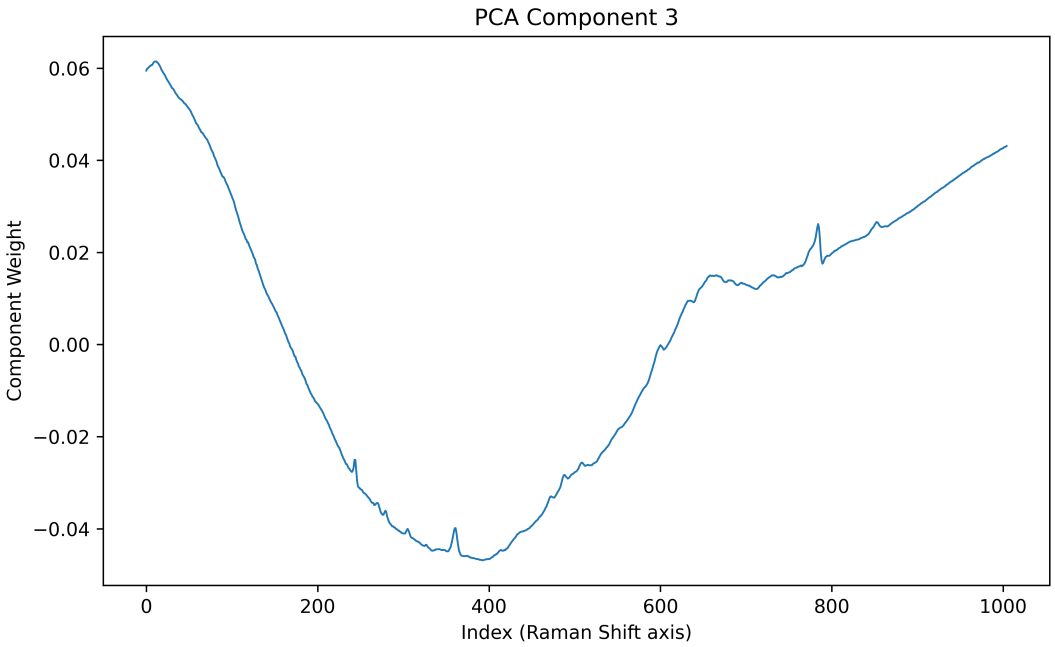

Figure 29: Plot of the third component of the PCA

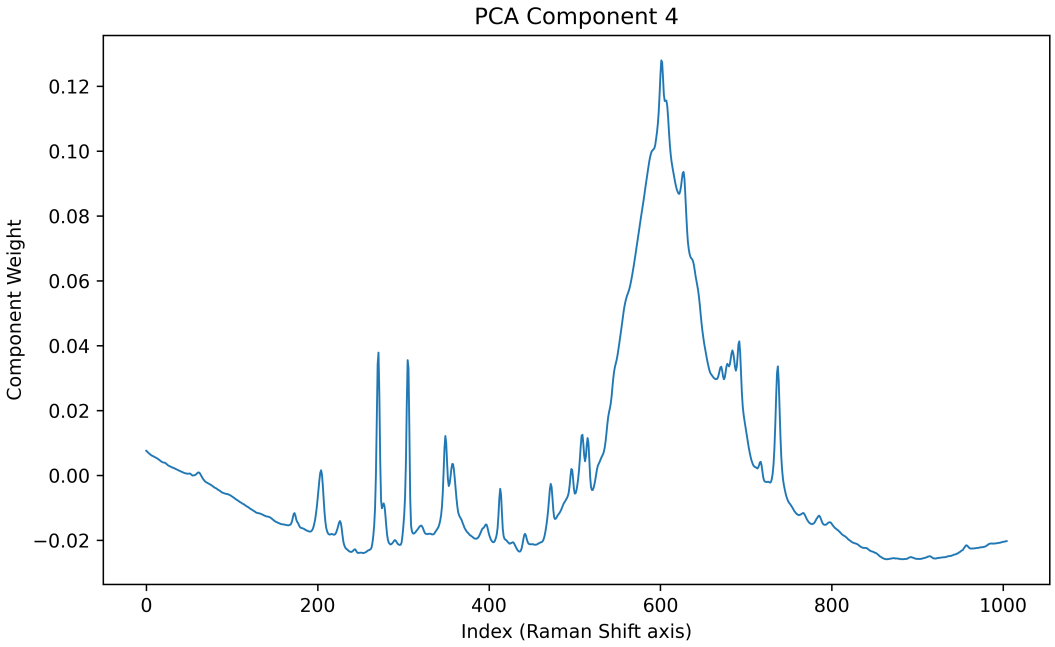

Figure 30: Plot of the fourth component of the PCA

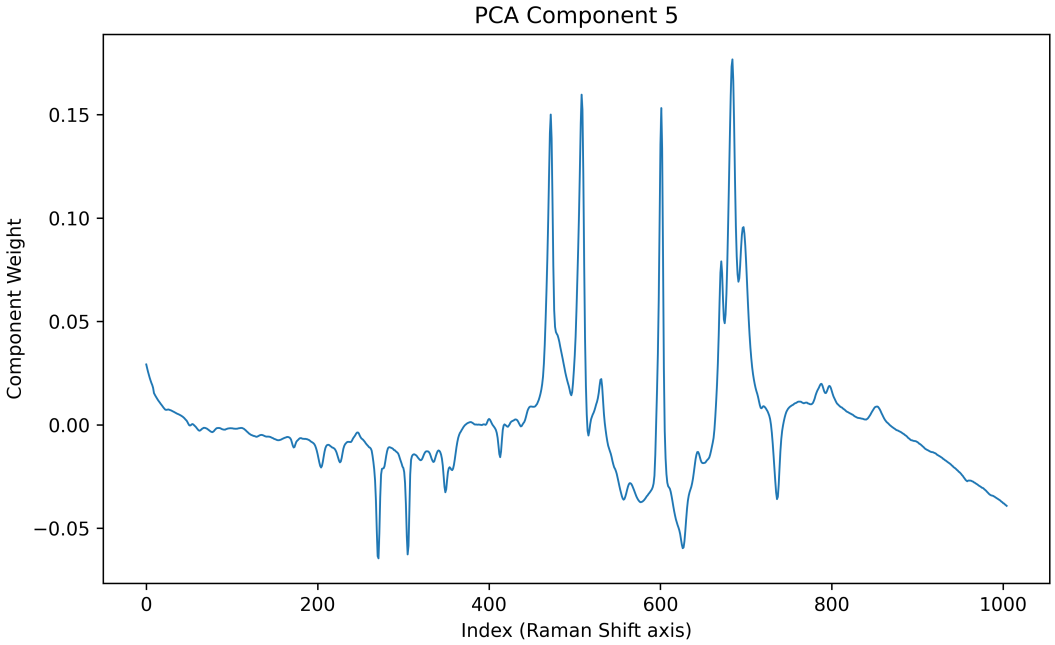

Figure 31: Plot of the fifth component of the PCA

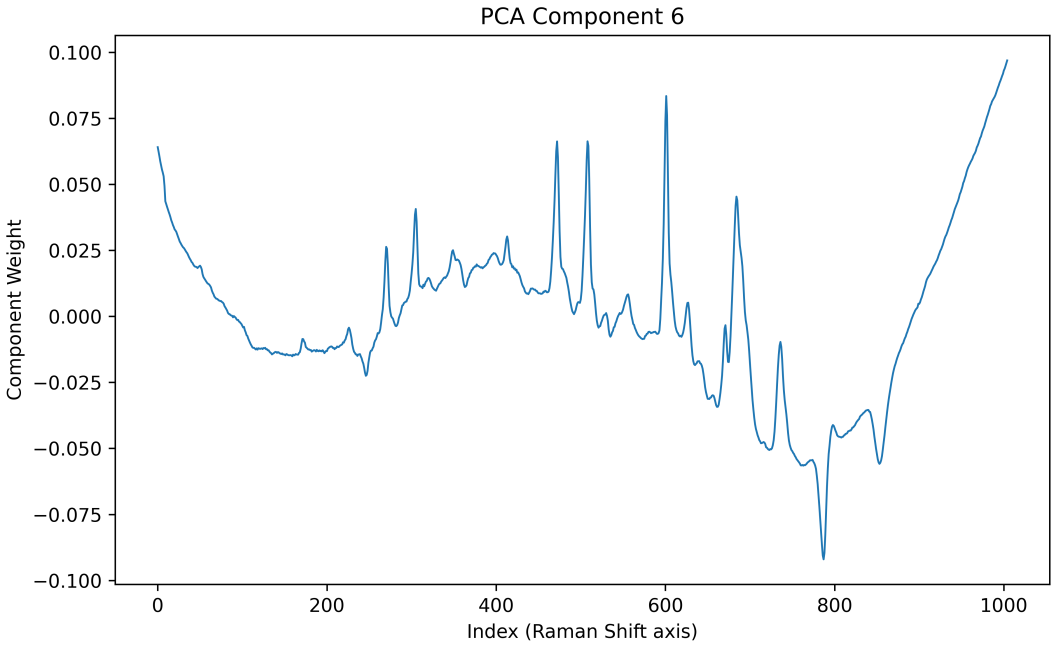

Figure 32: Plot of the sixth component of the PCA

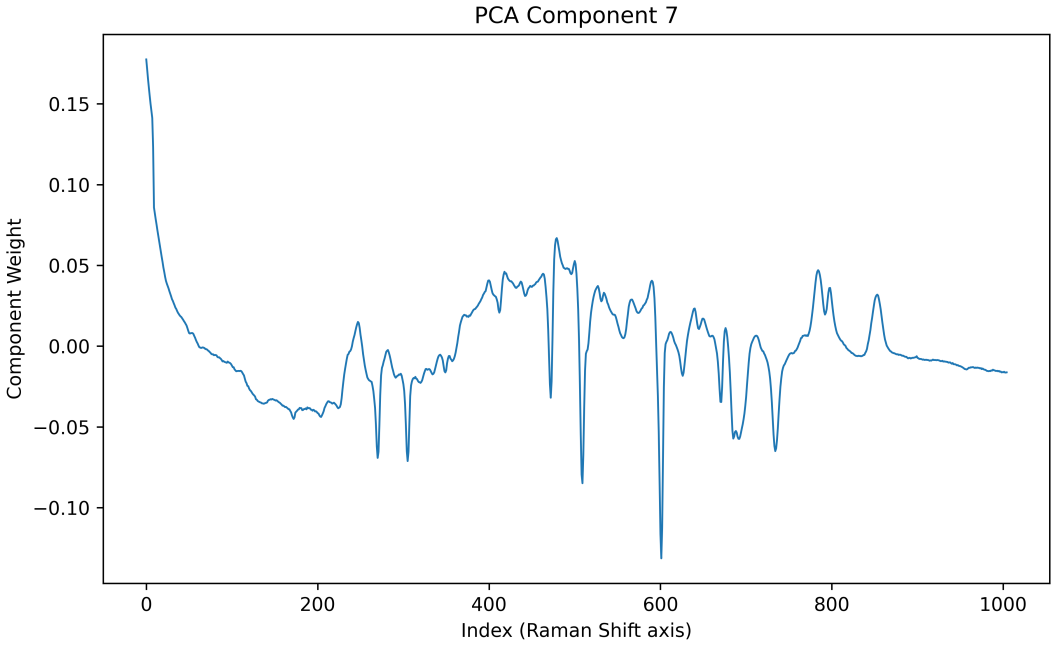

Figure 33: Plot of the seventh component of the PCA

The PCA plots presented in Figures 25 and 26 reveal several meaningful physical and chemical characteristics of the dataset. In the full projection (Fig. 25), certain classes such as HDPE and LDPE exhibit clear displacement along the first principal component, indicating a dominant spectral variation that strongly distinguishes them from other polymers. This separation may be attributed to broad differences in polymer backbone composition or crystalline content, which affect the Raman scattering profile in ways that PCA captures efficiently.

Meanwhile, the zoomed view in Figure 26 highlights more subtle intra-cluster differences. For example, classes like PP and PVC form relatively compact clusters, suggesting good spectral consistency within those materials. Conversely, classes such as LDPE and class 07 appear more elongated and dispersed, likely due to higher variability in sample composition, pigmentation, or degradation. In particular, the spread of class 07 is consistent with its role as a heterogeneous "other plastics" category, where spectral signatures vary significantly.

Importantly, some classes exhibit partial overlap in the PCA space—for instance, PP and PS—suggesting shared spectral features such as CH bending or symmetric stretch modes in overlapping regions of the Raman spectrum. While PCA cannot fully resolve these overlaps, especially when distinctions are nonlinear, it still provides useful clues about which plastic types may be inherently more difficult to separate using linear decision boundaries.

It is also important to consider what types of physical phenomena underlie the variance captured by the PCA components. In the context of Raman spectroscopy of polymers, the first few principal components typically reflect dominant vibrational modes or compositional differences that are consistent across many samples—such as the intensity of CH_2 or C=O stretches, the presence of aromatic rings, or global baseline shifts caused by pigmentation or fluorescence.

The fact that some classes remain compact while others appear elongated or diffuse in PCA space may point to real physical heterogeneity: for example, class 07 includes

multilayer and blended plastics, leading to non-systematic shifts in peak position and intensity that produce high intra-class variance. Similarly, the broader spread of LDPE may stem from different processing histories (e.g., film vs. bottle) or additive content.

PCA, being a linear method, cannot capture more complex interactions such as slight but consistent peak shifts, localized distortions due to surface effects, or non-linear pigment interactions. These kinds of structure may be better revealed through non-linear methods such as t-SNE, which will be discussed in the following section.

17.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

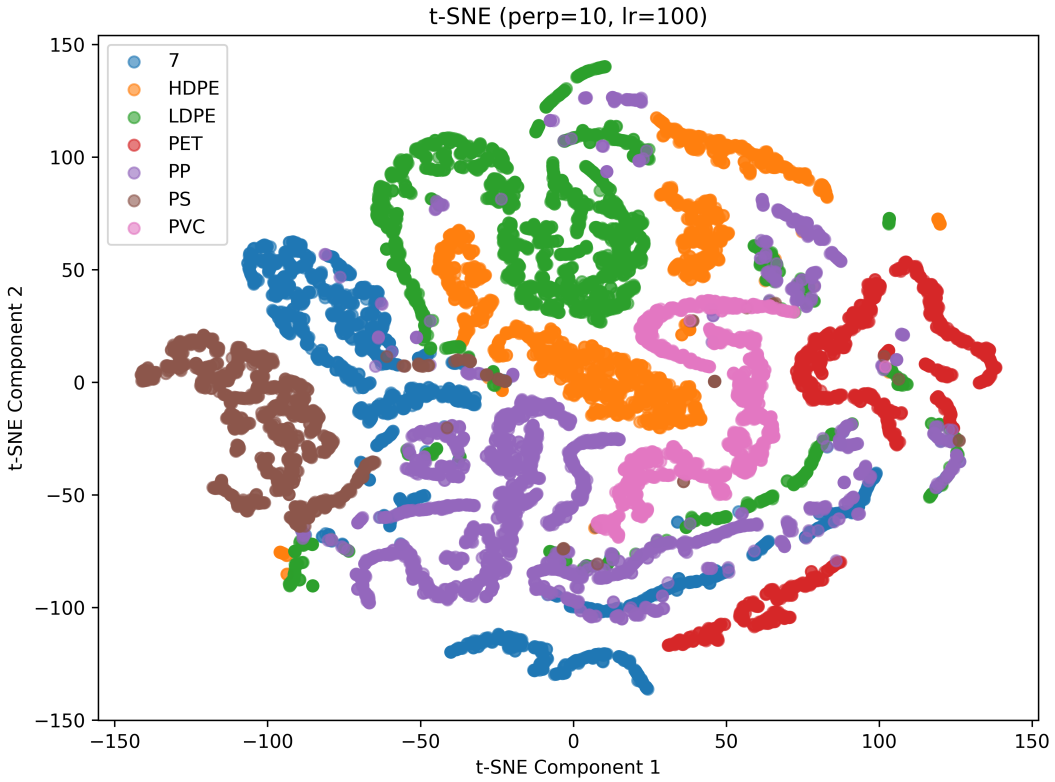

Figure 34: t-SNE projection of Raman spectra from the test set.

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique that is especially effective for visualizing local structure in complex datasets. It aims to preserve the similarity between close points in the high-dimensional space when projected onto two dimensions.

The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm involves several hyperparameters that influence the resulting 2D representation. In this work, we systematically explored the two most impactful parameters:

- **Perplexity** — values of 10, 30, and 50 were tested. This parameter controls the balance between local and global aspects of the data and effectively determines the number of nearest neighbors considered during embedding. Lower values tend to emphasize fine-grained local structure, whereas higher values promote more global coherence.

- **Learning rate** — values of 100, 200, and 500 were evaluated. The learning rate influences the optimization dynamics; inappropriate values may cause overlapping clusters or unstable embedding.

Other parameters were fixed to widely adopted defaults, based on literature and empirical practice:

- **Number of iterations** (`n_iter`) was set to 1500 to ensure stable convergence.

- **Initialization** (`init`) was set to `pca`, which typically improves stability across runs.
- **Random state** was fixed to 42 to ensure reproducibility.

After qualitative comparison of the resulting plots, the configuration with **perplexity = 10** and **learning rate = 100** was selected as it provided the clearest visual separation between spectral classes, with minimal distortion or overlap.

When reading a t-SNE plot:

- **Well-separated clusters** of the same color imply that the class is locally consistent — the model can likely distinguish it from others with high confidence.
- **Fragmented or diffuse clusters** may indicate significant intra-class variability or class ambiguity.
- **Overlapping clusters of different classes** suggest that those classes are spectrally similar in the raw data — this can stem from shared chemical structures, pigment-induced distortion, or environmental effects.

t-SNE does not preserve global geometry or distances between distant points, so interpretation should focus on local relationships — points that are close in 2D are similar in spectrum, but distant points may not be meaningfully distant in the original space.

17.3 Spectral Contribution Analysis

To gain insight into how individual spectral features contribute to classification decisions, a weight-based interpretability analysis was performed using the linear Support Vector Machine (SVM) model. Specifically, the learned model coefficients were analyzed to identify which wavenumbers (features) have the strongest influence on the decision boundaries for each plastic class.

Since a linear kernel was used, each class is associated with a specific set of weights applied to the 1005 Raman intensity features (i.e., wavenumbers). The magnitude of these weights reflects the relative importance of each feature in separating one class from the others. A higher absolute weight indicates a greater impact on the classification outcome.

The procedure involved:

- Standardizing all input spectra using z-normalization.
- Training a linear SVM on the full training set using the optimal hyperparameters obtained via grid search.
- Extracting and visualizing the learned model coefficients (weights) associated with each wavenumber.

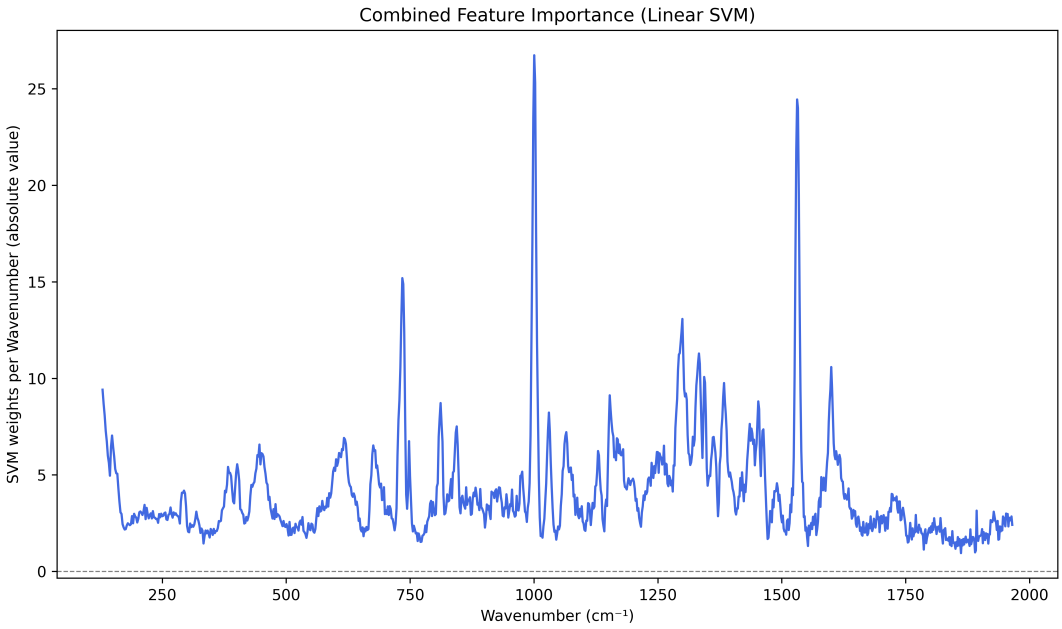

Figure 35: Overall feature importance: sum of absolute weights across all classes as a function of wavenumber.

Here we can also see an evaluation of the most significant wave numbers evaluated by the code before the plot creation:

```
Top 10 most important wavenumbers:
1000.5 cm^−1: 17.8409
1002.4 cm^−1: 16.2121
998.6 cm^−1: 15.6644
996.8 cm^−1: 11.9289
1004.2 cm^−1: 11.0196
1531.0 cm^−1: 8.8324
734.3 cm^−1: 8.6836
736.2 cm^−1: 8.6792
1532.7 cm^−1: 8.5342
994.9 cm^−1: 8.2648
```

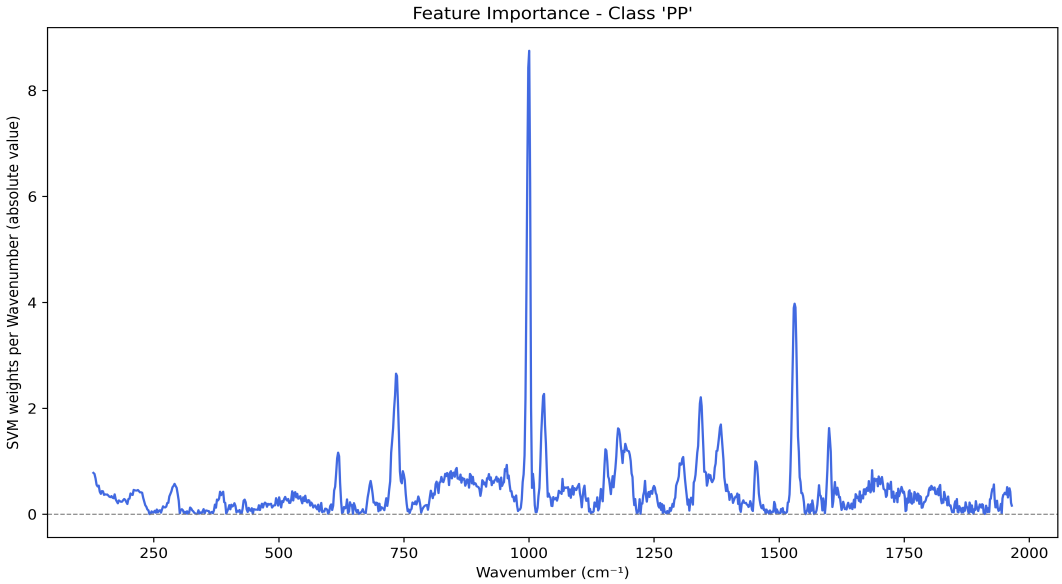

Figure 36: Feature importance for PP

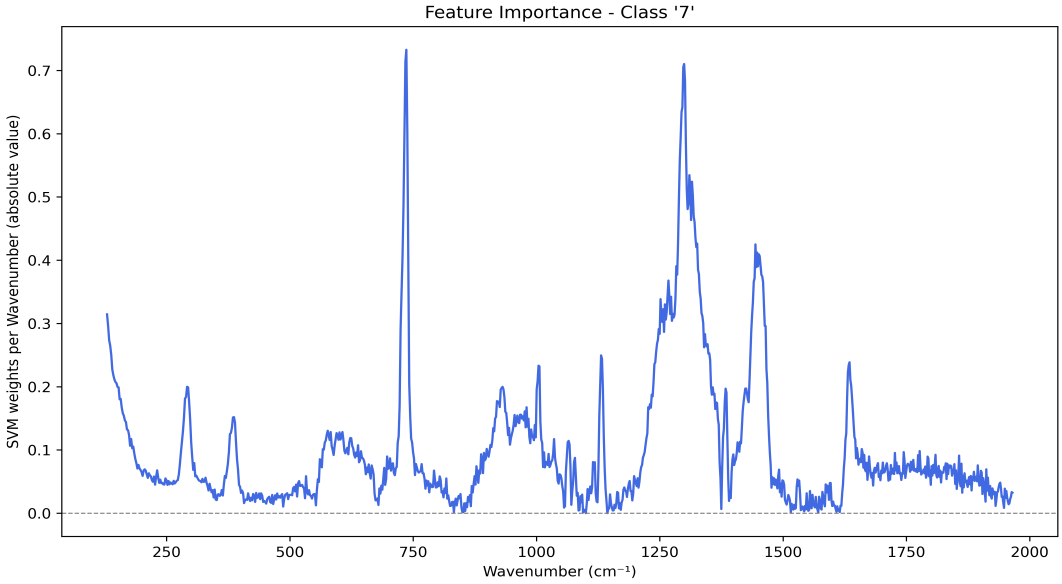

Figure 37: Feature importance for 7

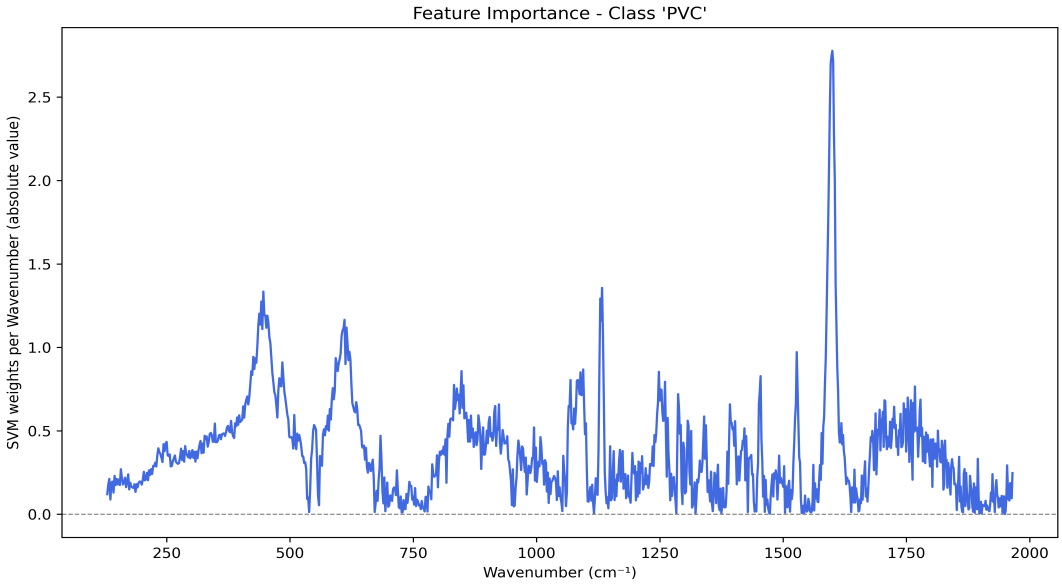

Figure 38: Feature importance for PVC

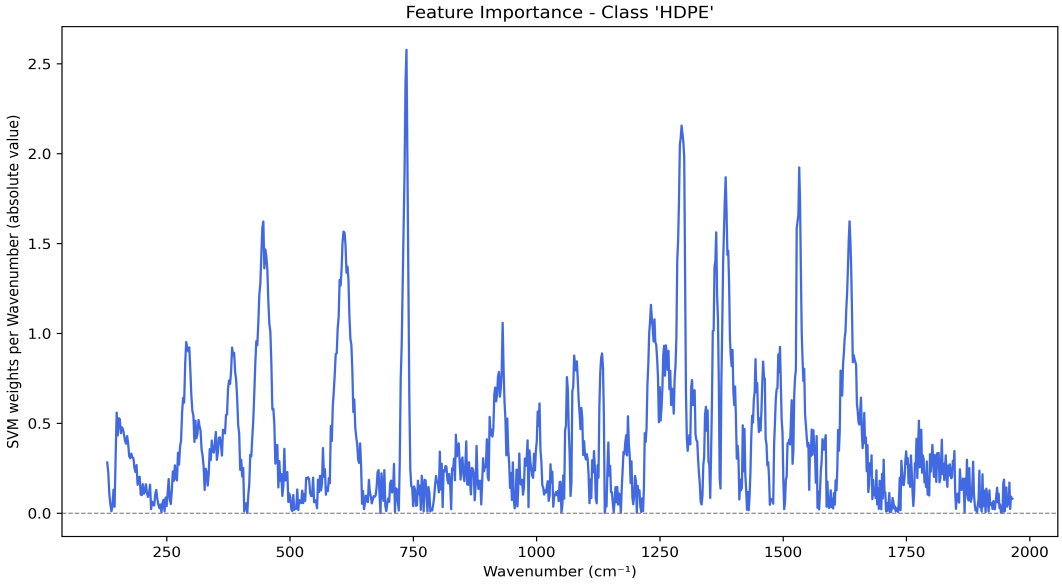

Figure 39: Feature importance for HDPE

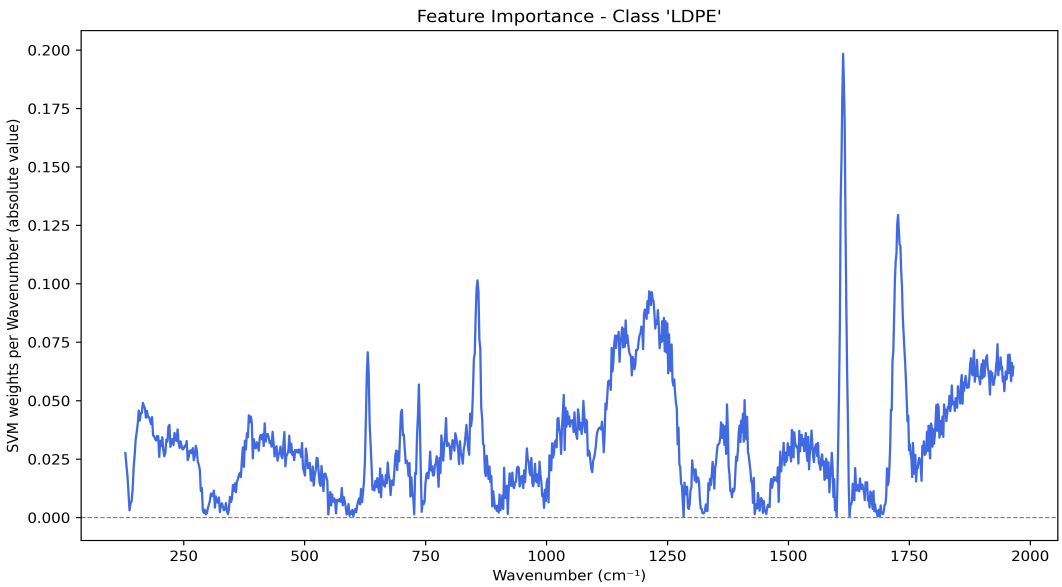

Figure 40: Feature importance for LDPE

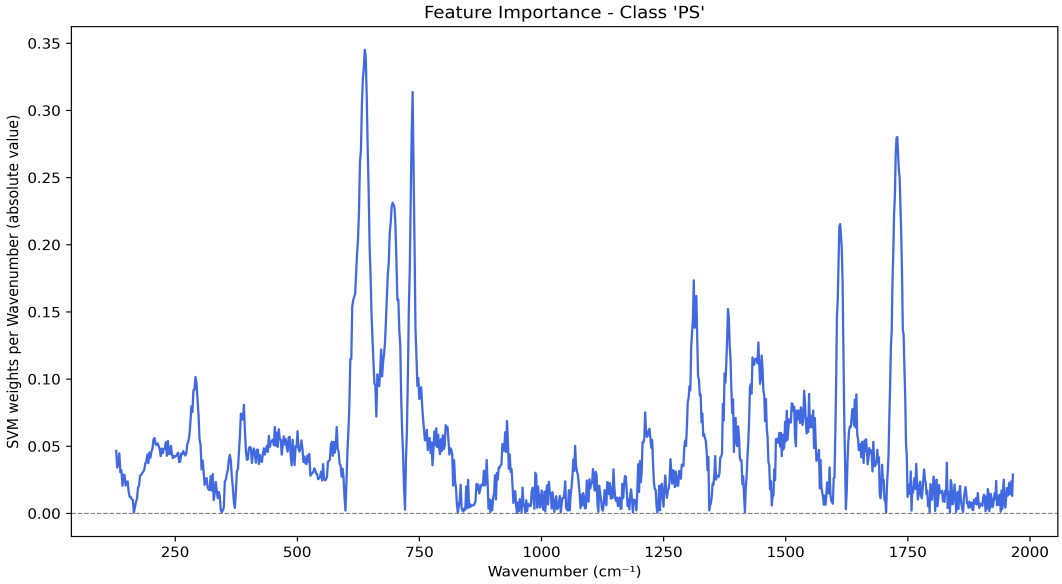

Figure 41: Feature importance for PS

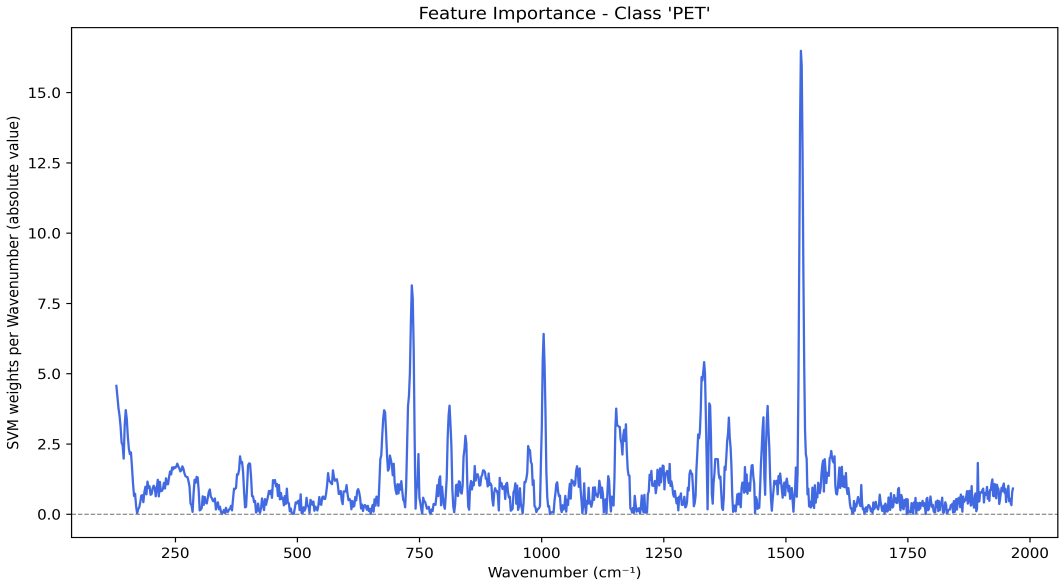

Figure 42: Feature importance for PET

In Figure 35, the wavenumbers with the highest cumulative influence across all classes are clearly visible as peaks. These may correspond to diagnostic spectral regions that are relevant for general plastic discrimination.

Finally, class-specific plots (e.g., Figure 40) provide a focused view of the most discriminative wavenumbers for each individual class. This kind of analysis is useful both for understanding the internal logic of the model and for validating that decisions are based on chemically meaningful spectral features.

This form of interpretability also supports potential physical validation — for instance, matching the peaks with known vibrational modes or confirming that color-related distortions do not dominate the classification logic.

18 Correlation Between Spectral Peaks and Model Weights

To better understand how the classification model makes its decisions, a set of visual analyses was performed comparing the average Raman spectra of each plastic type with the corresponding feature weights learned by the linear Support Vector Machine (SVM).

In a linear SVM, each input feature (in this case, each wavenumber) is associated with a weight in the vector w for each class. The magnitude and sign of this weight indicate how strongly the feature influences the classification outcome. On the other hand, the average spectrum represents the typical Raman signal associated with that plastic class, highlighting the main vibrational peaks.

By plotting both the average spectrum and the decision weights on the same graph (using two vertical axes), we aim to identify whether there is a correspondence between prominent spectral peaks and high-weight features in the SVM. Such correlations could validate the model's physical relevance and suggest that it leverages chemically meaningful signals to distinguish between polymers.

In the following figures, for each plastic type:

- The **right vertical axis** (blue curve) represents the mean Raman intensity across all training samples in that class.
- The **left vertical axis** (red curve) shows the absolute value of the SVM weight vector $|w|$ associated with that class.

All spectra and weight curves are plotted against the wavenumber axis (in cm^{-1}), enabling direct comparison between physical features and model contributions.

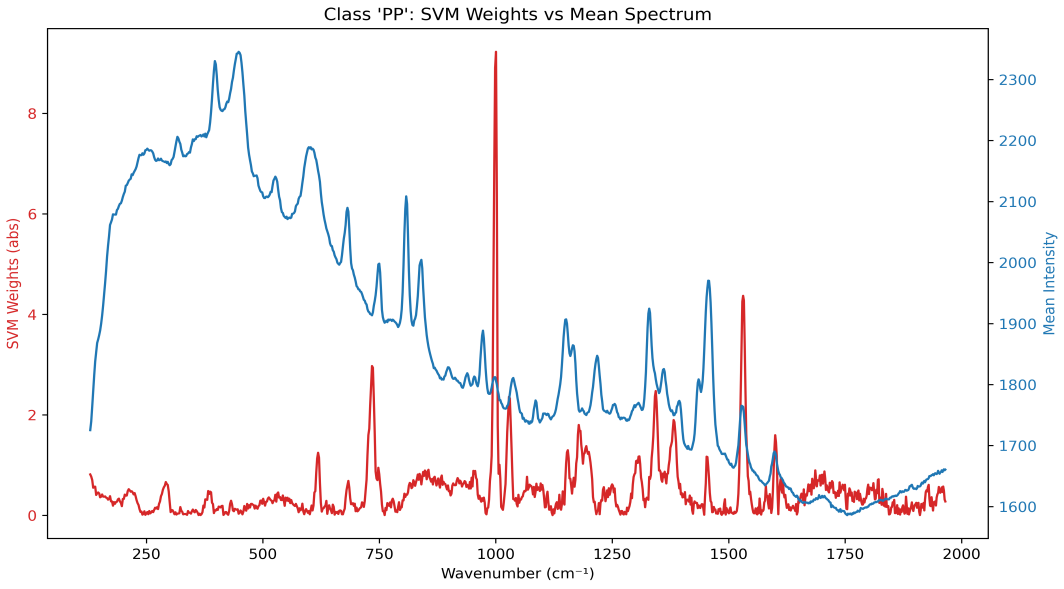

Figure 43: Comparison between mean spectra and SVM weights for PP

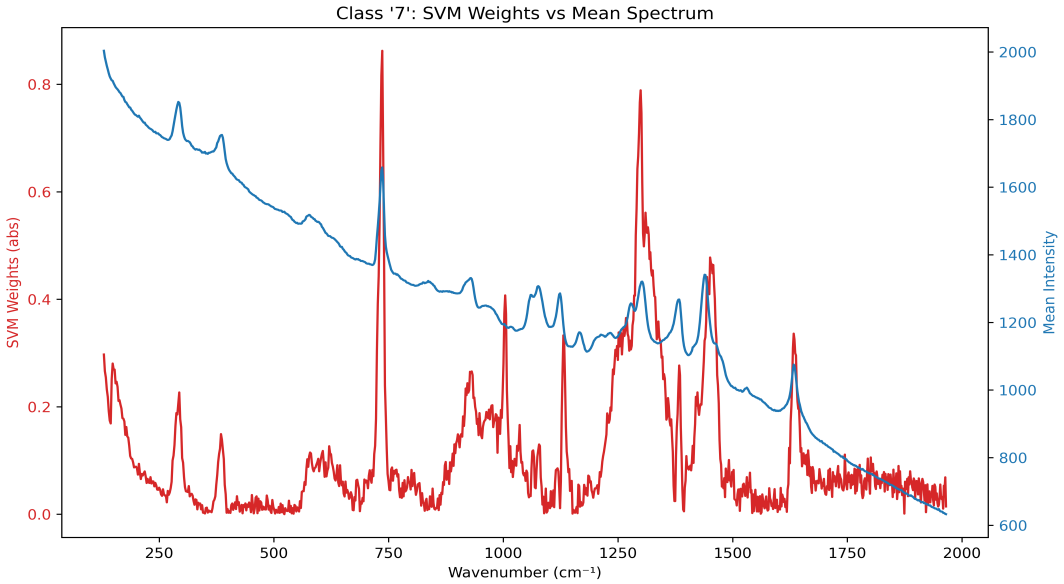

Figure 44: Comparison between mean spectra and SVM weights for 7

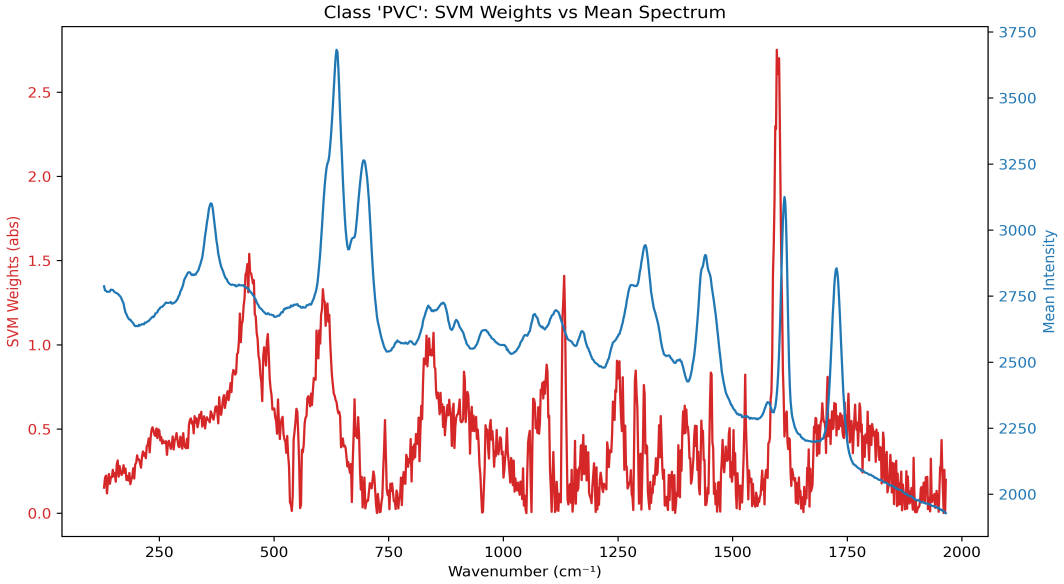

Figure 45: Comparison between mean spectra and SVM weights for PVC

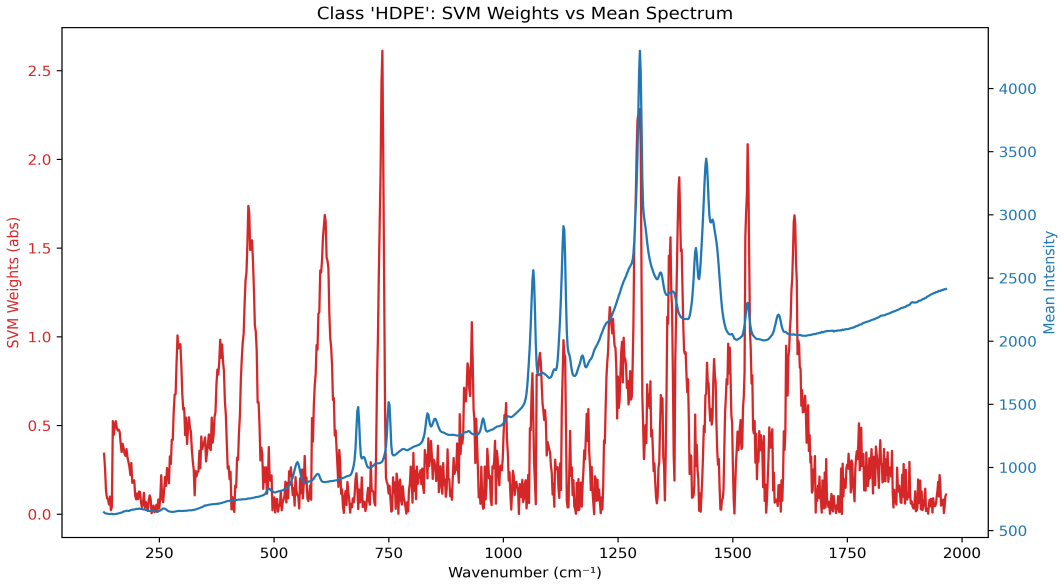

Figure 46: Comparison between mean spectra and SVM weights for HDPE

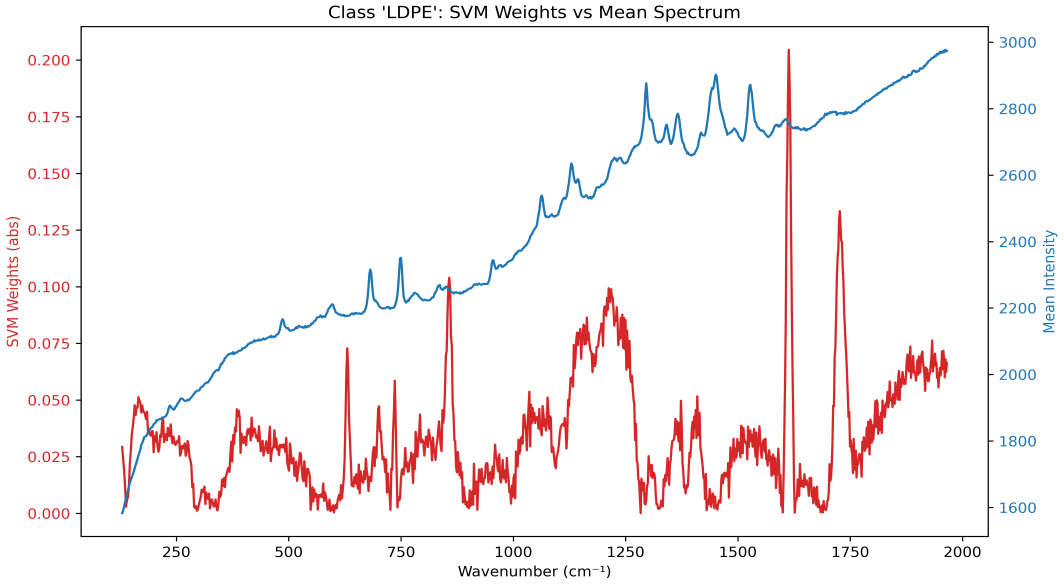

Figure 47: Comparison between mean spectra and SVM weights for LDPE

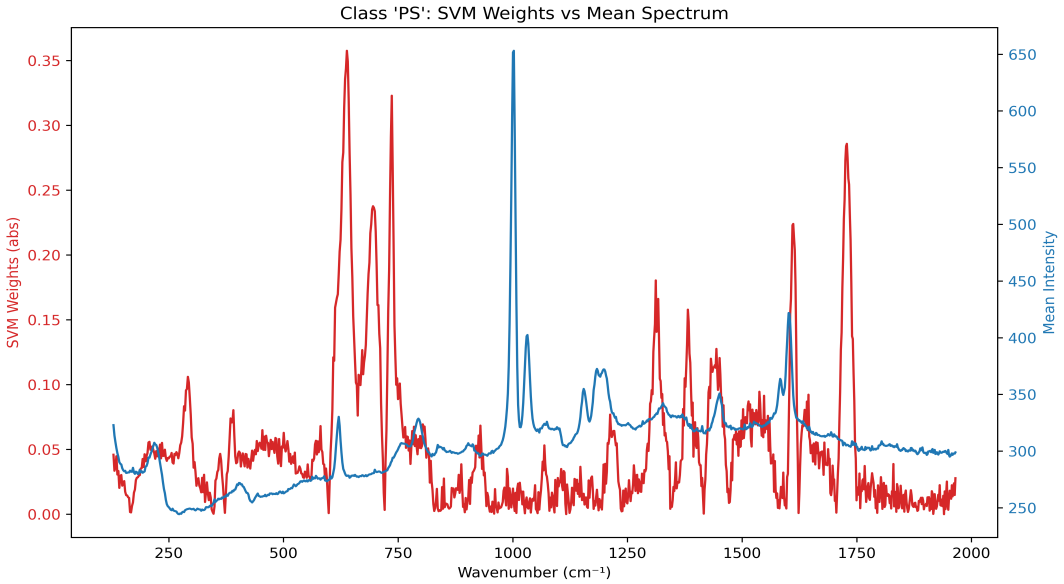

Figure 48: Comparison between mean spectra and SVM weights for PS

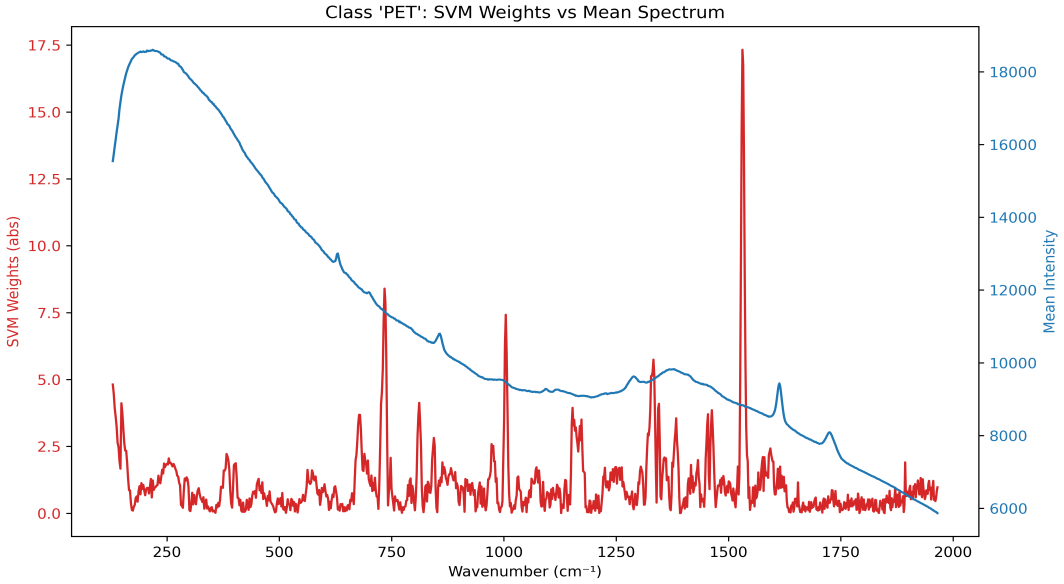

Figure 49: Comparison between mean spectra and SVM weights for PET

It is important to note, however, that Support Vector Machines are fundamentally *discriminative models*: their goal is not to reproduce the spectral structure of a class, but to find the most effective boundaries that separate it from the others. As a result, high SVM weights may correspond to features that are not necessarily the most intense or chemically significant within a spectrum, but rather those that best differentiate one class from its neighbors.

This distinction highlights the conceptual difference between a **characteristic peak**—a prominent vibrational feature typically associated with the identity of a polymer—and a **discriminative feature**, which is a spectral region that contributes strongly to the classification decision. In some cases, these may coincide, but in other cases, the most

discriminative regions may lie in less obvious parts of the spectrum, particularly where overlap with other classes is minimal.

Therefore, while the visual comparison between mean spectra and SVM weights is a powerful interpretability tool, it should not be taken as a direct one-to-one mapping of physical significance. Instead, it provides insight into how the model leverages the available information to distinguish between classes, which may reflect both chemical features and statistical structure in the data.

19 Real-World Case Study: Classification of Unknown Mixed Samples

In addition to the validation phase performed on a labeled and balanced dataset, the proposed classification pipeline was tested in a realistic context, simulating potential environmental and laboratory scenarios involving microplastic contamination. The primary objective of this section is to evaluate the model's generalization ability when applied to complex, unlabeled spectra collected from mixed or unknown samples, often contaminated with inorganic matter or embedded in foreign matrices.

Unlike the previous sections, where classification performance could be measured against ground truth labels, here the analysis is unsupervised and qualitative. Interpretation relies on consistency of predictions, confidence scores, and cluster separation in the embedded space (t-SNE). For all experiments, the model used is the same linear SVM trained on over 11,000 spectra from seven polymer classes. In this context, three realistic cases were designed:

1. Separation of ground soil from embedded plastic fragments.
2. Identification of polymer components in plastic-plastic mixtures.
3. Detection of plastics within a TiO_2-based matrix.

To support classification, the training set was extended by adding 900 spectra from pure ground samples, labeled as class "G". This allowed the model to explicitly recognize mineral material. Furthermore, a prediction confidence threshold of 0.6 was adopted to flag uncertain or low-quality assignments.

19.1 Case 1: Ground vs Plastics

Objective

The first test aims to simulate an environmental detection scenario, where plastic particles are embedded in soil or sediment. The model is required to distinguish between mineral spectra and plastic signals, classifying each measurement accordingly.

Sample Preparation

Plastic fragments (type unknown) were mechanically abraded obtaining powders with a broadened grain size ranging from tens of mcrons to hundreds of microns. Then the powders have been dispersed in the different matrices previously dissolved in isopropyl alcohol to form a paste. The weight ratio of plastic to soil was approximately 1 to 50. The so obtained paste was deposited on a solid substrate and left to dry before Raman acquisition. Spectra were then collected from various regions of the dried mixture.

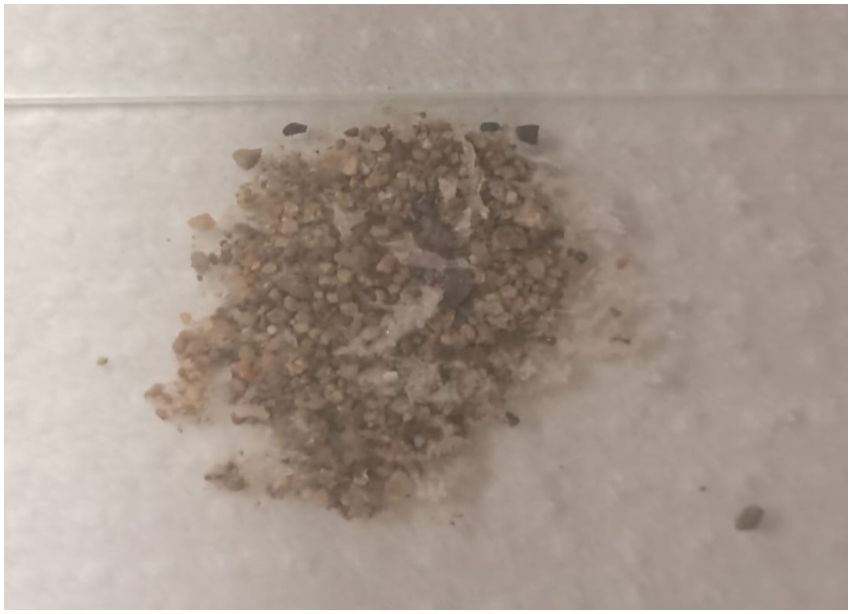

Figure 50: Pure ground sample

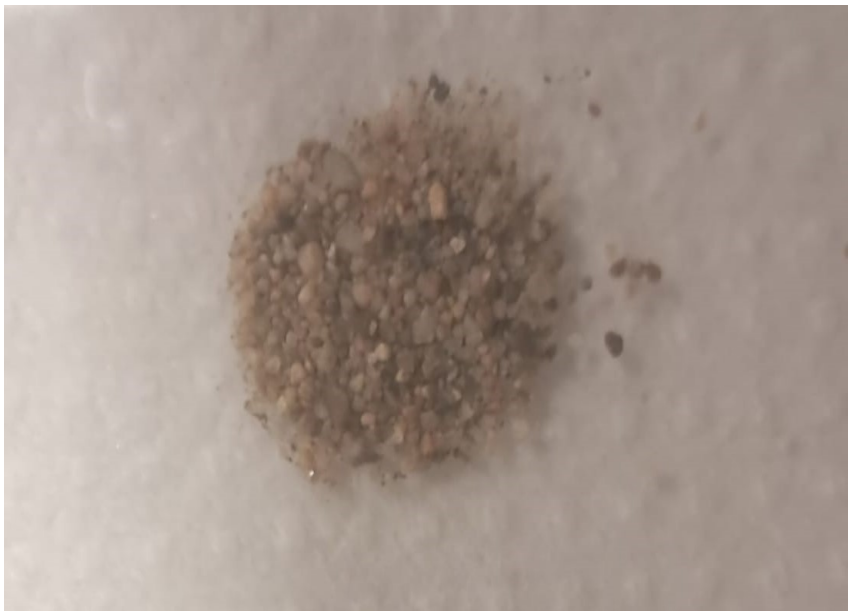

Figure 51: Ground and plastic mixtures sample

Challenges

Several obstacles complicate this classification scenario:

- The signal from ground material is highly variable and can partially overlap with plastic peaks.
- Surface roughness and drying artifacts can introduce noise or fluorescence.
- The presence of colorants or residues may distort the spectra, especially near the baseline.

Results

Figure 52 shows the distribution of predicted classes. Most samples were classified as "G", with a significant minority labeled as polypropylene (PP), and negligible predictions for other classes. This indicates that the model successfully identified the dominant mineral component while remaining sensitive to polymer inclusions.

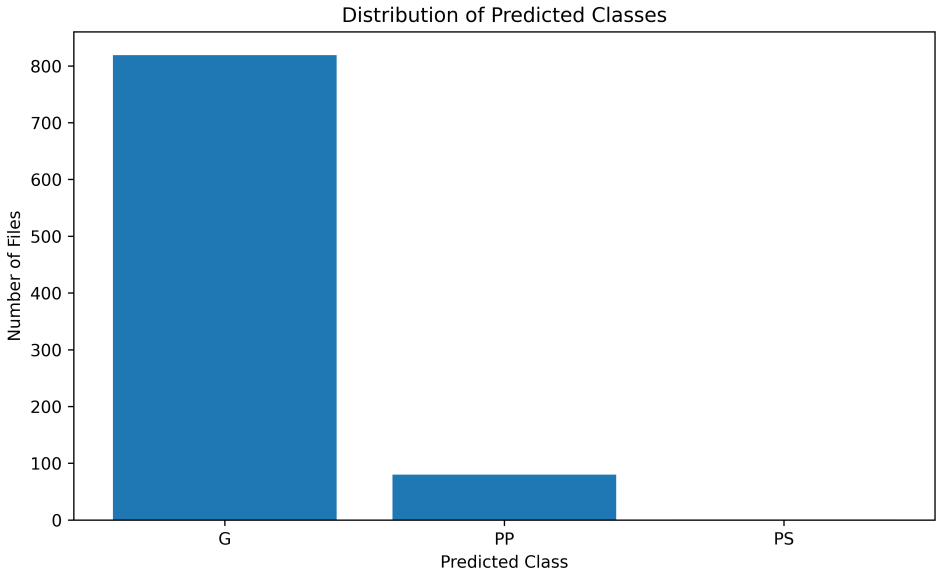

Figure 52: Predicted class distribution for the Ground vs Plastics test set.

The confidence distribution (Figure 53) shows a strong bias toward high-probability predictions, confirming the model's robustness in distinguishing polymeric signals from mineral ones.

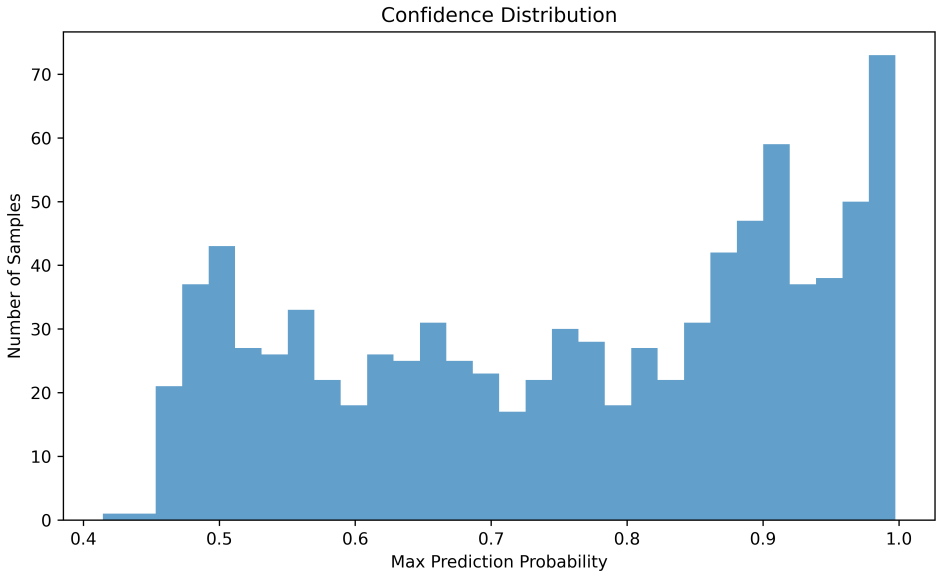

Figure 53: Confidence score distribution for Ground vs Plastics classification.

t-SNE visualization (Figure 54) reveals two main clusters, with minimal overlap between the predicted "G" and "PP" spectra. This validates the decision boundaries learned

during training and suggests that the model generalizes well even in noisy conditions.

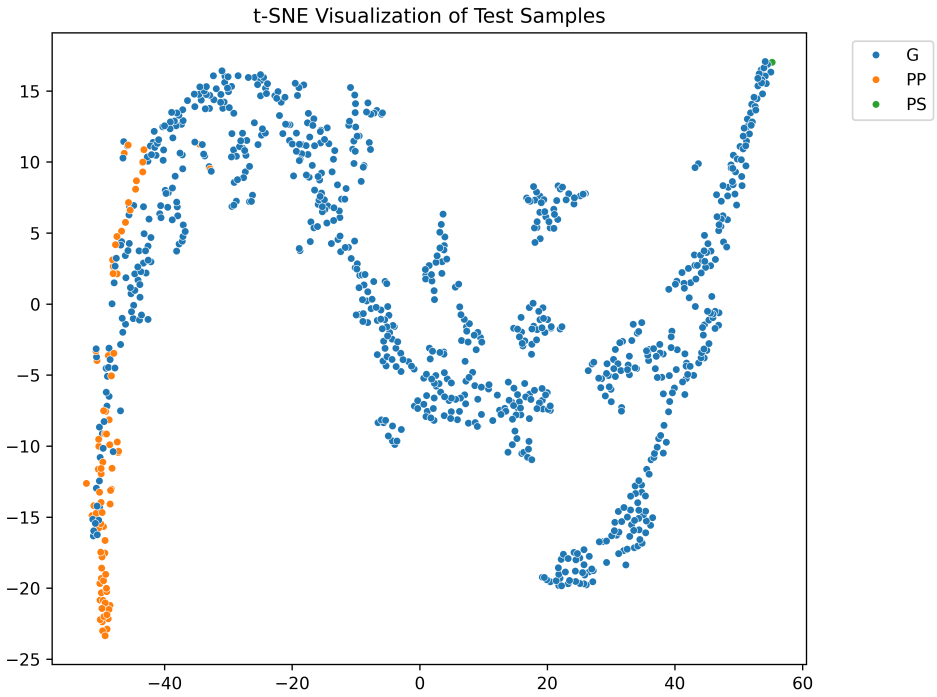

Figure 54: t-SNE embedding of test samples (Ground vs Plastics).

Interpretation

The model successfully distinguished plastic signals from ground material. PP was consistently identified as the most represented polymer in this scenario. This may reflect the actual presence of PP fragments or be due to spectral similarities with certain contaminated signals. Confidence metrics and cluster separation support the credibility of these predictions. This test confirms the practical feasibility of using Raman-based ML classification to screen real soil samples for microplastic inclusions.

19.2 Case 2: Plastic Mixtures (Two or More Classes)

Objective

The second scenario simulates a situation where two or more different plastic types are present in the same test sample. This case explores the model's ability to classify ambiguous or hybrid signals, as might occur in waste sorting facilities, multilayer packaging, or mixed environmental debris.

Sample Preparation

Abraded grains from different plastic sources were mixed to form a heterogeneous powder, then deposited on a glassy substrate to form a compact disk. The resulting spectra may represent clean signals from isolated fragments or overlapped signals from spatially mixed regions.

Challenges

The classification of polymer mixtures is inherently difficult due to:

- Spectral overlap between similar plastic types (e.g., LDPE and HDPE).
- Non-linear interactions in the signal if two polymers are physically close or layered.
- Absence of training data explicitly representing mixed spectra.

Results

The class distribution in Figure 55 indicates a strong dominance of polypropylene (PP) predictions, with smaller groups attributed to LDPE, PET, PS, and class 07 (Other). This suggests that PP-like features were prevalent across the dataset, or that the model defaulted to this class in cases of low separability.

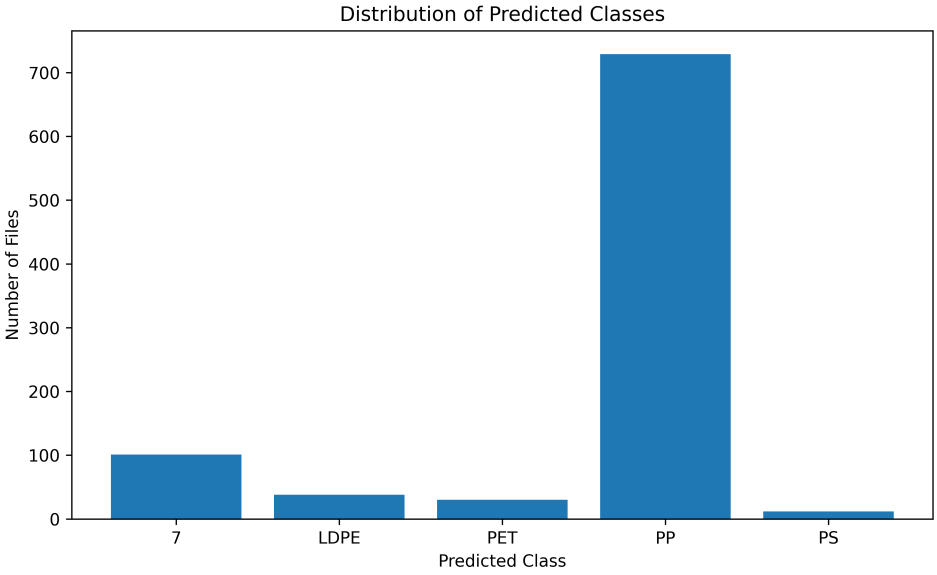

Figure 55: Predicted class distribution for the Plastic Mixtures test set.

The confidence distribution (Figure 56) shows a broader spread compared to Case 1, with a notable number of predictions falling below the confidence threshold. This supports the hypothesis that signal ambiguity and class overlap play a major role.

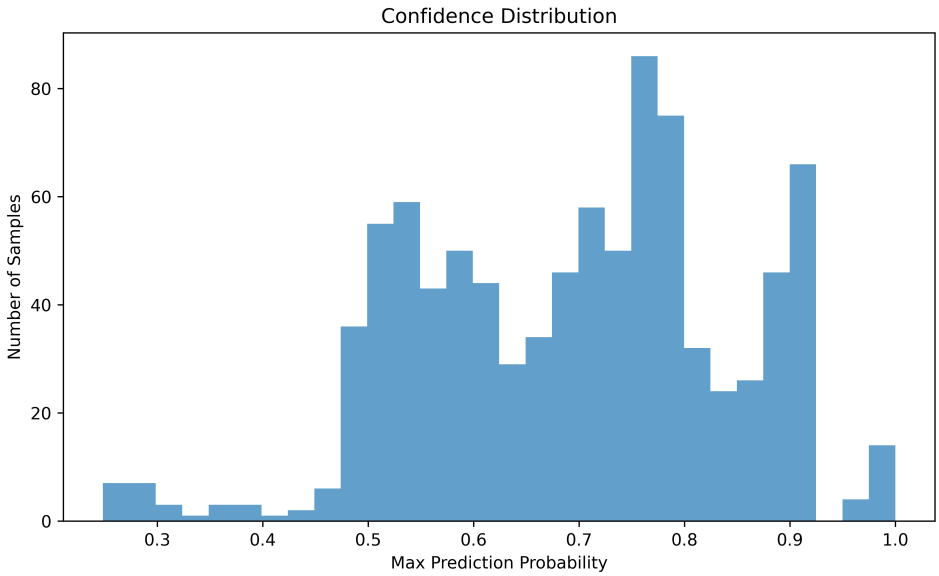

Figure 56: Confidence score distribution for Plastic Mixtures classification.

The t-SNE projection (Figure 57) reveals several partially overlapping clusters. PP and LDPE predictions form the most compact and isolated groups, whereas other classes are more dispersed or intermingled.

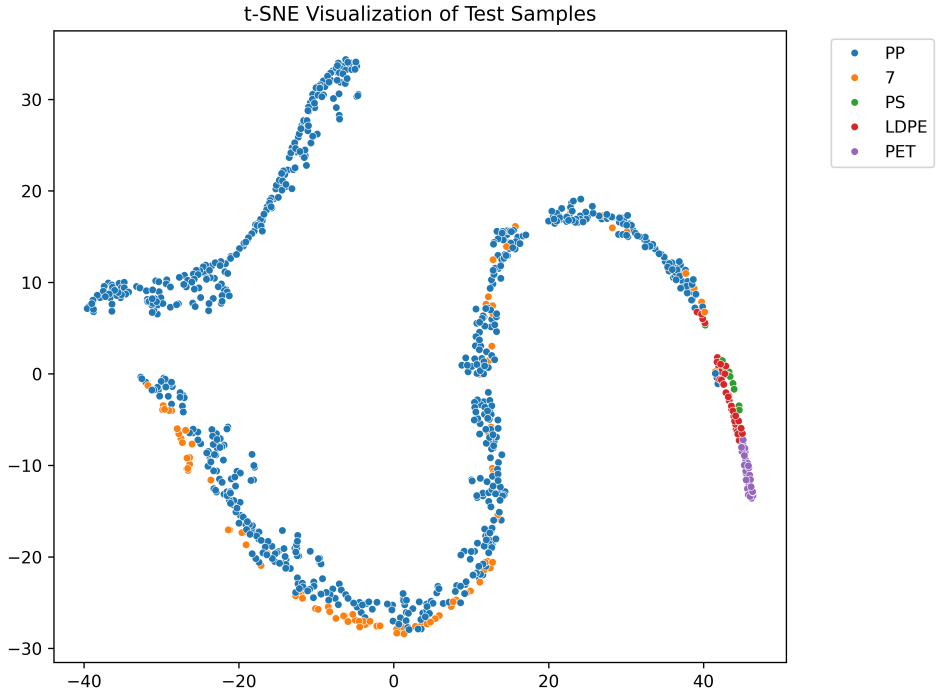

Figure 57: t-SNE embedding of test samples (Plastic Mixtures).

Interpretation

Despite the complexity of the input, the model managed to extract dominant class signals and separate them into coherent groups. The high number of PP predictions may reflect true prevalence or spectral bias, and the low-confidence predictions suggest borderline

spectra or mixed peaks. These results highlight the need for future model extensions to handle composite samples, possibly through ensemble approaches or mixture modeling.

19.3 Case 3: TiO_2 Matrix vs Plastics

Objective

The third test scenario explores the model's behavior when classifying spectra collected from plastic fragments embedded in a nanostructrured TiO2 (titanium dioxide) matrix. TiO2 was chosen fwith different motivation: from one side, differently from real ground, to represent an highly homogeneous matrix with an intense Raman signal mainly located at a lower energy region. From the other side this mixture mimics complex laboratory or industrial formulations, where the Raman signal of the target material may be distorted or partially masked by inorganic fillers or pigments.

Sample Preparation

Plastic powders were mixed into a TiO_2-based substrate and deposited as a uniform layer. After drying, Raman spectra were collected from multiple surface regions. No prior data on the composition of the plastic fragments was available.

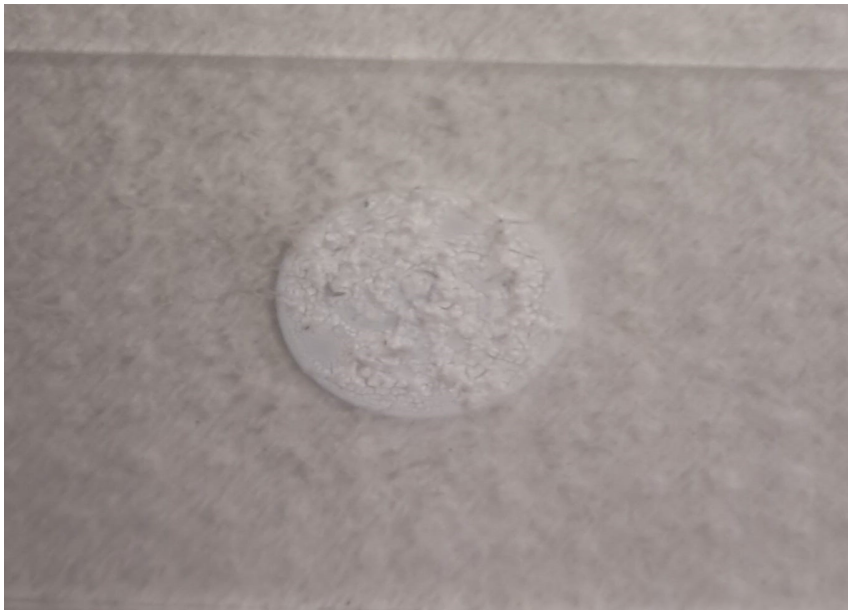

Figure 58: TiO_2 matrix sample

Challenges

This case introduces a number of critical issues:

- TiO_2 exhibits its own strong Raman features, which can overshadow the signal from microplastics.
- No TiO_2 spectra were included in the training set; the model must therefore rely on indirect learning or robustness to unknown interference.
- The possibility of non-linear signal mixing and high background noise makes classification uncertain.

Results

The histogram in Figure 59 shows that a substantial fraction of the samples was classified as class 07 (Other) or as PS and PP. This likely reflects the model's attempt to account for spectra that do not fit cleanly into standard plastic categories.

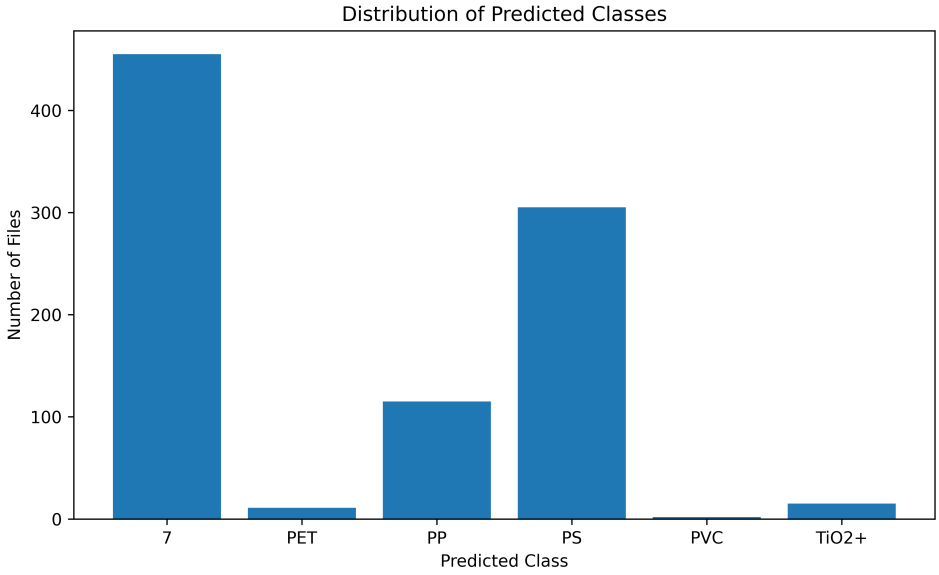

Figure 59: Predicted class distribution for the TiO_2 Matrix test set.

The confidence plot in Figure 60 shows a large number of predictions near or below the 0.6 threshold, reflecting the model's lower certainty in this scenario.

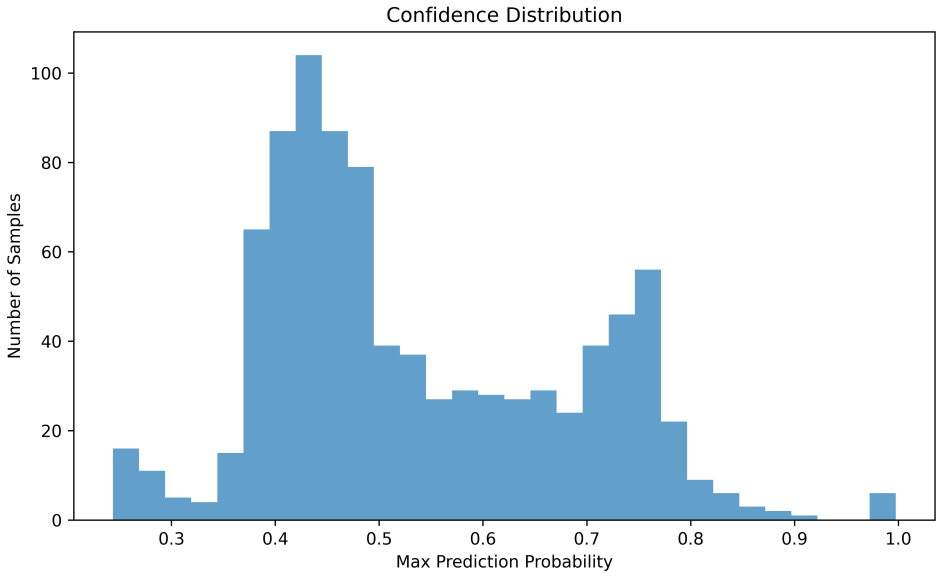

Figure 60: Confidence score distribution for TiO_2 Matrix classification.

Nonetheless, the t-SNE plot in Figure 61 reveals the emergence of coherent clusters—especially for samples classified as PS and PP—suggesting that, in spite of the matrix effects, some plastic signals remain distinguishable.

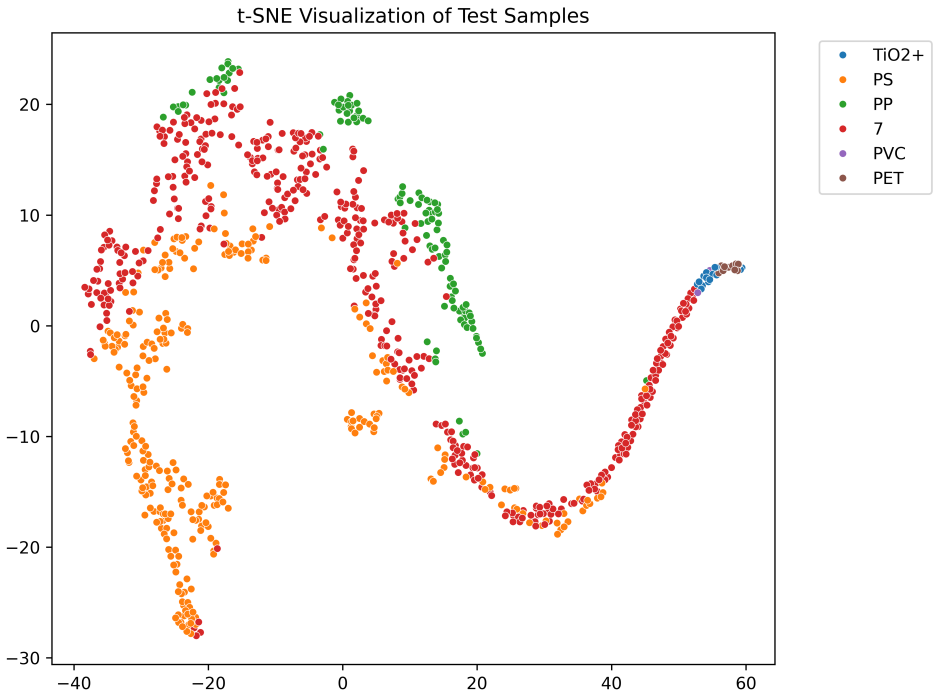

Figure 61: t-SNE embedding of test samples (TiO_2 Matrix).

Interpretation

This scenario highlights both the robustness and the limits of the current model. While classification performance degrades under matrix interference, the system remains capable of identifying structured patterns in the spectral data. The tendency to assign samples to class 07 may indicate a fallback behavior for unknown or complex signals, which could be improved by explicitly training the model on TiO_2-containing samples in future iterations.

20 Conclusion

This part presented the development, evaluation, and interpretability analysis of a machine learning-based classifier for identifying microplastics types from Raman spectra. The model was trained on a custom dataset of over 11,000 spectra, representing the seven standard recyclable plastic classes and including both transparent and colored samples.

A Support Vector Machine (SVM) with a linear kernel was selected as the baseline model, optimized through grid search. After applying z-normalization, the classifier achieved a test accuracy of 99.49%. Cross-validation confirmed the model's generalization ability, yielding a mean accuracy of 99.73%. The most challenging classes were LDPE, PS, and particularly class 07 (Other Plastics), which showed greater intra-class variability and overlapping spectral features.

Dimensionality reduction techniques such as PCA and t-SNE were employed to visualize class separability and support qualitative interpretation of spectral structures. Furthermore, an analysis of the SVM decision function revealed the spectral regions most influential in driving classification decisions, both globally and per class. This enhances the transparency of the model and opens the way for future integration with chemically meaningful peak attribution.

Despite the excellent performance, some limitations remain — including class imbalance in colored samples and heterogeneity in mixed plastic categories. These aspects point to future directions such as data augmentation, targeted collection of underrepresented sample types, and exploration of alternative or ensemble classifiers.

Overall, this project demonstrates the feasibility and potential of combining Raman spectroscopy with supervised learning for accurate, interpretable, and scalable microplastics classification. The proposed approach provides a foundation for future developments in automated, in situ analysis systems that can assist both environmental research and material diagnostics.

To further assess the applicability of the model, an exploratory analysis was conducted on unlabeled, real-world samples obtained from mixed and unknown plastic sources. These included mixtures of plastics and ground material, binary plastic blends, and samples embedded in TiO_2 matrices. By extending the training set with ground spectra and introducing a confidence-based prediction threshold, the model was able to infer plausible classifications and detect non-plastic regions (class G) in highly variable samples.

Results showed good agreement between confident predictions and expected plastic behavior, with polypropylene and ground material emerging as dominant classes in mixed samples. Dimensionality reduction via t-SNE confirmed meaningful clustering for certain categories, while the confidence score distributions provided insight into prediction uncertainty.

This final test reinforces the practical value of the proposed classification pipeline. It suggests that, despite inherent limitations in label quality and mixture complexity, the trained model retains discriminative power in field-like conditions. With further refinement, the approach could be integrated into semi-automated screening tools for environmental laboratories, recycling facilities, or quality control in plastic processing.

References

[1] Luca Prodi and Pietro Galinetto. Liasion - novel approaches to micro- and nanoplastics detection in water. prin 2022 [10/2023-10/2025] code: 2022waktfr. *PRIN Project Proposal*, 2022.

[2] B. Albini, P. Galinetto, S. Schiavi, and E. Giulotto. Food safety issues in the oltrepò pavese area: A sers sensing perspective. *Sensors*, 2023.

[3] Gyanendra Lamichhane, Ashok Acharya, Rajesh Marahatha, Sandeep Aryal, Nirmala Parajuli, Bijay Modi, Rajesh Paudel, Anil Adhikari, and Bhakta Kumar. Raut. Microplastics in the environment: Global concern, challenges, and controlling measures. *International Journal of Environmental Science and Technology*, 20:4673–4694, 2023.

[4] Yongjin Lee, Jaelim Cho, Jungwoo Sohn, and Changsoo Kim. Health effects of microplastic exposures: Current issues and perspectives in south korea. *Yonsei Medical Journal*, 2023.

[5] Ewa Winiarska, Marek Jutel, and Magdalena Zemelka-Wiacek. The potential impact of nano- and microplastics on human health: Understanding human health risks. *Environmental Research*, 2024.

[6] A. B. Silva, A. S. Bastos, C. I. L. Justino, J. P. Da Costa, A. C. Duarte, and T. A. P. Rocha-Santos. Microplastics in the environment: Challenges in analytical chemistry, a review. *Analytica Chimica Acta*, 1017:1–19, 2018.

[7] European Parliament and Council. Regulation (eu) no. 1025/2012 of the european parliament and of the council of 25 october 2012 on european standardisation. Official Journal of the European Union, L 316, 14.11.2012, p. 12.

[8] Young Kyoung Song, Sang Hee Hong, Soeun Eo, and Won Joon Shim. A comparison of spectroscopic analysis methods for microplastics: Manual, semi-automated, and automated fourier transform infrared and raman techniques. *Marine Pollution Bulletin*, 173:113101, 2021.

[9] Ana Isabel Pérez-Jiménez, Danya Lyu, Zhixuan Lu, Guokun Liu, and Bin Ren. Surface-enhanced raman spectroscopy: benefits, trade-offs and future developments. *Chemical Science*, 11:4563–4577, 2020.

[10] S. Hayashi, R. Koh, Y. Ichiyama, and K. Yamamoto. Evidence for surface-enhanced raman scattering on nonmetallic surfaces: Copper phthalocyanine molecules on gap small particles. *Physical Review Letters*, 1988.

[11] B Albini. Raman spectroscopy and nanostructured complex systems: A satisfactory win to win game? *PhD thesis*, 2020.

[12] PhysicsOpenLab. Polymer analysis using raman spectroscopy, 2022. Accessed: 2024-03-03.

[13] Claudio Cusano. Notes based on the lectures of the course "machine learning", 2024. University of Pavia.

[14] Megha Sunil, Nazreen Pallikkavaliyaveetil, N Mithun, Anu Gopinath, Santhosh Chidangil, Satheesh Kumar, and Jijo Lukose. Machine learning assisted raman spectroscopy: A viable approach for the detection of microplastics. *Journal of Water Process Engineering*, 60:105150, 2024.

[15] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. Available at: `https://www.csie.ntu.edu.tw/~cjlin/papers/svm_tutorial.pdf`.

[16] W. M. Tolles, J. W. Nibler, J. R. McDonald, and A. B. Harvey. A review of the theory and application of coherent anti-stokes raman spectroscopy (cars). *Applied Spectroscopy*, 1977.

[17] SGS DigiComply.