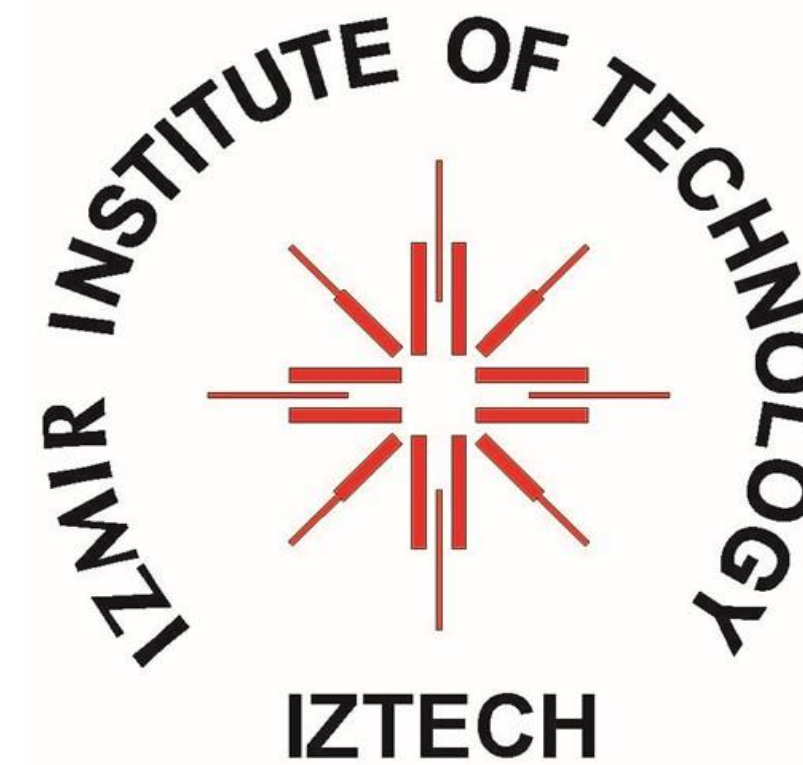# DECODING HONEY ORIGINS: ML AND DL METHODS FOR BOTANICAL CLASSIFICATION

Nathan Mekuria SOLOMON, Cansu TURAN, Durmuş ÖZDEMİR

İzmir Institute of Technology, Department of Chemistry , Gülbahçe-URLA 35430 / İZMİR

INTRODUCTION

Honey composition varies with its floral origin, influencing flavor, color, aroma, and nutritional value. Accurate identification of botanical origin is essential for quality control, authenticity verification, traceability in the supply chain, and protection against adulteration.

Traditional classification methods, such as pollen analysis (melissopalynology), remain the gold standard but are labor-intensive, require specialized expertise, and may take days to weeks per sample [1,2]. These limitations hinder large-scale and routine testing in the honey industry.

Fourier-transform infrared (FTIR) spectroscopy provides a rapid, non-destructive alternative by capturing the unique chemical "fingerprint" of honey with minimal sample preparation. Recent studies have demonstrated that FTIR, combined with chemometrics and machine learning, can effectively classify honey according to botanical origin [3]. Building on this, our study integrates FTIR with modern Machine Learning (ML) and Deep Learning (DL) methods to achieve scalable, reproducible, and automated classification of honey floral origins.

OBJECTIVES

- Apply ML and DL approaches to classify honey into five botanical classes.
- Compare classification performance across algorithms.
- Explore class separability using dimensionality reduction and visualization.
- Identify key spectral bands contributing to classification.
- Provide a framework for rapid, accurate, and reproducible botanical classification.

METHODS

- Honey samples from five botanical origins were analyzed using FTIR spectroscopy. Spectra were normalized and used as input features. Model training and evaluation employed 5-fold cross-validation.
- Classical machine learning algorithms (SVM, Decision Tree, Ensembles) and a 1D CNN were applied. Random Forest feature selection identified the most discriminative spectral bands.
- A deep autoencoder reduced dimensionality, and t-SNE visualized class separability.
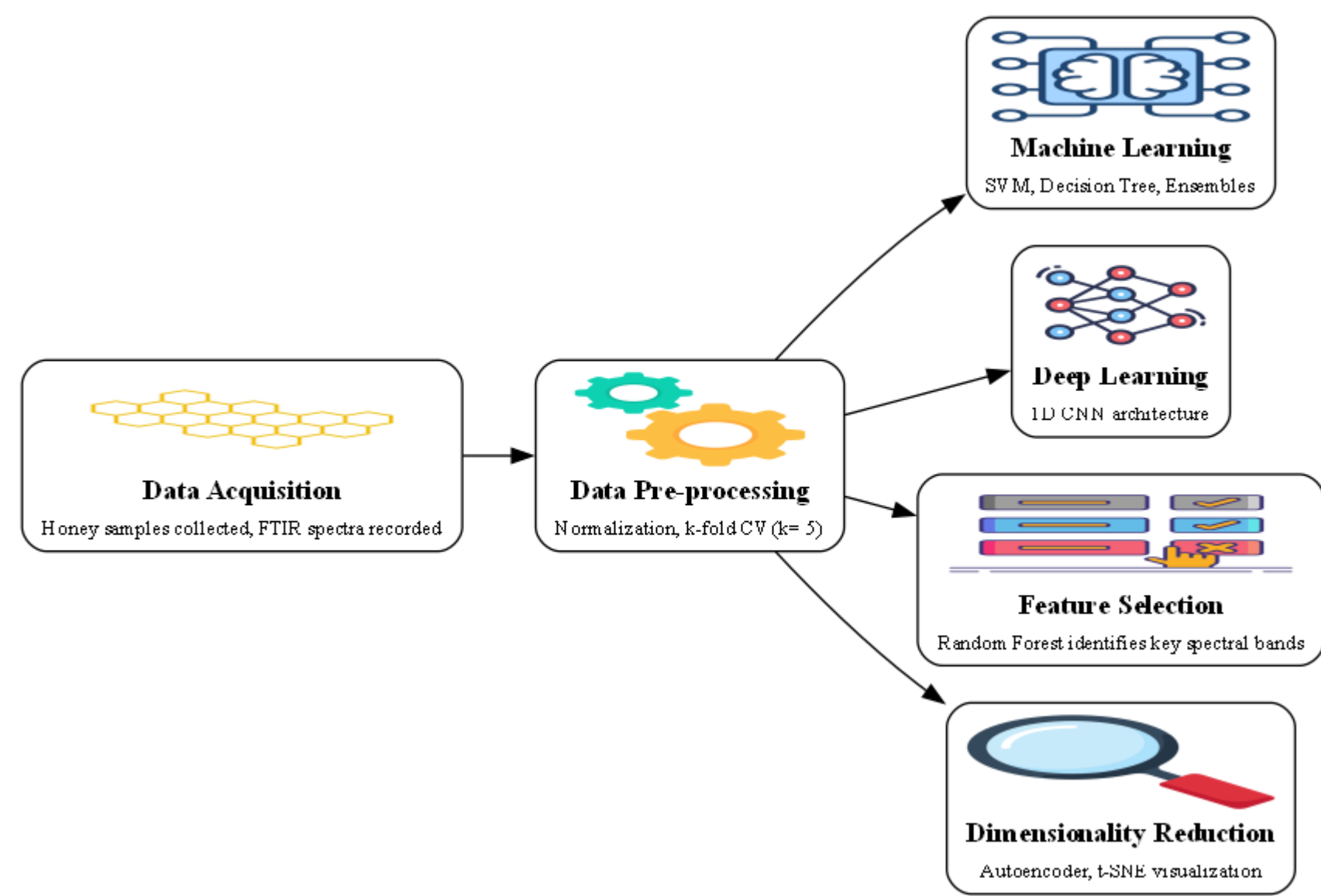- All analyses were performed in Python using scikit-learn and TensorFlow/Keras.

Figure 1. Workflow of FTIR-based Honey Classification Using Machine Learning and CNN

RESULTS

Dataset & Samples

Table 1. Sample Distribution by Botanical Class

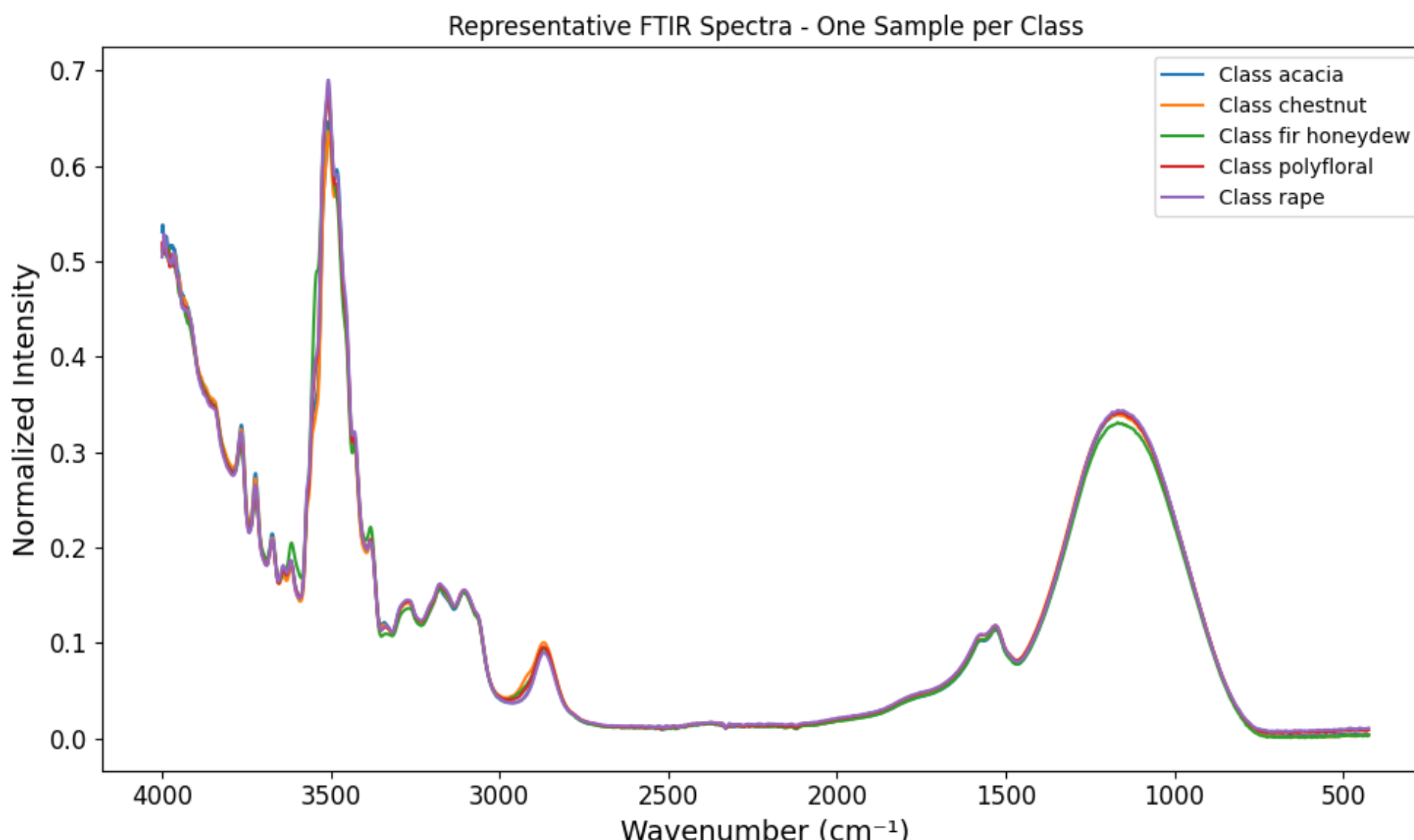| Botanic Name	Number of Samples
polyfloral	121
fir honeydew	50
chestnut	31
acacia	26
rape	24
Total	252

Figure 2. Representative FTIR Spectra of Honey Samples

Model Performance

Figure 3. Classification Performance of ML and DL Models

Feature Importance & Interpretable ML

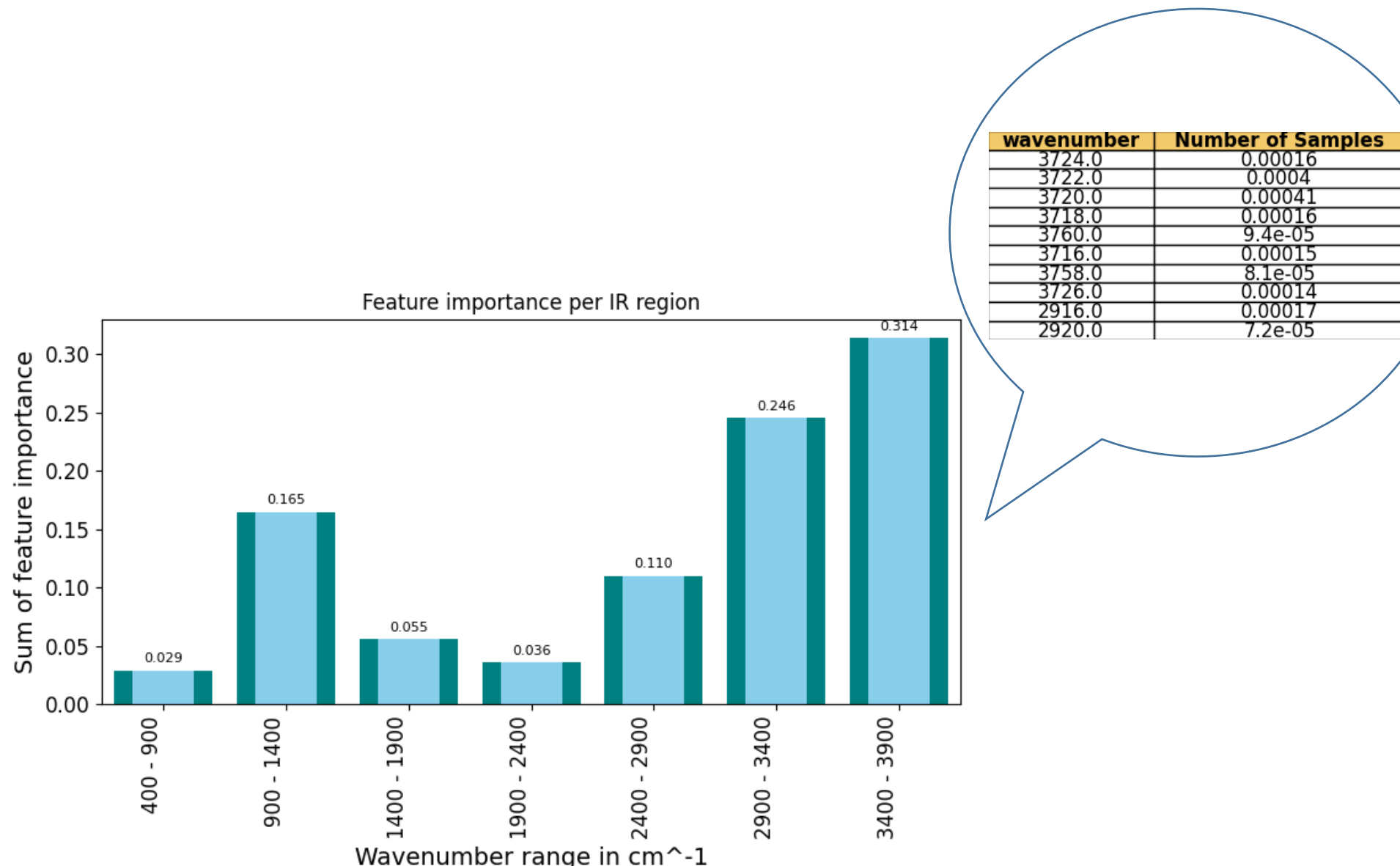| wavenumber	Number of Samples
3724.0	0.00016
3722.0	0.0004
3720.0	0.00041
3718.0	0.00016
3760.0	9.4e-05
3716.0	0.00015
3758.0	8.1e-05
3926.0	0.00014
2916.0	0.00017
2920.0	7.2e-05

Figure 3. Importance of FTIR spectral bands in classifying honey botanical origin using Random Forest. The top-10 most important wavenumbers with their corresponding importance values are shown in the inset table next to the bar graph.
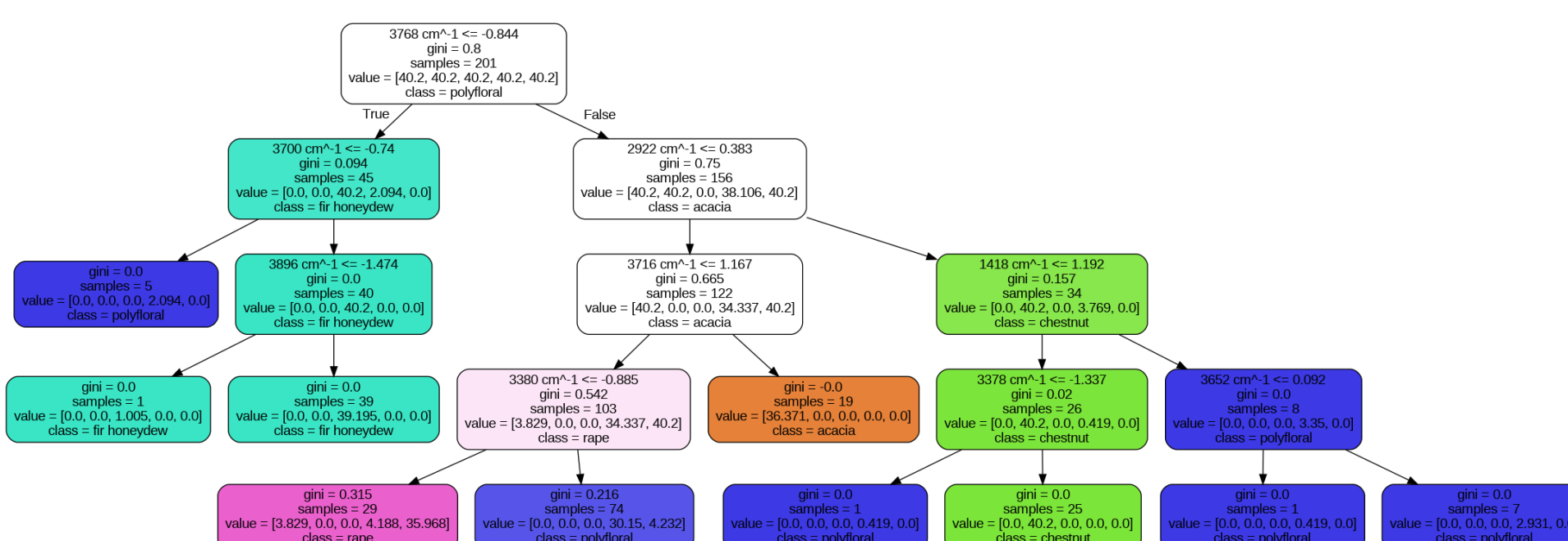

Figure 4: Interpretable tree showing feature-driven classification
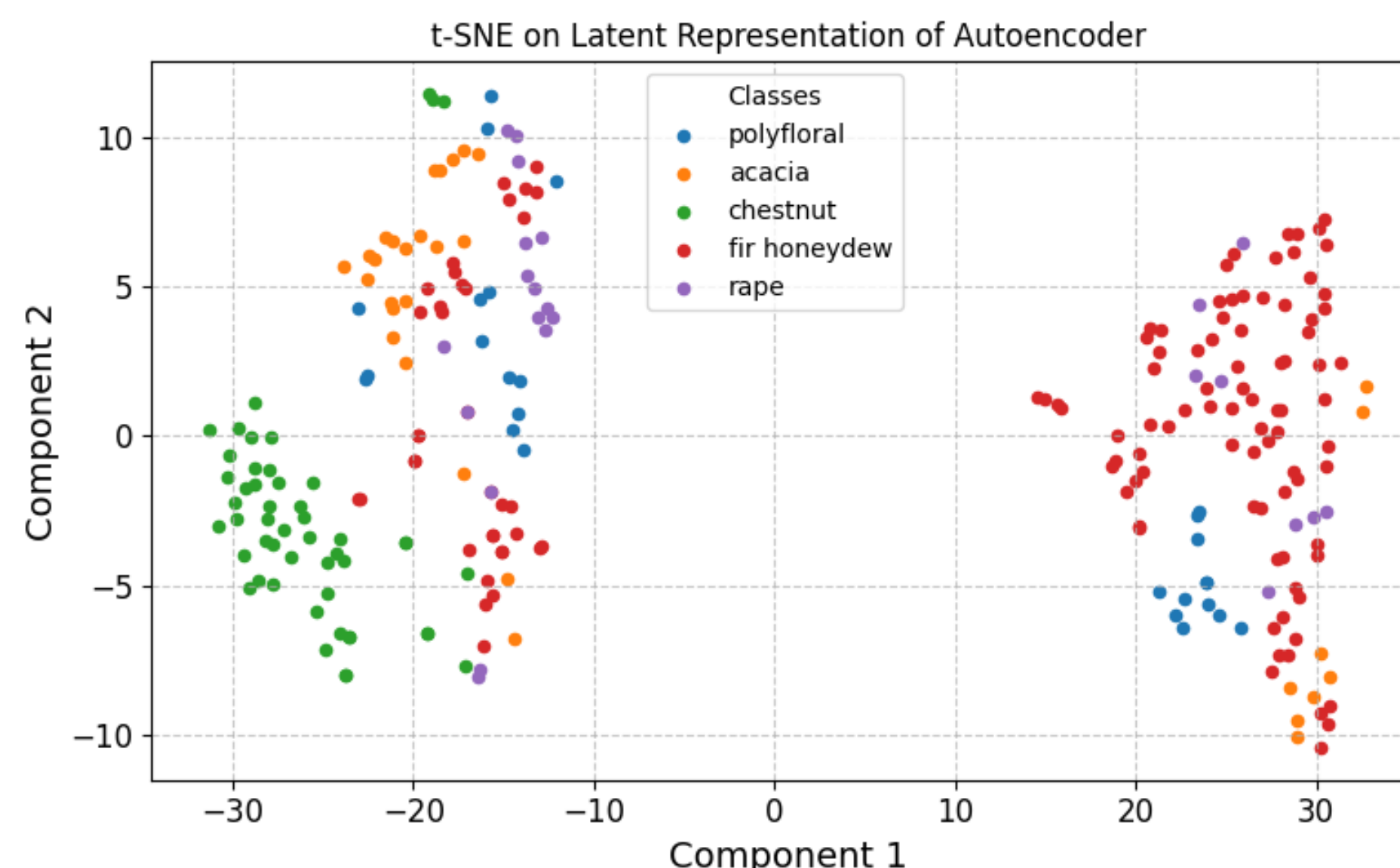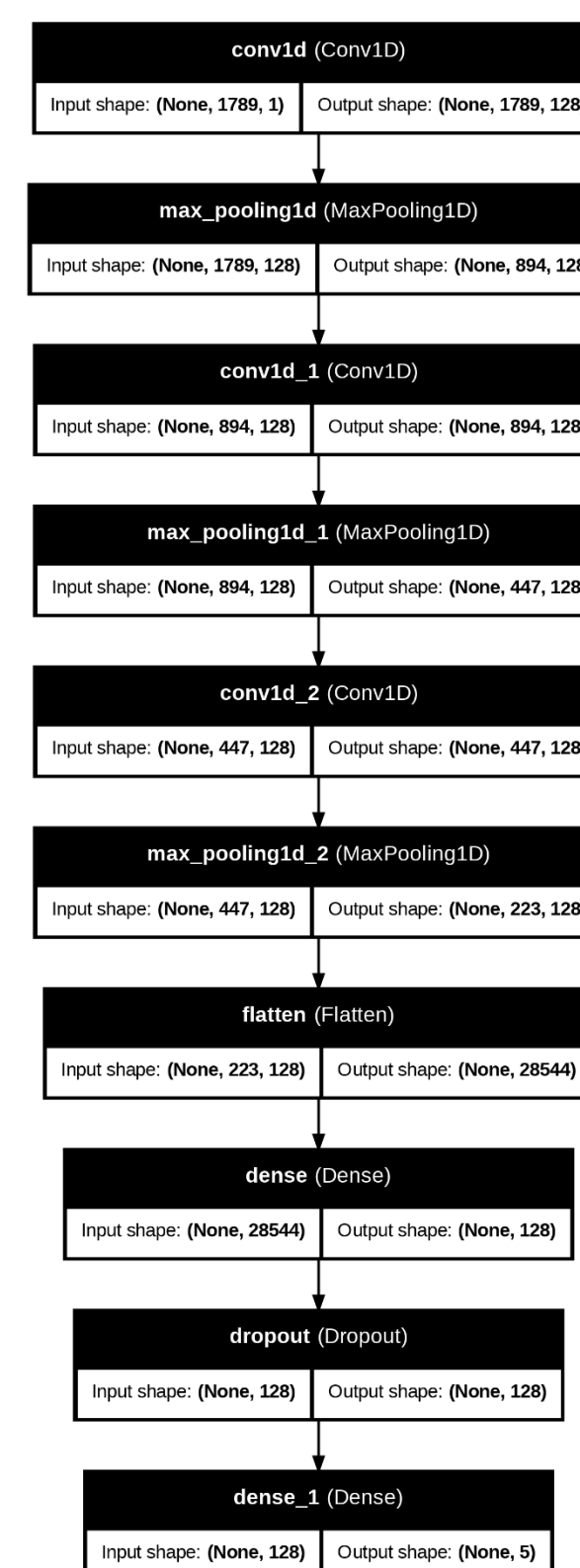
Dimensionality Reduction

Figure 5. t-SNE Visualization of Honey Samples

Deep Learning Architecture
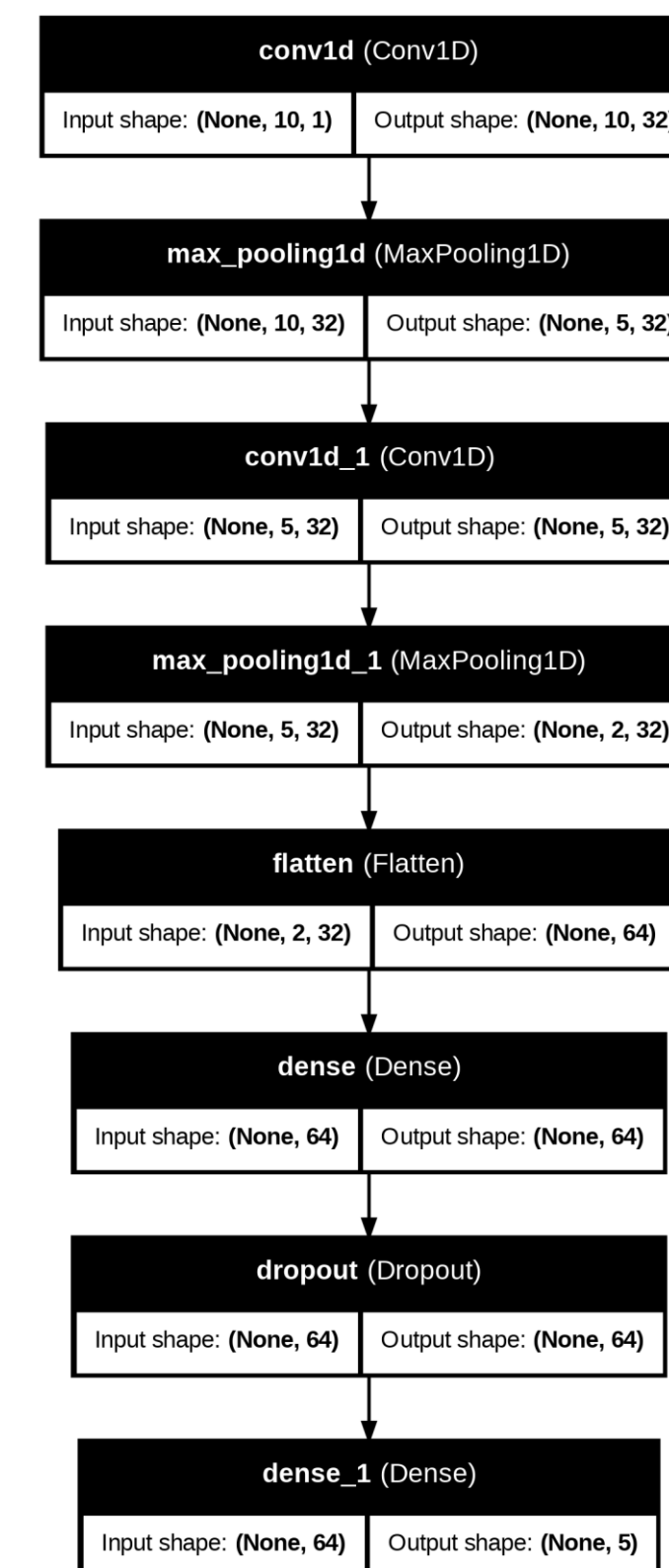
Original Data

PCA-Reduced Data

Figure 6: 1D CNN architectures trained on original spectra (left) and PCA-reduced spectra (right) for honey classification. PCA reduction slightly lowers dimensionality while maintaining high performance.

DISCUSSION AND CONCLUSION

Performance:
- Classical models (SVM/XGBoost 94.1%, Logistic/Voting 96.1%) performed well, but CNN reached 100% (raw) and 96.1% (PCA), showing stronger generalization.

Feature importance:
- Top bands (e.g., 3720–3760, 2916–2920 cm^{-1}) had very low individual weights, while summed regions (900–1400, 2900–3900 cm^{-1}) drove classification → supports distributed spectral contribution rather than single peaks.

Interpretability:
- Decision Tree (86.3%) revealed useful cut-points but underperformed, reflecting loss of broad spectral information.

Latent structure:
- t-SNE on autoencoder embeddings showed class clustering with some overlap, confirming that deep models extract discriminative, though not perfectly separable, patterns.

Conclusion:
- Spectroscopy + AI enables rapid, non-destructive honey authentication. Deep learning provides the most robust features, while interpretable ML highlights relevant spectral regions.

REFERENCES

[1] Persano Oddo, L. et al. (2004). Food Chemistry, 85, 305–312.
[2] Anupama, D. et al. (2019). J. Apic. Res., 58, 555–563.
[3] Ruoff, K. & Bogdanov, S. (2004). Apidologie, 35, S57–S65.

ACKNOWLEDGEMENT