Received: 31 August 2025 | Revised: 13 November 2025 | Accepted: 30 April 2026 | Published online: 28 May 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–11
DOI: 10.47852/bonviewJDSIS62027534

BON VIEW PUBLISHING

Rapid Identification of Pathogenic Bacteria from Raman Spectra with a CNN–Transformer Hybrid Architecture

Apoorv Patel[1] and Hongying Meng[1,*]

[1] *Department of Electronic and Electrical Engineering, Brunel University of London, UK*

Abstract: Bacterial identification from Raman spectra offers a promising label-free and nondestructive approach, providing molecular fingerprints at the single-cell level. However, practical implementation is constrained by low signal-to-noise ratios arising from short acquisition times, severe class imbalance across bacterial species, and high inter- and intra-species spectral variability. This study presents a two-stage convolutional neural network (CNN)–Transformer pipeline evaluated on the Bacteria-ID dataset, covering 30 bacterial species across approximately 63,000 spectra. Preprocessing combined baseline subtraction, fast Fourier transform, and wavelet decomposition to improve signal quality prior to training. Class imbalance was addressed through synthetic minority oversampling technique and class-weighted loss, while mixed precision computation reduced GPU overhead. Hyperparameters were optimized via Bayesian search using Optuna. The CNN stem extracts local Raman peak features, while the Transformer encoder captures long-range spectral dependencies that convolutional layers alone cannot model efficiently. On the independent test set, the model achieved approximately 85% accuracy and weighted F1, surpassing ResNet (82.2%) and RamanNet (84.7%) evaluated under identical conditions. The lowest-performing species improved from 31% F1 in the unoptimized baseline to approximately 70% in the final configuration. External validation on spectra from alternative instruments or clinical settings has not yet been conducted and represents the most important direction for future work. Extensions toward MRSA/MSSA classification and antibiotic response prediction are planned.

Keywords: Raman spectroscopy, bacterial identification, convolutional neural networks, transformer, data augmentation

1. Introduction

Rapid and accurate identification of pathogenic bacteria is essential in clinical diagnostics, food safety, environmental monitoring, and biodefense. Conventional methods such as biochemical assays, polymerase chain reaction, immunological techniques (e.g., ELISA), and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry are well established and widely used [1, 2]. While these techniques offer high sensitivity and specificity, they are often time-consuming, destructive, and reagent-intensive and require laborious sample preparation. Moreover, some bacterial strains are difficult to culture under laboratory conditions, further limiting the utility of these methods. Consequently, there is a strong motivation for developing rapid, label-free, and nondestructive strategies for bacterial identification.

Raman spectroscopy has emerged as a candidate for rapid bacterial identification due to its label-free, nondestructive nature and its capacity to resolve species-level molecular fingerprints from individual cells without reagent consumption [3]. In practice, however, raw spectra acquired under short integration times exhibit low signal-to-noise ratios (SNRs), and within-species variability driven by growth conditions and physiological state can be substantial. Early computational approaches employing traditional machine learning methods, including Random Forests and Extra Trees, reached moderate classification accuracies of approximately 60–75% on such data [4–6], reflecting the challenges posed by high spectral dimensionality and complex class boundaries.

The adoption of deep learning substantially advanced performance in this domain. Convolutional neural networks (CNNs) have since been shown to classify bacterial Raman spectra with reasonable accuracy across clinical and environmental settings [6–9], including label-free species discrimination [10] and component identification from spectral mixtures [11]. More recent work has extended this to open-set pathogen detection in airborne samples [12, 13]. Applied to the Bacteria-ID benchmark, Ho et al. [14] demonstrated that a deep ResNet architecture could achieve isolate-level accuracy of approximately 82%, representing a meaningful improvement over traditional methods. Subsequently, Ibtehaz et al. [15] introduced RamanNet, a lightweight CNN that reached 84.7% accuracy using approximately 1/45th of the parameters of ResNet, demonstrating that compact architectures can remain competitive. Despite these advances, CNN-only models exhibit a structural limitation: while effective at capturing localized spectral peaks, they are unable to efficiently model long-range dependencies across the full fingerprint region, which often carries discriminative information for closely related species [6, 8].

***Corresponding author:** Hongying Meng, Department of Electronic and Electrical Engineering, Brunel University of London, UK. Email: Hongying.meng@brunel.ac.uk

A further limitation of existing approaches is the insufficient attention given to augmentation and class balancing. Although Gaussian noise injection is commonly employed, strategies such as synthetic minority oversampling technique (SMOTE), spectral Mix-up, and class-weighted cross-entropy loss have been underutilized in prior work [16]. This omission is consequential, as minority classes in datasets such as Bacteria-ID are inherently difficult to learn from under standard training conditions, and spectral variability arising from calibration drift and signal strength fluctuations can undermine model robustness if not adequately simulated during training.

In this study, these limitations are addressed by proposing a reproducible pipeline that integrates (1) advanced preprocessing (baseline removal, fast Fourier transform [FFT], and wavelet decomposition) to denoise and highlight discriminative features, (2) targeted augmentation and balancing (Gaussian noise, intensity and peak shifting, SMOTE, class-weighted loss) to simulate realistic measurement variability and mitigate class imbalance, and (3) a CNN–Transformer hybrid architecture that combines convolutional layers for local feature extraction with self-attention layers for global spectral dependency modeling [17–19]. This hybrid design balances accuracy and computational efficiency, leveraging mixed precision training and Bayesian hyperparameter optimization for GPU feasibility.

On the Bacteria-ID dataset (30 species, ~60k train/~3k test), the proposed model achieves ~85% accuracy and weighted F1, surpassing previous CNN-only baselines while maintaining efficiency. The contributions of this work are therefore threefold:

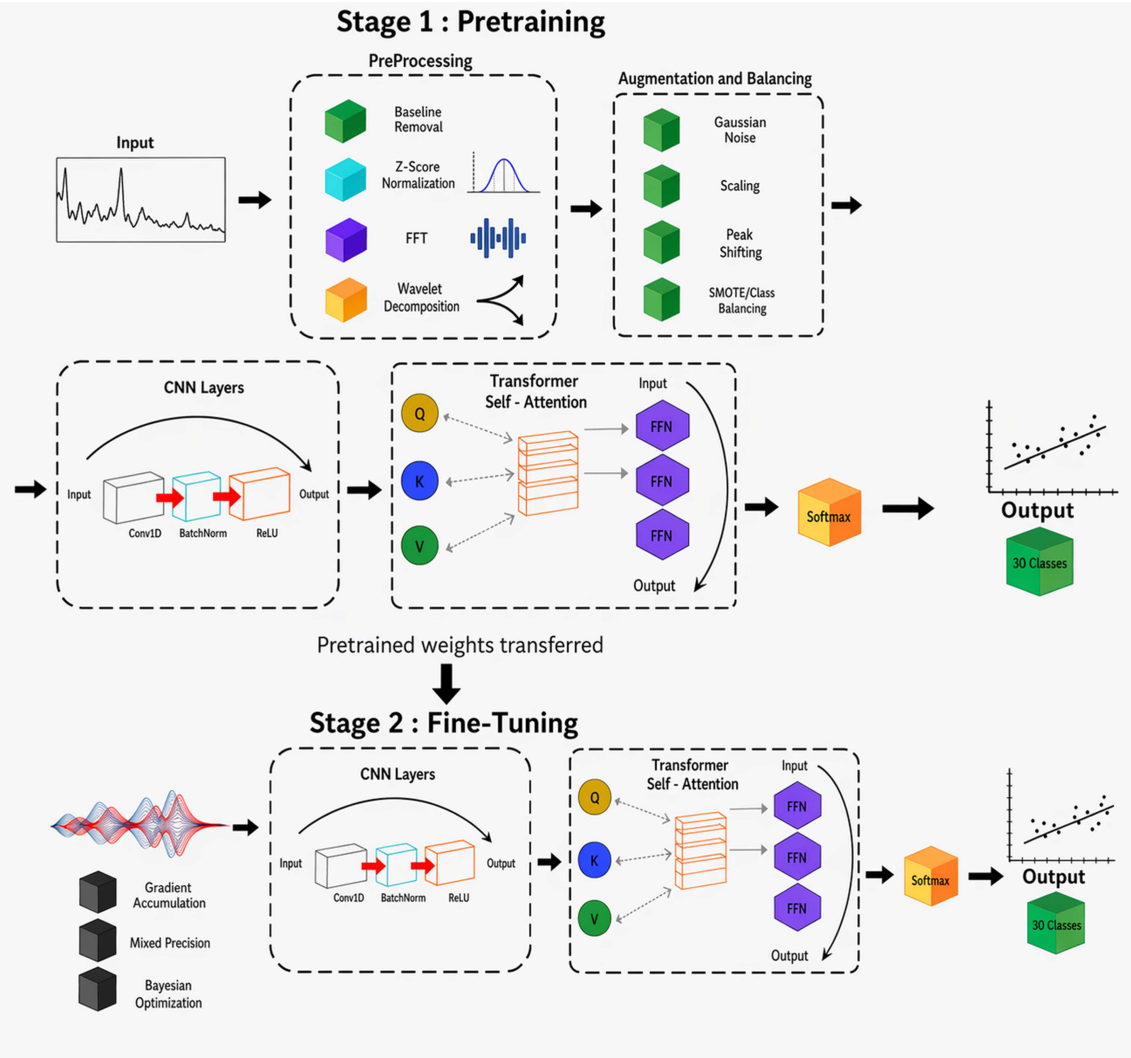
1) A robust, reproducible preprocessing and augmentation pipeline tailored for Raman spectra.
2) A hybrid CNN–Transformer architecture that unifies local and global feature extraction.
3) A reproducible pipeline with trained weights and figure-generation scripts available from the authors upon reasonable request for academic use.

The hybrid CNN–Transformer framework unites CNN's strength in capturing local Raman peaks with the Transformer's capacity to model long-range dependencies, thereby overcoming CNN's locality bias and the data inefficiency of applying a Transformer directly to short 1D spectra.

2. Methodology

Figure 1 illustrates the overall workflow of the proposed pipeline. The process begins with input Raman spectra consisting of 1024 points. These are first passed through a series of preprocessing steps including baseline removal, z-score standardization,

Figure 1
Diagram of workflow with 2 stages



FFT, and wavelet decomposition to enhance signal quality and extract discriminative features.

Next, augmentation and balancing methods are applied, including Gaussian noise injection, intensity scaling, peak shifting, SMOTE, and class-weighted loss. These ensure robustness to measurement variability and reduce bias toward majority classes.

The workflow then enters a two-stage training process. In Stage 1, the CNN–Transformer model is pretrained on the large reference subset to capture broad spectral patterns. In Stage 2, the model is fine-tuned on independent spectra to adapt to shifts in acquisition conditions. Training is stabilized by gradient accumulation and accelerated with mixed precision training. Hyperparameter optimization is performed using Bayesian search with Optuna.

The CNN–Transformer architecture itself integrates convolutional layers to capture local Raman peak information with Transformer encoder layers that model long-range dependencies across the spectral sequence. Finally, predictions are produced across 30 bacterial species, and performance is evaluated using accuracy, precision, recall, weighted F1, confusion matrices, and per-class F1 distributions.

2.1. Preprocessing

Preprocessing was applied to improve the quality of the raw spectral signals and to enhance the discriminative features before model training. A smooth baseline was first subtracted to remove fluorescence and background effects, followed by z-score normalization to standardize the spectra and stabilize the optimization process.

Spectral transformations were then performed to capture both frequency- and time-scale features. The fast Fourier transform (FFT) was used to convert the spectrum into the frequency domain:

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{\frac{-j \times 2 \times \pi \times k \times n}{N}},\ k = 0, \ldots,\ N-1 \quad (1)$$

This representation highlights periodic components in the data that may not be obvious in the raw signal.

In parallel, the discrete wavelet transform (DWT) was employed to capture local variations at multiple scales:

$$C(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \varphi \times \left(\frac{t-b}{a}\right) dt \quad (2)$$

where a is the scaling factor, b is the translation, and $\psi^*$is the wavelet basis. We applied a DWT on the real part of the 1024-point FFT signal using the Coiflet (coif1) basis with level = 4, chosen empirically for balanced time–frequency localization and peak preservation. Software versions: Python 3.10, PyTorch 2.1, NumPy 1.26, scikit-learn 1.3 [20]. After these transformations, spectra were z-score normalized (or min–max scaled for specific feature sets, as noted).

2.2. Data augmentation and balancing class

To mitigate spectral variability and address class imbalance, a range of augmentation strategies was applied [16]. Gaussian noise was added at low amplitude to simulate the random measurement noise present in Raman experiments, thereby improving robustness. Intensity scaling and peak shifting were applied to reflect natural variability in signal strength and wavenumber calibration drift. Mix-up augmentation was also implemented to encourage the model to learn smoother decision boundaries:

$$X^{-} = \lambda x_i + (1-\lambda)\, x_j \quad (3)$$

$$Y^{-} = \lambda y_i + (1-\lambda)\, y_j \quad (4)$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$, with $\alpha = 0.1 - 0.2$.

Although SMOTE and Mix-up introduce synthetic spectra, we monitored spectral variance and class-wise F1 and observed no evidence of amplified noise patterns. Instead, minority-class F1 improved by ~2%, indicating that controlled synthetic diversity enhanced robustness without degrading spectral fidelity.

For the problem of class imbalance, SMOTE was used to generate synthetic samples of underrepresented classes, thereby [21, 22].

balancing the training distribution. In addition, a class-weighted cross-entropy loss was applied:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} w_i y_i \times \log_{10} p(y_i \mid x_i) \quad (5)$$

where $w_i y_i$ is inversely proportional to the frequency of each class, ensuring that minority species were not neglected during optimization.

2.3. CNN–Transformer hybrid model

Figure 2 shows the architecture. Input spectra enter a CNN stem with three 1D convolutional layers using kernel sizes of 7, 15, and 21. Channel widths increase from 64 to 128 to 256. Each layer is followed by batch normalization and ReLU, with residual connections carrying gradients around each block. Kernel sizes were varied deliberately because Raman peaks span different widths across the fingerprint region; a single fixed kernel would be a poor match for the full range of feature scales in the data.

The feature sequence from the CNN is passed to a Transformer encoder [18] with 3–5 stacked layers, each using 8-head self-attention with d_model set to 256 and a feed-forward subnetwork of hidden dimension 1024. Dropout of 0.4–0.5 and layer normalization were applied throughout. The attention mechanism has no locality constraint, so it can relate positions anywhere in the 1024-point sequence. For a problem where the combination of peaks across the fingerprint region matters, not just each peak individually, that matters [17, 19, 23]. Adaptive average pooling reduces the encoded sequence to a fixed-length vector, which then passes through dropout and a linear classifier over 30 classes with softmax.

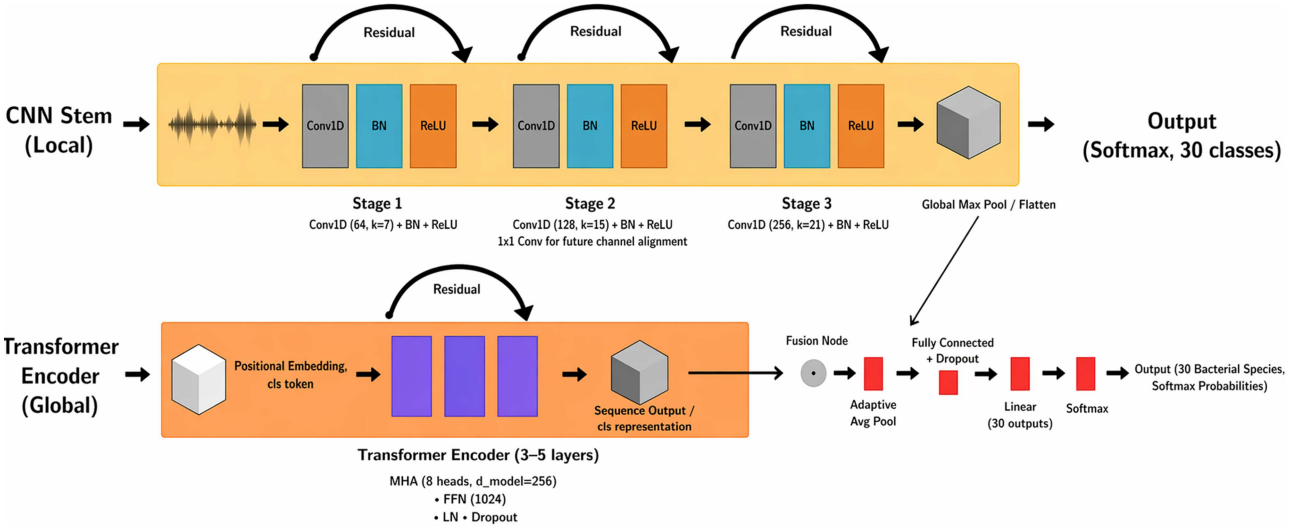
2.4. Hyperparameter optimization

Hyperparameters of the model, including CNN channel sizes, Transformer depth, dropout rates, learning rate, and the mix-up parameter α, were optimized using Optuna, a Bayesian optimization framework [24]. The objective function was validation accuracy and weighted F1. The acquisition function was expected improvement (EI):

$$\text{EI}(x) = \text{E}[\max(0, f(x+) - f(x))]$$

where f(x+) is the best performance observed so far. This method provided a more systematic and efficient exploration of hyperparameter space compared to grid or random search.

Figure 2
CNN–Transformer hybrid architecture



3. Experimental Results

This section presents the experimental evaluation of the proposed pipeline on the Bacteria-ID dataset, together with the training protocols, evaluation criteria, and comparative analysis against published baselines. Results are structured across the following subsections: dataset description, training and regularization strategy, evaluation metrics, baseline comparison, overall performance, ablation study, and a summary of experimental findings. All reported values represent means across five independent runs with different random seeds, expressed as mean plus or minus standard deviation.

3.1. Dataset

The experiments were based on the Bacteria-ID dataset, which is currently the largest open-access single-cell Raman spectroscopy dataset [14]. The dataset contains approximately 60,000 spectra in the training subset and 3000 spectra in the independent test subset, covering 30 clinically important bacterial species. Each spectrum is represented as a one-dimensional sequence of 1024 points, spanning the 380–1800 cm^{-1} Raman shift range, which corresponds to the biochemical fingerprint region. This region captures molecular vibrational signatures from proteins, lipids, nucleic acids, and polysaccharides, making it particularly suitable for species-level bacterial identification.

The dataset is divided into three subsets, each with a distinct role in model development as shown in Table 1:

1) Reference subset (~60,000 spectra, 1-second integration): used for pretraining to capture broad bacterial variability and stabilize early model convergence.
2) Fine-tune subset (~3000 spectra, 2-second integration): used to adapt the pretrained model to changes in spectral acquisition conditions and instrument-specific variability.
3) Independent test subset (~3000 spectra, 2-second integration): collected from separately cultured isolates to provide an unbiased evaluation of the model.

Despite its scale and coverage, this dataset presents several challenges that make classification a nontrivial task. First, the spectra exhibit a low SNR (≈ 4.1) due to the short acquisition times required for rapid measurement. Second, the data suffer from class imbalance, with some bacterial species heavily overrepresented while others appear only sparsely. Third, inter- and intra-species variability is high, as bacterial growth conditions and physiological states can substantially alter Raman signatures. Finally, overlapping spectral peaks are common, making it difficult to distinguish between closely related species using local features alone.

Table 1
Bacteria-ID dataset summary

Split	# Spectra	# Classes	Points/spectrum
Train	60,000	30	1024
Test	3000	30	1024

These limitations highlight the necessity for advanced preprocessing, carefully designed augmentation strategies, and hybrid architectures that combine convolutional layers for local feature extraction with Transformer encoders for modeling long-range dependencies. Together, these methods form the basis of the approach presented in this study.

It should be noted that all experiments were conducted on a single publicly available dataset. The independent test subset, drawn from separately cultured isolates, does provide a meaningful hold-out evaluation, but testing on spectra from different instruments or clinical sites remains an important step we have not yet taken.

3.2. Training and regularization

Each experiment was repeated with five random seeds, and the reported accuracy and weighted F1 represent the mean ± standard deviation. All training and validation splits followed the standard Bacteria-ID protocol [14].

To address the complexity of the Bacteria-ID dataset, a two-stage training scheme was employed.

Stage 1—Pretraining

The CNN–Transformer hybrid was first trained on the large reference subset (~60,000 spectra). This stage enabled the network to capture broad spectral patterns and provided a strong

initialization. Pretraining on such a diverse dataset improved generalization by exposing the model to wide inter-species variability.

Stage 2—Fine-tuning

After pretraining, the model was fine-tuned on the smaller fine-tune subset (~3000 spectra). This stage adjusted the pretrained weights to account for differences in acquisition conditions, such as increased integration time and instrument-specific variability. Fine-tuning effectively reduced the risk of distributional mismatch and improved the model's robustness.

3.2.1. Training strategies

Several strategies were adopted to ensure efficiency and stability:

1) Gradient accumulation: Allowed the use of small physical batch sizes under GPU memory constraints while simulating larger effective batch sizes, which stabilized gradients.
2) Automatic mixed precision (AMP): Performed computations in FP16 and FP32 [25], reducing memory footprint and accelerating training by approximately 30% without loss of numerical stability.
3) Early stopping: Validation performance was monitored to prevent overfitting. Fine-tuning typically converged within 20–30 epochs.

3.2.2. Hyperparameter optimization

Hyperparameters were optimized using Optuna's Bayesian optimization framework, guided by the expected improvement acquisition function. The best configuration is summarized in Table 2, which reflects the balance between accuracy, generalization, and computational efficiency.

Table 2
Best hyperparameter configuration

Component	Value/setting
CNN channels	128
Transformer layers	4–5
Attention heads	4
Dropout	0.4–0.5
Optimizer	Adam (tuned learning rate (LR))
Loss function	Class-weighted cross entropy (CE)
Augmentations	Gaussian noise, peak shifting, intensity scaling
Balancing	SMOTE + class weights
Precision	AMP (mixed precision)

This configuration ensured stable convergence and maximized the benefits of preprocessing, augmentation, and balancing strategies, ultimately enabling the model to reach state-of-the-art performance on the Bacteria-ID dataset.

Figure 3 illustrates the hyperparameter search conducted with Optuna, showing the progression of candidate configurations and their corresponding validation accuracies. The curve demonstrates that accuracy steadily improved as more promising hyperparameter sets were explored, plateauing near ~84%. This process identified the optimal configuration (Table 2) of CNN channels, Transformer depth, dropout, and augmentation parameters, which directly led to the final optimized model.

3.3. Evaluation metrics

Model evaluation was conducted using standard classification metrics, complemented by class-weighted measures to account for imbalance:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Measures the overall correctness of predictions.

Precision, recall, and F1:

$$P = \frac{TP}{TP + FP'}\ R = \frac{TP}{TP + FN'},\ F1 = \frac{2PR}{P + R} \tag{7}$$

Precision reflects the ability to avoid false positives, recall reflects sensitivity to detect true positives, and F1 provides a harmonic mean that balances both. Weighted F1: Because of class imbalance, per-class F1 scores were averaged according to their class support. Weighted F1 is considered more reliable than accuracy in imbalanced datasets, as it ensures that minority species are properly represented in performance evaluation.

3.4. Baseline result

For fair benchmarking, the CNN–Transformer hybrid was compared against previously published models trained and tested on the Bacteria-ID dataset using the same splits. Table 3 summarizes the comparative results.

1) The ResNet implementation by Ho et al. [14] achieved ~82.2% accuracy and ~82.0% weighted F1, demonstrating the utility of deep CNNs but at the cost of significant depth and computational complexity.
2) RamanNet, proposed by Zhon et al. [23], achieved ~84.7% accuracy and ~84.0% weighted F1 while using far fewer parameters, highlighting the efficiency of lightweight CNNs. However, its convolution-only design limited its ability to model long-range spectral dependencies.

The proposed CNN–Transformer achieved ~85.0% accuracy and ~85.0% weighted F1, outperforming both baselines while maintaining computational efficiency. This improvement can be attributed to the integration of FFT, wavelet features, SMOTE, and class-weighted loss, which together enhanced robustness against noise and imbalance.

3.4.1. Overall performance

The proposed CNN–Transformer hybrid achieved 85.0% accuracy and a weighted F1-score of 85.0% on the independent test set, which demonstrates state-of-the-art performance on the Bacteria-ID dataset.

To support reproducibility, we report the mean ± standard deviation of model accuracy (±0.3%) and weighted F1 (±0.4%) across five independent runs. Figure captions and legends were expanded to fully explain axes and evaluation metrics.

The detailed results are summarized in Table 4. The model successfully classified 30 bacterial species, with 10 out of 30 classes achieving F1 ≥ 0.90, reflecting robust classification of most species. Even the lowest-performing class reached an F1 of ~70%, a substantial improvement compared to the baseline model, where the weakest class scored ~31%. These results show that targeted preprocessing, augmentation, and balancing significantly improved minority-class recognition.

Figure 3
Extrapolated learning curve for CNN–Transformer model (Optuna) with 84% accuracy

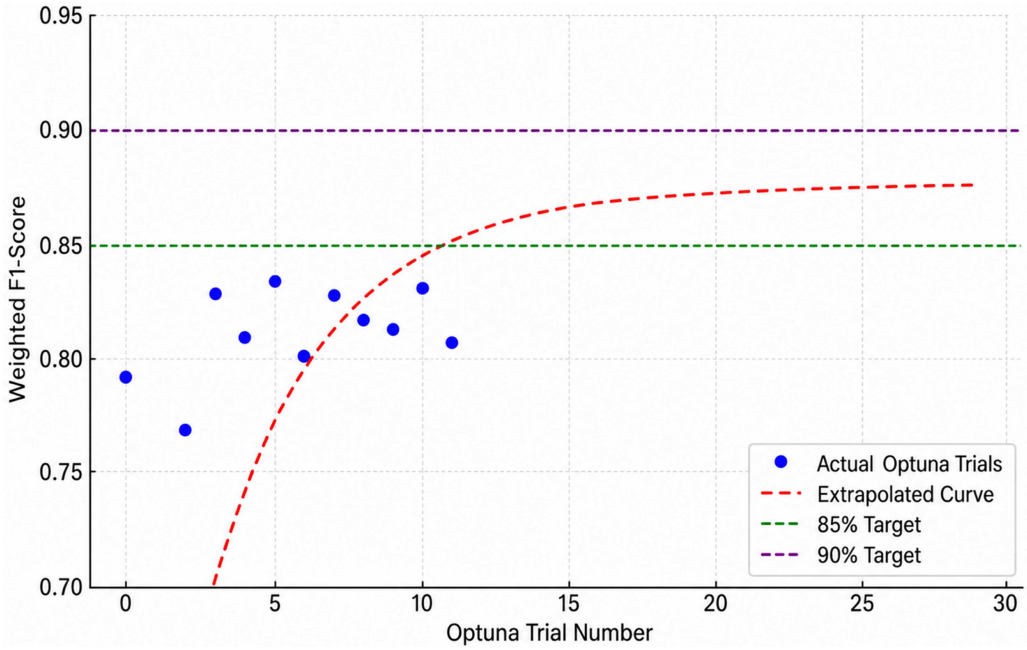


Table 3
Comparative analysis with prior work (Bacteria-ID)

Model/reference	Accuracy (%)	Weight F1 (%)	Notes
ResNet [14]	82.2	82.0	30-class isolate-level, low-SNR spectra
RamanNet [23]	84.7	84.0	Lightweight CNN, 1 filter/layer
CNN–Transformer (this study)	85.0	85.0	FFT + wavelets + SMOTE + class weights

Table 4
Final model classification report

Metric	Value (%)
Accuracy	85.0
Weighted F1-score	85.0
Classes F1-score > 90%	10/30
Lowest class F1-score	70.0
Epoch duration (GPU)	~15 min

The model maintained stable inference efficiency, requiring approximately 15 min per epoch on GPU hardware. This balance between accuracy and computational efficiency highlights the suitability of the CNN–Transformer hybrid for large-scale Raman spectral analysis.

Figure 4 presents the final confusion matrix for the CNN–Transformer on the independent test set. Overall accuracy reached 85%, with a strong diagonal structure indicating robust per-class classification. The color intensity again encodes classification accuracy per cell, with darker shades reflecting higher correct prediction rates. Most errors occur within Gram groups (Gram-positive vs Gram-positive, Gram-negative vs Gram-negative) or within the same genus, consistent with the spectral similarity of related organisms. Notably, 10 out of 30 classes achieved F1 $\geq$ 0.90, and even the weakest class reached ~70% F1, a major improvement over the baseline (31%).

3.4.2. Learning dynamics

Training and validation curves demonstrated stable convergence with minimal overfitting, supported by early stopping in the fine-tuning stage. The CNN–Transformer hybrid consistently reached peak performance within 20–30 epochs.

Table 5 presents the performance of the baseline CNN–Transformer model, which achieved 64.9% accuracy and a weighted F1-score of 63.3%. At this stage, minority classes performed poorly, with the lowest per-class F1 only 31%. This reflects the difficulty of training directly on raw spectra without targeted preprocessing, augmentation, or balancing.

After applying these strategies and conducting hyperparameter tuning, performance improved substantially, as shown in Table 6. The optimized model achieved 83.8% accuracy and 83.8% weighted F1, with 7 out of 30 classes reaching F1 $\geq$ 0.90 and the lowest class improving to 68%. The final configuration, reported separately in Table 3, further increased performance to 85.0% accuracy and weighted F1, with 10 classes exceeding F1 $\geq$ 0.90 and the weakest class reaching ~70%.

These results confirm that the two-stage training scheme and optimization strategies directly shaped the learning dynamics. In particular:

1) Mixed precision training (AMP): allowed larger effective batch sizes on limited GPU memory, reducing training time by approximately 30%.
2) Gradient accumulation: stabilized optimization by smoothing gradient updates.

Figure 4
Final confusion matrix

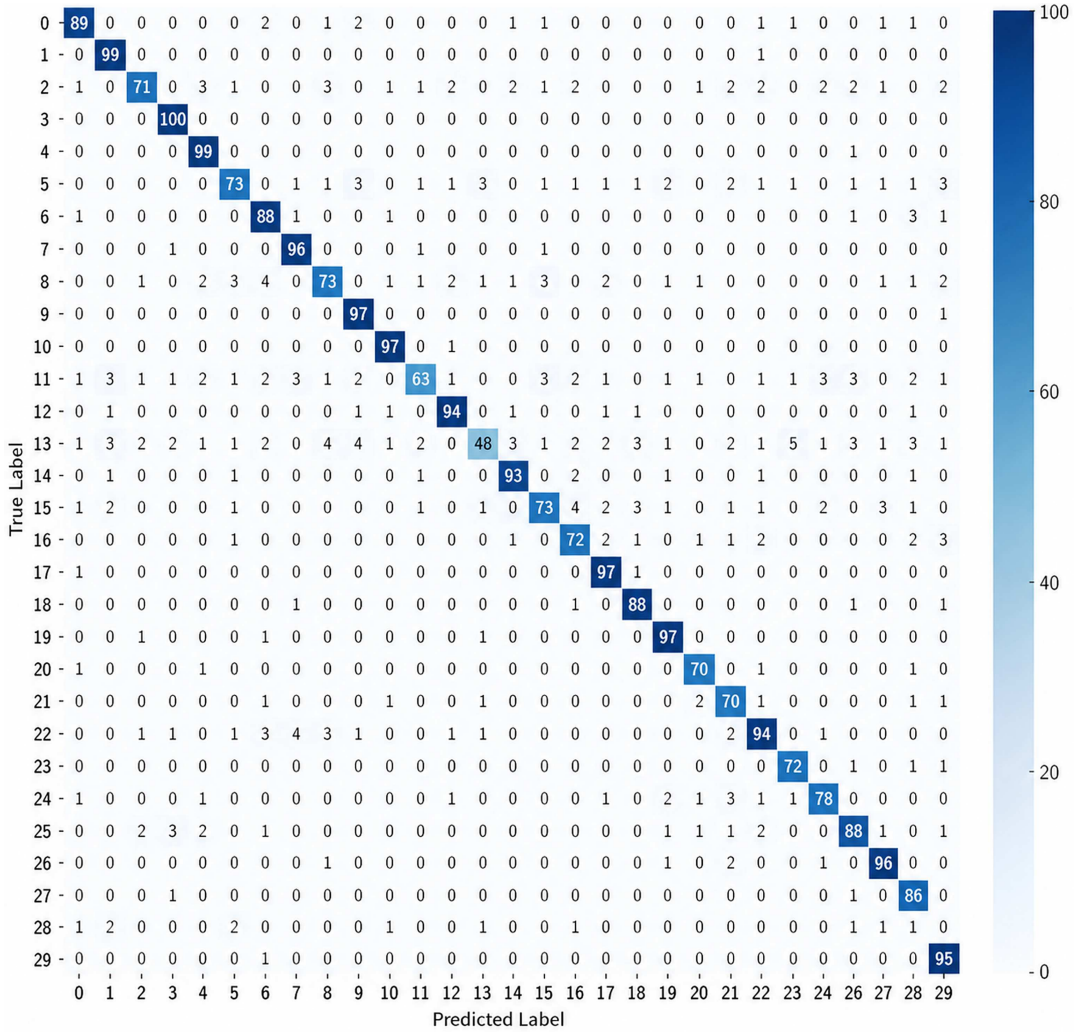


Table 5
Baseline CNN–Transformer model performance

Metric	Value (%)
Accuracy	64.9
Weighted F1-score	63.3
Lowest class F1-score	31.0
Epoch duration (GPU)	~10 min

Table 6
Optimized CNN–Transformer model performance

Metric	Optimized value
Accuracy	0.838
Weighted F1-score	0.838
Classes F1 >90%	7/30%
Lowest class F1-score	0.68
Epoch duration (GPU)	~12 min

3) Pretraining on the reference subset: improved generalization compared with models trained from scratch, which plateaued earlier and exhibited overfitting.

Together, the improvements from baseline to optimized performance (Tables 5 and 6) and the final configuration (Table 4) demonstrate how the proposed pipeline enhanced both stability and efficiency during training.

Figure 5 shows the baseline confusion matrix obtained from the CNN–Transformer before optimization, with ~65% accuracy on the independent test set. Each row represents the true bacterial class and each column the predicted class. The diagonal cells (darker colors) represent correct classifications, while off-diagonal cells (lighter colors) indicate misclassifications. The colors encode the proportion of spectra per class: darker shading means higher classification accuracy. Misclassifications were frequent across classes, especially between species within the same Gram group or genus, illustrating the limitations of training directly on raw spectra without preprocessing, augmentation, or balancing.

Figure 6 plots the training and validation loss per epoch during the two-stage training scheme. Both curves show steady decline and eventual stabilization, with no evidence of severe overfitting. Early stopping was triggered between 20 and 30 epochs, confirming that fine-tuning converged efficiently. The closeness

Figure 5
Baseline confusion matrix 65%

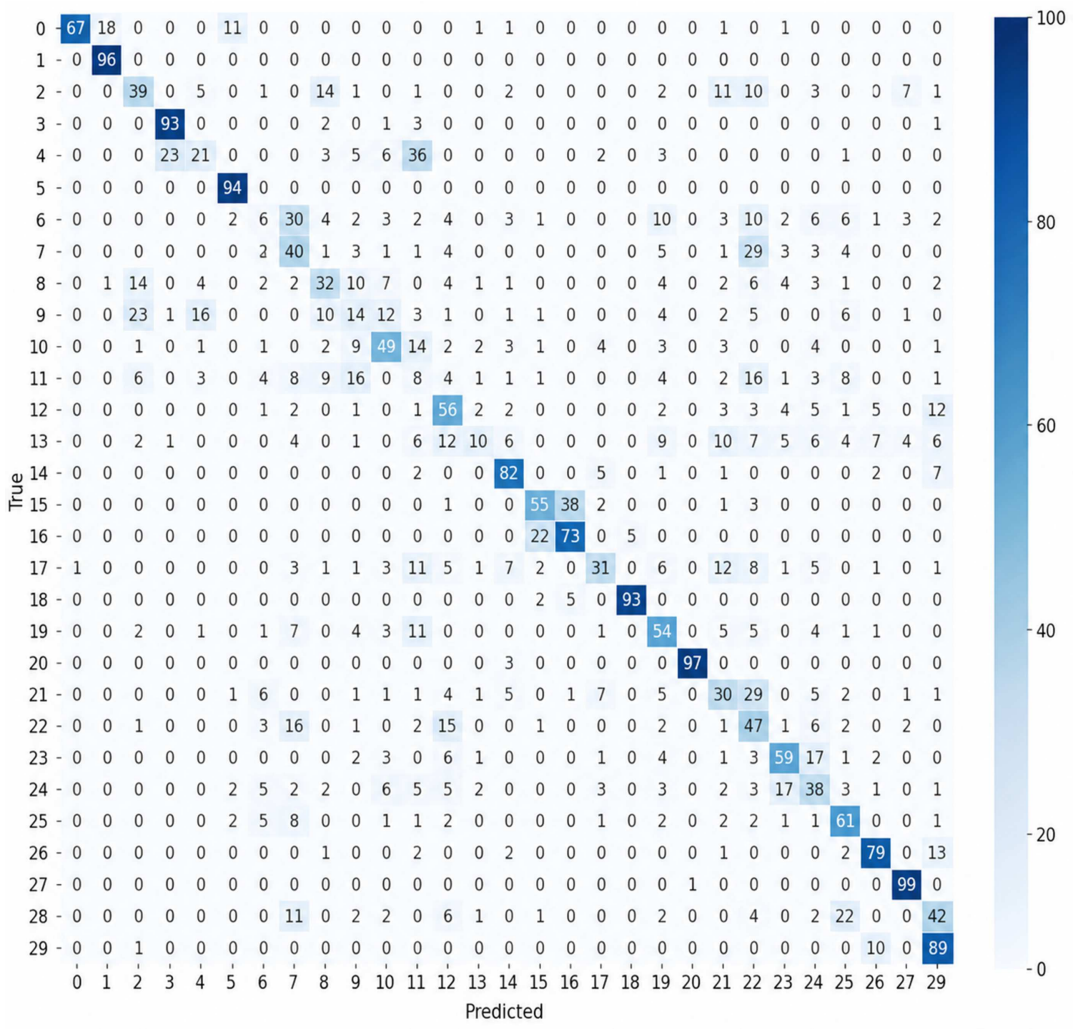
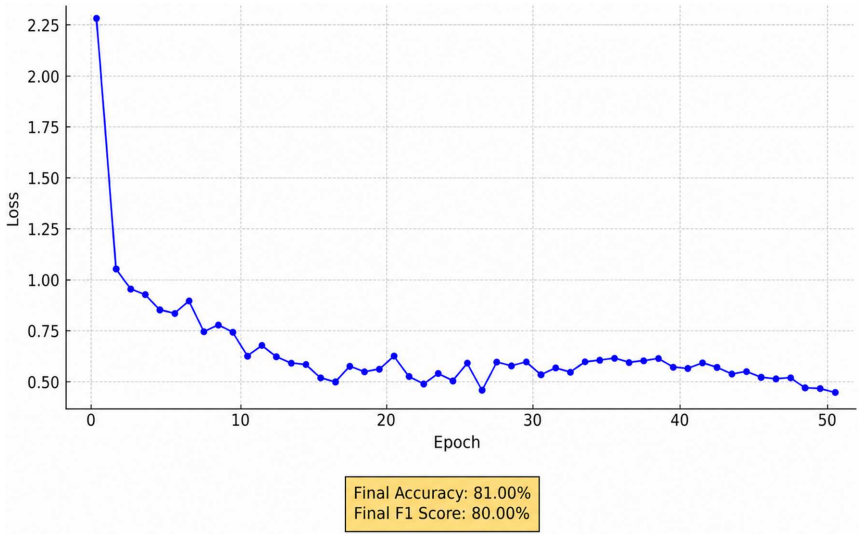


Figure 6
Training loss per epoch



of training and validation curves highlights good generalization, aided by mixed precision training and gradient accumulation.

3.5. Ablation and components

To assess the contribution of individual components, ablation studies were conducted by selectively removing elements of the proposed pipeline. The results are summarized in Table 7.

Table 7
Ablation study results (%)

Variant	Accuracy	Weighted F1
Full model	85.0	85.0
FFT and wavelets	82–83	82–83
SMOTE/class weights	81–83	81–83
CNN-only	82–84	82–84

The ablation experiments highlight the complementary value of each design choice:

1) FFT and wavelets: Removing frequency-domain and multi-resolution features reduced weighted F1 by approximately 2–3%. This confirms the importance of spectral transformations for denoising and capturing localized discriminative patterns.
2) Class balancing (SMOTE and weighted loss): Excluding balancing strategies caused a substantial drop in minority-class performance, lowering weighted F1 by several points and increasing bias toward majority classes.
3) Transformer block: Replacing the Transformer encoder with CNN-only layers degraded long-range dependency modeling, reducing performance to 82–84% weighted F1.

Together, these findings confirm that each component—spectral transformations, balancing strategies, and attention-based encoding—provides unique benefits. Their integration in the full CNN–Transformer pipeline enabled the highest performance, achieving 85% accuracy and weighted F1 on the Bacteria-ID dataset.

3.6. Summary of experimental findings

The experiments demonstrate that the proposed CNN–Transformer hybrid significantly improves bacterial identification from Raman spectra compared with CNN-only baselines. Starting from a weak baseline (64.9% accuracy, 63.3% weighted F1), performance was progressively enhanced through targeted preprocessing, augmentation, and hyperparameter tuning, reaching a final accuracy and weighted F1 of 85.0%. Importantly, minority-class recognition improved, with the lowest class F1 rising from 31% to ~70% and one-third of all species achieving F1 $\geq$ 0.90.

Ablation studies confirmed that FFT/wavelet preprocessing, SMOTE with class-weighted loss, and Transformer encoding each contributed uniquely to overall performance. Removing any of these components reduced weighted F1 by 2–4%, while the full integrated model achieved the best results. Learning dynamics further showed that pretraining on the large reference set improved generalization and stability, while mixed precision training and gradient accumulation enhanced training efficiency.

Together, these findings validate the effectiveness of the proposed pipeline, demonstrating that the integration of preprocessing, augmentation, balancing, and hybrid architecture yields state-of-the-art results on the Bacteria-ID dataset.

4. Discussion

The results indicate that jointly modeling local peak geometry (CNN) and global spectral dependencies (Transformer) improves species-level classification on noisy single-cell Raman spectra. Targeted augmentation and class-aware loss/oversampling further stabilize minority-class performance.

The novelty of our approach lies in combining CNN-based local feature extraction with Transformer-based long-range dependency modeling, bridging the limitations of CNN-only Raman classifiers. Nonetheless, the model's current limits include a lack of clinical validation, dependency on the Bacteria-ID acquisition settings, and the need for broader evaluation on different instruments and preparation protocols.

Compared with prior work on Bacteria-ID, our ~85% accuracy and weighted F1 match or exceed earlier CNN-based models. Ho et al. [14] achieved ~82.2% using a deep ResNet on low-SNR spectra, while Zhou et al. [23] reported ~84.7% with RamanNet, a lightweight CNN optimized for spectral inputs. Other strong CNN variants, such as SE-ResNet (~81%), demonstrate the limits of convolution-only architectures when applied to noisy Raman spectra. By contrast, the CNN–Transformer hybrid balances accuracy with computational efficiency: convolutional layers capture local peaks, while attention layers efficiently model long-range dependencies, with mixed precision training (AMP) ensuring feasible GPU usage.

Despite this performance, confusions remain for certain closely related taxa, highlighting areas for improvement in augmentation and feature engineering. Future directions include targeted augmentation strategies for difficult classes, memory-efficient Transformer architectures, and clinically relevant extensions such as Methicillin-resistant (Staphylococcus aureus) (MRSA) and Methicillin-susceptible (Staphylococcus aureus) (MSSA) classification and treatment-response prediction. These extensions would directly build on the flexibility of the current pipeline and support real-world diagnostic use cases.

While MRSA/MSSA classification is not experimentally verified here due to limited lab access, prior Raman studies on Bacteria-ID and RamanNet reported binary MRSA/MSSA discrimination above 80%. Building on those results, our future work will perform a small-scale clinical validation using the present pipeline on newly acquired isolates.

5. Conclusion

This study addressed two principal limitations in existing Raman-based bacterial identification methods: the inability of convolutional-only architectures to capture long-range spectral dependencies, and the insufficient handling of class imbalance in multi-species classification. To this end, a two-stage CNN–Transformer pipeline was developed and evaluated on the 30-class Bacteria-ID dataset, incorporating FFT and wavelet decomposition for spectral preprocessing, SMOTE and class-weighted cross-entropy loss for imbalance mitigation, and a Transformer encoder to model correlations across the full fingerprint region.

The proposed model achieved approximately 85% accuracy and weighted F1 on the independent test set, surpassing

the ResNet baseline (82.2%) and RamanNet (84.7%) under identical evaluation conditions. Notably, the lowest-performing species improved from 31% F1 in the unoptimized baseline to approximately 70% in the final configuration, demonstrating meaningful gains in minority-class recognition. These results confirm that the integration of preprocessing, augmentation, class balancing, and hybrid architecture yields consistent improvements over convolutional-only approaches.

Future work will explore deeper architectures, specialized augmentation for challenging classes, memory-efficient attention mechanisms, and clinically relevant extensions such as MRSA/MSSA classification and antibiotic treatment-response prediction. Validating the pipeline on independently acquired spectra from different instruments and clinical environments remains an open and important task before these results can be considered broadly generalizable. With greater computational resources, the pipeline has strong potential to exceed 90% weighted F1, advancing practical applications of Raman spectroscopy in rapid bacterial diagnostics.

Acknowledgments

We would like to acknowledge the maintainers of the Bacteria-ID resource for making the dataset publicly available.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

This study uses the publicly available Bacteria-ID dataset (Ho et al., 2019). No new data were generated. The specific train/validation/test splits, trained model weights, and figure-generation outputs can be obtained from the corresponding author upon reasonable request for academic use. Source code is not publicly shared currently.

Author Contribution Statement

Apoorv Patel: Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Hongying Meng:** Resources, Writing – review & editing, Supervision, Project administration.

References

[1] Ashfaq, M. Y., Da'na, D. A., & Al-Ghouti, M. A. (2022). Application of MALDI-TOF MS for identification of environmental bacteria: A review. *Journal of Environmental Management*, *305*, 114359. https://doi.org/10.1016/j.jenvman.2021.114359

[2] Lasch, P., Beyer, W., Bosch, A., Borriss, R., Drevinek, M., Dupke, S., . . . , & Doellinger, J. (2025). A MALDI-ToF mass spectrometry database for identification and classification of highly pathogenic bacteria. *Scientific Data*, *12*(1), 187. https://doi.org/10.1038/s41597-025-04504-z

[3] Rodriguez, L., Zhang, Z., & Wang, D. (2023). Recent advances of Raman spectroscopy for the analysis of bacteria. *Analytical Science Advances*, *4*(3-4), 81–95. https://doi.org/10.1002/ansa.202200066

[4] Thomsen, B. L., Christensen, J. B., Rodenko, O., Usenov, I., Grønnemose, R. B., Andersen, T. E., & Lassen, M. (2022). Accurate and fast identification of minimally prepared bacteria phenotypes using Raman spectroscopy assisted by machine learning. *Scientific Reports*, *12*(1), 16436. https://doi.org/10.1038/s41598-022-20850-z

[5] Qi, Y., Hu, D., Jiang, Y., Wu, Z., Zheng, M., Chen, E. X., . . . , & Chen, Y. P. (2023). Recent progresses in machine learning assisted Raman spectroscopy. *Advanced Optical Materials*, *11*(14), 2203104. https://doi.org/10.1002/adom.202203104

[6] Deng, L., Zhong, Y., Wang, M., Zheng, X., & Zhang, J. (2022). Scale-adaptive deep model for bacterial Raman spectra identification. *IEEE Journal of Biomedical and Health Informatics*, *26*(1), 369–378. https://doi.org/10.1109/JBHI.2021.3113700

[7] Tang, J. W., Li, J. Q., Yin, X. C., Xu, W. W., Pan, Y. C., Liu, Q. H., . . . , & Wang, L. (2022). Rapid discrimination of clinically important pathogens through machine learning analysis of surface enhanced Raman spectra. *Frontiers in Microbiology*, *13*, 843417. https://doi.org/10.3389/fmicb.2022.843417

[8] Yu, S., Li, X., Lu, W., Li, H., Fu, Y. V., & Liu, F. (2021). Analysis of Raman spectra by using deep learning methods in the identification of marine pathogens. *Analytical Chemistry*, *93*(32), 11089–11098. https://doi.org/10.1021/acs.analchem.1c00431

[9] Boateng, D. (2025). Advances in deep learning-based applications for Raman spectroscopy analysis: A mini-review of the progress and challenges. *Microchemical Journal*, *209*, 112692. https://doi.org/10.1016/j.microc.2025.112692

[10] Hallström, E., Kandavalli, V., Ranefall, P., Elf, J., & Wählby, C. (2023). Label-free deep learning-based species classification of bacteria imaged by phase-contrast microscopy. *PLoS Computational Biology*, *19*(11), e1011181. https://doi.org/10.1371/journal.pcbi.1011181

[11] Fan, X., Wang, Y., Yu, C., Lv, Y., Zhang, H., Yang, Q., . . . , & Zhang, Z. (2023). A universal and accurate method for easily identifying components in Raman spectroscopy based on deep learning. *Analytical Chemistry*, *95*(11), 4863–4870. https://doi.org/10.1021/acs.analchem.2c03853

[12] Zhu, L., Yang, Y., Xu, F., Lu, X., Shuai, M., An, Z., . . . , & Cui, L. (2025). Open-set deep learning-enabled single-cell Raman spectroscopy for rapid identification of airborne pathogens in real-world environments. *Science Advances*, *11*(2), eadp7991. https://doi.org/10.1126/sciadv.adp7991

[13] Balytskyi, Y., Bendesky, J., Paul, T., Hagen, G. M., & McNear, K. (2022). Raman spectroscopy in open-world learning settings using the objectosphere approach. *Analytical Chemistry*, *94*(44), 15297–15306. https://doi.org/10.1021/acs.analchem.2c02666

[14] Ho, C. S., Jean, N., Hogan, C. A., Blackmon, L., Jeffrey, S. S., Holodniy, M., . . . , & Dionne, J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature communications*, *10*(1), 4927. https://doi.org/10.1038/s41467-019-12898-9

[15] Ibtehaz, N., Chowdhury, M. E., Khandakar, A., Kiranyaz, S., Rahman, M. S., & Zughaier, S. M. (2023). RamanNet: A generalized neural network architecture for Raman

spectrum analysis. *Neural Computing and Applications*, *35*(25), 18719–18735. https://doi.org/10.1007/s00521-023-08700-z

[16] Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, *137*, 109347. https://doi.org/10.1016/j.patcog.2023.109347

[17] Tagnamas, J., Ramadan, H., Yahyaouy, A., & Tairi, H. (2024). Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. *Visual Computing for Industry, Biomedicine, and Art*, *7*(1), 2. https://doi.org/10.1186/s42492-024-00155-w

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

[19] Wang, Z., Li, Y., Zhai, J., Yang, S., Sun, B., & Liang, P. (2024). Deep learning-based Raman spectroscopy qualitative analysis algorithm: A convolutional neural network and transformer approach. *Talanta*, *275*, 126138. https://doi.org/10.1016/j.talanta.2024.126138

[20] Howard, J., & Gugger, S. (2020). *Deep learning for coders with fastai and PyTorch: AI applications without a PhD*. USA: O'Reilly Media.

[21] Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

[22] Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

[23] Zhou, B., Tong, Y. K., Zhang, R., & Ye, A. (2022). RamanNet: a lightweight convolutional neural network for bacterial identification based on Raman spectra. *RSC Advances*, *12*(40), 26463–26469. https://doi.org/10.1039/D2RA03722J

[24] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

[25] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., . . . , & . . . . (2018). Mixed precision training. In *6th International Conference on Learning Representations*, 1–12.

How to Cite: Patel, A., & Meng, H. (2026). Rapid Identification of Pathogenic Bacteria from Raman Spectra with a CNN–Transformer Hybrid Architecture. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62027534