Comparison of CNN and Transformer Models for Optical Filter Classification

Pharadee Tawansangaram
Harbin Institute of Technology, Shenzhen
pharadeeonepiece@gmail.com

Abstract

Automatic classification of optical filters from spectral measurements is important for optical system design and characterization. This project compares convolutional neural networks (CNNs) and Transformer-based models for classifying one-dimensional absorption spectra. A synthetic dataset of four filter types—long-pass, short-pass, band-pass, and neutral-density—was generated over 400–800 nm. Both models were trained under identical conditions and evaluated using accuracy, confusion matrices, ROC curves, PCA, confidence histograms, and training behavior. Results show that both architectures achieve high accuracy, with CNNs slightly outperforming in final accuracy and confidence, while Transformers converge faster and effectively capture global spectral dependencies. These findings highlight a trade-off between inductive bias and representational flexibility: CNNs excel at precise spectral classification with moderate datasets, whereas Transformers offer rapid convergence and global feature modeling, with potential for improved performance on larger or more diverse data.

1. Introduction

Machine learning has become an effective approach for analyzing one-dimensional spectral signals in optoelectronic systems. Optical filters are commonly characterized by their absorption spectra, which contain both local features, such as peaks and cutoff edges, and global patterns spanning wide wavelength ranges. This makes spectral data well suited for comparing convolution-based and attention-based learning models.

Convolutional neural networks (CNNs) are widely used for spectral classification due to their ability to capture local, physically meaningful features through convolutional filters. In contrast, Transformer models use self-attention mechanisms to model global dependencies across the entire spectrum, enabling the identification of long-range relationships and informative wavelength regions.

In this project, CNN and Transformer models are compared for optical filter classification using absorption spectra as input. A synthetic dataset comprising long-pass, short-pass, band-pass, and neutral-density filters is generated, and both models are evaluated under identical training conditions to assess their suitability for medium-sized spectral datasets and the interpretability benefits of attention mechanisms.

2. Background

2.1. Optical Spectra and Filter Characteristics

An optical spectrum describes the wavelength-dependent response of an optical component and is typically characterized in terms of transmission, reflection, or absorbance. Different types of optical filters exhibit distinct spectral shapes that reflect their underlying physical behavior. A neutral-density filter exhibits a nearly flat spectral profile, uniformly attenuating light across a broad wavelength range. Long-pass filters transmit wavelengths above a defined cutoff, while short-pass filters transmit wavelengths below the cutoff. Band-pass filters allow light to pass within a limited wavelength interval bounded by two cutoff regions. These distinctive spectral patterns form the basis for the classification of optical filters using data-driven learning models.

2.2. Machine Learning for Spectral Classification

Machine learning has been extensively applied to spectral data analysis in various fields, including spectroscopy, remote sensing, and material identification. Unlike the traditional method, which relies on manually extracted features, machine learning models can automatically learn informative features from raw spectral measurements. Convolutional Neural Networks (CNNs) have been widely applied to one-dimensional spectral classification because of their ability to capture local patterns. However, capturing relationships between distant wavelength regions across the entire spectrum can be a challenge for CNNs.

2.3. Transformer models and Self-attention

Transformers are a class of neural network architectures. With a self-attention mechanism, transformers can model relationships between different parts of a sequence. Unlike CNNs, which use fixed-size kernels, self-attention allows transformers to capture relationships across the entire spectrum. This makes them more suitable for analyzing spectral data, where important features may be distributed over a wide range of wavelengths. By assigning attention weight to different spectral regions, transformers can focus on the informative part of the wavelength that is relevant to filter classification.

3. Methodology

The system follows a structured pipeline consisting of data generation, preprocessing, feature learning, and classification, as illustrated in Fig. 1. First, optical spectra are repre-

sented as one-dimensional wavelength–response sequences and normalized to ensure a consistent input scale. The processed spectra are then fed into a Transformer-based model, where self-attention layers learn global spectral relationships across the entire wavelength range. Finally, a classification head maps the learned representations to output the predicted optical filter type.

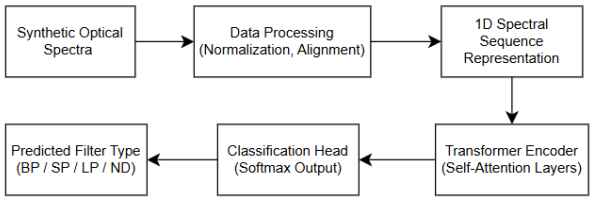

Figure 1. Overall technical scheme

3.1. Data Generation and Dataset Description

Since publicly available optical spectral datasets are limited, a synthetic dataset was generated to support supervised model training. A total of 12,000 optical spectra covering the wavelength range 400–800 nm were generated. Each spectrum was sampled on a fixed wavelength grid consisting of 401 wavelength points. For each optical filter class, 3,000 spectra were produced by randomly varying physically meaningful parameters such as absorption coefficient magnitude, cutoff slope, and center wavelength. Parameter ranges for optical filter spectra generation are summarized in Tab. S1 of the supplementary material.

Beer-Lambert Law

In this project, the Beer–Lambert law is formulated in terms of absorbance rather than transmitted intensity, as absorbance provides a linear relationship between spectral features and parameters, improves numerical stability during training, and matches the output format of common spectroscopic instruments. The absorbance at each wavelength is given by:

$$A(\lambda) = \varepsilon(\lambda)\, cL \tag{1}$$

Where:

- $\varepsilon(\lambda)$ is the wavelength-dependent absorption coefficient,
- c is the concentration (or an equivalent scaling factor),
- L is the optical path length.

For simplification, both c and L are normalized to 1, so classification depends only on spectral shape. Absorption coefficients were defined uniquely for each filter type to generate characteristic responses, with distinct mathematical forms for long-pass, short-pass, band-pass, and neutral-density filters (see Tab. S2 in the supplementary material for details). These formulations capture key features—cutoff behavior, passband regions, and uniform attenuation—providing the basis for model training.

To simulate realistic measurement conditions, additive Gaussian noise and a small baseline offset were applied to the ideal spectra, and absorbance values were clipped to remain non-negative, enhancing generalization. Examples of the generated spectra for each filter type are shown in Figs. S1 to S4 of the supplementary material.

3.2. Data representation and processing

Each optical spectrum was represented as a one-dimensional sequence of 401 wavelength-response values on a common grid. No additional alignment algorithm was required since all spectra already shared the same wavelength sampling. Spectra were normalized to the range [0, 1] to ensure consistent feature scaling across samples (see Eq. (S1) of supplementary for details). Categorical labels corresponding to each filter type were assigned, as summarized in Table 2, and applied consistently across the training, validation, and test datasets.

Table 1. Filter type mapping

Index	Filter Type
0	Long-pass (LP)
1	Short-pass (SP)
2	Band-pass (BP)
3	Neutral-density (ND)

Duplicate spectra were removed to prevent redundancy, and checks were conducted to avoid data leakage between training, validation, and test sets, ensuring that no identical or highly similar spectra appeared across subsets.

The dataset's discriminative characteristics were verified by computing the mean normalized spectrum for each filter class. As shown in Fig. 2, each class exhibits distinct spectral features, with characteristic shapes and cutoff regions preserved. Intra-class variability, assessed via standard deviation, is provided in Fig. S5 of the supplementary material. This variability reflects added noise and parameter perturbations but remains smaller than inter-class differences, ensuring realistic variations without obscuring class-defining features.

3.3. Model Architecture

This project implements two deep learning architectures for optical filter classification: a 1D CNN and a Transformer. Both models take 1D absorption spectra as input and output probabilities for four filter classes—band-pass, short-pass, long-pass, and neutral-density—allowing comparison of local feature extraction versus global attention mechanisms.

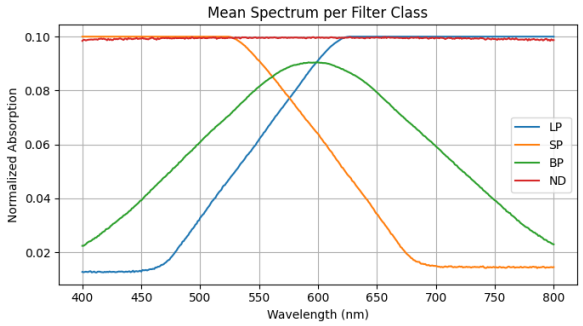

Figure 2. Mean spectral profiles for each optical filter class

3.3.1. CNN Architecture

A one-dimensional CNN is used as the baseline model to assess the relative advantages of the Transformer-based approach. CNNs are well suited for spectral analysis as they effectively capture local features such as absorption peaks and cutoff edges. However, their limited receptive field can restrict the modeling of long-range spectral dependencies, which may be important for distinguishing certain filter types.

The CNN is trained using the same dataset, preprocessing steps, and data splitting strategy as the Transformer to ensure a fair comparison. The network consists of stacked one-dimensional convolutional and pooling layers followed by fully connected layers for classification. A detailed layer configurations are summarized in Tab. S3 of the supplementary material, with an overview of CNN architecture shown in Fig. 3.

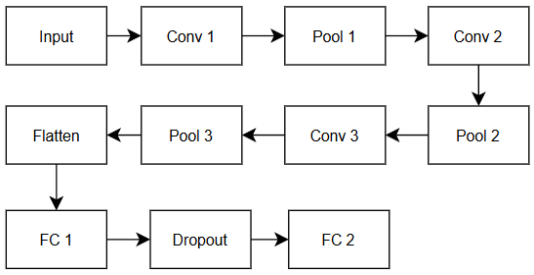

Figure 3. CNN model schematic

3.3.2. Transformer Architecture

The Transformer-based model is designed to overcome the limitations of local feature extraction by capturing global spectral dependencies. Each input spectrum (length 401) is projected into a higher-dimensional feature space using a linear embedding layer. Positional encoding is added to preserve the wavelength order information, resulting in an embedded feature sequence of shape $N \times d_{model}$, where $N = 401$ and $d_{model} = 64$.

The embedded sequence is then processed by a stacked Transformer encoder composed of self-attention layers and feed-forward networks. Self-attention allows the model to learn interactions between distant wavelength points, which is particularly valuable for distinguishing filter types with subtle, distributed spectral differences.

After feature extraction, the sequence output is summarized into a single spectrum-level representation using global average pooling (see Eq. (S2) of the supplementary for the full equation). This pooled vector is then fed into a fully connected classification head with a softmax activation to produce the final four-class probability vector. A schematic of the Transformer architecture is shown in Fig. 4, with detailed model parameters provided in Tab. S4 of the supplementary material.

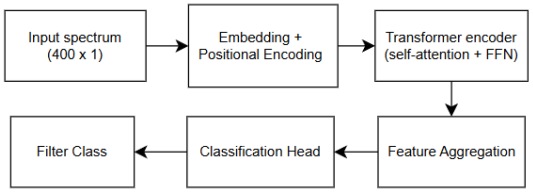

Figure 4. Transformer model schematic

3.3.3. Comparative Considerations

CNN and Transformer models process spectral data differently. CNNs extract local features, efficiently identifying filters with sharp or distinctive absorption, but they struggle with long-range dependencies. Transformers use self-attention to capture global relationships across the spectrum, detecting patterns spanning multiple wavelengths. Training both models under identical conditions ensures observed differences reflect architectural design, allowing evaluation of the trade-offs between local feature specialization in CNNs and global modeling with interpretability in Transformers.

3.4. Training Procedure

The dataset was randomly split into training, validation, and test sets (70/15/15) using stratified sampling to maintain balanced class representation. The training set was used to optimize model parameters, the validation set guided hyperparameter tuning and early stopping, and the test set evaluated final performance. All weights were initialized using Xavier initialization. Training parameters for both models are provided in Tab. S5 of the supplementary material.

Models were trained with the Adam optimizer, processing spectra in mini-batches. Forward propagation computed predicted probabilities and cross-entropy loss (see Eq. (S3) of the supplementary for details), and parameters were updated via backpropagation. Validation performance was monitored each epoch to prevent overfitting, with training

and validation loss and accuracy recorded to generate convergence curves.

Regularization techniques included 10% dropout, early stopping, and data augmentation, improving generalization on unseen spectra. Final model evaluation used classification accuracy, confusion matrices, and class-wise recall and F1-scores, enabling quantitative and qualitative comparisons. Loss and accuracy curves further illustrate training stability and efficiency. Figure 5 summarizes the model training process.

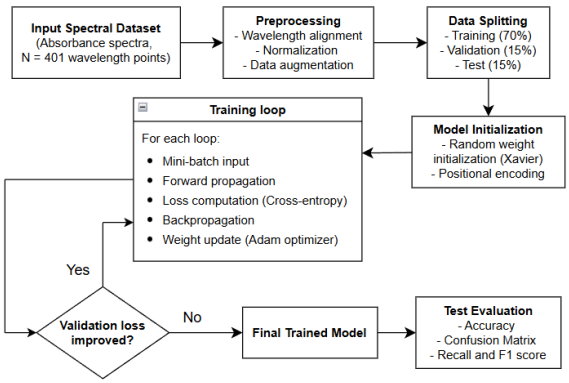

Figure 5. Training model schematic

4. Result

This section presents a comprehensive evaluation of the convolutional neural network (CNN) and Transformer-based models for optical filter classification. Both qualitative and quantitative analyses are conducted to assess classification performance, robustness, and feature representation capability. All results are obtained using the same test dataset to ensure a fair comparison.

4.1. Quantitative Performance

4.1.1. Classification Performance

Both CNN and Transformer models achieve high classification performance on the test set, with the CNN slightly outperforming (99.44% vs. 97.98%). The CNN shows consistently strong precision, recall, and F1-scores across all classes, with perfect classification of neutral-density filters and only minor confusion among long-pass, short-pass, and band-pass types. The Transformer also performs robustly, though recall is slightly lower for band-pass filters, reflecting sensitivity to subtle inter-class similarities. Macro-averaged metrics are high for both models, indicating balanced performance. Overall, the CNN demonstrates a slight advantage in accuracy and class-wise consistency, highlighting the effectiveness of convolutional architectures for localized spectral features. Full metrics are provided in Tabs. S6 and S7 of the supplementary material.

4.1.2. Confusion Matrix Analysis

Confusion matrices (Figs. 6 and 7) provide a detailed class-wise evaluation, revealing specific misclassification patterns beyond overall accuracy. For the CNN, the confusion matrix shows near-perfect performance across all four filter types. Short-pass and neutral-density filters are classified with particularly high reliability, while minor confusion occurs between long-pass and band-pass filters due to similarities in their cutoff regions and overlapping transition slopes. This reflects the CNN's ability to capture localized spectral features such as absorption edges and abrupt transitions, leading to highly accurate predictions.

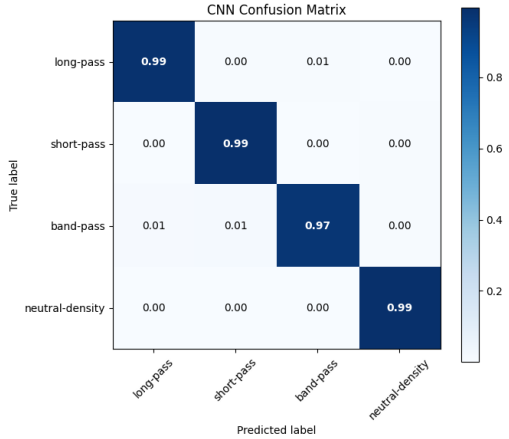

Figure 6. CNN confusion matrix

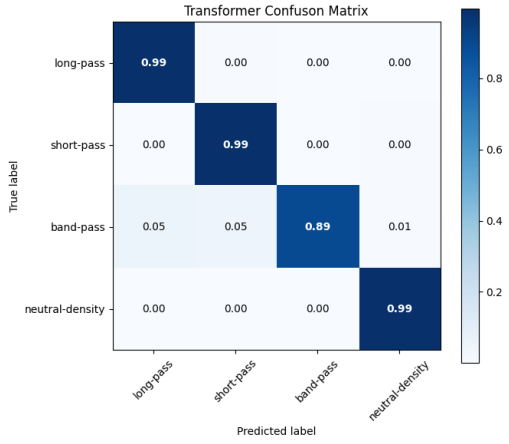

Figure 7. Transformer confusion matrix

The Transformer also performs strongly but shows slightly higher misclassification rates, especially for band-pass filters, which are occasionally confused with long-pass or short-pass types. This aligns with the lower recall observed for band-pass in quantitative metrics and suggests

that subtle local features may be less emphasized when attention is distributed globally. Both models, however, achieve high true positive rates for neutral-density filters, indicating that spectra with uniform attenuation are easily separable from wavelength-dependent filters.

4.2. Qualitative Performance

4.2.1. ROC Curve Visualization

ROC curves evaluate class-wise discrimination in a threshold-independent manner, using one-vs-rest curves for each filter class. Both models achieve near-perfect AUC values, indicating strong separability. As shown in Figs. 8 and 9, the CNN's ROC curves remain closer to the top-left corner, reflecting robust class separation due to its ability to capture localized spectral features such as sharp cutoffs and absorption transitions.

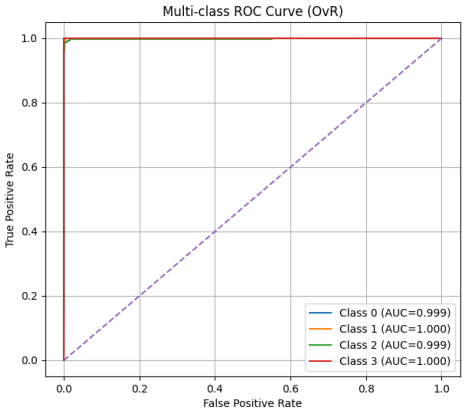

Figure 8. CNN ROC curve

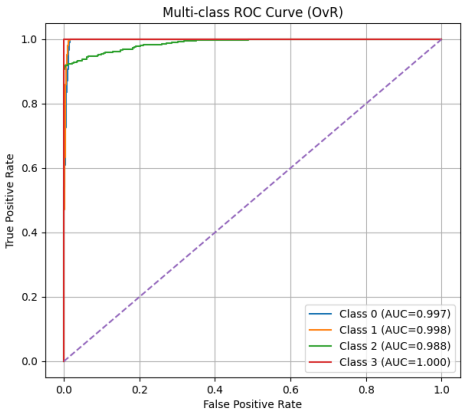

Figure 9. Transformer ROC curve

The Transformer also shows strong ROC performance, indicating that self-attention captures global spectral patterns. Its AUC values are slightly lower than the CNN, especially for band-pass filters, reflecting increased ambiguity near class boundaries. Overall, ROC–AUC analysis confirms strong class separability for both models, with the CNN exhibiting more stable discrimination while the Transformer remains highly competitive.

4.2.2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is used to visualize the distribution of optical spectra in a two-dimensional feature space and to assess the inherent separability between different optical filter classes. Applied to the combined training and test sets, the first two components capture 54.3% and 24.2% of the variance. The cumulative explained variance of around 78.5% indicates that the two-dimensional projection preserves the dominant spectral variations. As shown in Fig. 10, neutral-density filters form compact, well-separated clusters, while long-pass and short-pass filters partially overlap with band-pass filters near class boundaries. This overlap explains the minor misclassifications observed in the confusion matrices and highlights the need for models that capture subtle spectral differences.

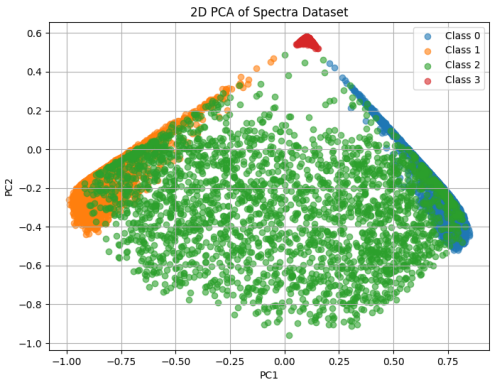

Figure 10. 2D PCA of the optical spectra dataset by filter class

4.2.3. Confidence Histogram Analysis

Confidence histograms assess the distribution of predicted class probabilities and the certainty of each model's decisions. CNN predictions are tightly concentrated near 1.0, indicating highly confident classification for nearly all spectra, reflecting its strength in capturing discriminative local features. In contrast, the Transformer shows a slightly broader distribution from 0.7–1.0, with most predictions between 0.9 and 1.0 as well as minor misclassifications near 0.8–1.0. This suggests the Transformer accounts for global spectral context and subtle ambiguities, particularly near class boundaries. Prediction confidence histograms for both

models are shown in Figs. S6 and S7 of the supplementary material.

4.3. Training Behavior

Training and validation loss and accuracy curves were used to assess model convergence and generalization. The Transformer (Fig. 11) converges rapidly, reaching stable loss in fewer epochs, reflecting its ability to capture global spectral dependencies early. In contrast, the CNN (Fig. 12) exhibits more gradual convergence, consistent with progressive hierarchical feature learning, but achieves strong final performance. For both models, validation closely follows training, indicating minimal overfitting. Overall, the results highlight a trade-off between fast global representation learning in Transformers and steadier convergence in CNNs. Additional accuracy curves are provided in the supplementary material (Figs. S8 and S9).

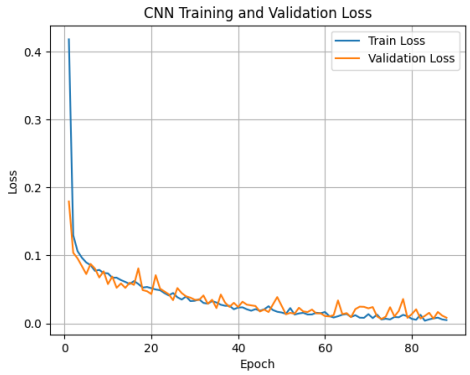

Figure 11. CNN training and validation loss curves

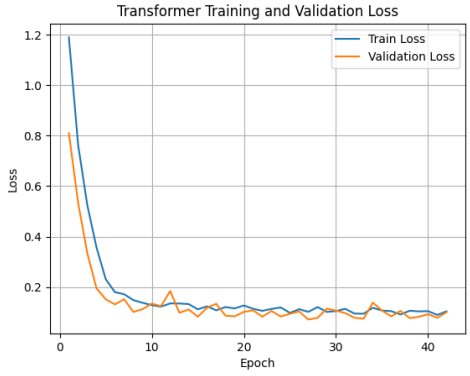

Figure 12. Transformer training and validation loss curves

4.4. Overall Model Comparison and Discussion

Quantitative evaluation shows that both CNN and Transformer models achieve strong optical filter classification, with the CNN slightly outperforming (99.44% vs. 97.98%). CNNs excel at capturing localized spectral features, while Transformers leverage global dependencies and benefit from larger datasets. ROC–AUC and confidence analyses confirm robust class separability, with CNN predictions more tightly clustered and the Transformer showing slightly higher uncertainty near class boundaries. PCA reveals partial overlap among long-pass, short-pass, and band-pass filters, which CNNs resolve locally while Transformers capture broader spectral context. In training, the Transformer converges faster, whereas the CNN reaches higher final accuracy. Both models exhibit minimal overfitting, highlighting a trade-off between precision and global feature modeling, with CNNs offering higher accuracy and Transformers providing interpretability and rapid convergence.

5. Conclusion

This study compared CNN and Transformer models for classifying four types of optical filters using synthetic absorption spectra. Both models achieved high accuracy, demonstrating the feasibility of deep learning–based spectral analysis. The CNN attained slightly higher final accuracy and more confident predictions, reflecting its strength in capturing localized spectral features, while the Transformer converged faster and effectively modeled global dependencies. Overall, CNNs excel at high-precision tasks with moderate data, whereas Transformers offer efficient global feature learning and potential gains with larger datasets.

References

[1] Chroma Technology Corp. Filter characteristics – about fluorescence. https://www.chroma.com/support/knowledge-center/about-fluorescence/filter-characteristics. Accessed: Jan. 2026.

[2] Edmund Optics. Optical filters. https://www.edmundoptics.com/knowledge-center/application-notes/optics/optical-filters/. Accessed: Jan. 2025.

[3] Rui Hang, Qian Liu, Danfeng Hong, Pedram Ghamisi, and Siddhartha Bhattacharyya. Hyperspectral image classification with attention-aided cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2281–2293, 2020.

[4] Eugene Hecht. *Optics*. Pearson, Boston, MA, USA, 5 edition, 2017.

[5] Danfeng Hong et al. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

[6] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015:1–12, 2015. Online; accessed 2026.

Comparison of CNN and Transformer Models for Optical Filter Classification

Supplementary Material

1. Figures

The following figures provide additional visualization of the optical spectra and model predictions. Each figure references the relevant section in the main paper.

1.1. Examples of Generated Spectra for Each Filter Type

Figs. S1-S4 show representative examples of the generated spectra for each filter type. The data generation procedure is described in Sec. 3.1.

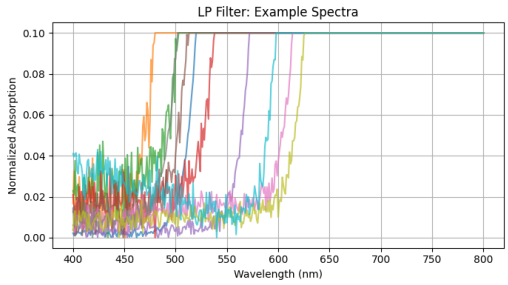

Figure S1. Spectral shapes of long-pass filters.

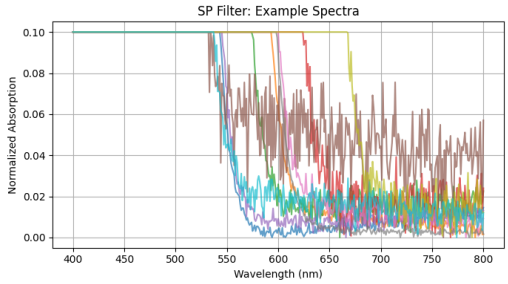

Figure S2. Spectral shapes of short-pass filters.

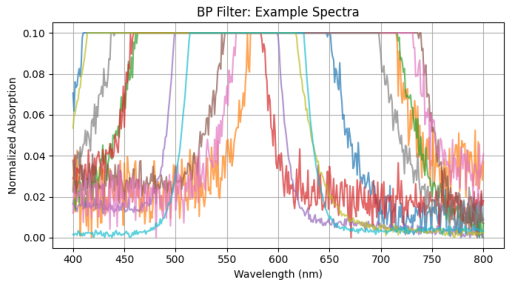

Figure S3. Spectral shapes of band-pass filters.

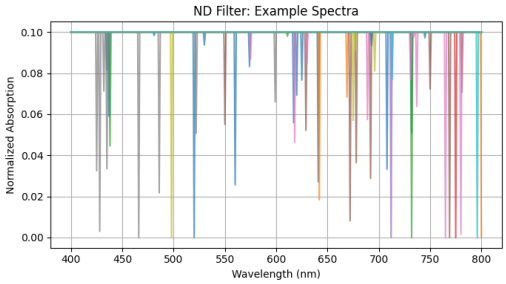

Figure S4. Spectral shape of a neutral density filter.

1.2. Intra-Class Variation of Optical Spectra

The intra-class variability, discussed in Sec. 3.2, is shown in Fig. S5 below.

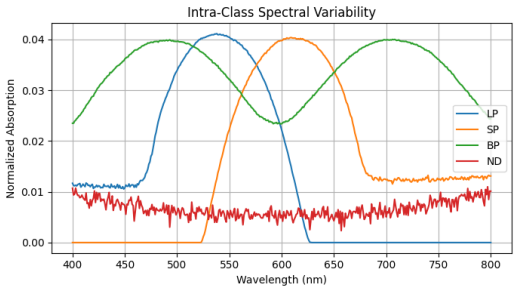

Figure S5. Intra-class variation of optical spectra

1.3. Prediction Confidence Distribution

The distribution of prediction confidence scores for both models (from Sec. 4.2.3) is shown in Figs. S6 and S7.

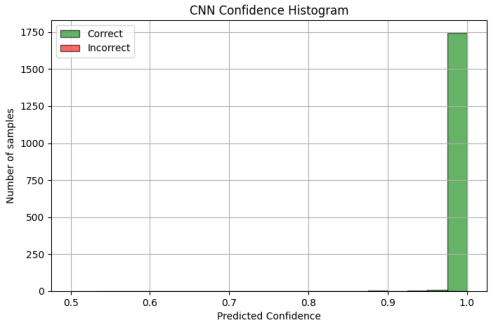

Figure S6. CNN prediction confidence distribution on the test dataset.

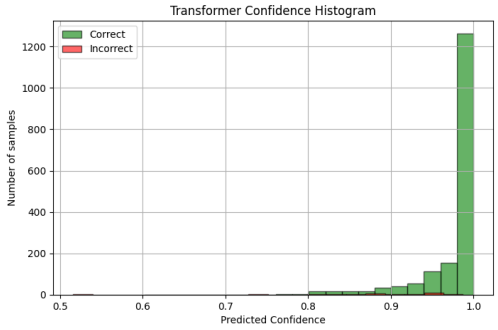

Figure S7. Transformer prediction confidence distribution on the test dataset.

1.4. Training and Validation Accuracy Curves

Training and validation accuracy curves for both models are shown in Figs. S8 and S9 , corresponding to the training setup described in Sec. 3.4.

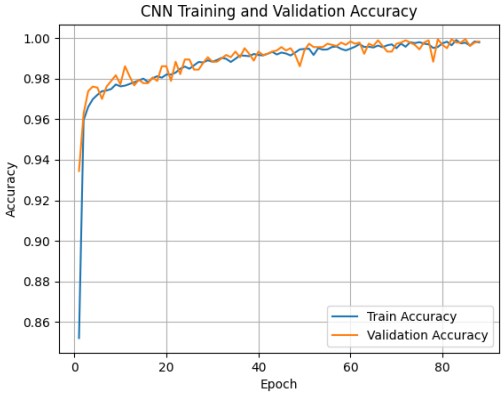

Figure S8. CNN training and validation accuracy curves

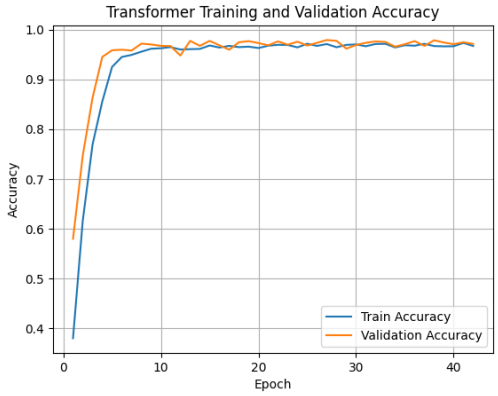

Figure S9. Transformer training and validation accuracy curves

2. Tables

The following tables provide additional details on data generation, model configurations, training settings, and classification performance. Each table is linked to the relevant section in the main paper.

2.1. Spectral Generation Model

Table S1 and S2 summarize the parameter ranges and absorption coefficient formulations used to generate the synthetic optical filter spectra, as described in Section Sec. 3.1 of the main paper.

Table S1. Parameter ranges for optical filter spectra

Parameter	Value
Wavelength range	400–800 nm
Sampling resolution	1 nm
Optical path length (L)	1.0
Concentration (c)	1.0
Base absorption (ε_0)	1.0
Center wavelength (λ_c)	500–650 nm
Cutoff slope (k)	10 nm
Band center (λ_0)	500–700 nm
Bandwidth (σ)	20–60 nm
ND absorption	0.3–0.9
Noise standard deviation (σ_n)	0.002–0.01
Baseline offset (b)	-0.02 – 0.02

Table S2. Absorption coefficient equations used to generate synthetic spectra for each optical filter type.

Filter Type	Absorption Coefficient
Long-pass (LP)	$\varepsilon_{LP}(\lambda) = \frac{1}{1+e^{-(\lambda-\lambda_c)/\sigma}}$
Short-pass (SP)	$\varepsilon_{SP}(\lambda) = \frac{1}{1+e^{+(\lambda-\lambda_c)/\sigma}}$
Band-pass (BP)	$\varepsilon_{BP}(\lambda) = e^{-(\lambda-\lambda_0)^2/(2\sigma^2)}$
Neutral-density (ND)	$\varepsilon_{ND}(\lambda) \approx$ constant

2.2. Model Architectures

The CNN and Transformer architectures used in this study are detailed in Tables S3 and S4, which complement the methodology described in Section Secs. 3.3.1 and 3.3.2 of the main paper.

2.3. Training Configuration

Training hyperparameters shared by both models are listed in Table S5, corresponding to the setup described in Section Sec. 3.4 of the main paper.

Table S3. CNN Layer Configuration

Layer	Type	Kernel	Filters	Activation
Conv1	1D Conv	7	16	ReLU
Conv2	1D Conv	5	32	ReLU
Conv3	1D Conv	3	64	ReLU
Max-Pooling	Pool	–	–	–
Flatten / GAP	–	–	–	–
FC1	Dense	–	128	ReLU
FC2	Dense	–	4	Softmax

Table S4. Transformer model parameter

Parameter	Value
Number of wavelength points	401
Embedding dimension (d_{model})	64
Number of encoder layers	4
Number of attention heads	8
Classification head	Dense, 4 units, Softmax

Table S5. Training parameters

Parameter	Value
Batch size	32
Learning rate	5×10^{-4}
Number of epochs	100
Early stopping patience	10

2.4. Classification Performance

Detailed class-wise performance metrics for the CNN and Transformer models are reported in Tables S6 and S7, extending the results presented in Section Sec. 4.1.1 of the main paper.

Table S6. CNN classification performance metrics

Class	Precision	Recall	F1-score	Support
Long-pass	0.99	1.00	0.99	450
Short-pass	1.00	0.99	0.99	450
Band-pass	0.99	0.99	0.99	451
Neutral-density	1.00	1.00	1.00	429
Overall Accuracy	–	–	0.99	1780
Macro Average	0.99	0.99	0.99	1780
Weighted Average	0.99	0.99	0.99	1780

3. Equations

The following equations are used in different parts of our workflow. Each is referenced with the relevant section in the main paper.

Table S7. Transformer classification performance metrics

Class	Precision	Recall	F1-score	Support
Long-pass	0.95	1.00	0.98	450
Short-pass	0.97	1.00	0.99	450
Band-pass	1.00	0.92	0.96	451
Neutral-density	1.00	1.00	1.00	429
Overall Accuracy	–	–	0.98	1780
Macro Average	0.98	0.98	0.98	1780
Weighted Average	0.98	0.98	0.98	1780

3.1. Min-Max Normalization

As described in Section Sec. 3.2, min-max normalization was applied to the input spectra:

$$A_{\text{norm}}(\lambda) = \frac{\tilde{A}(\lambda) - A_{\min}}{A_{\max} - A_{\min}} \tag{S1}$$

where $A_{\min}$ and $A_{\max}$ denote the minimum and maximum absorbance values of the spectrum.

3.2. Global Average Pooling

Following the feature extraction stage discussed in Section Sec. 3.3.2, the sequence of outputs is summarized into a single spectrum-level representation using global average pooling:

$$h_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} h_i \tag{S2}$$

where N is the number of wavelength points in the input spectrum, and h_i is the feature vector corresponding to wavelength i.

3.3. Cross-Entropy Loss

The classification loss, as introduced in Section Sec. 3.4, is evaluated using the cross-entropy function:

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{S3}$$

where C is the number of filter classes, y_i is the ground-truth label, and $\hat{y}_i$ is the predicted class probability.