Data Augmentation for Small-Scale Raman Spectroscopy Dataset in Breast Cancer Cell Classification

Xin Li [1], Jiajun Sun [1] , Pengju Yin [1, 2] *

[1] School of Mathematics and Physics, Hebei University of Engineering, Handan 056038, China;

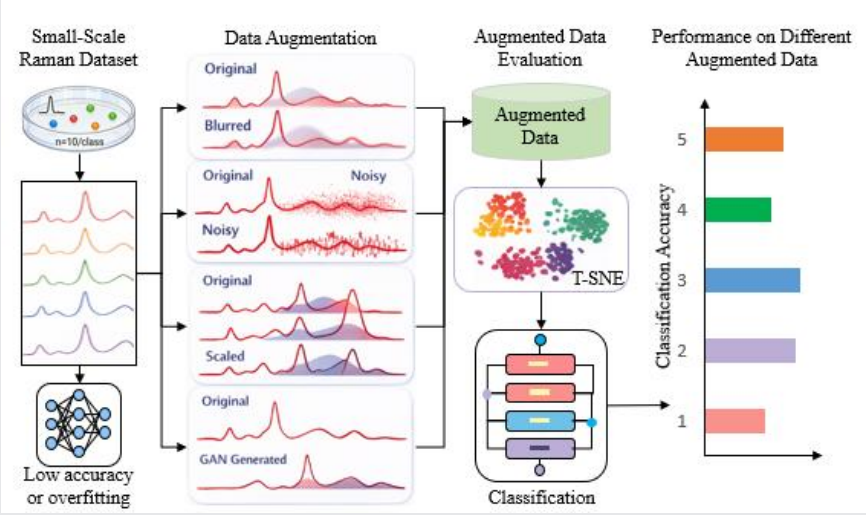
[2] Natural Sciences and Science Education, National Institute of Education, Nanyang Technological University, Singapore, 637616, Singapore

* Correspondence: yinpengju@hebeu.edu.cn (P.Y.)

Abstract

The application of machine learning to Raman spectral analysis is limited by the scarcity of labeled data, particularly in single-cell studies where large datasets are difficult to obtain. Under such small-sample conditions, the reliability of different data augmentation strategies remains unclear. This study systematically evaluates four data augmentation methods—localized blurring, Gaussian noise addition, random amplitude scaling, and generative adversarial network (GAN) - based synthesis—for small-scale Raman spectral classification of breast cells, focusing on a training set containing 10 samples per class. Distributional similarity between original and augmented data was assessed using Fréchet Inception Distance and t-distributed stochastic neighbor embedding, and classification performance was evaluated using a one-dimensional ResNet model. The results show that augmentation effectiveness depends on both the augmentation strategy and the number of synthetic samples. Gaussian noise augmentation achieved the highest distributional similarity and improved classification accuracy from 92.45% to 95.35%, while localized blurring also yielded consistent improvements, with accuracies exceeding 94.30%. GAN-based augmentation enhanced performance at suitable augmentation sizes, reaching peak accuracies above 94.10%, but showed greater sensitivity to parameter selection. In contrast, random amplitude scaling provided little improvement across most settings. Additional parameterization experiments were conducted by changing the original training set to investigate extremely data-scarce scenarios, including 1 or 2 spectra per class. In these cases, data augmentation alleviated severe overfitting, with Gaussian and blurring methods providing the most stable gains, whereas GAN-based augmentation showed variable effectiveness and scaling remained ineffective. These results offer practical, quantitative guidance for selecting appropriate data augmentation strategies in small-scale Raman spectral analysis.

Keywords: Raman spectroscopy; data augmentation; GAN; machine learning; classification; breast cell; small-scale dataset



1. Introduction

Breast cancer remains the most frequently diagnosed cancer and the leading cause of cancer-related deaths among women worldwide, with millions of new cases reported each year[1]. Early and accurate characterization of breast cancer cells is therefore of critical importance for both fundamental research and clinical diagnosis. In this context, Raman spectroscopy has emerged as a powerful analytical tool for the study of breast cells[2-5]. As a label-free and non-destructive technique, Raman spectroscopy provides high-resolution molecular fingerprint information, enabling the detection of biochemical alterations associated with cancer development, including changes in proteins, nucleic acids, lipids, and other biomolecular components [6-8].

In recent years, the integration of machine learning techniques has substantially enhanced the analysis of Raman spectral data. Traditional machine learning algorithms, such as Support Vector Machines (SVM)[9], have been widely adopted due to their effectiveness in spectral classification tasks[10-12]. More recently, deep learning architectures have demonstrated strong capabilities in automatically extracting high-level features from complex Raman spectral signals. Representative models include one-dimensional Convolutional Neural Network (1D CNN)[13-18], Recurrent Neural Network (RNN)[19, 20], one-dimensional Residual Neural Network (1D ResNet)[21-24], as well as Transformer-based approaches developed for Raman spectroscopy[25-27]. These deep models often achieve superior performance in challenging classification scenarios; however, their success typically relies on the availability of large-scale, high-quality training datasets.

A major obstacle to fully exploiting these machine learning and deep learning models in Raman-based breast cell analysis lies in the limited availability of data. The acquisition of Raman spectra from biological samples is frequently constrained by clinical and ethical considerations, high experimental costs, time-consuming measurement procedures, and the requirement for expert annotation. As a result, available datasets are often small in size, particularly at the single-cell level. Insufficient training data not only increases the risk of overfitting but also restricts the development of robust and generalizable models, especially for data-intensive deep learning architectures[28-31].

To alleviate the challenges associated with small datasets, data augmentation has been widely adopted as a practical strategy to artificially increase the size and diversity of training data. By generating realistic synthetic samples from existing data, augmentation methods encourage models to learn more generalizable patterns. Conventional augmentation techniques for spectral data include the addition of Gaussian noise, baseline shifting, amplitude scaling, localized blurring, and related transformations[32-36]. More recently, neural network–based generative approaches, particularly Generative Adversarial Network (GAN)[37], have been introduced to learn underlying data distributions and generate highly realistic synthetic spectra[38, 39]. Despite their success in various applications, GAN-based methods typically require a substantial amount of training data to achieve stable and reliable performance. This requirement presents an inherent contradiction when GAN are applied to small-scale Raman datasets, where limited data may hinder accurate distribution learning and compromise generation quality.

Although numerous augmentation strategies have been proposed, the comparative effectiveness of simple transformation-based

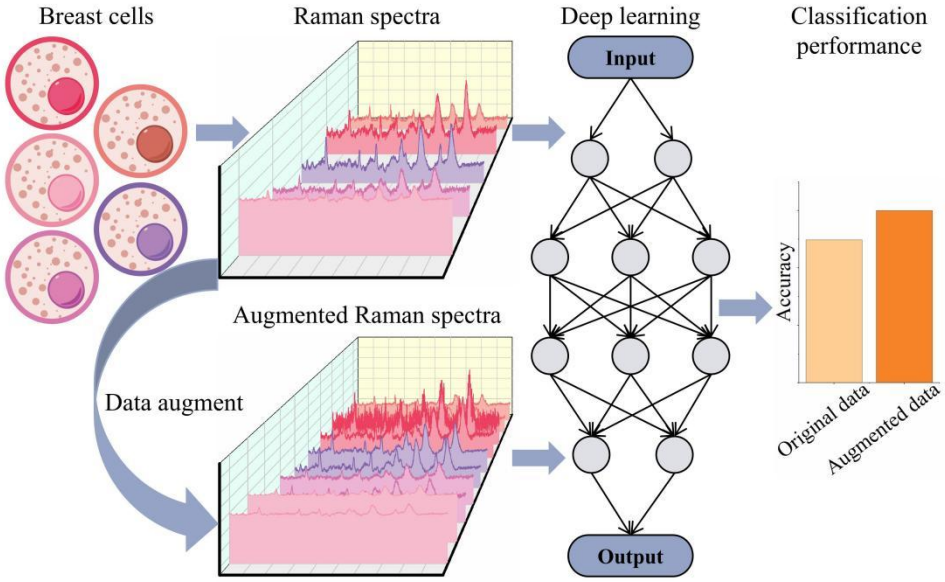


Figure 1. Schematic overview of the study workflow for evaluating data augmentation strategies on a small breast cell Raman spectroscopy dataset. The process begins with breast cell samples, from which original Raman spectra are collected. These raw spectral data then enter a data augmentation module, where synthetic spectra are generated using various augmentation techniques. Both the original and augmented spectra are subsequently used as input to identical deep learning models for classification. The output classification performance of models trained on original versus augmented data is quantitatively compared using evaluation metrics such as accuracy.

methods and complex generative models has not been systematically evaluated in the context of small-scale Raman spectroscopy of breast cells. In particular, it remains unclear whether data-hungry generative models such as GAN can outperform simpler, physics-inspired perturbation methods under severe data scarcity. Addressing this gap is essential for guiding the selection of appropriate augmentation strategies in practical Raman spectroscopy applications, where collecting large datasets is often infeasible.

To address this issue, a systematic evaluation was conducted on a small-scale Raman spectral dataset of breast cells. Four representative data augmentation methods were investigated, including localized blurring, Gaussian noise addition, random amplitude scaling, and GAN-based synthesis. The schematic of this work is illustrated in **Figure 1**. A one-dimensional Residual Neural Network (1D ResNet) was employed as a unified classification framework to ensure fair comparison across different augmentation strategies. The original dataset was first used to establish a baseline classification performance. Each augmentation method was then applied independently to expand the training data. The quality of the augmented spectra was preliminarily assessed using Fréchet Inception Distance (FID)[40] and t-distributed Stochastic Neighbor Embedding (t-SNE)[41] visualization. Subsequently, classification performance was evaluated using Accuracy, Precision, Recall, and F1-score. In addition, a parameterization study was carried out to investigate the influence of augmentation size and original training set size on model robustness. The findings aim to provide practical guidance for handling small-scale Raman spectral datasets and to support the development of more reliable machine learning models for breast cancer research and diagnostic applications.

2. MATERIALS AND METHODS

2.1 Dataset: Small-scale Breast Cell Raman Spectra

The dataset employed in this study comprised a collection of single-cell Raman spectra acquired from five established human breast cell lines, namely MCF-10A, MDA-MB-231, BT-474, SK-BR3, and T-47D. The average Raman spectra of the five breast cell categories, along with their corresponding characteristic peak positions, are presented in **Figure 2a**. All cell lines were obtained from the Cell Resource Centre, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (China Union Medical University). Polymerase chain reaction (PCR) and microbial culture assays confirmed that the cell lines were free of mycoplasma contamination. Additional details regarding the dataset can be found in the published work[42].

The dataset was partitioned into three subsets: a training set, a validation set, and a test set. To simulate a data-scarce scenario, the training set contained only 50 Raman spectra, with 10 spectra assigned to each breast cell line. The validation set followed the same

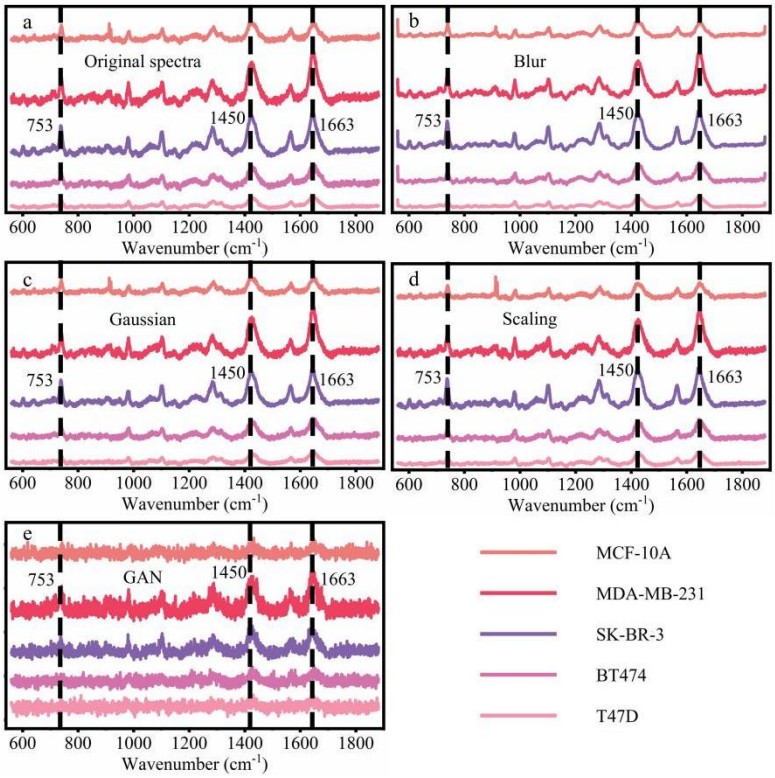


Figure 2. Average Raman spectroscopy and peak position for original data and four augmented data. a). Original spectra's average Raman spectra. Position of the peaks are 753 cm-1, 1450 cm-1 and 1663 cm-1. b). Blur spectra's average Raman spectra. Position of the peaks are 753 cm-1, 1450 cm-1 and 1663 cm-1. c). Gaussian spectra's average Raman spectra. Position of the peaks are 753 cm-1, 1450 cm-1 and 1663 cm-1. d). Scaling spectra's average Raman spectra. Position of the peaks are 753 cm-1, 1450 cm-1 and 1663 cm-1. e). GAN spectra's average Raman spectra. Position of the peaks are 753 cm-1, 1450 cm-1 and 1663 cm-1.

class distribution and size, also comprising 50 Raman spectra in total. The test set consisted of 400 Raman spectra, including 80 spectra per breast cell line. After the initial data split, the validation and test sets remained unchanged and were consistently used in all subsequent experiments. A detailed overview of the dataset partition is provided in the table below.

2.2 Data Augmentation Strategies for Small Dataset

When working with small-scale Raman spectroscopy datasets, overfitting represents a major challenge, as models tend to memorize limited training samples rather than learning generalizable patterns. Data augmentation is commonly employed to mitigate this issue by artificially increasing the size and diversity of the training set. By generating plausible new spectral samples through controlled modifications of the original data, augmentation techniques help models focus on robust and invariant spectral features, thereby improving generalization performance on unseen data.

In this study, four widely used data augmentation methods were applied to the breast cell Raman training set. These methods included localized blurring, Gaussian noise addition, random amplitude scaling, and Generative Adversarial Network (GAN)-based synthesis, which are referred to as Blur, Gaussian, Scaling, and GAN, respectively. Each augmentation method was used to generate 100 additional Raman spectra per class based on the original training data. The average spectra of the augmented data produced by these four methods, together with their characteristic peak positions, are shown in **Figure 2b**, **2c**, **2d**, and **2e**. These were then combined with the original 10 training spectra, resulting in four newly augmented datasets. All augmented datasets were kept separate to avoid contamination of the validation and test sets. Next, the details of the four augmentation methods are introduced.

2.2.1 Blur

Localized blurring was applied to simulate minor variations in spectral acquisition that may arise from factors such as instrumental resolution or focal plane fluctuations. This method employs a one-dimensional average blur filter applied along the spectral axis. For each augmented sample, a kernel size of either 2 or 3 spectral points was randomly selected. The convolution operation was performed using same padding to preserve the original spectral length. After blurring, the resulting spectrum was passed through a hyperbolic tangent (tanh) function, constraining the signal values to the range (−1, 1) and ensuring consistency with the normalization applied to the original data. This augmentation strategy encourages the model to learn features that are insensitive to small smoothness variations along the spectral dimension.

2.2.2 Gaussian

Gaussian noise addition was employed to mimic the stochastic noise commonly observed in real Raman measurements, which may originate from electronic instrumentation or environmental fluctuations. For each spectrum, zero-mean Gaussian noise was added, with the standard deviation uniformly sampled from the range of 0.02 to 0.08. This procedure introduces controlled random perturbations while preserving the overall spectral structure. The noisy spectrum was subsequently passed through a tanh function to restrict the amplitude values to the interval (−1, 1), maintaining consistency with the normalization scheme of the original dataset. This augmentation method promotes robustness to random noise and enhances the model's ability to generalize to experimentally acquired spectra.

2.2.3 Scaling

Random amplitude scaling was used to emulate variations in experimental conditions, such as changes in laser power or sample concentration. In this approach, each spectrum was multiplied by a random scaling factor uniformly sampled from the interval [0.85, 1.15]. This operation modifies the absolute intensity of the spectrum while preserving the relative shapes and positions of spectral features. The scaled spectrum was then transformed using a tanh function to ensure that all augmented samples remained within the normalized range of (−1, 1). This augmentation strategy enables the model to focus on intrinsic spectral patterns rather than absolute intensity levels.

2.2.4 GAN

GAN-based data augmentation was employed to generate synthetic Raman spectra by learning the underlying data distribution of the original training set. A conditional GAN architecture was adopted, in which class label information was incorporated through embedding layers to guide the generation of class-specific spectra. The network architecture of the conditional GAN is illustrated in **Figure 3**. The model consists of three components: a conditional encoder, a conditional generator, and a conditional discriminator. The encoder processes input spectra together with their corresponding class labels and maps them into a latent representation. The generator combines random noise vectors with class embeddings to produce synthetic spectra, using a tanh activation function at the output layer to constrain the generated signals to the range (−1, 1). The discriminator evaluates both real and generated spectra while taking class labels into account, distinguishing between authentic and synthetic samples. Through adversarial training, the generator is encouraged to produce increasingly realistic spectra that capture class-specific characteristics, providing a data-driven augmentation approach for expanding the training set.

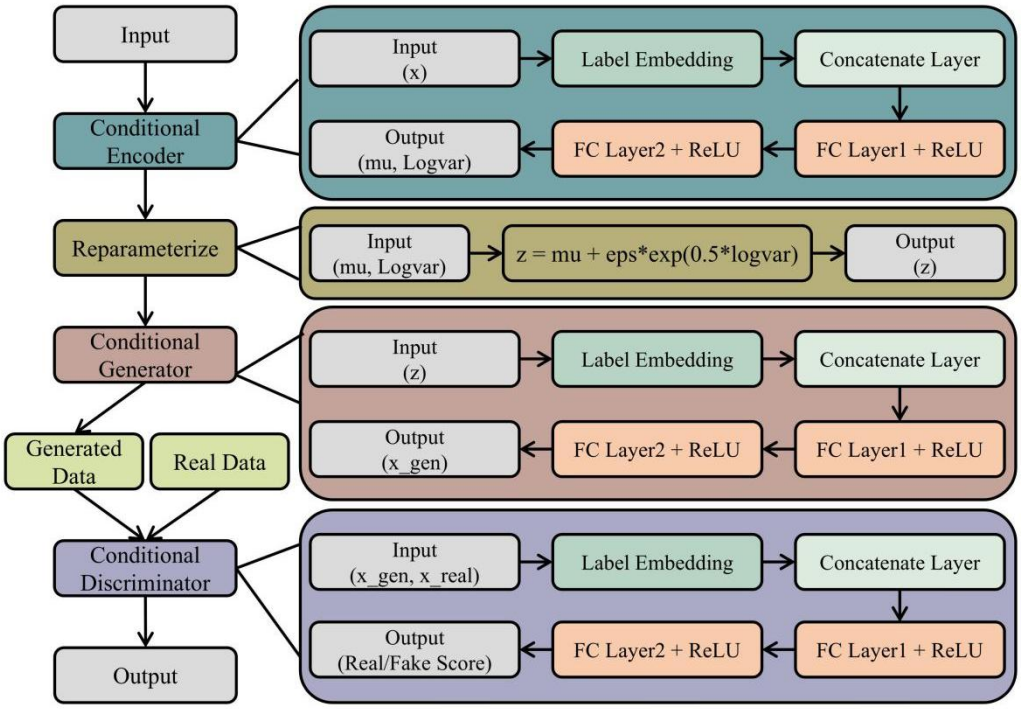


Figure 3. The architecture of the conditional Generative Adversarial Network (cGAN) employed in this study. The generator takes an input and processes it through a conditional encoder, followed by a reparameterization module that produces the latent variable z weing the mean (μ) and log variance (logvar). The conditioned generator then utilizes label embeddings and fully connected (FC) layers with ReLU activation to produce synthetic Raman spectra. The discriminator evaluates both real and generated data by concatenating input spectra with their corresponding label embeddings, processing them through FC layers, and ultimately outputting a realism score.

2.3. Augmented Data Evaluation Metrics

After applying four distinct augmentation methods (Blur, Gaussian, Scaling, and GAN) to the original training set, the quality of the augmented Raman spectra was evaluated using a dual strategy that combined the Fréchet Inception Distance (FID) and t-distributed Stochastic Neighbor Embedding (t-SNE). This combined approach enables a comprehensive assessment by providing both a quantitative measure of global distribution similarity and a qualitative visualization of the underlying feature space structure.

FID is a widely used metric for quantitatively evaluating the quality of generated data by measuring the similarity between the distributions of two datasets. The calculation is based on high-level feature representations extracted from real and augmented data. These features are modeled as multivariate Gaussian distributions, and the FID score is computed as the Fréchet distance between them. Lower FID values indicate a smaller statistical divergence between the augmented and original datasets, suggesting that the generated spectra more closely resemble real samples. In this study, the FID score between the original training set and each augmented dataset was used as the primary quantitative indicator of augmentation quality.

To complement the quantitative evaluation provided by FID, t-SNE was employed for qualitative visualization. While FID summarizes distribution similarity using a single scalar value, t-SNE reveals the internal structure of high-dimensional feature representations by projecting them into a two-dimensional space. This visualization allows for an intuitive inspection of whether augmented samples are well integrated with real data clusters or form isolated groups, which would indicate limited distributional similarity. In addition, t-SNE helps assess whether the augmented data exhibit comparable coverage and structural organization to the original data, reflecting both sample diversity and preservation of intrinsic feature relationships.

2.4. Classification Pipeline and Performance Metrics

Following data augmentation, the classification pipeline and evaluation metrics are described. Classification was performed using a one-dimensional Residual Neural Network (ResNet), which is well suited for the analysis of Raman spectral data. This architecture was selected due to its public availability, extensive validation in previous studies, and demonstrated effectiveness in spectral classification tasks. Using a standardized model framework allows performance differences to be attributed primarily to the data augmentation strategies rather than to variations in model design.

2.4.1. Classification algorithm

The classification task was conducted using a one-dimensional Residual Neural Network (1D ResNet), a deep learning architecture designed for sequential data analysis. The ResNet framework effectively alleviates the vanishing gradient problem

through the use of residual connections, enabling stable training of deeper networks. The overall architecture of the employed 1D ResNet is illustrated in **Figure 4**.

The network consists of an initial one-dimensional convolutional layer followed by batch normalization, after which multiple residual blocks form the main feature extraction module. Each residual block contains two convolutional layers with batch normalization, along with an identity skip connection that bypasses the convolutional operations. This design facilitates efficient gradient propagation and improves training stability. In this study, the number of convolutional filters in successive stages was set to 16, 32, and 64, while the number of residual blocks per stage was fixed at 2. The extracted feature representations were subsequently flattened and passed to a fully connected linear layer to produce the final classification output. The use of a consistent and well-established ResNet architecture ensures that observed performance variations can be reasonably attributed to the effects of data augmentation.

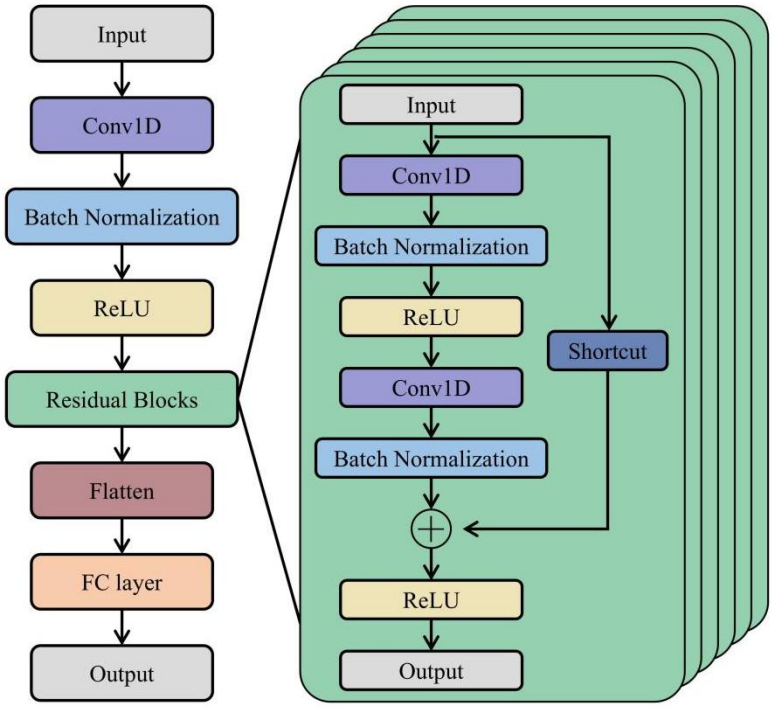


Figure 4. The architecture of the 1D ResNet model used in this study. The network begins with an input layer followed by an initial 1D convolution, batch normalization, and ReLU activation. The core component consists of several residual blocks, which are detailed in the magnified section. Each residual block contains two convolutional layers, each followed by batch normalization and ReLU activation. A key feature is the identity shortcut connection that bypasses these layers, enabling smooth gradient flow and stabilizing deep network training. After the residual blocks, the output is flattened and processed by a fully connected layer to produce the final classification.

2.4.2. Training and testing protocol

The ResNet model was trained from scratch on the original dataset as well as on each augmented dataset independently. Model optimization was performed using the Adam optimizer with a learning rate of 1e-4 and a batch size of 5. To reduce the risk of overfitting and to select the most generalizable model, an early stopping strategy was employed with a patience of 10 epochs. Training was terminated if the validation loss did not improve for 10 consecutive epochs, and the model parameters corresponding to the lowest validation loss were retained.

To account for stochastic effects arising from factors such as random weight initialization, the entire training procedure was repeated five times for each dataset. This repetition enables a more reliable evaluation by reducing the influence of random variability. Model performance was subsequently assessed on a held-out test set that was not used during training or validation. For each training run, the model achieving the lowest validation loss was evaluated on the test set. Performance metrics were then aggregated across the five independent runs and reported as mean values with corresponding standard deviations.

2.4.3. Performance metrics

Classification performance was evaluated using four complementary metrics: Accuracy, Precision, Recall, and F1-score. Although Accuracy provides an overall measure of correct predictions, it does not distinguish between different types of classification errors. Precision and Recall offer additional insights into false positive and false negative behavior, while the F1-score summarizes their balance. The combined use of these metrics enables a comprehensive and balanced assessment of model performance and ensures a fair comparison among different data augmentation strategies.

2.5. Parameterization Study: Impact of Training Set Size and Augmentation Size

To enhance the robustness of the experimental conclusions and to reduce potential bias associated with a single parameter setting, a parameterization study was conducted. This study focused on two key factors that may influence classification performance: the number of augmented samples generated per class and the size of the original training set used for data augmentation.

2.5.1. Variation of augmented data size per class

In the baseline experiment, the augmentation size was set to 100 samples per class. To evaluate the sensitivity of classification performance to this parameter, additional experiments were performed using augmented data sizes of 20, 50, 100, 200, and 500 samples per class. For each augmentation size, the data generation process was repeated independently, followed by dataset merging, model training, and evaluation using the same 1D ResNet architecture and training protocol described in Section 2.4. Throughout this analysis, the size of the original training set was fixed at 10 samples per class. This systematic variation enables a quantitative assessment of how the amount of augmented data influences the final classification performance.

2.5.2. Variation of original training set size per class

The baseline configuration employed an original training set size of 10 samples per class. To examine the effectiveness of data augmentation under varying degrees of data availability, this initial training size was further varied. Additional experiments were conducted using original training sets containing 1, 2, 5, 10, 20, and 50 samples per class. For each setting, the complete experimental pipeline was repeated. Data augmentation was performed with the augmentation size fixed at 100 samples per class for consistency, followed by dataset construction and ResNet classification. This design allows for a detailed analysis of how the amount of available real data affects the performance gains achievable through data augmentation.

3. RESULTS

3.1. Evaluation of Augmented Data Quality

After applying the four data augmentation methods, namely Blur, Gaussian, Scaling, and GAN, the average Raman spectra of the augmented datasets are presented in **Figure 2**. As shown in the figure, the major characteristic peak positions of the augmented spectra remain consistent with those of the original spectra. Specifically, prominent peaks are observed at approximately 753 cm^{-1}, 1450 cm^{-1}, and 1663 cm^{-1}, with no noticeable peak shifts introduced by any augmentation method. However, direct visual inspection of the average spectra reveals only subtle differences between the original and augmented data, making it difficult to objectively assess augmentation quality based solely on spectral appearance. Therefore, quantitative and feature-based evaluation methods were employed.

The FID scores comparing the distributions of the original and augmented datasets are summarized in **Table 1**. FID values were first computed separately for each breast cell class and then averaged across all five classes to obtain an overall metric for each augmentation method. Lower FID values indicate smaller distributional divergence between the augmented and original data. As reported in **Table 1**, the Gaussian augmentation method achieved the lowest average FID score, followed by Blur and Scaling, whereas the GAN-based method produced the highest FID score. These results indicate that Gaussian augmentation generates spectra that are most similar to the original data in terms of global feature distribution, while GAN-based augmentation exhibits the largest distributional deviation.

Table 1. The Fréchet Inception Distance (FID) scores used to evaluate the quality of data generated by the four augmentation methods. A lower FID value indicates a higher similarity between the augmented spectra and the original data.

Metric	Class	Blur	Gaussian	Scaling	GAN
FID	T47D	0.29	1.19	0.66	13.64
	BT474	3.19	1.89	2.61	16.47
	SK-BR-3	3.93	3.52	5.29	29.12
	MDA-MB-231	5.30	2.40	4.95	36.44
	MCF-10A	6.99	6.27	10.20	14.86
	Average	3.94	3.05	4.74	22.11

To further examine the feature-space characteristics of the augmented data, t-SNE visualization was performed, as shown in **Figure 5**. The t-SNE plots reveal distinct behaviors for different augmentation methods. Features generated using the Blur method are distributed closely around the original data clusters, indicating a high degree of similarity while preserving class separability. Gaussian-augmented features exhibit substantial overlap with the original features and maintain clear inter-class boundaries, consistent with the low FID scores. In comparison, Scaling-augmented features also overlap with the

original data but show reduced separation between different cell classes. In contrast, features generated by the GAN method form clusters that are clearly separated from the original data distribution, indicating lower distributional similarity. Nevertheless, these GAN-generated features exhibit strong inter-class separation, suggesting enhanced class discriminability within the generated feature space.

3.2. Classification Performance

While FID and t-SNE analyses evaluate the similarity between original and augmented spectral distributions, the practical effectiveness of data augmentation must ultimately be assessed through classification performance. To this end, a one-dimensional ResNet classifier was employed, and performance was evaluated using four metrics: Accuracy, Precision, Recall, and F1-score. The model was first trained and tested using the original small-scale dataset. The same procedure was then applied independently to each of the four augmented training sets, while the validation and test sets were kept identical across all experiments. In addition, further analyses were conducted to examine the influence of augmentation size and original training set size on classification outcomes.

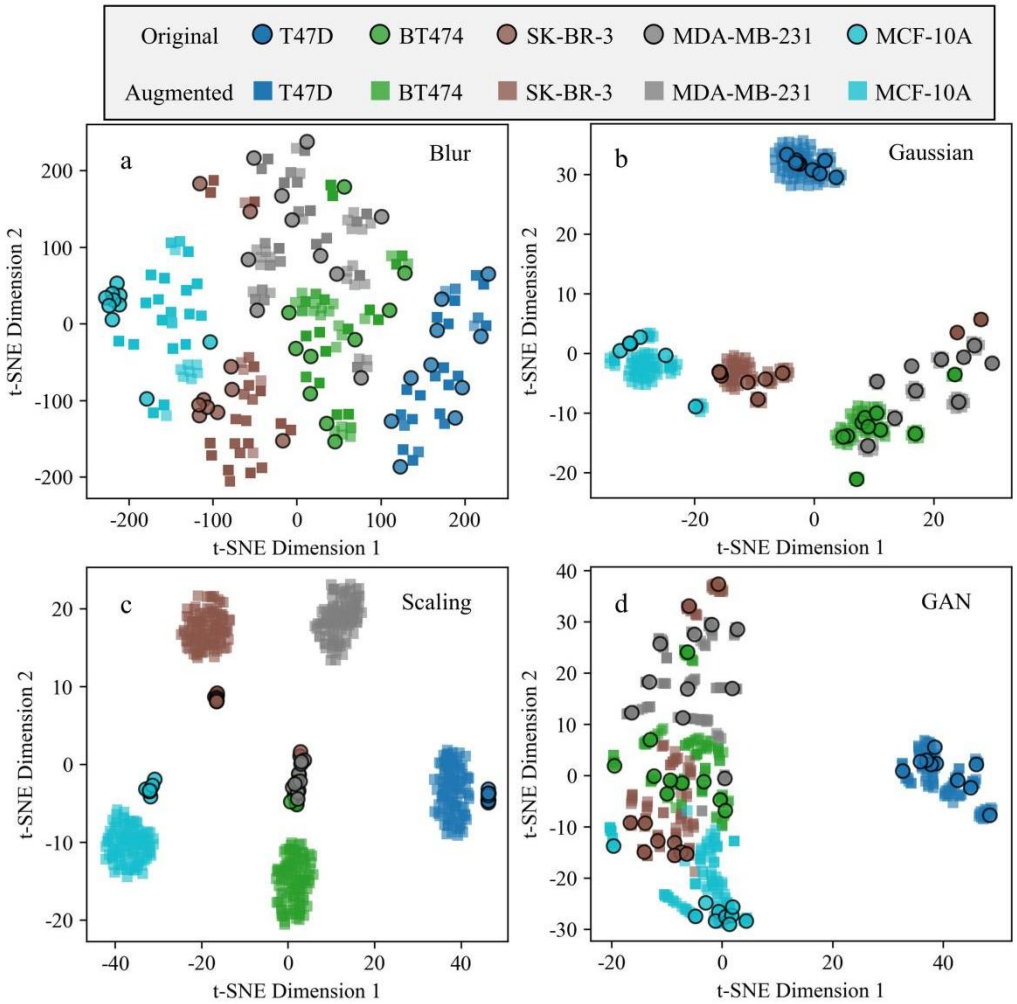


Figure 5. Feature visualization using t-SNE. a) t-SNE visualization of data augmented using the Blur method. b) t-SNE visualization of data augmented using the Gaussian method. c) t-SNE visualization of data augmented using the Scaling method. d) t-SNE visualization of data augmented using the GAN method.

3.2.1. Classification results of original small-scale dataset and augmented dataset

The original training set consisted of 50 Raman spectra, with 10 spectra per breast cell class. The corresponding classification performance is summarized in **Table 2**. The ResNet model achieved an average accuracy of 92.45% ± 1.05%, along with a precision of 92.90% ± 1.07%, a recall of 92.45% ± 1.05%, and an F1-score of 92.42% ± 1.05%.

Each augmented training set contained 550 Raman spectra, corresponding to 110 spectra per class. The classification results obtained using these augmented datasets are also reported in **Table 2**. Among the four augmentation strategies, Gaussian augmentation yielded the highest performance, achieving an accuracy of 95.00% ± 1.03%, a precision of 95.17% ± 1.03%, a recall of 95.00% ± 1.03%, and an F1-score of 94.99% ± 1.05%. The Blur-augmented dataset ranked second, with an accuracy of 93.90% ± 1.39% and consistently high values across the

remaining metrics. In contrast, the GAN-augmented dataset produced performance comparable to the original dataset, while the Scaling-augmented dataset resulted in the lowest scores among all evaluated configurations. Relative to the original dataset, only the Gaussian and Blur augmentation methods led to consistent improvements across all evaluation metrics.

Table 2. The classification accuracy, precision, recall and F1 score obtained on the same test set after training with the original dataset and the four augmented datasets.

	Accuracy	Precision	Recall	F1 score
Original data	92.45%±1.05%	92.90%±1.07%	92.45%±1.05%	92.42%±1.03%
Blur	93.90%±1.39%	94.08%±1.35%	93.90%±1.39%	93.89%±1.40%
Gaussian	95.00%±1.03%	95.17%±1.03%	95.00%±1.03%	94.99%±1.05%
Scaling	91.20%±1.12%	91.40%±1.11%	91.20%±1.12%	91.19%±1.14%
GAN	92.30%±1.65%	92.65%±1.41%	92.30%±1.65%	92.32%±1.62%

Table 3. The classification accuracy obtained on the same test set after training with different augmented sizes of 20, 50, 100, 200 and 500.

	20	50	100	200	500
Blur	93.20%±1.27%	94.30%±0.98%	93.90%±1.39%	91.85%±1.05%	93.55%±0.99%
Gaussian	93.25%±1.21%	93.65%±2.70%	95.00%±1.03%	95.35%±1.83%	94.55%±3.06%
Scaling	91.70%±3.11%	89.90%±2.18%	91.20%±1.12%	91.10%±0.91%	90.10%±1.33%
GAN	93.85%±1.93%	94.10%±2.42%	92.30%±1.65%	90.35%±2.17%	87.95%±3.22%

3.2.2. Influence of augmentation size on classification performance

To assess the sensitivity of classification performance to the number of augmented samples per class, additional experiments were conducted using augmentation sizes of 20, 50, 100, 200, and 500 spectra per class. Accuracy was used as the primary metric for comparison, with results summarized in **Table 3**. Detailed results for Precision, Recall, and F1-score are provided in **Table 2** and **Supplementary Tables S1-S4**.

For the Blur augmentation method, the highest accuracy of 94.30% was achieved with 50 augmented samples per class. Gaussian augmentation exhibited optimal performance at 200 samples per class, reaching an accuracy of 95.35%. The GAN-based method attained its maximum accuracy of 94.10% at 50 samples per class. For these three methods, classification accuracy consistently exceeded the baseline accuracy of 92.45% obtained using the original dataset alone. In contrast, the Scaling method failed to surpass the baseline accuracy at most augmentation sizes, achieving its best result of 91.70% at 20 samples per class. These results demonstrate that the effectiveness of data augmentation depends on both the augmentation strategy and the selected augmentation size.

3.2.3. Influence of original training set size on classification performance

The robustness of the augmentation methods under varying degrees of data scarcity was further examined by varying the size of the original training set. Training sets containing 1, 2, 5, 10, 20, and 40 samples per class were evaluated. Accuracy was again adopted as the primary metric, with averaged results summarized in **Table 4.** The relative impact of data augmentation is visualized in **Figure 6**, which depicts the accuracy difference between augmented and original datasets across different combinations of training set size and augmentation size. Complete results for Precision, Recall, and F1-score are provided in **Table 2** and **Supplementary Tables S5-S9**.

When the original training set was extremely limited (1 or 2 samples per class), models trained solely on the original data exhibited severe overfitting, resulting in a baseline accuracy of approximately 20.00%. Under these conditions, all augmentation methods substantially improved classification performance. Gaussian augmentation achieved the highest accuracies, reaching 70.90% and 72.70% for 1 and 2 samples per class, respectively, followed by the Blur method with accuracies of 58.00% and 69.45%. With an increase to 5 samples per class, the baseline performance remained unstable, yielding an average accuracy of 58.90%, whereas the augmented datasets produced more consistent results, particularly for Gaussian and Blur augmentation. As the original training set size increased to 10, 20, and 40 samples per class, the classification performance across all methods gradually converged toward higher

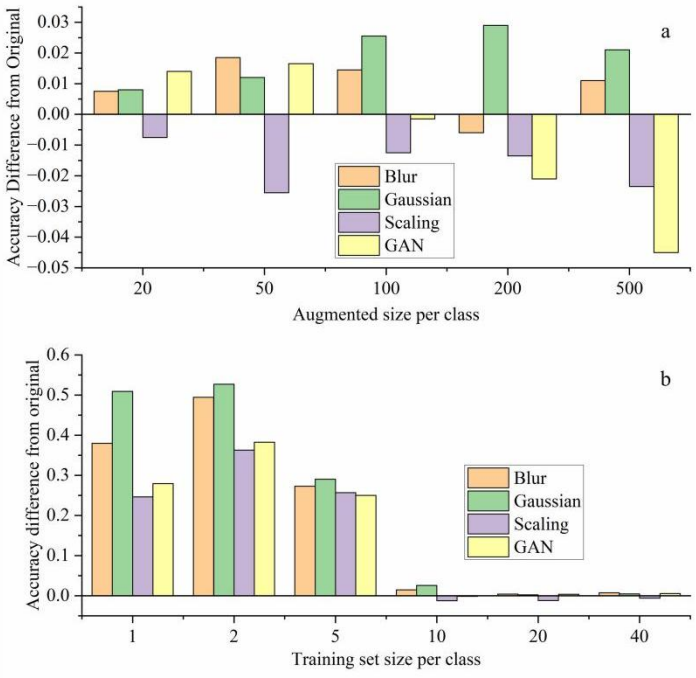


Figure 6. Comparison of classification accuracy between original and augmented datasets. (a) Accuracy difference when varying the number of augmented samples per class (20, 50, 100, 200, 500), with augmentation initially set at 100 samples per class. (b) Accuracy difference when varying the number of original training samples per class (1, 2, 5, 10, 20, 40), with the initial training set size set at 10 samples per class. All models were evaluated on the same test set.

Table 4. The classification accuracy achieved on an identical test set following training using both the original dataset and the four augmented datasets, each evaluated at varying training set sizes of 1, 2, 5, 10, 20 and 40 per class.

	1	2	5	10	20	40
Original data	20.00% ±0.00%	20.00% ±0.00%	58.90% ±35.69%	92.45% ±1.05%	94.80% ±1.39%	96.80% ±0.65%
Blur	58.00% ±7.12%	69.45% ±9.09%	86.20% ±4.78%	93.90% ±1.39%	95.20% ±1.08%	97.55% ±0.65%
Gaussian	70.90% ±2.73%	72.70% ±6.09%	87.95% ±2.24%	95.00% ±1.03%	95.05% ±2.56%	97.25% ±0.31%
Scaling	44.65% ±4.75%	56.30% ±6.51%	84.60% ±7.37%	91.20% ±1.12%	93.60% ±1.76%	96.20% ±0.93%
GAN	47.95% ±12.69%	58.25% ±9.90%	83.90% ±9.58%	92.30% ±1.65%	95.15% ±0.60%	97.35% ±0.55%

accuracy levels. In this regime, most augmentation strategies achieved performance comparable to or slightly exceeding the baseline, while the Scaling method consistently underperformed relative to the other techniques. Overall, these results indicate that data augmentation is most beneficial under conditions of limited training data, with Gaussian and Blur methods showing the strongest and most consistent performance gains.

4. DISCUSSION

This study systematically examined the effectiveness of four representative data augmentation strategies (localized blurring, Gaussian noise addition, random amplitude scaling, and GAN-based synthesis) under small-sample conditions in Raman spectral classification of breast cells. By jointly analyzing distributional similarity and downstream classification performance across multiple parameter settings, this work provides insight into how augmentation mechanisms interact with data scarcity and model learning behavior, and under which conditions different strategies are most suitable.

The similarity analysis based on FID and t-SNE revealed clear and consistent differences among the augmentation methods. Gaussian noise addition produced augmented spectra that most closely matched the statistical distribution of the

original data, followed by localized blurring, whereas GAN-based synthesis exhibited the largest distributional deviation. These observations can be explained by the nature of the perturbations introduced. Gaussian noise and localized blurring impose small, smooth, and physically plausible variations that resemble experimental uncertainty commonly present in Raman measurements, such as detector noise or minor instrumental fluctuations. Consequently, the global spectral structure and characteristic peak positions are preserved. In contrast, GAN-based augmentation relies on learning the underlying data distribution from a very limited number of samples, which constrains its ability to accurately capture the true spectral manifold and may result in distributional shifts.

The classification results demonstrate that distributional similarity is closely linked to practical performance gains. Augmentation methods that maintained high similarity to the original data, particularly Gaussian noise and localized blurring, consistently improved classification accuracy relative to the baseline model trained on the original dataset. Random amplitude scaling, despite preserving spectral shape, generally failed to improve performance. This behavior suggests that uniform intensity scaling may distort intensity-dependent biochemical information and weaken class-discriminative features. Although GAN-generated data formed well-separated class clusters in feature space, their limited overlap with real data indicates a mismatch between synthetic and authentic spectra, which may hinder generalization when models are evaluated on real-world measurements.

The parameterization study further highlights that the effectiveness of data augmentation depends on both the number of synthetic samples and the size of the original training set. For augmentation size, an optimal range was observed for Gaussian, Blur, and GAN methods, beyond which performance gains saturated or declined. This trend indicates that excessive synthetic data may introduce redundancy or amplify biases inherent in the small original dataset, rather than providing additional informative variability. In contrast, increasing the amount of scaled data did not compensate for the limitations of the scaling transformation itself, reinforcing that augmentation quality is more critical than quantity.

Varying the size of the original training set provides additional insight into the role of augmentation under different data availability scenarios. When the number of real training samples was extremely limited, models trained without augmentation exhibited severe overfitting and unstable performance. Under these conditions, all augmentation methods mitigated overfitting to some extent, with Gaussian noise and localized blurring producing the most stable and substantial improvements. GAN-based augmentation showed variable effectiveness, reflecting its sensitivity to both data availability and augmentation size, while scaling remained largely ineffective. As the number of original samples increased, the relative benefit of augmentation gradually diminished, indicating diminishing returns once the training data became sufficiently representative. Nevertheless, Gaussian and Blur methods continued to provide modest improvements across a wide range of training set sizes, demonstrating their robustness.

Overall, these findings indicate that simple, physics-inspired augmentation strategies are generally more reliable than complex generative models for small-scale Raman spectral datasets. Gaussian noise addition is particularly effective in extremely data-scarce settings, while localized blurring performs well when a moderate amount of real data is available. GAN-based augmentation can enhance performance when sufficient initial data and appropriate augmentation sizes are used, but its effectiveness is more sensitive to parameter selection. Random amplitude scaling does not appear to be a suitable augmentation strategy for Raman-based breast cell classification. These results emphasize that data augmentation should be treated as a data-dependent and parameter-sensitive process rather than a fixed preprocessing step.

Several limitations of this study should be noted. First, all experiments were conducted using a single deep learning architecture; extending the analysis to other classifiers, such as support vector machines, alternative convolutional networks, or Transformer-based models, would help assess the generality of the conclusions. Second, the experiments focused on a single breast cell Raman dataset, and validation on additional biomedical Raman datasets, including those involving microorganisms or viral samples, would further strengthen the findings. Finally, only a basic conditional GAN architecture was explored. Future work incorporating more advanced generative models or physics-informed constraints may improve synthetic data quality under limited data conditions.

Data availability

The training and test database were set up on the cell line samples. Raw data to reproduce **Figure 2** can be shared upon request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supporting Information

The Supporting Information is available free of charge at .

REFERENCES

1. Sharma, M. P.; Shukla, S.; Misra, G., Recent advances in breast cancer cell line research. **2024,** *154* (10), 1683-1693.

2. Yin, P.; Lian, X.; Wu, X.; Xiao, Y.; Feng, C.; Lv, Y.; Yi, L.; Liang, M.; Ge, G.; Dmitriy, K.; Hu, B., Raman Peak Features Matching: Enhancing Spectral Analysis through Feature Augmentation. *Analytical Chemistry* **2025,** *97* (16), 8801-8812.

3. Hajab, H.; Anwar, A.; Nawaz, H.; Majeed, M. I.; Alwadie, N.; Shabbir, S.; Amber, A.; Jilani, M. I.; Nargis, H. F.; Zohaib, M.; Ismail, S.; Kamal, A.; Imran, M.,

Surface-enhanced Raman spectroscopy of the filtrate portions of the blood serum samples of breast cancer patients obtained by using 30 kDa filtration device. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2024,** *311*, 124046.

4. Ma, M.; Zhang, J.; Liu, Y.; Wang, X.; Han, B., Advances in the clinical application of Raman spectroscopy in breast cancer. *Applied Spectroscopy Reviews* **2024,** *59* (10), 1459-1493.

5. Spaziani, S.; Esposito, A.; Barisciano, G.; Quero, G.; Elumalai, S.; Leo, M.; Colantuoni, V.; Mangini, M.; Pisco, M.; Sabatino, L.; De Luca, A. C.; Cusano, A., Combined SERS-Raman screening of HER2-overexpressing or silenced breast cancer cell lines. *Journal of Nanobiotechnology* **2024,** *22* (1), 350.

6. Li, J.; Wang, X.; Min, S.; Xia, J.; Li, J., Raman spectroscopy combined with convolutional neural network for the sub-types classification of breast cancer and critical feature visualization. *Computer Methods and Programs in Biomedicine* **2024,** *255*, 108361.

7. Liu, X.; Jia, Y.; Zheng, C., Recent progress in Surface-Enhanced Raman Spectroscopy detection of biomarkers in liquid biopsy for breast cancer. **2024,** *Volume 14 - 2024*.

8. Wang, M.; Zhang, K.; Yue, L.; Liu, X.; Lai, Y.; Zhang, H., Robust Diagnosis of Breast Cancer Based on Silver Nanoparticles by Surface-Enhanced Raman Spectroscopy and Machine Learning. *ACS Applied Nano Materials* **2024,** *7* (11), 13672-13680.

9. Yin, P.; Li, G.; Zhang, B.; Farjana, H.; Zhao, L.; Qin, H.; Hu, B.; Ou, J.; Tian, J., Facile PEG-based isolation and classification of cancer extracellular vesicles and particles with label-free surface-enhanced Raman scattering and pattern recognition algorithm. *Analyst* **2021,** *146* (6), 1949-1955.

10. Liu, T.; Chen, J.; Kong, L.; Li, X.; Chen, X., Utilization of a portable Raman spectrometer combined with a PCA-SVM model for starch type differentiation. *Food Bioscience* **2024,** *57*, 103465.

11. Kang, S.; Kim, I.; Vikesland, P. J., Discriminatory Detection of ssDNA by Surface-Enhanced Raman Spectroscopy (SERS) and Tree-Based Support Vector Machine (Tr-SVM). *Analytical Chemistry* **2021,** *93* (27), 9319-9328.

12. Du, Y.; Han, D.; Liu, S.; Sun, X.; Ning, B.; Han, T.; Wang, J.; Gao, Z., Raman spectroscopy-based adversarial network combined with SVM for detection of foodborne pathogenic bacteria. *Talanta* **2022,** *237*, 122901.

13. Ouyang, Q.; Fan, Z.; Chang, H.; Shoaib, M.; Chen, Q., Analyzing TVB-N in snakehead by Bayesian-optimized 1D-CNN using molecular vibrational spectroscopic techniques: Near-infrared and Raman spectroscopy. *Food Chemistry* **2025,** *464*, 141701.

14. Lim, J.; Shin, G.; Shin, D., Fast Detection and Classification of Microplastics below 10 μm Using CNN with Raman Spectroscopy. *Analytical Chemistry* **2024,** *96* (17), 6819-6825.

15. Wan, Y.; Jiang, Y.; Zheng, W.; Li, X.; Sun, Y.; Yang, Z.; Qi, C.; Zhao, X., Rapid and high accuracy identification of culture medium by CNN of Raman spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2025,** *329*, 125608.

16. Kang, Z.; Li, Y.; Liu, J.; Chen, C.; Wu, W.; Chen, C.; Lv, X.; Liang, F., H-CNN combined with tissue Raman spectroscopy for cervical cancer detection. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2023,** *291*, 122339.

17. Zhang, Y.; Li, Z.; Li, Z.; Wang, H.; Regmi, D.; Zhang, J.; Feng, J.; Yao, S.; Xu, J., Employing Raman Spectroscopy and Machine Learning for the Identification of Breast Cancer. *Biological Procedures Online* **2024,** *26* (1), 28.

18. Zeng, Q.; Chen, C.; Chen, C.; Song, H.; Li, M.; Yan, J.; Lv, X., Serum Raman spectroscopy combined with convolutional neural network for rapid diagnosis of HER2-positive and triple-negative breast cancer. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2023,** *286*, 122000.

19. Nunekpeku, X.; Zhang, W.; Gao, J.; Adade, S. Y.-S. S.; Li, H.; Chen, Q., Gel strength prediction in ultrasonicated chicken mince: Fusing near-infrared and Raman spectroscopy coupled with deep learning LSTM algorithm. *Food Control* **2025,** *168*, 110916.

20. Wu, X.; Du, Z.; Ma, R.; Zhang, X.; Yang, D.; Liu, H.; Zhang, Y., Qualitative and quantitative studies of phthalates in extra virgin olive oil (EVOO) by surface-enhanced Raman spectroscopy (SERS) combined with long short term memory (LSTM) neural network. *Food Chemistry* **2024,** *433*, 137300.

21. Chen, X.; Shen, J.; Liu, C.; Shi, X.; Feng, W.; Sun, H.; Zhang, W.; Zhang, S.; Jiao, Y.; Chen, J.; Hao, K.; Gao, Q.; Li, Y.; Hong, W.; Wang, P.; Feng, L.; Yue, S., Applications of Data Characteristic AI-Assisted Raman

Spectroscopy in Pathological Classification. *Analytical Chemistry* **2024,** *96* (16), 6158-6169.

22. Chen, T.; Baek, S.-J., Library-Based Raman Spectral Identification Using Multi-Input Hybrid ResNet. *ACS Omega* **2023,** *8* (40), 37482-37489.

23. Ho, C.-S.; Jean, N.; Hogan, C. A.; Blackmon, L.; Jeffrey, S. S.; Holodniy, M.; Banaei, N.; Saleh, A. A. E.; Ermon, S.; Dionne, J., Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature Communications* **2019,** *10* (1), 4927.

24. Xie, Y.; Yang, S.; Zhou, S.; Liu, J.; Zhao, S.; Jin, S.; Chen, Q.; Liang, P., SE-ResNet-based classifier for highly similar mixtures based on Raman spectrum: Classification for alcohol systems as an example. **2023,** *54* (2), 191-200.

25. Chang, M.; He, C.; Du, Y.; Qiu, Y.; Wang, L.; Chen, H., RaT: Raman Transformer for highly accurate melanoma detection with critical features visualization. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2024,** *305*, 123475.

26. Wang, Z.; Li, Y.; Zhai, J.; Yang, S.; Sun, B.; Liang, P., Deep learning-based Raman spectroscopy qualitative analysis algorithm: A convolutional neural network and transformer approach. *Talanta* **2024,** *275*, 126138.

27. Zhou, X.; Chen, C.; Lv, X.; Zuo, E.; Li, M.; Wu, L.; Chen, X.; Wu, X.; Chen, C., CMACF: Transformer-based cross-modal attention cross-fusion model for systemic lupus erythematosus diagnosis combining Raman spectroscopy, FTIR spectroscopy, and metabolomics. *Information Processing & Management* **2024,** *61* (6), 103804.

28. Ozer, I.; Cetin, O.; Gorur, K.; Temurtas, F., Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset. *Neural Computing and Applications* **2021,** *33* (21), 14975-14989.

29. Li, G.; Li, C.; Wang, C.; Wang, Z., Suboptimal capability of individual machine learning algorithms in modeling small-scale imbalanced clinical data of local hospital. *PLOS ONE* **2024,** *19* (2), e0298328.

30. Shimakawa, H.; Kumada, A.; Sato, M., Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *npj Computational Materials* **2024,** *10* (1), 11.

31. Yan, C.; Feng, X.; Wick, C.; Peters, A.; Li, G., Machine learning assisted discovery of new thermoset shape memory polymers based on a small training dataset. *Polymer* **2021,** *214*, 123351.

32. Zhao, J.; Lui, H.; Kalia, S.; Lee, T. K.; Zeng, H., Improving skin cancer detection by Raman spectroscopy using convolutional neural networks and data augmentation. **2024,** *Volume 14 - 2024*.

33. Qi, Y.; Hu, D.; Zheng, M.; Jiang, Y.; Chen, Y. P., Deep learning assisted Raman spectroscopy for rapid identification of 2D materials. *Applied Materials Today* **2024,** *41*, 102499.

34. Luo, J.; Wu, Q.; Cao, J.; Fang, H.; Xu, C.; He, D., Comparison of data augmentation and classification algorithms based on plastic spectroscopy. *Analytical Methods* **2025,** *17* (6), 1236-1251.

35. Deng, L.; Zhong, Y.; Wang, M.; Zheng, X.; Zhang, J., Scale-Adaptive Deep Model for Bacterial Raman Spectra Identification. *IEEE Journal of Biomedical and Health Informatics* **2022,** *26* (1), 369-378.

36. Flanagan, A. R.; Glavin, F. G., A Comparative Analysis of Data Synthesis Techniques to Improve Classification Accuracy of Raman Spectroscopy Data. *Journal of Chemical Information and Modeling* **2024,** *64* (7), 2311-2322.

37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y., Generative adversarial networks. **2020,** *63* (11 %J Commun. ACM), 139–144.

38. Gracia Moisés, A.; Vitoria Pascual, I.; Imas González, J. J.; Ruiz Zamarreño, C., Data Augmentation Techniques for Machine Learning Applied to Optical Spectroscopy Datasets in Agrifood Applications: A Comprehensive Review. **2023,** *23* (20), 8562.

39. Wu, M.; Wang, S.; Pan, S.; Terentis, A. C.; Strasswimmer, J.; Zhu, X., Deep learning data augmentation for Raman spectroscopy cancer tissue classification. *Scientific Reports* **2021,** *11* (1), 23842.

40. Pavlou, E.; Kourkoumelis, N., Deep adversarial data augmentation for biomedical spectroscopy: Application to modelling Raman spectra of bone. *Chemometrics and Intelligent Laboratory Systems* **2022,** *228*, 104634.

41. Safir, F.; Vu, N.; Tadesse, L. F.; Firouzi, K.; Banaei, N.; Jeffrey, S. S.; Saleh, A. A. E.; Khuri-Yakub, B. T.; Dionne, J. A., Combining Acoustic Bioprinting with AI-Assisted Raman Spectroscopy for High-Throughput Identification of Bacteria in Blood. *Nano Letters* **2023,** *23* (6), 2065-2073.

42. Yu, Q.; Shen, X.; Yi, L.; Liang, M.; Li, G.; Guan, Z.; Wu, X.; Castel, H.; Hu, B.; Yin, P.; Zhang, W., Fragment-Fusion Transformer: Deep Learning-Based Discretization Method for Continuous Single-Cell Raman Spectral Analysis. *ACS Sensors* **2024,** *9* (8), 3907-3920.