A hybrid approach of vision transformers and CNNs for classification of Nanoplastics from Raman spectrum signals

Sumedh Deepak Kudale[1], Shubham Kumar Sinha[1], Md Arshad Nesar[1], Sridevi S[1*†], P Anandan[1†], Gorkem Kar[2†]

[1*]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600048, India.
[2]Computer Science & Cybersecurity, University of Central Missouri, , Warrensburg, 64093, USA.

*Corresponding author(s). E-mail(s): sridevi.s@vit.ac.in;
Contributing authors: sumedhkudale2@gmail.com; shubhamkrsinha11111@gmail.com; mdarshadnesar@gmail.com; anandan.p@vit.ac.in; kar@ucmo.edu;
†These authors contributed equally to this work.

Abstract

Accurate identification of nanoplastics in environmental samples is required to establish their distribution and effects. Raman spectroscopy is a strong, label-free tool for the detection of nanoplastics, but the produced spectral images are intricate and need sophisticated computational techniques to classify them with precision. In this article, we introduce a hybrid deep learning architecture that incorporates Vision Transformer (ViT) and Data-efficient Image Transformer (DeiT) models for effective feature extraction from high-resolution Raman spectrogram images. The 1D spectrogram is transformed to 2D scaleogram by employing QFT (Quantum Fourier Transform) and CWT (Continuous Wavelet Transform). ViT and DeiT transformer model features are concatenated and normalized to generate richer representations of every sample. In order to acquire relational information between spectra, we build a graph based on cosine similarity between the feature vectors and each node in the graph is a scaleogram and edges are similarity-based relationships. This graph is then passed into a Spectral-based Graph Convolutional Network (GCN) classifier. The introduced pipeline, as illustrated in the accompanying flowchart, shows an average accuracy

of 96.46% compared to standalone ViT's 93.55% and DeiT's 95.63%, providing an efficient solution for automated nanoplastics classification from Raman spectrum signals. CWT transformed images showed higher accuracy than QFT Transformed Images. The model was evaluated on 4 metrics and furthermore, Confusion Matrix, Feature Spaces & Hyperparameters were also analyzed to draw conclusions.

Keywords: ViT, DeiT, Swin, GCN, Nanoplastics, CWT, QFT, Raman spectroscopy

1 Introduction

The rapid accumulation of nanoplastics in the environment has caused a lot of alarm because of their potential impact on the environment and human health. Raman spectroscopy has been noted as a powerful, non-destructive technique for identification and characterization of nanoplastics with high specificity by molecular vibrations. However, the resulting spectral information are often complex and high-dimensional and hence challenging for traditional analysis and classification.

Recent advancements in deep learning have also impacted spectral analysis significantly, particularly through the use of convolutional neural networks (CNNs) and more recently, transformer models. Vision Transformers (ViT), Shifted Window Transformers (Swin), and Data-efficient Image Transformers (DeiT) have been highly effective in retrieving global contextual cues from images and outperforming classic CNNs in most vision tasks. Yet, the problem of effectively retrieving both the higher-level patterns and the subtle interdependencies between samples remains, especially for high-level complex spectral data such as those collected by Raman spectroscopy.

To surpass the above limitations, the current study introduces a hybrid deep learning model that combines the advantages of transformer models with those of graph-based learning. By comparing the Vision Transformer (ViT), Swin Transformer, and Data-efficient Image Transformer (DeiT) models for feature extraction applications, as well as the use of a Spectral-based Graph Convolutional Network (GCN) to extract the interrelationships in spectra, our approach attempts to develop a hybrid model that learns the optimal accuracies of these models for the classification of nanoplastics from Raman spectrogram images. Apart from the incorporation of the global feature extraction intrinsic in transformers, this combined model also involves relational knowledge through the construction of graphs, thus improving the representation of data.

The novel framework is evaluated based on comparative analysis with other models with the aim of demonstrating its superiority over single models and traditional methods. The results highlight the importance of hybrid transformer-GCN models in boosting automated detection and classification of nanoplastics in complex spectral environments.

2 Related Work

Development of machine learning applications from Raman spectroscopy data has been shown in various fields. Berghian-Grosan et al.[1] investigated food safety through five predictive modeling methods on Raman spectroscopy of fruit distillates, while Khan et al.[2] showed 85% accuracy of dengue detection with Raman-SVM integration. Biomedical applications have also advanced, with Ryzhikova et al.[3] diagnosing Alzheimer's disease through Raman analysis of varying algorithms and Du et al.[4] identifying foodborne pathogens with adversarial network-augmented Raman-SVM systems. Tian et al.[5] also showed the applicability of Raman-ML platforms by identifying rice origins with 94.68% accuracy using SPA-LS-SVM model combinations. These studies collectively confirm the applicability of the Raman spectroscopy technique for ML-based classification tasks on natural and synthetic materials.

Microplastic identification studies have propelled methodological advances in spectroscopic analysis. Luo et al.[6] employed CNN-SVM hybrids for aquatic microplastic identification on FTIR-Raman fusion, whereas Back et al.[7] compared SVM, random forests on ocean microplastic vibrational spectra. Supervised classification methods by Shan et al.[8] facilitated soil microplastic monitoring with variable accuracy for variable sizes, whereas Zhu et al.'s[9] tested the accuracy of Deep Learning using holographic deep learning system for marine microplastics. Other techniques are Ishmukhametov et al.'s[10] dark-field microscopy-ResNet pipeline for intracellular microplastics and Stefas et al.'s[11] LIBS-PCA/LDA framework for polymer identification. Michel et al.[12] attained marine plastic identification on multitype spectroscopic data i.e., ATR–FTIR, NIR, LIBS and XRF.

Recent studies have demonstrated the applicability of Raman spectroscopy in identifying and detecting microplastics in water-based systems. Chakraborty et al.[13] reviewed its application to determine the types of polymers in various aquatic sources, while Neo et al.[14] focused on the use of Raman spectra coupled with chemometric approaches to enable automatic sorting of plastics. Lin et al.[15] employed surface-enhanced Raman spectroscopy for the identification of plastic particles leached from packaging material into drinking water effectively. Schymanski et al.[16] employed micro-Raman spectroscopy to quantify and identify microplastics in bottled water, which could detect particles of 1 µm effectively. Overall, these studies confirm the sensitivity and applicability of Raman spectroscopy in the detection of microplastics.

Our methodology attempts to make use of Raman Spectra for Microplastic classification and builds majorly upon four machine learning advancements: Vision Transformers (ViTs) introduced by Dosovitskiy et al.[17] and Data-efficient distillation (DeiT) introduced by Touvron et al.'s[18] for feature extraction. Graph-based learning integrates Kipf et al.'s[19] Graph Convolutional Networks (GCNs) for modeling extracted feature relationships, while optimization leverages Kingma et al.'s[20] Adam algorithm. This paper aims to address nanoplastic-identification challenges through ViT-DeiT feature fusion, GCN-based relational refinement, and adaptive learning rate strategies.

Our approach tries to leverage Raman Spectra for Microplastic categorization and relies on four machine learning developments: Vision Transformers (ViTs) of Dosovitskiy et al.[17] and Data-efficient distillation (DeiT) of Touvron et al.'s[18]

for feature extraction. Graph-based learning incorporates Kipf et al.'s[19] Graph Convolutional Networks (GCNs) for modeling relations of features extracted, and optimization leverages Kingma et al.'s[20] Adam algorithm. This work tries to solve nanoplastic-identification problems using ViT-DeiT feature fusion, GCN-based relational enhancement, and adaptive learning rate methods.

3 Datasets

The experimental dataset employed in this study was specifically processed and created by [6]. The dataset consists of multiple samples of Raman spectral intensity captured with a Raman spectroscopy instrument, where each spectrum corresponds to varying molecular vibrational modes. Table I illustrates the sample size of CWT and QFT spectra of 5 plastic classes and 1 Blank class, describing that the data was evenly distributed. The data collection contains a total of 1200 samples of CWT and QFT each of the Raman scaleogram, 200 samples for each class of plastics and 200 blank samples.

For the internal Raman spectral database generation, an experiment was performed on 1,000 different nanoparticles. Five common types of plastic waste most frequently found in environmental sample analyses are some of these nanoparticles: Polyethylene (PE), Polytetrafluoroethylene (PTFE), Polystyrene (PS), Polymethyl methacrylate (PMMA), and Polyvinyl chloride (PVC). We also divided the data with the 80-20% ratio to generate two sets for training and testing.

The data set was stored in an organized hierarchical directory system, with separate directories for every type of plastic and a special folder by the name "BLANK." Each directory contains single Raman spectra stored as text files. In every text file, two columns exist: the X-coordinate (wavenumber) in the first column and the Y-coordinate (Raman signal intensity) in the second column. Fig. 1 provides examples of the Raman spectra scaleograms of the entire data set, showing their sensitivity and the scaleogram complexity.

Table I: Raman Scaleogram dataset Description

Class	**Training set**		**Testing set**		**Total**	
	CWT	**QFT**	**CWT**	**QFT**	**CWT**	**QFT**
BLANK Samples	160	160	40	40	200	200
PE Samples	160	160	40	40	200	200
PMMA Samples	160	160	40	40	200	200
PS Samples	160	160	40	40	200	200
PTFE Samples	160	160	40	40	200	200
PVC Samples	160	160	40	40	200	200
Total Samples	960	960	240	240	1200	1200

4 Dataset Pre-processing

Raman spectral signals, which are inherently one-dimensional and non-stationary, were first transformed into two-dimensional time-frequency representations using Wavelet Transforms. Previous works [21] had proved the necessity of converting 1D scaleogram to 2D scaleogram for Machine Learning applications.

The CWT is calculated by applying the Morlet wavelet across relevant scales, and the CWT generates a scaleogram that shows the time-frequency relationships in the signal. The Morlet wavelet was selected for converting Raman spectra in the proposed model due to its proven efficacy in capturing the oscillatory and localized spectral features innate to Raman signals. The Morlet wavelet offers a balance between time and frequency resolutions, since is it a complex exponential modulated by a Gaussian envelope, making it well-suited for spectroscopic data. Eq. 1 shows the function for CWT while Eq. 2 shows the function for Morlet Wavelet.

$$W_j(a,b) = \int_{-\infty}^{+\infty} cA_j(t) \cdot \psi^* \left(\frac{t-b}{a} \right) dt \tag{1}$$

where, $W_j(a,b)$ is the wavelet coefficient at scale a and translation b, $cA_j(t)$ is the input signal, ψ is the Morlet wavelet function, defined by Eq. 2

$$\psi(t) = e^{i\omega_0 t} e^{-\frac{t^2}{2}} \tag{2}$$

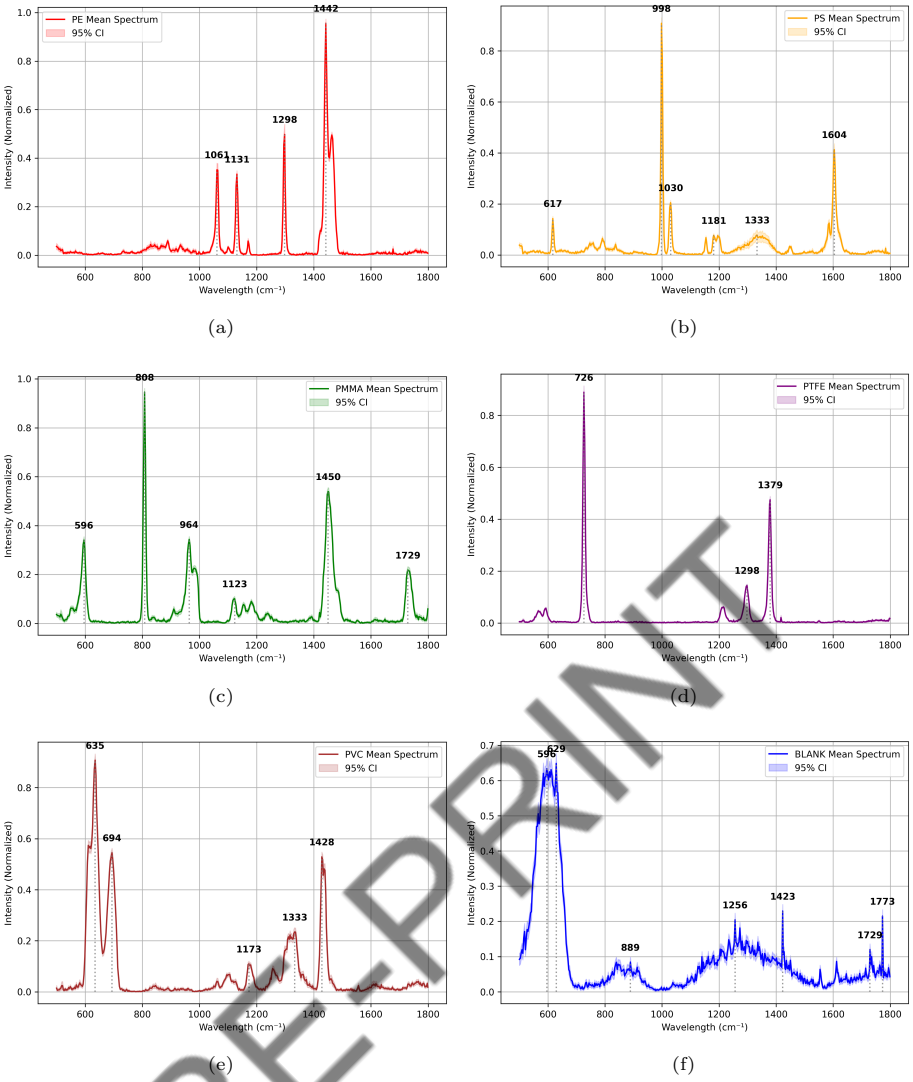
where ω_0 denotes the central frequency of the Morlet wavelet, typically chosen to balance time and frequency localization, x controls the scale (frequency), with larger x corresponding to lower frequencies, y controls the translation (position in time).

The Discrete Wavelet Transform (DWT) is calculated by decomposing the normalized Raman spectrogram using Biorthogonal wavelet function denoted by Eq. 3. Fig. 5 also shows a circuit diagram of QFT for Spectrogram conversion. This will split the signal into approximation and detailed coefficient components at several scales, thus isolating low-frequency approximations from high-frequency details to facilitate further analysis. Eq. 3 shows the function used.

$$cX_j[n] = \sum_k r[k] \cdot \varphi_{j,n[k]} \tag{3}$$

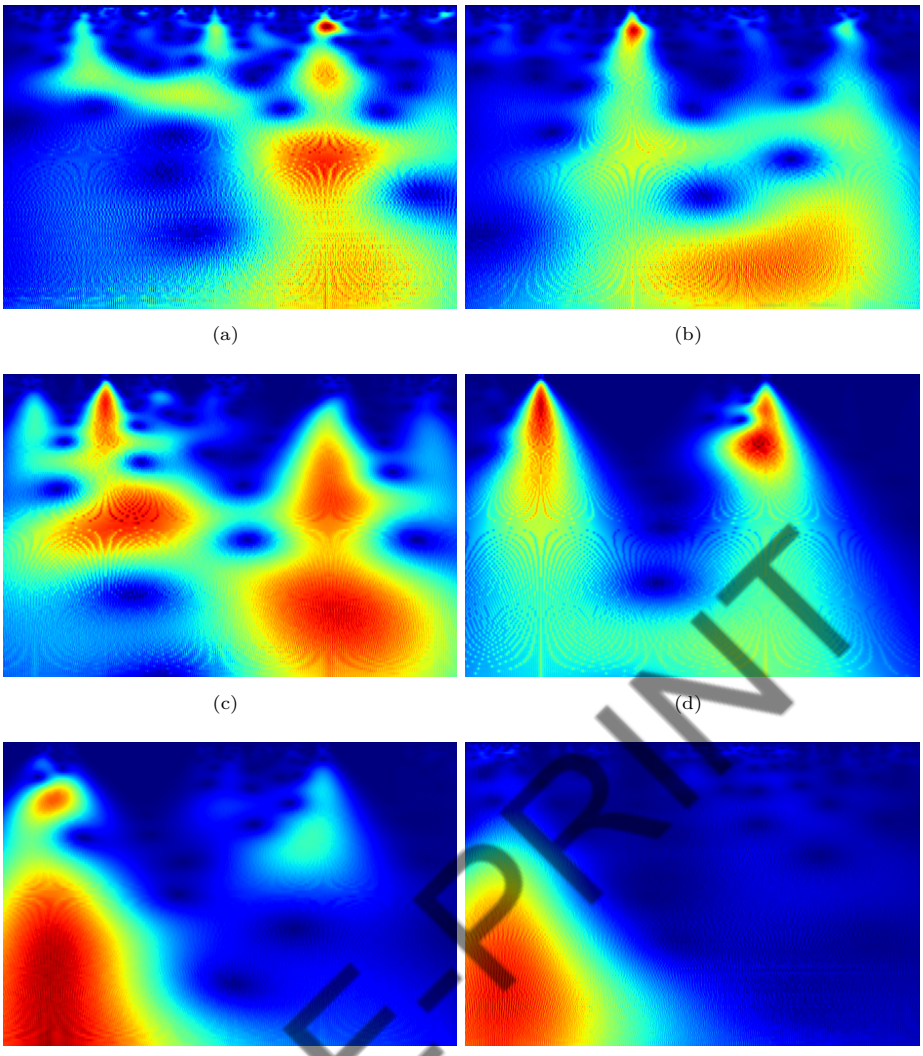
where, $cX_j[n]$ represents combination of $cA_j[n]$ the biorthogonal scaling approximation coefficients at low pass filter path at level j, and $cD_j[n]$ detailed coefficients at high pass filter path at level j, $\varphi_{j,n[k]}$ represents the biorthogonal wavelet function.

Each scaleogram is made at an resolution of 56 $\times$ 56 pixels, capturing necessary spatial and spectral information extracted from the Raman spectra. This method ensures that crucial features within the spectroscopic data are properly represented for machine learning applications. Fig. 2 depict the final 2D images obtained after transforms.



(a) (b) (c) (d) (e) (f)

Fig. 1: Raman spectrum 1D signal samples belonging to class labels of (a)PE Nanoplastic (b)PS Nanoplastic (c)PMMA Nanoplastic (d)PTFE Nanoplastic (e)PVC Nanoplastic (f)BLANK sample

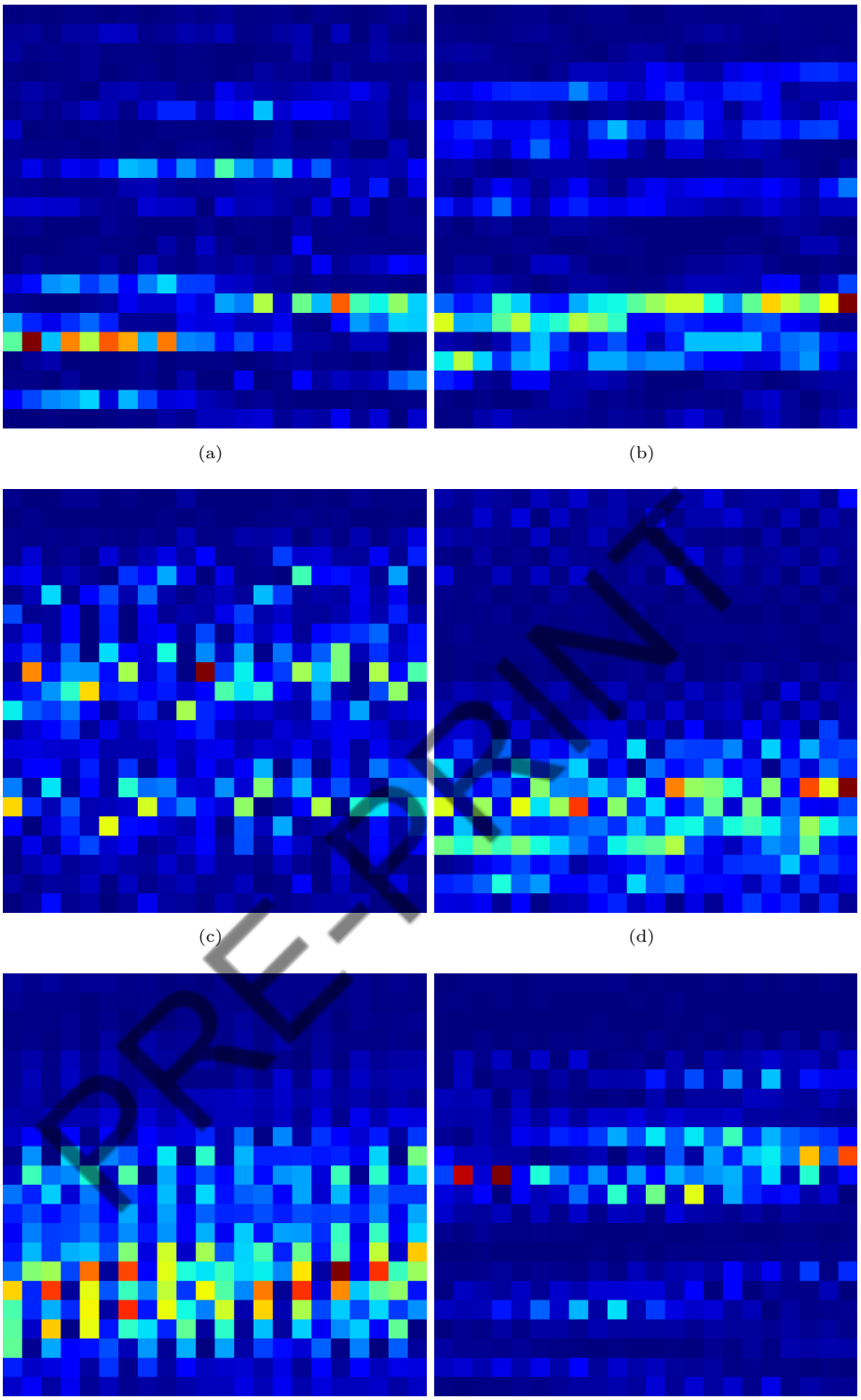


(a) (b)

(c) (d)

(e) (f)

Fig. 2: Wavelet transformed 2D Spectrogram image samples belonging to class labels of (a)PE Nanoplastic (b)PS Nanoplastic (c)PMMA Nanoplastic (d)PTFE Nanoplastic (e)PVC Nanoplastic



(a) (b) (c) (d) (e) (f)

Fig. 3: (a)BLANK Raman QFT spectrogram (b)PE Raman QFT spectrogram (c)PMMA Raman QFT spectrogram (d)PS Raman QFT spectrogram (e)PTFE Raman QFT spectrogram (f) PVC Raman QFT spectrogram

5 Methodology

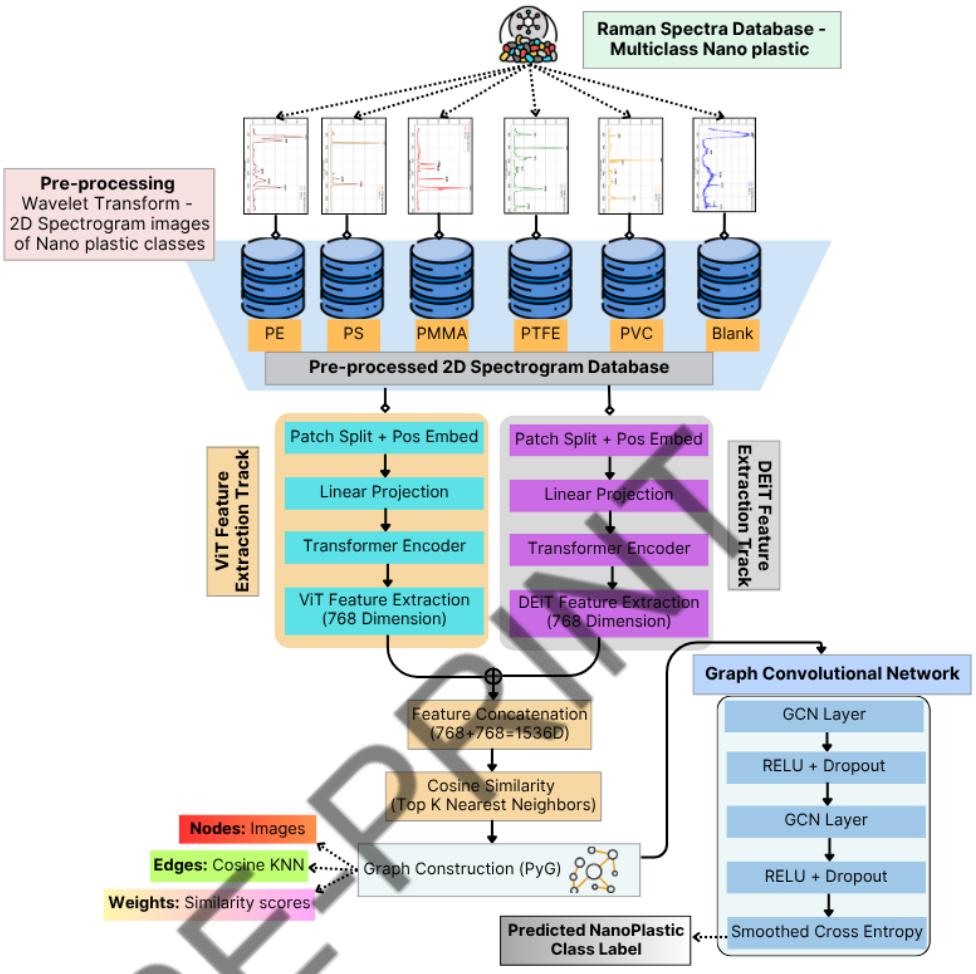


Fig. 4: Proposed Hybrid Model

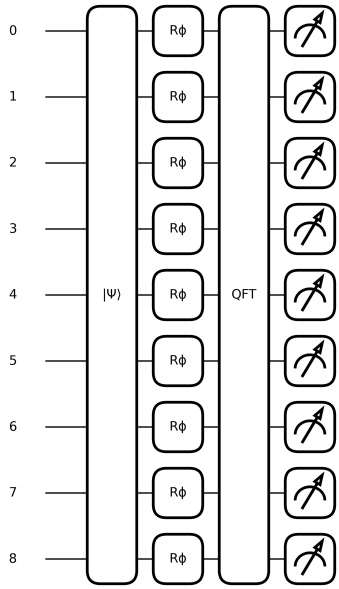


Fig. 5: Quantum Fourier Transform Based circuit for Spectrogram conversion

5.1 Quantum Fourier Transform Computation

Alternate time-frequency representations of the Raman signals were obtained using a specially designed Quantum Fourier Transform (QFT) circuit in addition to wavelet-based spectrogram transformations. Eight qubits were used to implement the QFT circuit, which is depicted in Fig. 5. This was sufficient to encode amplitude information from padded Raman spectra of dimension $2^8 = 256$. This procedure was essential for transforming 1D spectral data into 2D image representations appropriate for quantum-assisted frequency domain analysis. The output achieved can be seen in Fig. 3.

The following crucial steps were taken in order to perform the computation using PennyLane, a quantum machine learning library:

1. `AmplitudeEmbedding`, a variational template that converts classical data into quantum amplitudes, was used to embed the input signal into quantum states after normalization.
2. The quantum state was subjected to a standard Quantum Fourier Transform (QFT) template, which used interference patterns to identify frequency components in the data.
3. The 2D QFT spectrogram image was created by reshaping the spectrum obtained from the final measurement of the state vector's probability distribution into a 16×16 matrix.

The code for the entire QFT processing pipeline is as follows:

```
dev = qml.device("default.qubit", wires=8)

@qml.qnode(dev)
def qft_circuit(input_signal):
    qml.templates.AmplitudeEmbedding(input_signal, wires=range(8), normalize=True)
    qml.templates.QFT(wires=range(8))
    return qml.probs(wires=range(8))
```

For use in subsequent classification tasks, the QFT-based spectrograms were then resized and displayed as false-color heatmaps. The QFT-transformed spectrograms showed higher inter-class feature overlap in dimensionality reduction visualizations (Fig. 7), indicating relatively lower robustness compared to CWT images, even though they achieved competitive accuracy. However, the quantum-generated spectral patterns enrich the diversity of spectrogram representations and offer a complementary viewpoint in frequency space.

5.2 Overview

In this current research, we propose a hybrid machine learning model that combines feature extraction with transformers and classification with graph convolutional networks (GCN) for enabling the detection of nanoplastics from Raman spectrogram images. A depiction of the proposed model is given in Fig. 4. The process starts with the collection of high-resolution images of spectrograms of an input nanoplastic that can be PE, PS, PMMA, PTFE, PVC, or a control sample. Wavelet transformations are used for image pre-processing, and the images are resized to 224×224 pixel size. The images are then converted to tensor form in order to make them compatible for use in deep learning models without losing spectral data for subsequent processing.

Feature extraction is performed using a two-pathway transformer model in which two pre-trained models process each image independently: the Vision Transformer (ViT) and the Data-efficient Image Transformer (DeiT). Both the models split the input image into patches of fixed size, add positional embeddings, and project the patches through a linear transformation before passing them through a transformer encoder. The built-in self-attention mechanism of these transformers is instrumental in capturing long-range and local dependencies, and this is particularly important when classifying spectrograms. This is achieved by projecting the embedded patch sequence X into three distinct representational spaces using learnable weight matrices:

$$Q = XW^Q \tag{4}$$

$$K = XW^K \tag{5}$$

$$V = XW^V \tag{6}$$

where W^Q,W^K and W^V are the weight matrices for queries, keys, and values, respectively. The transformer's self-attention layer then produces an output Z given by

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{7}$$

This formulation[22] captures the relationships between all the elements of X, thereby capturing the global context of every single element. Every path in the transformer produces a feature vector of dimensionality 768, and these are combined to produce a 1536-dimensional high-level representation per image.

To ensure consistency in the feature space, normalization of all the concatenated feature vectors is done via a standard scaler. Cosine similarity between all pairs of the normalized feature vectors is computed, enabling the computation of the top-K nearest neighbors for every image based on the similarity scores computed. This forms a relational graph where every image is projected onto its most similar counterparts. A graph is formed where each node is every single spectrogram image, edges are between every node and its K nearest neighbors, and edge weights are cosine similarity. The resulting graph is encoded using the PyTorch Geometric (PyG) library where node attributes are the normalized hybrid feature vectors.

Classification is carried out by a two-layer Spectral-based GCN with ReLU (Rectified Linear Unit) activation functions and dropout regularization strategies. The GCN aggregates neighborhood information of each node, tapping into both the intrinsic attributes of each spectrogram and the relational graph structure of the data. The layer-wise propagation rule for the GCN is such that[23]

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{8}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix with added self-loops, $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$, $H^{(l)}$ is the matrix of activations at layer l, $W^{(l)}$ is the trainable weight matrix for that layer, and σ denotes the non-linear activation function, such as ReLU. For a two-layer GCN, the final output is produced as

$$Z = \text{softmax}\left(\hat{A}\,\text{ReLU}\left(\hat{A}XW^{(0)}\right)W^{(1)}\right) \tag{9}$$

where X is the matrix of node feature vectors, and $W^{(0)}$,$W^{(1)}$ are the input-to-hidden and hidden-to-output weight matrices, respectively. The softmax activation yields class probabilities for each node, enabling multiclass classification.

Training of the model employs a smoothed cross-entropy loss function to promote generalization and prevent overconfidence. Network optimization is performed using the Adam optimizer in conjunction with a StepLR learning schedule that reduces the learning rate in a systematic manner in a manner that promotes convergence. Model evaluation is performed by comparing predicted class outputs to ground truth labels in the test set using classification accuracy as the primary metric. This combined strategy is shown to take advantage of the global feature extraction capabilities inherent in transformer models combined with relational learning capabilities of graph

neural networks, thereby promoting effective nanoplastic classification from Raman spectrogram data.

The proposed Hybrid Model pseudocode is given as follows.

```
Initialize ViT, DeiT, and GCN model parameters
Initialize dataset paths for CWT/QFT spectrogram images
Initialize k=10 for k-nearest neighbors, hidden_dim=256
Load pre-trained ViT base model (patch16_224) and DeiT base model
Remove classification heads, set to feature extraction mode
For each image I in training dataset
    Resize image I to (224, 224) and convert to tensor
    Extract ViT features: f_vit = ViT(I) in R^768
    Extract DeiT features: f_deit = DeiT(I) in R^768
    Concatenate hybrid features: f_hybrid = [f_vit; f_deit] in R^1536
Store all hybrid features F and corresponding labels Y
Normalize features using StandardScaler: F_norm = scale(F)
Compute cosine similarity matrix: S = cosine_similarity(F_norm)
For each sample i = 1, 2, ..., N
    Find k nearest neighbors using S[i,:]
    Create edges (i,j) for top-k neighbors with weights S[i,j]
Construct graph G = (X, E, W) where X=F_norm, E=edges, W=weights
Initialize GCN with input_dim=1536, hidden_dim=256, output_dim=num_classes
For each epoch t = 1, 2, ..., T
    Forward pass: H1 = ReLU(GCN1(X, E, W))
    Apply dropout: H1 = Dropout(H1, p=0.5)
    Output layer: Y_hat = GCN2(H1, E, W)
    Compute loss using Label Smoothing: L = LabelSmoothingLoss(Y_hat, Y)
    Backpropagate and update parameters using Adam optimizer
    Update learning rate using StepLR scheduler
Evaluate on test set using cosine k-NN graph construction
Compute final metrics: Accuracy, Precision, Recall, AUC-ROC
End
```


5.3 Feature Extraction

After transforming the datasets into 2D spectrograms, the datasets were input into three separate transformer-based models—Vision Transformer (ViT), Shifted window Transformer (Swin), and Data-efficient Image Transformer (DeiT)—that were separately trained on each of the two types of spectrograms. The models make use of the self-attention mechanism to learn contextual patterns and long-range dependencies among the spectrogram patches. To ensure consistency in model training and feature extraction, all transformer models were configured to default hyperparameters.

An overview of architectural configurations and feature extraction layers borrowed in ViT, Swin Transformer, and DeiT models is summarized in Table II.

Parameter	ViT	Swin Transformer	DeiT
Input Resolution	224 × 224	224 × 224	224 × 224
Patch Size	16 × 16	4 × 4	16 × 16
Number of Layers	12	12	12
Number of Heads	12	3 per stage	12
Hidden Dimension	768	768	768
MLP Dimension	3072	3072	3072
Dropout Rate	0.1	0.1	0.1
Pretraining Dataset	ImageNet-1k	ImageNet-1k	ImageNet-1k
Feature Layer Used	Penultimate	Penultimate	Penultimate
Output Feature Dimension	768	768	768

Table II: Comparison of Vision Transformer (ViT), Swin Transformer, and DeiT architectures.

The ViT and DeiT models divide the spectrogram image into patches of fixed size, flatten them, and linearly project them to embeddings before going through a series of transformer encoder layers.

Each encoder block in a multi-head self-attention mechanism is succeeded by a feedforward network. During training, the models acquire complex spectral and spatial relationships that are embedded in the Raman spectrograms. After training is complete, features were extracted from the final transformer encoder layer as these embeddings retain full high-level representations free from the distorting effect of the last classification head. The ViT and DeiT features were combined to achieve a common representation. The combined feature vector was fed into a Graph Convolutional Network (GCN). In the approach, each sample is represented as a node in the graph, and edges are created on the basis of the similarity of their transformer-derived feature vectors. The GCN updates each node's attributes by diffusing information from the adjacent nodes, allowing the model to learn relational structures and inter-sample relationships.

6 Results and Discussion

Using the `lightning.qubit` simulator for Quantum Fourier Transform (QFT) computation, we used the PennyLane quantum machine learning library with 500 shots for quantum spectral transformation. The QFT-derived spectrograms were then resized to 16 × 16 matrices. For improved time-frequency resolution, the Morlet wavelet was used to create representations based on the Continuous Wavelet Transform (CWT). Using the `timm` (PyTorch Image Models) library, pretrained Vision Transformer (ViT) and Data-efficient Image Transformer (DeiT) models were used for feature extraction. The `PyTorch Geometric (PyG)` framework was used to implement graph construction and GCN-based classification. The `Matplotlib`, `Seaborn`, and `Scikit-learn`

libraries were used to visualise the data. Using a `StepLR` scheduler to control learning rate decay, all models were trained using the `Adam` optimiser.

To evaluate the effectiveness and generalization capability of the proposed hybrid architecture, the model was run on both types of data, and the mean of accuracies and standard deviation over multiple epochs were calculated. The average of those results provides a less biased estimate of the model's performance, while the standard deviation reflects the stability of the method. Graphs were obtained using python's matplotlib and seaborn libraries.

The mean and standard deviation for each model were computed using the following expressions:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{11}$$

For measuring the performance of the model in classification, we employed 4 performance metrics: Accuracy, Precision, Recall and AUC-ROC (Area Under the Curve-Receiver Operating Characteristic). Following formula were used:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx \tag{15}$$

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{16}$$

where TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative, TPR is True Positive Rate and FPR is False Positive Rate.

6.1 Feature Space Visualization Analysis

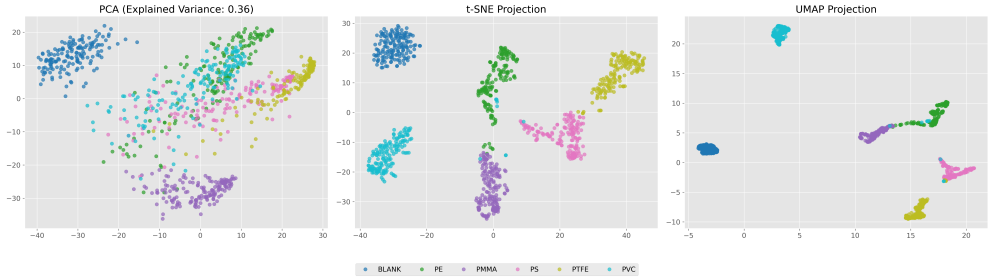


Fig. 6: Dimensionality reduction visualizations of hybrid ViT-DeiT features extracted via CWT.

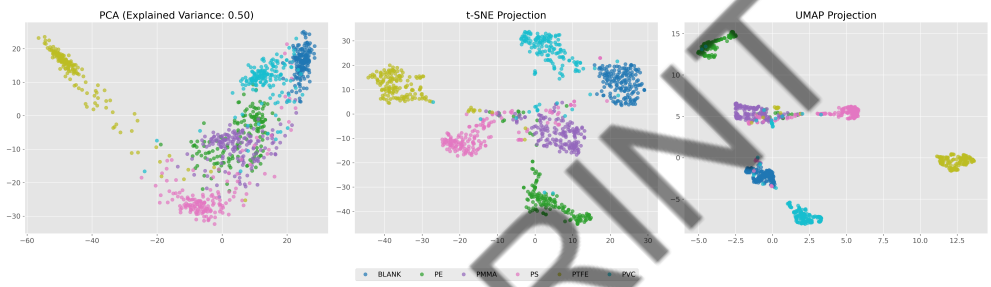


Fig. 7: Dimensionality reduction visualizations of hybrid ViT-DeiT features extracted via QFT.

The quality of the feature space derived from the hybrid ViT-DeiT embeddings was assessed using dimensionality reduction visualizations with PCA, t-SNE, and UMAP projections as shown in Figure 6 (CWT) and Figure 7 (QFT). Features derived from CWT show improved separability of the six classes of microplastics (BLANK, PE, PMMA, PS, PTFE, and PVC) with well-defined and distinct clusters observed with all three projection methods. Interestingly, the t-SNE and UMAP projections of the CWT features show little overlap between classes, which suggests that the continuous wavelet transform successfully extracts discriminative spectroscopic features of different microplastic species in the hybrid feature space.

On the other hand, the QFT-derived features, despite their superior PCA explained variance score, show greater cluster dispersion and inter-class overlap, though seen in the PCA projection where some of the classes seem to be spatially near. The t-SNE projection of QFT features shows less dense clustering compared to CWT, and this implies that Fourier-based transformations might introduce spectral artifacts that reduce the specificity of microplastic signatures. This visual evidence supports

the conjecture that CWT's superior time-frequency localization ability preserves the fine spectroscopic features important for accurate microplastic classification better.

The dimensionality reduction analysis supports the effectiveness of the hybrid ViT-DeiT model in learning informative features from spectrogram representations. The observable class discrimination observed in CWT representations is due to improved classification performance, and it signifies that the model is able to learn discriminative feature representations effectively that improve the inherent spectroscopic properties of different microplastic materials within the reduced-dimensional space.

6.2 Graph Construction Parameter Analysis

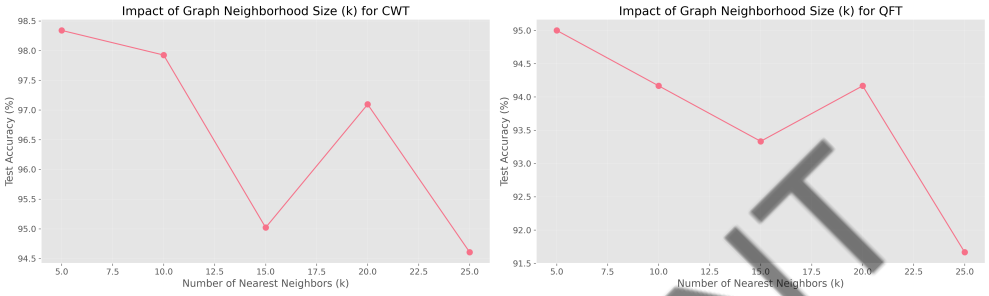


Fig. 8: Hyperparameter sensitivity analysis showing the impact of graph neighborhood size (k) on CWT and QFT-based hybrid model performance.

The graph neighbor parameter k sensitivity analysis shows evident optimization patterns between CWT and QFT-based methods, as shown in Figure 8 . For CWT-based features, the best accuracy is obtained when k=5, which indicates that sparsity in graph connectivity well captures the most informative feature relationships. Performance loss at larger k values (especially the dip at k=15) implies that too many neighborhood connections add noise and weaken the discriminative strength of the cosine similarity-based graph structure.

The QFT model displays oscillating sensitivity patterns with the best performance at k=5 and k=15, reflecting a less robust optimization behavior than CWT. The entire range of overall accuracy observed and the more unstable performance oscillations at varying k values further illustrate the intrinsic instability of QFT features in graph-based learning settings.

This analysis reaffirms that k=5, used in our final models, is an acceptable trade-off between connectivity and specificity, and also illustrates the higher robustness of features pooled from CWT compared to QFT graph construction parameters.

6.3 Model Evaluation

To analyze the convergence behavior of the individual and hybrid models, we have plotted their training loss curves. Figure 9 presents the loss over 40 training epochs

for each of the three individual transformer models—ViT, Swin, and DeiT—trained individually on the CWT and QFT spectrograms.

Figure 10 shows the loss curves of the proposed approach for 50 epochs in both the CWT and QFT domains. The proposed approach converges faster and has less final loss than single models, again showing the advantage of feature fusion and graph refinement.

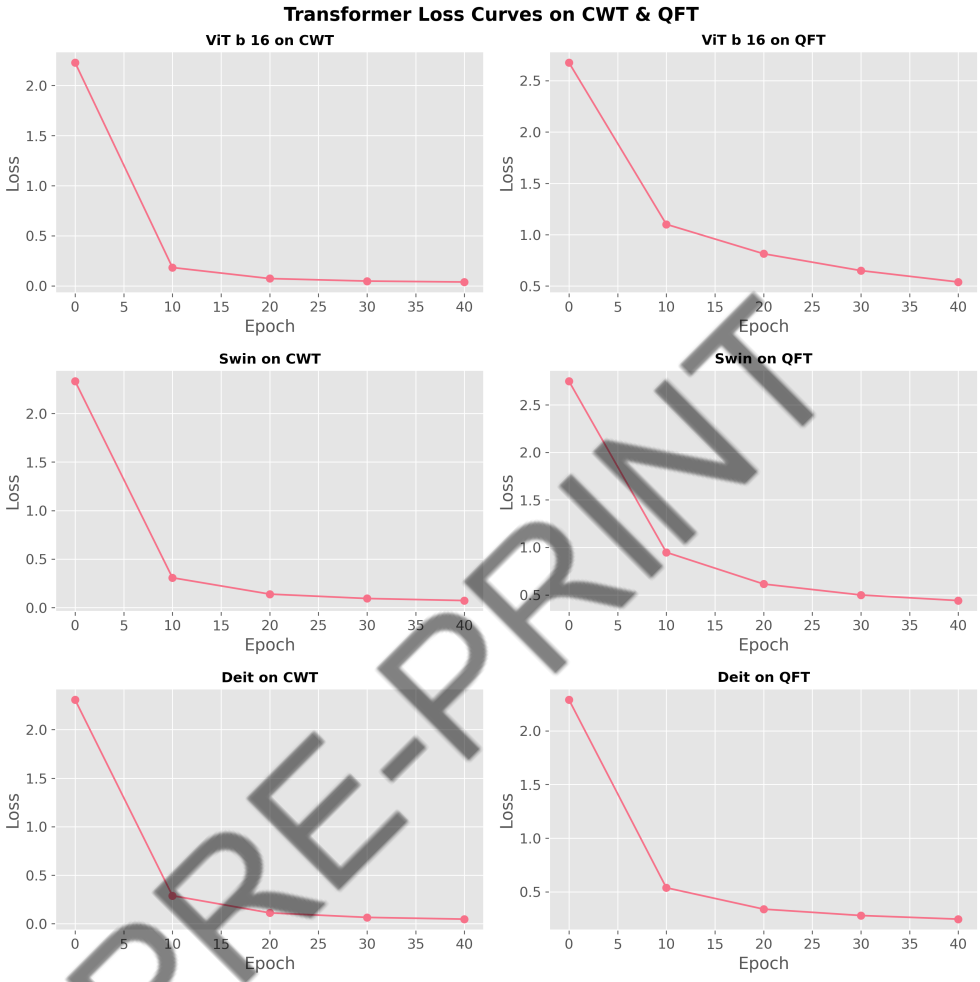


Fig. 9: Loss curves of standalone transformer models (ViT, Swin, DeiT) trained on CWT and QFT spectrograms.

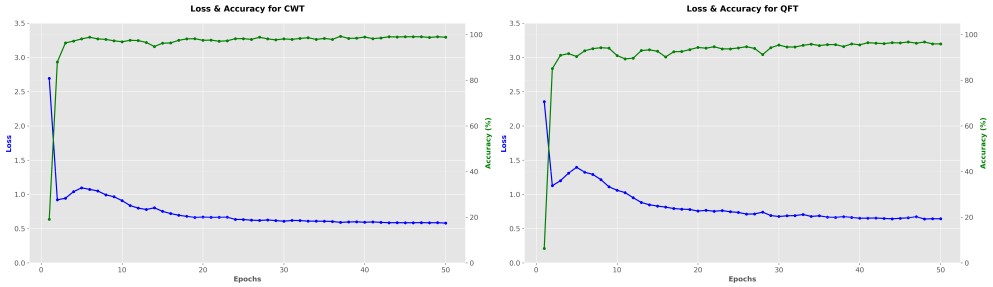


Fig. 10: Accuracy vs Loss curves of the Hybrid model trained on CWT and QFT spectrograms.

The relative performance of isolated transformer models compared to their hybrid models is depicted in Table III. For both CWR and QFT spectrograms, the DeiT model recorded the maximum solo accuracy, with the ViT model being in second place.

To enhance these results, a hybrid method was employed. The two best-performing transformer models (ViT and DeiT) for each type of spectrogram (CWT and QFT) were selected, and their penultimate layer's outputs were concatenated to form integrated feature vectors. After preprocessing, these integrated vectors were input to a Graph Convolutional Network. The hybrid model had great improvements in classification accuracy.

The final result across 4 metrics can be seen in Table IV.

Model Type	**Accuracy (%)**			**Standard Deviation (%)**
	CWT	**QFT**	**Average**	
ViT (standalone)	95.02	92.08	93.55	±0.34
DeiT (standalone)	96.27	95.00	95.63	±0.29
Swin (standalone)	92.95	91.25	92.10	±0.32
Proposed Model	98.34	94.58	96.46	±0.27

Table III: Comparative classification accuracy of standalone and hybrid models on CWT and QFT spectrogram datasets.

The Hybrid CWT model achieved a peak accuracy of 98.34% on CWT and a remarkable 94.58% on QFT, yielding an overall average accuracy best among all the models compared at 96.46%. The hybrid models' Precision, Recall and AOC-RUC provide further evidence of strong class-level performance with minimum misclassification. Those improvements justify the effectiveness of the transformer-GCN hybridization approach for Raman signal classification.

6.4 Confusion Matrix

To more effectively examine the performance of the transformer models on the polymorphism classification of the spectrograms produced by CWT and QFT, confusion

Metric	Dataset		
	CWT	**QFT**	**Average**
Accuracy (%)	98.34	94.58	96.46
Precision (Macro)	0.9848	0.9480	0.9664
Recall (Macro)	0.9833	0.9458	0.96455
AUC-ROC (OvR)	0.9962	0.9978	0.997

Table IV: Multimetric analysis of hybrid model on CWT and QFT spectrogram datasets

matrices are presented in Figure 11 and Figure 12. These matrices provide class-wise comparison of the prediction results to enable examination of the ability of the model to accurately differentiate between various polymers.

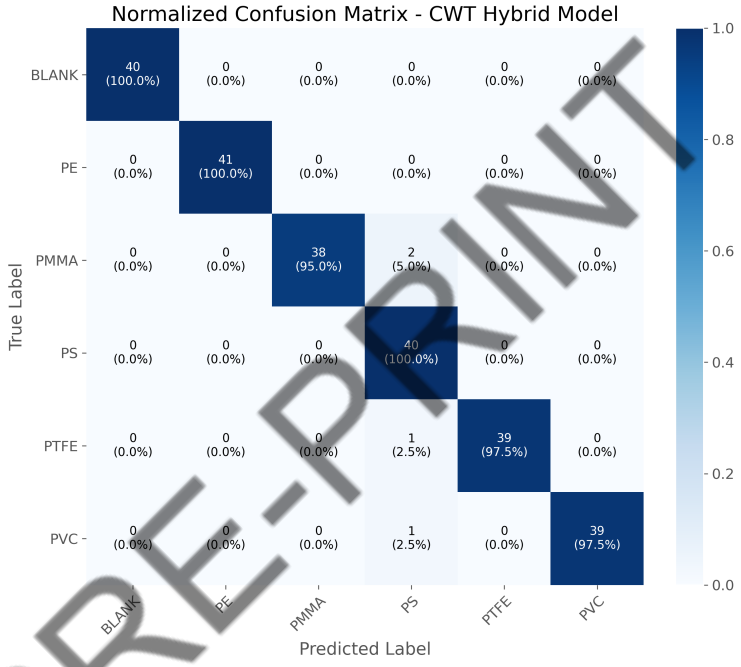


Fig. 11: Confusion matrix for the transformer model trained on CWT spectrograms. High classification accuracy is observed across all classes, with minimal off-diagonal misclassifications.

Figure 11 shows the confusion matrix of the CWT-transformed spectrogram-trained classifier. It illustrates excellent classification performance for all classes with ideal accuracy for PE and PMMA, and minimal minor misclassifications for

PS, PTFE, and PVC. This verifies the success of CWT in preserving local and discriminative features in Raman spectra.

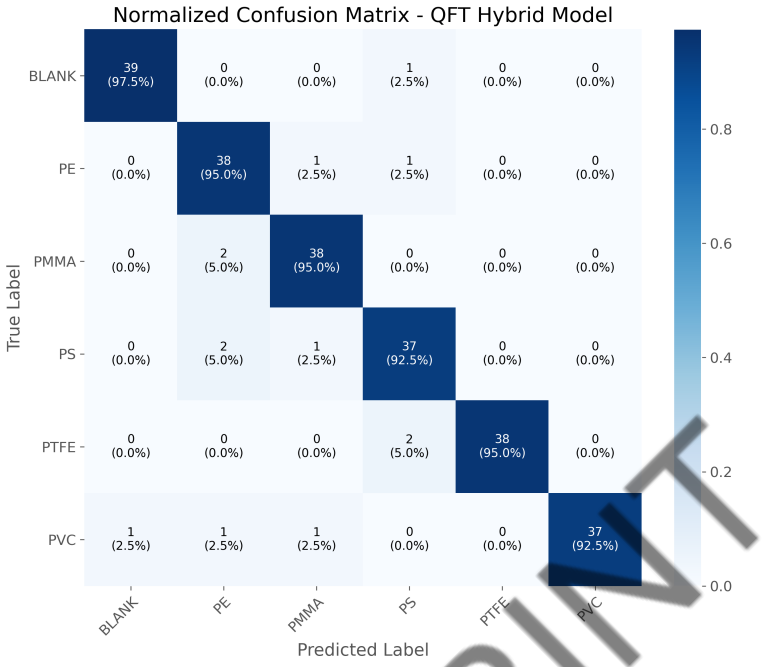


Fig. 12: Confusion matrix for the transformer model trained on QFT spectrograms. Slightly higher misclassification rates are noted, particularly among spectrally similar polymer classes.

Figure 12 displays the confusion matrix for the QFT-transformed spectrogram-trained model. Overall performance continues to be high, but misclassification happens a bit more frequently than in the case of the CWT-based model. Confusion among highly similar classes like PE–PMMA and PTFE–PVC indicates that QFT, although successful at maintaining global phase information, potentially introduces minor class feature overlaps.

These confusion matrices graphically verify the said above quantitative results and show that CWT-based spectrograms provide marginally improved class separability for the classification of Raman signals.

7 Conclusion

This study proposed a novel hybrid deep learning architecture for the classification of Raman spectral signals using spectrogram representations. The methodology involves transforming the spectral signals into two-dimensional spectrogram images using Continuous Wavelet Transform (CWT) and Quantum Fourier Transform (QFT), which

are then used as input for feature extraction to ViT and DeiT. Feature vectors extracted from the penultimate transformer encoder layers were fused to generate a representation of each sample. These fused embeddings were passed through a Spectral-based Graph Convolutional Network (GCN), which modeled the inter-sample relationships in a graph structure to learn class-discriminative features.

The combination of attention-based global context modeling and graph-based local structure modeling enabled the proposed method to outperform standalone transformer networks. The Hybrid model achieved the highest classification overall accuracy of 97.93% using CWT. The results demonstrated that this hybrid fusion of spectral features leads to significant improvements in classification performance, particularly by enhancing minority class recall and improving class separability.

References

[1] Berghian-Grosan, C., Magdas, D.A.: Application of raman spectroscopy and machine learning algorithms for fruit distillates discrimination. Scientific Reports **10**, 21152 (2020) https://doi.org/10.1038/s41598-020-78159-8

[2] Khan, S., Ullah, R., Khan, A., Wahab, N., Bilal, M., Ahmed, M.: Analysis of dengue infection based on raman spectroscopy and support vector machine (svm). Biomedical Optics Express **7**(6), 2249–2256 (2016) https://doi.org/10.1364/BOE.7.002249

[3] Ryzhikova, E., Ralbovsky, N.M., Sikirzhytski, V., Kazakov, O., Halamkova, L., Quinn, J., Zimmerman, E.A., Lednev, I.K.: Raman spectroscopy and machine learning for biomedical applications: Alzheimer's disease diagnosis based on the analysis of cerebrospinal fluid. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **248**, 119188 (2021) https://doi.org/10.1016/j.saa.2020.119188

[4] Du, Y., Han, D., Liu, S., Sun, X., Ning, B., Han, T., Wang, J., Gao, Z.: Raman spectroscopy-based adversarial network combined with svm for detection of foodborne pathogenic bacteria. Talanta **237**, 122901 (2022) https://doi.org/10.1016/j.talanta.2021.122901

[5] Tian, F., Tan, F., Li, H.: An rapid nondestructive testing method for distinguishing rice producing areas based on raman spectroscopy and support vector machine. Vibrational Spectroscopy **107**, 103017 (2020) https://doi.org/10.1016/j.vibspec.2019.103017

[6] Luo, Y., Wang, Y., Li, B., Zhang, Y., Chen, Z., Wang, X., Wang, Y., Xu, Z., Wang, X., Liu, C.: Raman spectroscopy and machine learning for microplastics identification and classification in water environments. IEEE Journal of Selected Topics in Quantum Electronics **29**(4), 1–8 (2023) https://doi.org/10.1109/JSTQE.2022.3222065

[7] Medeiros Back, H., Junior, E.C.V., Alarcon, O.E., Pottmaier, D.: Training and evaluating machine learning algorithms for ocean microplastics classification through vibrational spectroscopy. Chemosphere **287**(Part 1), 131903 (2022) https://doi.org/10.1016/j.chemosphere.2021.131903

[8] Shan, J., Zhao, J., Liu, L., Zhang, Y., Wang, X., Wu, F.: A novel way to rapidly monitor microplastics in soil by hyperspectral imaging technology and chemometrics. Environmental Pollution **238**, 121–129 (2018) https://doi.org/10.1016/j.envpol.2018.03.026

[9] Zhu, Y., Yeung, C., Lam, E.: Microplastic pollution monitoring with holographic classification and deep learning. Journal of Physics: Photonics **3**(2) (2021) https://doi.org/10.1088/2515-7647/abf250

[10] Ishmukhametov, I., Nigamatzyanova, L., Fakhrullina, G., *et al.*: Label-free identification of microplastics in human cells: dark-field microscopy and deep learning study. Analytical and Bioanalytical Chemistry **414**, 1297–1312 (2022) https://doi.org/10.1007/s00216-021-03749-y

[11] Stefas, D., Gyftokostas, N., Bellou, E., Couris, S.: Laser-induced breakdown spectroscopy assisted by machine learning for plastics/polymers identification. Atoms **7**(3), 79 (2019) https://doi.org/10.3390/atoms7030079

[12] Michel, A.P.M., Morrison, A.E., Preston, V.L., Marx, C.T., Colson, B.C., White, H.K.: Rapid identification of marine plastic debris via spectroscopic techniques and machine learning classifiers. Environmental Science & Technology **54**(17), 10630–10637 (2020) https://doi.org/10.1021/acs.est.0c02099

[13] Chakraborty, I., Banik, S., Biswas, R., *et al.*: Raman spectroscopy for microplastic detection in water sources: a systematic review. International Journal of Environmental Science and Technology **20**, 10435–10448 (2023) https://doi.org/10.1007/s13762-022-04505-0

[14] Neo, E.R.K., Yeo, Z., Low, J.S.C., Goodship, V., Debattista, K.: A review on chemometric techniques with infrared, raman and laser-induced breakdown spectroscopy for sorting plastic waste in the recycling industry. Resources, Conservation and Recycling **180**, 106217 (2022) https://doi.org/10.1016/j.resconrec.2022.106217

[15] Lin, P.-Y., Wu, I.-H., Tsai, C.-Y., Kirankumar, R., Hsieh, S.: Detecting the release of plastic particles in packaged drinking water under simulated light irradiation using surface-enhanced raman spectroscopy. Analytica Chimica Acta **1198**, 339516 (2022) https://doi.org/10.1016/j.aca.2022.339516

[16] Schymanski, D., Goldbeck, C., Humpf, H.-U., Fürst, P.: Analysis of microplastics in water by micro-raman spectroscopy: Release of plastic particles from different packaging into mineral water. Water Research **129**, 154–162 (2018) https://doi.

org/10.1016/j.watres.2017.11.011

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) https://doi.org/10.48550/arXiv.2010.11929

[18] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). https://doi.org/10.48550/arXiv.2012.12877 . PMLR

[19] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) https://doi.org/10.48550/arXiv.1609.02907

[20] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) https://doi.org/10.48550/arXiv.1412.6980

[21] Qi, Y., Yang, L., Liu, B., Liu, L., Liu, Y., Zheng, Q., Liu, D., Luo, J.: Accurate diagnosis of lung tissues for 2d raman spectrogram by deep learning based on short-time fourier transform. Analytica Chimica Acta **1179**, 338821 (2021) https://doi.org/10.1016/j.aca.2021.338821

[22] P.K., N.H.K..A.A..G.P..R.: Vista: vision transformer-attention enhanced cnn ensemble for optimized classification of acute lymphoblastic leukemia benign and progressive malignant stages. International Journal of Information Technology (2024) https://doi.org/10.1007/s41870-024-02126-z . https://doi.org/10.1007/s41870-024-02126-z

[23] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv (2017) 1609.02907