A Suitable and Interpretable Methodology for FTIR Spectral Classification

Thomas Hartigan
Department of Physics
University of Cambridge
tjh200@cam.ac.uk

Tiago Azevedo, Pietro Liò
Department of Computer Science and Technology
University of Cambridge
{tmla2, pl219}@cam.ac.uk

Abstract

We propose a suitable and interpretable methodology for FTIR spectral classification when using weakly-labelled data. A multi-scale CNN is implemented with first layer kernel widths chosen to match common FTIR peak widths, before an ensemble of these models (EMSCNN) is constructed using validation set voting. In the context of cancerous tissue classification from FTIR spectra, EMSCNN achieves a weighted mean of per-class sensitivities of $83 \pm 6\%$ and F1 score of $67 \pm 7\%$, beating all other models tested. A new semi-supervised VAE version of the CRIME framework is implemented to interpret the model, elucidating distinct pathways to each spectral classification. Multiple VAE architectures are investigated using convolutional or transformer encoders and linear or transformer decoders. Finally, linear weighted cosine similarity models are constructed using the VAE latent space and achieve similar performance to direct classification methods. Our code is available here.

1 Introduction

The scientific community has long appreciated the potential of Machine Learning (ML) for predictive tasks using spectroscopic data [Zhang et al., 2022, Contreras and Bocklitz, 2025]. Fourier Transform Infrared Spectroscopy (FTIR) is one of the most widely used techniques for determining sample composition, and unlike most spectrometers, FTIR spectrometers are generally fast, inexpensive, and robust. Over the past few years, multi-scale convolutional neural networks (MSCNNs) have been shown to perform exceptionally in many spectroscopic fields ranging from acute neonatal quiet sleep detection using EEG signals to Salmonella serovar identification from surface-enhanced Raman spectra [Ansari et al., 2022, Ding et al., 2021, Yu et al., 2021, Cai et al., 2022, Tang et al., 2023]. Despite this, the application of MSCNNs to FTIR spectra has been limited, with most analyses instead applying linear models, CNNs, random forests, or gradient boosting to a small number of manually-selected integrated spectral peaks or PCA components [Haghi et al., 2021, Zhang et al., 2025]. Three previous works [Luo et al., 2025, Leng et al., 2023, Shuai et al., 2024] have applied MSCNNs to FTIR spectra, yet none of them use a physics-informed kernel design.

Work on explainable FTIR systems has also been limited, with previous works either focussing on a small number of input features [Ceran and Gurbanov, 2025, Zhang et al., 2025], genetic programming (GP) [Goodacre, 2003], or LIME and SHAP weighting analysis [Zhang et al., 2025, Haghi et al., 2021]. Of these approaches, limiting features and genetic programming both risk producing over-simplified models, whilst LIME and SHAP weightings are difficult to interpret. Consequently, the existing FTIR explainability methods are unlikely to simultaneously satisfy both clinical explainability and accuracy requirements [Slack et al., 2020, Aboy et al., 2024, Farah et al., 2023, Contreras and Bocklitz, 2025].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

To overcome some of these challenges, this work provides a framework for utilising weakly-labelled (one label per image) FTIR hypercubes to extract multiple interpretable pathways per classification. This is achieved in three steps. First, we implement a new MSCNN architecture with kernel sizes chosen to match the widths of common FTIR features. Second, we apply a modified CRIME [Zaki et al., 2024] framework to this model with a physics-motivated semi-supervised variational autoencoder (VAE) architecture, generating clusterings of LIME [Ribeiro et al., 2016] explanations. We then hypothesise that separate CRIME VAE latent-space clusterings corresponding to the same classification can represent different paths to that classification. Based on this, we use the CRIME latent space clusters as the basis for a cosine-based spectral-similarity model to approximate the original MSCNN.

2 Methodology

Model Architecture The relatively simple MSCNN architecture with 74866 trainable parameters implemented in this work is detailed in Figure 1. The first layer incorporates multiple convolutional heads with different kernel sizes chosen to span the same widths as common FTIR spectral features. The choice of relatively large kernel sizes here is essential, as it provides a pathway for the model to easily differentiate between slightly translated or skewed absorption peaks - features that are characteristic of environments with differing levels of acidity or hydrogen bonding [Wolpert and Hellwig, 2006]. The remaining structure of the architecture is a compacted version of that proposed by Zaki et al. [2024], but with significantly increased dropout frequency to prevent overfitting to weakly-labelled data. To further prevent overfitting, we construct a voting ensemble (EMSCNN) of models which converge well based on the validation set.

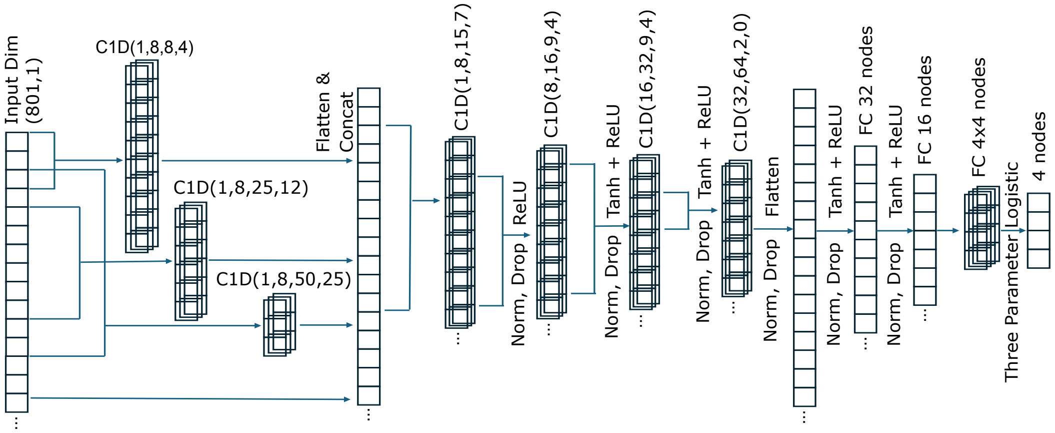

Figure 1: The MSCNN architecture implemented in this work. C1D(input channels, output channels, kernel size, stride) denotes a 1D convolutional layer and FC denotes a fully connected layer.

Modified CRIME Framework We modified the CRIME framework proposed by Zaki et al. [2024] in two main ways. First, we implemented the framework using a semi-supervised M2 VAE [Kingma et al., 2014], allowing us to investigate the effect of incorporating model classification predictions directly into the latent space clustering. Second, we implemented a range of encoder-decoder pairs designed to capture the interactions across different wavelengths. The input to each VAE is the concatenation of a FTIR spectrum and its LIME weights. These encoder-decoder pairs had convolutional-perceptron (CP), convolutional-transformer (CT) and transformer-transformer (TT) architectures. The convolutional encoder incorporates a MSCNN approach with kernel sizes of 3 and 300 in the first layer, before concatenation and fully-connected (FC) projection. The perceptron decoder has a FC layer mapping to the output features through a scaled and biased $\tanh$ (SBT) projection. The transformer encoder copies the first convolutional encoder layer, but projects each convolved feature to 4 dimensions after concatenation. It then uses a transformer encoder framework with 3 layers, 4 heads, 64 hidden dimensions, and sinusoidal positional encoding, followed by a FC projection. The transformer decoder uses a FC projection followed by a causal transformer decoder framework with 3 layers, 4 heads, feed-forward dimension 256, and output dimension 64, before FC reduction to one dimension per output feature, and SBT projection. The classifier head after each feature extractor has FC layers mapping from $64 \rightarrow 128 \rightarrow 4$ dimensions with ReLU activation. Latent embeddings were calculated by concatenating the classifier probabilities and the encoder outputs, before FC projecting to 4 dimensions for both μ and σ.

Linearisation Process After constructing the VAE latent space, we again follow the standard CRIME framework and apply K-means clustering to the embeddings, generating contexts. To construct a maximally interpretable model from this, we first calculate the average LIME weightings and average spectrum for each context. We then label these with the most frequent MSCNN classification prediction for the spectra within that context. To classify a spectrum, we calculate the normalised-LIME-weighted cosine similarity between each normalised average spectra and the normalised test spectrum, classifying the test spectrum according to the label of the context with which it best matches. Here, normalisations are linear mappings of the range to $[-1, 1]$.

Experimental Method An FTIR dataset recorded and pre-processed as described by Nallala et al. [2016] was provided by the University of Exeter. This dataset consisted of 90 FTIR spectral images (1.19 M spectra) with 5 μm spatial resolution and intensities at 801 wavenumbers from 1000–1800 cm^{-1}. Of these images, 31 were classified as normal, 17 as hyperplastic, 27 as adenomatous, and 15 as cancerous, with the latter classifications taking precedence if multiple classes of tissue were present. The spectra within each image were therefore weakly-labelled.

Using the stratified group K-fold method, the dataset was split into four independent sets of spectral images, whilst maximising the cross-set class balance. All experiments were performed with 4-fold cross-validation by permuting which dataset splits constituted the training, validation and test sets in a 2:1:1 ratio respectively. MSCNN models were trained on the training set, and tuned using the validation set, whilst VAEs were trained and tuned using the validation set. All results presented show test-set performance. The models and VAEs were trained for four-class prediction. However, it is often more convenient to present results as non-cancerous (normal, hyperplastic and adenomatous tissues) vs cancerous tissues. To account for the varying number of spectra with each classification label, we define the sample weighted sensitivity (SWS) as the weighted mean of per-class sensitivities, where the weights are proportional to the number of spectra in each class. Unless otherwise stated, F1 denotes the mean across folds value in binary cancerous vs non-cancerous classification.

3 Results and Discussion

MSCNN Performance As the spectra utilised here are weakly labelled, and the number of independent samples in an evaluation is relatively small (~20), we expect performance to vary depending on which features a model learns. To ensure reliable ensemble performance, we therefore set the threshold for EMSCNN inclusion at the cross-fold MSCNN validation set median SWS (> 0.55).

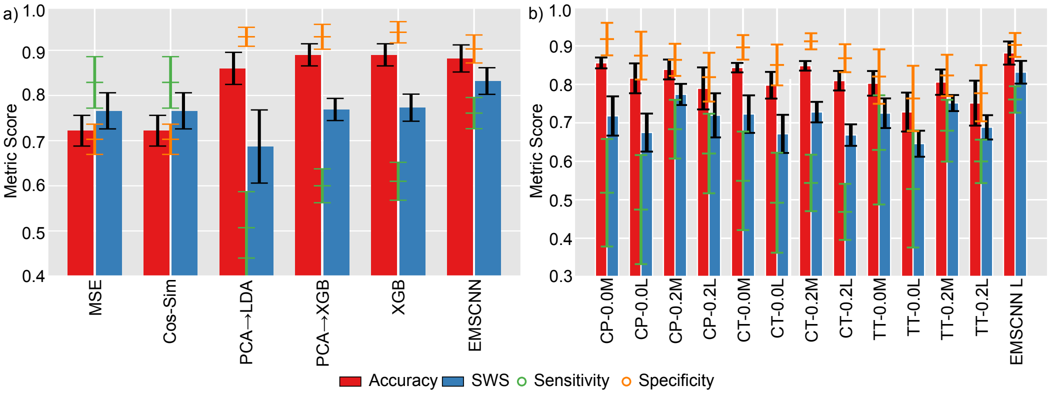

Figure 2: Binary cancer classification performance by model type. a) Direct predictions b) Linearised CRIME predictions. M and L denote EMSCNN model reproduction, and label reproduction performance respectively. The model suffix denotes the semi-supervision level, where $0.2 \rightarrow 20\%$ supervision.

To evaluate EMSCNN performance, we compared it with mean square error, cosine similarity, PCA fed linear discriminant analysis (PCA→LDA), PCA fed eXtreme gradient boosting (PCA→XGB), and direct eXtreme gradient boosting (XGB). In each case, the PCA dimension was 25, and default hyperparameters were used. The results in Figure 2a show that our EMSCNN model achieves the best SWS and F1 scores (0.83 ± 0.06 and 0.67 ± 0.07 respectively), and is the only model to perform well in all metrics simultaneously. For comparison, XGB achieves SWS and F1 scores of 0.77 ± 0.03 and 0.62 ± 0.07 respectively. This good cross-metric performance, and the ability for the EMSCNN to produce qualitatively excellent spatially reconstructed predictions gives us confidence that EMSCNN is the most suitable model for weakly-labelled FTIR spectral classification.

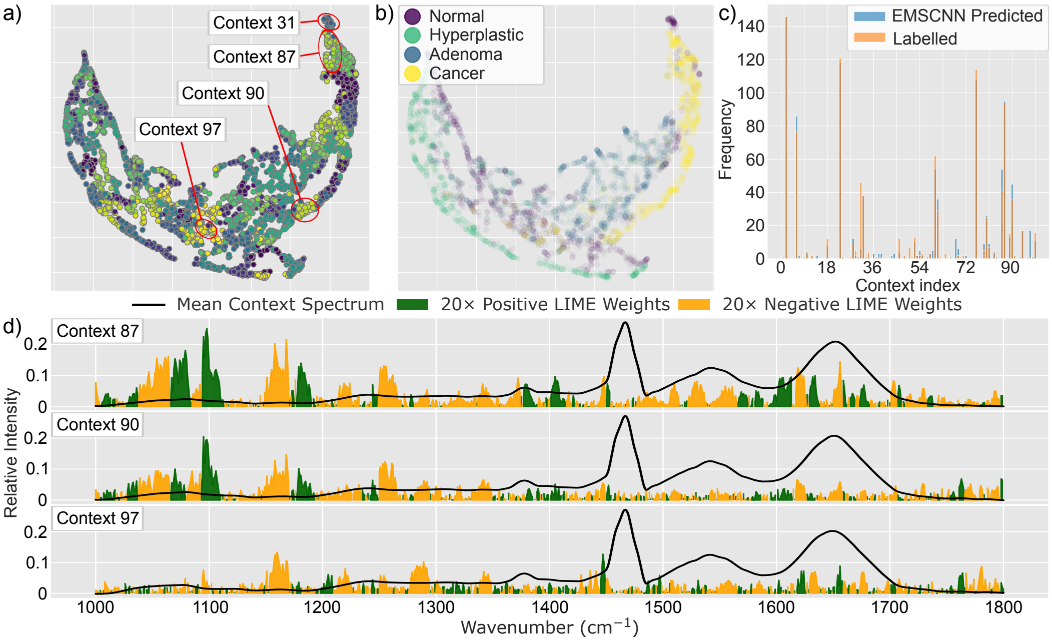

Figure 3: UMAP representations of a 20% supervised CT VAE latent space after CRIME clustering (a) and coloured by label (b). c) The number of predicted and labelled cancerous entries within each context. d) The mean spectra and LIME weightings for cancerous labelled contexts at extreme ends of the UMAP arm (87, 90), and one which correctly opposes the target labels (97).

Explainability and Conclusions Although the CRIME framework proposed by Zaki et al. can produce interpretable latent spaces, we found that the addition of low-level supervision improved interpretability, as shown in Figures 3a-b. Compared with the unsupervised approach, which produces the latent space shown in Figure 4a, the latent space in Figure 3b much better separates the non-cancerous classifications. These figures also demonstrate how the semi-supervised VAE has learnt to produce two wings; one with normal and hyperplastic tissues, and one with adenomatous and cancerous tissues, with all tissues present in the centre. This corresponds well with the weakly-labelled nature of the dataset, and the physical similarities between tumour (adenomatous and cancerous) and non-tumour (normal and hyperplastic) tissues. The cluster of normal labelled spectra in context 31 corresponds to necrotic tissue, which is physiologically similar to cancerous tissue. Remarkably, the VAE well separates this context, suggesting it could be classified well if necrotic labels were also available for training.

The spectra for contexts 87 and 90 in Figure 3d both correspond to cancer-dominated clusters, yet their LIME weightings differ significantly, demonstrating the ability of this methodology to extract different pathways to the same classification. For example, the weightings which are heavier in context 87 between 1300–1700 cm^{-1} are clustered around spectral peaks and troughs - indicative of combined imperceptible shifts due to the environmental pH [Wolpert and Hellwig, 2006]. The majority of spectra in context 97 are labelled as cancerous, but the EMSCNN has correctly learnt that these were mislabelled, as is particularly evident when comparing the spatial locations of these clustered spectra with the H&E stain of the sample as shown in Figure 4b-c. Our claim is further

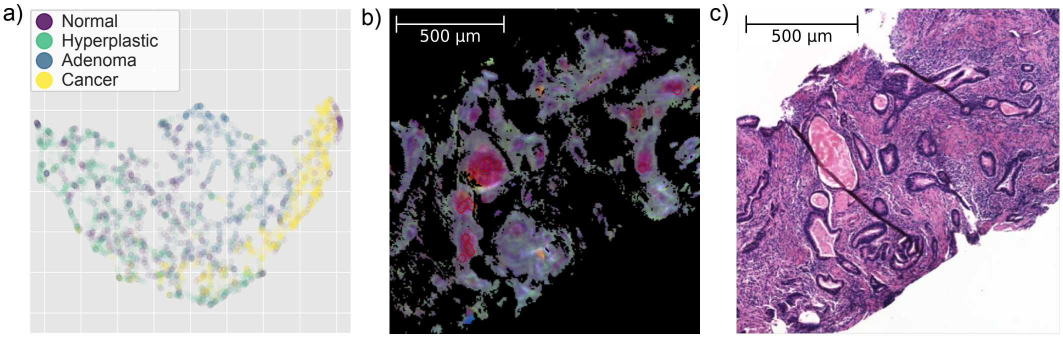

Figure 4: a) The UMAP representation of an unsupervised CT VAE latent space after CRIME, coloured by label. b) A false-colour FTIR image for a cancerous sample, showing the spatial locations (dots within red circles) of spectra within non-cancerous CRIME context 97. c) The H&E image corresponding to b).

supported by the major differences between LIME weightings in the range 1000–1300 cm^{-1} for context 97 and those in contexts 87 and 90. This region alone could be sufficient to inform cancer / non-cancer classification, but further work is required to confirm this.

The results in Figure 3c demonstrate that context-wise EMSCNN and label classifications often agree well, and that different classifications are primarily clustered in small, independent subsets of all contexts. This assists the linearisation methodology, enabling the performance seen for many of the VAE architectures detailed in Figure 2b. It is generally observed that the VAEs recreate the model predictions better than the label predictions, and that enabling supervision improves prediction performance in all cases. The 20% supervised CP and TT models performed especially well, achieving performance comparable to the direct classification methods shown in Figure 2a. Our methodology therefore increases the interpretability of CRIME in weakly-labelled FTIR spectral classification problems, and enables the generation of performant and interpretable linearised models.

References

M. Aboy, T. Minssen, and E. Vayena. Navigating the EU AI Act: implications for regulated digital medical products. *npj Digital Medicine*, 7(1):237, Sept. 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01232-3.

A. H. Ansari, K. Pillay, A. Dereymaeker, K. Jansen, S. Van Huffel, G. Naulaers, and M. De Vos. A Deep Shared Multi-Scale Inception Network Enables Accurate Neonatal Quiet Sleep Detection With Limited EEG Channels. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1023–1033, Mar. 2022. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2021.3101117.

Y. Cai, D. Xu, and H. Shi. Rapid identification of ore minerals using multi-scale dilated convolutional attention network associated with portable Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 267:120607, Feb. 2022. ISSN 13861425. doi: 10.1016/j.saa.2021.120607.

N. Ceran and R. Gurbanov. A proof-of-concept diagnostic platform for neonatal calf diarrhea using serum infrared spectroscopy and predictive analytics. *Analytical Biochemistry*, 705:115924, Oct. 2025. ISSN 00032697. doi: 10.1016/j.ab.2025.115924.

J. Contreras and T. Bocklitz. Explainable artificial intelligence for spectroscopy data: a review. *Pflügers Archiv - European Journal of Physiology*, 477(4):603–615, Apr. 2025. ISSN 0031-6768, 1432-2013. doi: 10.1007/s00424-024-02997-y.

J. Ding, Q. Lin, J. Zhang, G. M. Young, C. Jiang, Y. Zhong, and J. Zhang. Rapid identification of pathogens by using surface-enhanced Raman spectroscopy and multi-scale convolutional neural network. *Analytical and Bioanalytical Chemistry*, 413(14):3801–3811, June 2021. ISSN 1618-2642, 1618-2650. doi: 10.1007/s00216-021-03332-5.

L. Farah, J. M. Murris, I. Borget, A. Guilloux, N. M. Martelli, and S. I. Katsahian. Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know. *Mayo Clinic Proceedings: Digital Health*, 1(2):120–138, June 2023. ISSN 29497612. doi: 10.1016/j.mcpdig.2023.02.004.

R. Goodacre. Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrational Spectroscopy*, 32(1):33–45, Aug. 2003. ISSN 09242031. doi: 10.1016/S0924-2031(03)00045-6.

R. Haghi, E. Pérez-Fernández, and A. Robertson. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma*, 396:115071, Aug. 2021. ISSN 00167061. doi: 10.1016/j.geoderma.2021.115071.

D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-Supervised Learning with Deep Generative Models. (arXiv:1406.5298), Oct. 2014. doi: 10.48550/arXiv.1406.5298.

H. Leng, C. Chen, C. Chen, F. Chen, Z. Du, J. Chen, B. Yang, E. Zuo, M. Xiao, X. Lv, and P. Liu. Raman spectroscopy and FTIR spectroscopy fusion technology combined with deep learning: A novel cancer prediction method. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 285:121839, Jan. 2023. ISSN 13861425. doi: 10.1016/j.saa.2022.121839.

J. Luo, W. Zhao, F. Ouyang, K. Sheng, and S. Wang. Research on High-Precision Gas Concentration Inversion for Imaging Fourier Transform Spectroscopy Based on Multi-Scale Feature Attention Model. *Applied Sciences*, 15(5):2438, Feb. 2025. ISSN 2076-3417. doi: 10.3390/app15052438.

J. Nallala, G. R. Lloyd, N. Shepherd, and N. Stone. High-resolution FTIR imaging of colon tissues for elucidation of individual cellular and histopathological features. *Analyst*, 141(2):630–639, Jan. 2016. ISSN 1364-5528. doi: 10.1039/C5AN01871D.

M. T. Ribeiro, S. Singh, and C. Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. (arXiv:1602.04938), Aug. 2016. doi: 10.48550/arXiv.1602.04938.

W. Shuai, X. Wu, C. Chen, E. Zuo, X. Chen, Z. Li, X. Lv, L. Wu, and C. Chen. Rapid diagnosis of rheumatoid arthritis and ankylosing spondylitis based on Fourier transform infrared spectroscopy and deep learning. *Photodiagnosis and Photodynamic Therapy*, 45:103885, Feb. 2024. ISSN 15721000. doi: 10.1016/j.pdpdt.2023.103885.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. Feb. 2020. doi: 10.48550/arXiv.1911.02508.

J.-W. Tang, J.-W. Lyu, J.-X. Lai, X.-D. Zhang, Y.-G. Du, X.-Q. Zhang, Y.-D. Zhang, B. Gu, X. Zhang, B. Gu, and L. Wang. Determination of Shigella spp. via label-free SERS spectra coupled with deep learning. *Microchemical Journal*, 189:108539, June 2023. ISSN 0026265X. doi: 10.1016/j.microc.2023.108539.

M. Wolpert and P. Hellwig. Infrared spectra and molar absorption coefficients of the 20 alpha amino acids in aqueous solutions in the spectral range from 1800 to 500cm^{-1}. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 64(4):987–1001, July 2006. ISSN 13861425. doi: 10.1016/j.saa.2005.08.025.

G. Yu, B. Ma, J. Chen, X. Li, Y. Li, and C. Li. Nondestructive identification of pesticide residues on the Hami melon surface using deep feature fusion by Vis/NIR spectroscopy and 1D-CNN. *Journal of Food Process Engineering*, 44(1):e13602, Jan. 2021. ISSN 0145-8876, 1745-4530. doi: 10.1111/jfpe.13602.

J. K. Zaki, J. Tomasik, J. A. McCune, S. Bahn, P. Liò, and O. A. Scherman. Explainable Deep Learning Framework for SERS Bio-quantification. (arXiv:2411.08082), Nov. 2024. doi: 10.48550/arXiv.2411.08082.

F. Zhang, A. Cannone Falchetto, D. Wang, Z. Li, Y. Sun, and W. Lin. Prediction of asphalt rheological properties for paving and maintenance assistance using explainable machine learning. *Fuel*, 396:135319, Sept. 2025. ISSN 00162361. doi: 10.1016/j.fuel.2025.135319.

W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin. A Review of Machine Learning for Near-Infrared Spectroscopy. *Sensors*, 22(24):9764, Dec. 2022. ISSN 1424-8220. doi: 10.3390/s22249764.