Applied Materials Today 41 (2024) 102499

Contents lists available at ScienceDirect

Applied Materials Today

journal homepage: www.elsevier.com/locate/apmt

Deep learning assisted Raman spectroscopy for rapid identification of 2D materials

Yaping Qi [a,b,c,1,*], Dan Hu [b,1], Ming Zheng [d], Yucheng Jiang [e,*], Yong P. Chen [a,b,c,f,*]

[a] Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Sendai 980-8577, Japan
[b] Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, Av. Wai Long, Macau SAR, 999078, China
[c] Institute of Physics and Astronomy and Villum Center for Hybrid Quantum Materials and Devices, Aarhus University, Aarhus-C, 8000 Denmark
[d] School of Materials Science and Physics, China University of Mining and Technology, Xuzhou 221116, China
[e] Jiangsu Key Laboratory of Micro and Nano Heat Fluid Flow Technology and Energy Application, School of Physical Science and Technology, Suzhou University of Science and Technology, Suzhou, Jiangsu 215009, China
[f] Department of Physics and Astronomy and Elmore Family School of Electrical and Computer Engineering and Birck Nanotechnology Center and Purdue Quantum Science and Engineering Institute, Purdue University, West Lafayette, IN 47907, United States

ARTICLE INFO

Keywords:
Deep learning
Raman spectroscopy
Convolutional neural network
2D materials
Data augmentation
Denoising diffusion probabilistic models

ABSTRACT

Two-dimensional (2D) materials have attracted extensive attention due to their unique characteristics and potential applications. Raman spectroscopy, as a rapid and non-destructive probe, exhibits distinct features and holds notable advantages in the characterization of 2D materials. However, traditional data analysis of Raman spectra relies on manual interpretation and feature extraction, which are both time-consuming and subjective. In this work, we employ deep learning techniques, including classificatory and generative deep learning, to assist the analysis of Raman spectra of representative 2D materials. For the limited and unevenly distributed Raman spectral data, we propose a data augmentation approach based on Denoising Diffusion Probabilistic Models (DDPM) to augment the training dataset and construct a four-layer Convolutional Neural Network (CNN) for 2D material classification. The proposed CNN model achieves an impressive accuracy of 98.8 % on the original dataset. Experiments illustrate the effectiveness of DDPM in addressing data limitations and significantly improving the performance of the classification model. Notably, when enhanced with DDPM-augmented data, the DDPM-CNN method shows high reliability, with 100 % classification accuracy. Our work demonstrates the practicality of deep learning-assisted Raman spectral analysis for high-precision recognition and classification of 2D materials, presenting a promising avenue for rapid and automated materials analysis via spectroscopy.

1. Introduction

Since the discovery of graphene in 2004, an ever-expanding family of two-dimensional (2D) materials has been discovered and explored [1,2]. Due to their unique physical and chemical properties, 2D materials have garnered significant attention in the scientific community and [3] have exhibited tremendous potential in an extensive range of applications [4-6]. To investigate the diverse properties of 2D materials, it is essential to characterize their basic structures and compositions.

Raman spectroscopy is commonly employed as a measurement technique for identifying and analyzing 2D materials. It has been widely used in analytical sciences due to its sensitivity and non-invasive nature [7-9]. However, conventional Raman spectroscopy analysis often involves laborious efforts and human intervention for data interpretation [10-12]. While manual identification by visual inspection of Raman spectra may be feasible for small datasets, a limited number of possible choices of 2D materials, and relatively simple heterostructures, it becomes less practical for increased choices and varieties of 2D materials, and more complex, multi-layered stacked heterostructures. For example, as the number of constituent materials and stacking sequences increases in heterostructures involving three or more distinct layers, the Raman spectra can become more complex with peak overlaps and even become significantly altered due to interlayer interactions, easily obscuring the unique spectral signatures of individual materials and making it very

* Corresponding authors.
E-mail addresses: qi.yaping.a2@tohoku.ac.jp (Y. Qi), jyc@usts.edu.cn (Y. Jiang), yongchen@purdue.edu (Y.P. Chen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.apmt.2024.102499
Received 1 August 2024; Received in revised form 2 October 2024; Accepted 26 October 2024

hard to resolve them using manual inspection and conventional methods. Additionally, in high throughput or industrial applications, quality control processes, or other situations where large-scale experimental datasets with numerous spectra must be analysed rapidly and accurately, manual inspection is neither scalable nor reliable. Therefore, automated solutions, such as deep learning-based methods, are essential to systematically extract subtle spectral features and achieve efficient and precise classification in these challenging scenarios. To address these challenges, there has been a growing interest in integrating machine learning techniques with Raman spectroscopy [13-16]. In the high-throughput or industrial applications, efficient analysis through Raman spectroscopy is crucial, particularly in areas such as quality control in product manufacturing, where rapid and accurate assessment of a large quantity of materials is necessary [17,18]. Moreover, the complexity of material structures in industrial applications, involving multiple materials, makes the use of Raman spectroscopy combined with machine learning highly valuable for rapid analysis of materials, interfaces, and configurations.

Machine learning has attracted growing interest among researchers for its ability to analyze complex spectral data [13,19,20]. For instance, algorithms such as random forest, kernel ridge regression, and multi-layer perceptron have been applied to the study of 2D materials, such as identifying and characterizing monolayer MoS_2 as well as twist angles of twisted bilayer graphene [21-24]. Despite these advances, it is important to note that conventional machine learning methods largely depend on manual preprocessing and feature engineering of spectra to achieve optimal performance [25]. To address these limitations, researchers are applying deep learning to assist Raman spectroscopy analysis [13,25]. A prime example is the application of convolutional neural networks (CNN), which have been successfully employed in high-speed Raman imaging for the rapid identification of carbon nanotubes [26]. Furthermore, CNN has demonstrated its efficacy in accurately identifying hundreds of mineral categories and in distinguishing spectra of materials that are highly similar in subtly different environments [27,28].

Nevertheless, deep learning generally requires extensive training datasets to optimize the network parameters and mitigate overfitting risks [29]. Data augmentation methods such as generative adversarial networks (GAN), often applied when there may be insufficient number of datasets, have shown promise in reducing overfitting and improving the accuracy of classification algorithms [30-32]. However, some researchers observed that GAN trades off diversity for fidelity to produce high-quality data samples but cannot cover the whole distribution of features in abundant sample scenarios [33-35]. In response to this limitation, our study proposes a novel approach, integrating a denoising diffusion probabilistic model (DDPM) with a 1D CNN-based classifier [35]. This hybrid model aims to efficiently and accurately identify distinct types of 2D materials and their stacked combinations using Raman spectroscopy.

In this research, we explore the fusion of classification-focused deep learning and generative deep learning methodologies for the identification of various 2D materials through Raman spectroscopy. In response to the challenge of limited and non-uniformly distributed experimental Raman data of 2D materials, we implement advanced data augmentation strategies to substantially expand the number of training samples. The expansion is crucial for enhancing the performance of classification algorithms. Considering the characteristic diversity of Raman spectral features even for one material on varied substrates, we have constructed a DDPM based on ResNet for data augmentation. Subsequently, we developed a four-layer CNN for the automatic classification of spectra. This approach holds the potential to streamline the experimental process, reduce human intervention, and facilitate automated analysis of Raman spectra of 2D materials, particularly in complex situations such as multilayer stacked heterostructures or dealing with a large number of possible choices of constituent 2D materials.

2. Materials and overall framework

2.1. Raman spectral data

This article defines the task of identifying various categories of 2D materials as a multi-class classification problem. In this study, we utilize a dataset that comprises a total of 594 experimental Raman spectra for seven distinct 2D materials and three stacked combinations chosen as examples to demonstrate our methodology: Black phosphorus (BP), Graphene, Molybdenum disulfide (MoS_2), Rhenium disulfide (ReS_2), Tellurium (Te), Tungsten diselenide (WSe_2), Tungsten ditelluride (WTe_2), BP–WSe_2 stack (S_1), Te-ReS_2-WSe_2-Graphene stack (S_2), and Te-WSe_2-WTe_2 stack (S_3). It is noteworthy that the spectral features of these materials might exhibit a range of variations, as the spectra of each material are obtained on more than one type of substrate. The number of experimental Raman spectra for 2D materials studied in this work is summarized in Table 1.

2.2. Overall framework

In practical applications, many issues may limit the amount of acceptable spectral data obtained from experiments. For example, there may be limited experimental samples or insufficient measurement resources, or the weak Raman signals of substances measured may be difficult to separate from the background [36]. To address this issue, our framework introduces a novel data augmentation-based approach for Raman spectroscopy-based 2D material classification, as illustrated in Fig. 1. It primarily consists of the following two components:

1. Data Augmentation Module: To augment the limited training dataset (experimental spectra), we employ DDPM to generate synthetic (data) samples for each category of materials. This process generates a substantial number of independently and identically distributed data samples based on the original Raman spectral dataset. The objective is to assist the classification model in accurately and efficiently identifying various types of 2D materials.
2. Data Classification Module: Combining the original spectral data samples with those generated by DDPM, the data classifier learns to determine the category of each sample. In our study, we construct a four-layer CNN to classify each sample into their respective categories and compared it with other commonly used classification methods (comparison results are shown in the experiments and results section) such as Artificial Neural Networks (ANN), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR) [37,38].

Table 1
Statistics of Raman spectral dataset of 2D materials studied in this work. S_1 to S_3 refer to various heterostructure stacks (see text for details).

Materials	Quantity of spectra
BP	35
Graphene	209
MoS_2	8
ReS_2	15
Te	270
WSe_2	6
WTe_2	28
S_1	8
S_2	7
S_3	8
Total:	594

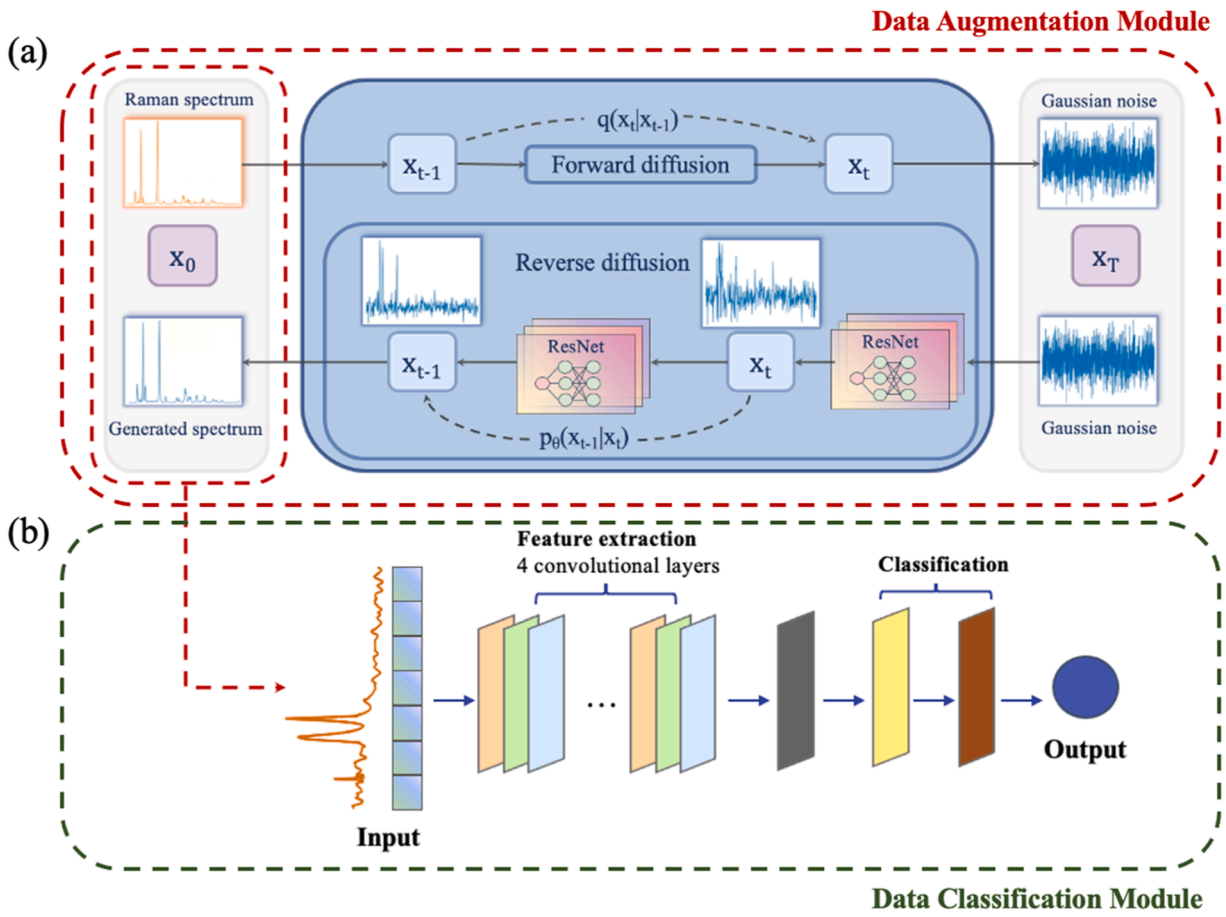

Fig. 1. Illustration of the DDPM-based data augmentation and classification framework for Raman Spectra of 2D materials. (a) Data augmentation module based on DDPM. (b) Spectral classification module based on 1D CNN.

3. Methodology

3.1. Data augmentation module

Firstly, this study employs DDPMs for data augmentation. Earlier studies have used DDPMs to synthesize high-quality data [39-42]. Diffusion probabilistic models were first introduced by Sohl-Dickstein et al. [43]. They defined a Markov chain of diffusion steps to construct desired data samples by adding random noise to data and then learning to reverse the diffusion process. Subsequently, Ho et al. (2020) [35] proposed DDPM, a simplified diffusion model driven by the connection between denoising diffusion models and denoising fractional matching. DDPM utilizes a multi-step Markov chain process to create synthetic data samples by initially adding Gaussian noise to the original data in a forward diffusion process, gradually transforming it into a noise distribution. The model then learns to reverse this diffusion process through a series of denoising steps, progressively removing the noise to reconstruct and generate new data samples that accurately reflect the features of the original distribution [44].

DDPM is composed of two processes: forward diffusion (left to right) and reverse diffusion (right to left), as shown in Fig. 1(a). Forward diffusion is a process of adding noise to the input data, represented by q, which is fixed to a Markov chain from data x_0 to the latent variables x_1, ……, x_T:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

The sampling noise latent based on the input x_0 at an arbitrary step t can be expressed by defining $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{t=0}^{T} \alpha_t$, where β_1, ……, β_T are the noise schedule consisting of a set of linearly increasing constants:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_1 \tag{2}$$

$$x_t = \sqrt{\overline{\alpha_t}} x_0 + \sqrt{1-\overline{\alpha_t}} z_t \quad z_t \sim N(0, I) \tag{3}$$

where $1 - \overline{\alpha_t}$ demonstrates the variance of noise for an arbitrary time step. Given sufficiently large time step T, the latent x_T tends to the standard normal distribution $x_T \sim \mathcal{N}(0, \boldsymbol{I})$.

The reverse diffusion process is also defined as a Markov chain from the Gaussian noise input x_T to x_{T-1}, ……, x_0. According to $q(x_T)$, we can sample the reverse steps $q(x_{t-1}|x_t)$. Here, we use p_θ to indicate the reverse process:

$$p_\theta(x_{0:T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{4}$$

Using Bayes theorem, the diffusion process can be represented by the known quantities from the forward process, and it can be proved that $p_\theta(x_{t-1}|x_t, x_0)$ is also a Gaussian distribution:

$$p_\theta(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0) \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{5}$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \widetilde{\mu}_t(x_t, x_0), \widetilde{\beta}_t \boldsymbol{I}), \quad \widetilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \beta_t \tag{6}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t \tag{7}$$

The relationship between x_0 and x_t is already obtained in the forward process:

$$\tilde{\mu}_t(x_t) = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\overline{z}_t\right) \tag{8}$$

Since the noise $\overline{z}_t$ at time step t depends on the entire forward training process, it is hard to estimate. Therefore, we constructed residual networks (ResNet) based on the DiffWave model presented by Kong et al. (2020) to approximate the distribution of $\overline{z}_t$ in the reverse process. The structure of the ResNet is illustrated in Fig. 2, it consists of eight residual layers and utilizes skip connections to connect the entire network.

The input of the model consists of two parts: input diffusion noise and step embedding, where the model generates different diffusion results for different values of step t. Step embedding is a positional embedding introduced by Vaswani et al. [45], and in this study, we utilize it for the time step. The diffusion noise is fed into a 1D convolutional layer, while step t is input into a two-layer fully connected layer, where the parameters of these two parts are shared. Subsequently, step t is mapped to an embedding vector through the third fully connected layer, and together with the diffusion noise, it is added to the input of each residual layer in the model. Each residual layer utilizes one convolutional layer for feature extraction. The obtained features are then activated by gated activation units and passed through a pointwise convolutional layer. The output of the pointwise convolutional layer is divided into two parts along the channel dimension: one part is the input of the next residual layer, while the other is directly output through a skip connection, where the output module consists of two convolutional layers.

3.2. Data classification module

From the data augmentation module, we can obtain a set of new samples for each class (material type) of the original spectral data, which will be utilized to enhance the performance of the classifier. In our classification module, we employ a 1D CNN as the core component. The neural network architecture constructed for classification is illustrated in Fig. 3. The convolutional layer is crucial in CNNs for feature extraction. It convolves input data with trainable filters, directly influencing model performance. More convolutional layers allow the learning of additional features but increase training time. Each convolutional layer consists of trainable filters (kernels) that slide over input data, performing convolution operations. The process can be denoted as:

$$c_j^l = f\left(\sum_{i\in E_j} c_j^{l-1} * k_{ij}^l + b_j^l\right) \tag{9}$$

where * represents the convolution operation, l denotes the current convolutional layer, c_j^l is the output of j_{th} feature map, k_{ij}^l is the convolutional kernel, E_j represents the input feature maps, f is the activation function, and b is the bias. The convolutional kernel hyperparameters are randomly initialized and optimized iteratively for optimal performance.

Our CNN model uses four convolutional layers to extract data features and uses Leaky ReLU as the activation function. Subsequently, a flattening layer is applied to transform the multi-dimensional input data into a set of 1D vectors, which is then fed into a fully connected layer. The fully connected layer receives the output from the convolutional and pooling layers and maps the learned features to a predefined vector space for feature classification. The expression for the fully connected layer is as follows:

$$h_{w,b}(x) = f(w^T x + b) \tag{10}$$

Where h represents the output of the current neuron, x denotes the 1D feature vector input, and w corresponds to the weight vector connected to the neuron. Finally, there is a Softmax function with an output dimension equal to the number of classes. The Softmax function takes a set of 1D vectors as input and normalizes them into a probability distribution.

The classifier uses sparse categorical cross-entropy to calculate loss, which is expressed as follows:

$$loss = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} y_{ij} log\hat{y}_{ij} \tag{11}$$

where m denotes the number of samples and k denotes the number of categories, y_{ij} means the real label of Raman data (if sample i belongs to class j then y_{ij} is 1, else 0) and $\hat{y}_{ij}$ means the probability of model predicts the sample i belongs to class j.

4. Experiments and results

4.1. Data preprocessing

To ensure consistent dimensionality for the input of the model, we employ a simple spline interpolation technique to convert each Raman spectrum into a vector of 851 intensity values, within the wavenumber range of 50–1750 cm $^{-1}$. This range is selected to maximize the

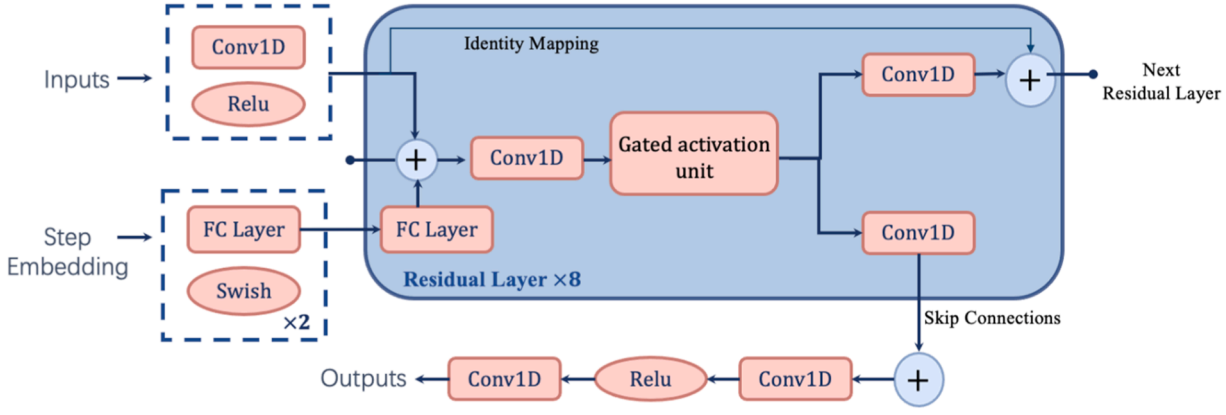

Fig. 2. An illustration of the ResNets architecture. In this schematic, "FC" denotes the fully connected layer. "Relu" represents the rectified linear unit (Relu) activation function. The "gated activation unit" consists of the hyperbolic tangent (tanh) activation function and the sigmoid activation function, it can be denoted as $\boldsymbol{tanh}(\boldsymbol{W}_{f,k}*\boldsymbol{x}) \odot \boldsymbol{\sigma}(\boldsymbol{W}_{g,k}*\boldsymbol{x})$, where W represents a convolutional filter, f and g represent the filter and gate, respectively, and k represents the layer index.

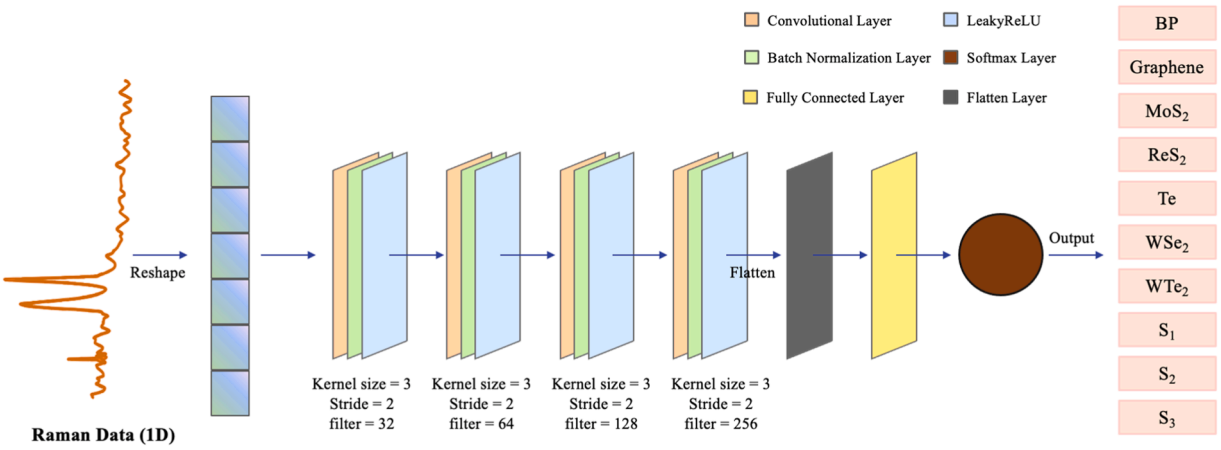

Fig. 3. The architecture of the four-layer CNN for Raman spectra-based classification.

information within each spectrum, and effectively encompass the characteristic peaks necessary for differentiating the Raman spectra of studied 2D materials. For spectra that do not cover the entire range of wavenumbers, the missing intensity values are padded with zeros. Finally, the dataset is normalized (to intensity values between 0 and 1) to establish consistent scaling across all features, thus preparing for model training and analysis.

4.2. Implementation settings

The experiments in this study are conducted on the Windows 11 operating system using Python 3.9 programming language. The hardware used for the experiments includes a 12th Gen Intel(R) Core (TM) i7–12,700 CPU and a Nvidia GeForce RTX 3060 Ti GPU.

In the data augmentation module during the training of ResNet, each

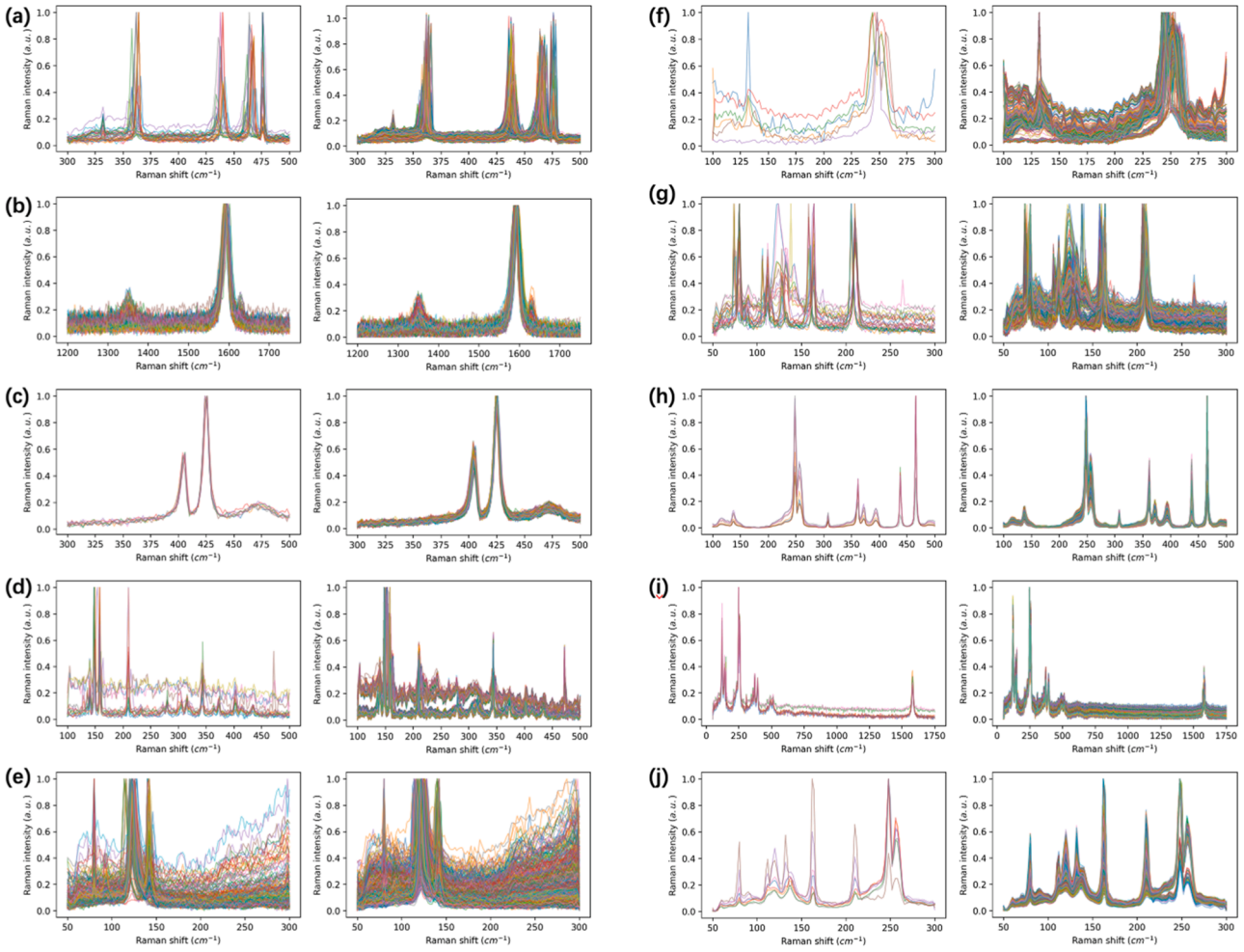

Fig. 4. The Raman spectra of various 2D materials before and after data augmentation using DDPM: (a) BP, (b) Graphene, (c) MoS_2, (d) ReS_2, (e) Te, (f) WSe_2, (g) WTe_2, (h) BP–WSe_2 stack (S_1), (i) Te-ReS_2-WSe_2-Graphene stack (S_2), and (j) Te-WSe_2-WTe_2 stack (S_3). The left side of each panel represents the original Raman spectra dataset, while the right represents the augmented Raman spectra dataset.

convolutional layer has a kernel size of 3, and the channel dimension within the residual blocks is set to 128. The generation of diffusion noise follows a linear schedule spanning 50 steps, with the range of β_t values set between 0.0001 and 0.02. For the proposed four-layer CNN in the classification module, the kernel size is set to 3, and a stride of 2 is applied. The number of filters is configured as 32, 64, 128, and 256 for the respective layers. All neural network models are optimized using the Adam optimizer with an initial learning rate of 0.0002. The model undergoes training for 100 epochs, utilizing a batch size of 32. The training and test datasets, consisting of 10,000 synthetic spectra combined with 594 experimental spectra, were divided in a 4:1 ratio, with 8475 spectra (80 %) allocated for training and 2119 spectra (20 %) reserved for testing.

4.3. Generated data

To ensure a comprehensive augmented dataset, the experiment utilizes the best-saved model to generate Raman spectra for each trained diffusion model. Fig. 4 shows the Raman spectra of ten categories of materials included in the dataset, as well as the generated Raman spectra. We employ DDPM to augment 1000 spectral data for each type of 2D material, generating a total of 10,000 Raman spectrum samples for further analysis. It can be observed that DDPM can generate diverse synthetic spectra that closely resemble the features of the original spectra. Additionally, DDPM exhibits the ability to fill in new data within a specific range based on original data (extrapolate data within predefined limits using the original data as a reference). This capability enables the comprehensive capture of all characteristics and improves the diversity of the dataset.

Furthermore, we employ the t-SNE dimensionality reduction technique to visualize the original and augmented data (Fig. 5), providing insight into the high-dimensional structure of the dataset in a lower-dimensional space. The proximity of data points in the t-SNE plot reflects their similarity in the original high-dimensional space, thus if two data points are close in the t-SNE visualization, their spectral features are likely to be similar. This is clearly demonstrated in Fig. 5(b), where the augmented data clusters closely with the original spectral data for each material category, suggesting that the augmented data preserves the intrinsic properties of the original spectra. Moreover, it can be observed that the features of augmented data for different categories exhibit more distinct boundaries in the low-dimensional space than original dataset. The clearer demarcation between clusters of different materials indicates that DDPM augmentation method does not compromise the inherent characteristics of each category but rather enriches the dataset in a way that can improve the performance of classification algorithms. Therefore, integrating the generated spectra into the dataset facilitates more diverse and comprehensive analysis, enabling a robust evaluation of deep learning-assisted methods.

4.4. Results and analysis

To assess the advantages of the proposed model, the experiments conduct the following baselines for comparison: RF, SVM, KNN, LR, and an ANN model with two hidden layers. For the multi-class classification task to identify 2D materials, we evaluated the performance of the model on the original dataset as well as augmented dataset respectively using the average accuracy, precision, and recall of ten-fold cross-validation as evaluation metrics. Table 2 reports the performance comparisons between the proposed method with the baselines (the DDPM prefix indicates the use of the enhanced dataset, otherwise the original dataset). It is worth noting that CNN and DDPM-CNN exhibit superior classification performance compared to other models in the evaluation. Specifically, CNN achieves an exceptional accuracy rate of 98.8 % without data augmentation, surpassing most other models. This indicates its ability to accurately classify data and exhibit good generalization.

Fig. 6 supplements this evaluation with an array of confusion matrices for different algorithms, where CNN displays higher accuracy in classifying most categories. The reported accuracy, precision, and recall metrics were obtained through a ten-fold cross-validation on a dataset comprising 10,000 synthetic spectra generated via DDPM and 594 experimental spectra. The combined dataset was divided into ten (nearly) equal-sized subsets, with nine used for training and remaining one for validation (alternating to a different subset for each iteration). The final metrics were averaged across all 10 iterations to provide a robust and unbiased assessment of the model's generalization performance. Such results corroborate that deep learning methods can

Table 2
The average performance of ten-fold cross-validation comparisons between the proposed methods (with DDPM) vs. baselines (no DDPM, using only original data).

Method	Accuracy	Precision	Recall
CNN	0.988	0.945	0.937
DDPM-CNN	1.000	1.000	1.000
ANN	0.946	0.658	0.646
DDPM-ANN	1.000	1.000	1.000
RF	0.906	0.566	0.574
DDPM-RF	1.000	1.000	1.000
SVM	0.966	0.829	0.786
DDPM-SVM	1.000	1.000	1.000
KNN	0.953	0.826	0.770
DDPM-KNN	0.988	0.989	0.988
LR	0.960	0.731	0.711
DDPM-LR	1.000	1.000	1.000

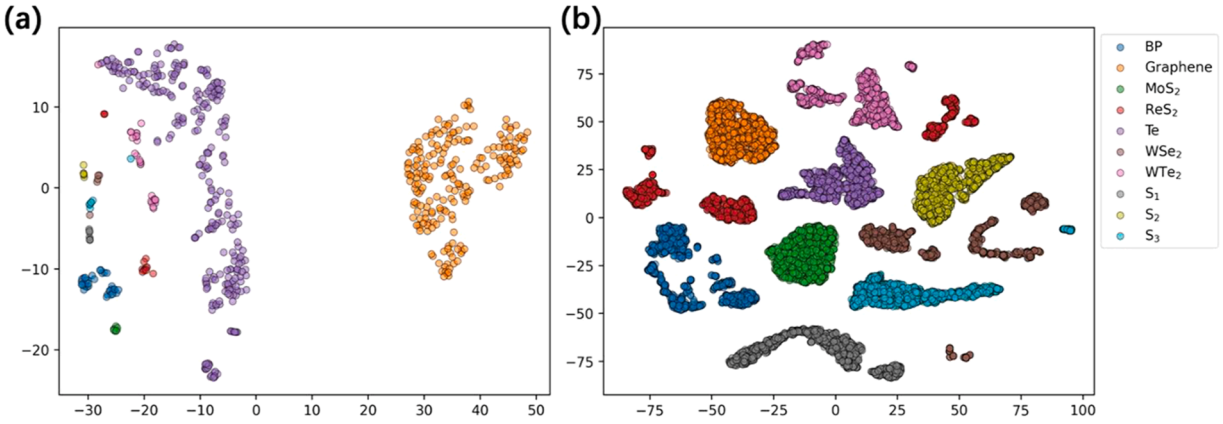

Fig. 5. t-SNE plots for (a) the original dataset and (b) the augmented dataset (including original spectral data) of different 2D materials. [S_1: BP–WSe_2 stack, S_2: Te-ReS_2-WSe_2-Graphene stack, S_3: Te-WSe_2-WTe_2 stack.].

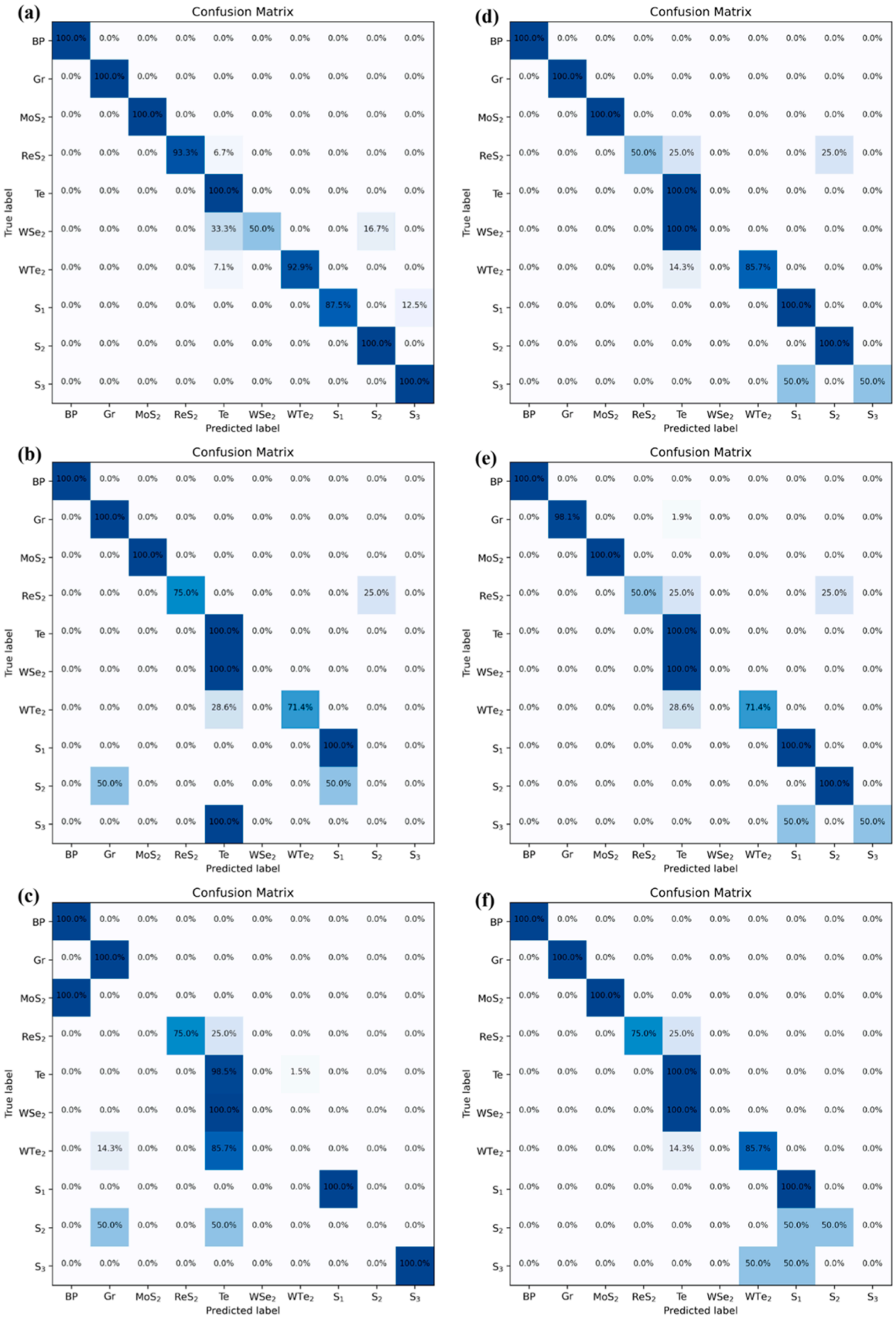

Fig. 6. Confusion matrices depicting the average accuracy of ten-fold cross-validation in the classification of each category by different algorithms: (a) CNN, (b) ANN, (c) RF, (d) SVM, (e) KNN, and (f) LR. The diagonal elements represent the percentage of true positives, which is a key indicator of the algorithm's ability to correctly identify each category. The off-diagonal elements represent misclassification rates.

effectively extract sample features even when training data is limited. Conversely, conventional machine learning techniques such as RF and KNN tend to underperform when relying on raw Raman data as input features, which may limit the exploitation of inter-feature correlations, thereby affecting classifier performance. Moreover, with a precision of 94.5 % and a recall of 93.7 %, it indicates that CNN still faces challenges in accurately identifying positives and capturing all True Positive instances.

In comparison, models such as ANN, RF, SVM, KNN, and LR, demonstrate varying levels of performance in accuracy, precision, and recall. Although these models may be slightly inferior to CNN, their performance nonetheless indicates their capabilities for Raman-based 2D material recognition. It is worth emphasizing that the incorporation of the DDPM for data augmentation method significantly enhances the performance across all evaluated models. Notably, the average accuracy in ten-fold cross-validation of DDPM-ANN and DDPM-RF models ascended from 94.6 % and 90.6 % to 100 %. This highlights the effectiveness of DDPM in refining algorithm performance in Roman-based 2D material recognition. The comparison results are shown in Fig. 7.

However, the advantages of DDPM-CNN are not prominent due to significant differences among the data categories used in this study. All DDPM-based conventional machine learning can achieve remarkable results. Typically, deep neural networks require larger datasets to reach their optimal performance, so further validation of its performance can be conducted on more complex datasets with smaller inter-category differences that are more difficult to distinguish.

The DDPM-based data augmentation module proposed in this study significantly enhances sample density and diversity, enabling the classifier to establish decision boundaries more effectively. As a result, it outperforms baseline models. Higher sample density, in comparison to sparse data, often allows classifiers to learn more precise boundaries. Overall, the utilization of DDPM-based data augmentation has the potential to be a valuable technique in materials science. It has ability to generate realistic spectra and improve the recognition capabilities of classification models. The findings underscore the effectiveness of leveraging data augmentation methods for more accurate and robust 2D material recognition, ultimately contributing to the progress and exploration of novel materials in the scientific community.

5. Conclusion

This study explores the application of deep learning techniques to assist in identifying different 2D materials based on Raman spectroscopy. In response to the challenge of limited data availability, we employ data augmentation techniques to substantially augment the training samples to improve the effectiveness of the classification. We have constructed a DDPM-based augmentation model with ResNet, which effectively addresses data distribution, promotes diversity, and boosts the performance of all classification models, including CNN, ANN, RF, SVM, KNN, and LR. The four-layer CNN model that we constructed demonstrates exceptional performance in this study, achieving classification accuracies of 98.8 % without data augmentation and a score of 100 % accuracy upon integrating DDPM-based data augmentation. These outcomes highlight the practicality of the proposed data augmentation approach, enabling high-precision identification of 2D materials even in small-scale data tasks. Furthermore, this study is the inaugural application of the DDPM in spectral generation, presenting a novel tool for data augmentation in Raman spectroscopy and other spectral analysis. It can simplify the experimental process, reduce human intervention, and facilitate automated analysis of spectroscopy, thus paving a new avenue for further research in this domain.

Researchers in this field can leverage our methodology to enhance the robustness and accuracy of material identification tasks, especially in instances where data is scarce or expensive to acquire. However, it's important to acknowledge that the current tool has certain limitations. While our approach significantly improves classification performance, it still relies on the quality and representativeness of the initial dataset. Future investigations could delve into refining and expanding upon our approach, exploring its applicability across diverse spectroscopic techniques and materials, thus advancing the capabilities of material characterization and analysis. Additionally, the computational resources required for training deep learning models with DDPM augmentation may pose challenges for researchers with limited access to high-performance computing infrastructure. Addressing these challenges and exploring additional avenues for improvement, such as transfer learning, multimodal integration, and real-time analysis, will be essential for realizing the full potential of deep learning in material science. Moreover, it would be worthwhile to extend the current approach to include layer numbers (thickness) and polytypes, and twist angle identification in complex heterostructures. This would further broaden the utility and impact of our approach, enabling more advanced structural characterizations and fostering deeper insights and new advancements in 2D materials research.

Data availability

The data are available from the corresponding authors upon reasonable request. https://github.com/Naduhi/Raman_DL

CRediT authorship contribution statement

Yaping Qi: Conceptualization, Investigation, Methodology, Project administration, Resources, Funding acquisition, Software, Validation, Supervision, Writing – original draft, Writing – review & editing. **Dan Hu:** Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Ming Zheng:** Resources, Writing – original draft, Writing – review & editing. **Yucheng Jiang:** Conceptualization, Project administration, Resources, Data curation, Writing – original draft, Writing – review & editing. **Yong P. Chen:** Conceptualization, Project administration, Resources, Supervision, Writing – review & editing.

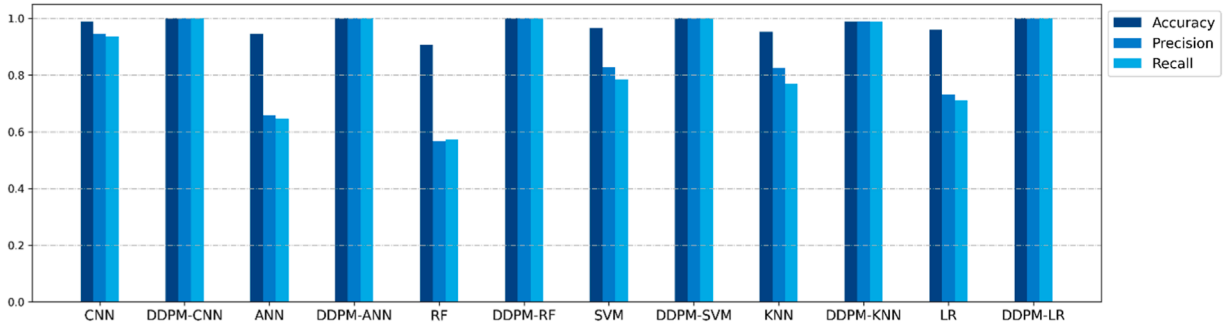

Fig. 7. Bar chart of the average performance of ten-fold cross-validation between the proposed methods (with DDPM) vs. baselines.

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.

Acknowledgments

We acknowledge partial support of this work by Macau Science and Technology Development Fund (FDCT Grants 0031/2021/ITP) and MUST-FRG, and by JSPS KAKENHI (no. 22H00278), Tohoku University TUMUG Startup Research Fund and FY2023 AIMR Fusion Research Fund, and FY2024 AIMR Fusion Research Fund. We also thank the support of Villum Foundation (Grant No. 25931). Y. Jiang acknowledges support by the National Natural Science Foundation of China (Grant No. 12274316 and the Natural Science Foundation of Jiangsu Province for Distinguished Young Scholar (BK20240048). We also thank Z. Wu, G. Cheng and R. Pasechnik for helpful discussions.

References

[1] K.S. Novoselov, A.K. Geim, S.V. Morozov, D.-e. Jiang, Y. Zhang, S.V. Dubonos, I. V. Grigorieva, A.A. Firsov, Electric field effect in atomically thin carbon films, Science 306 (5696) (2004) 666–669, https://doi.org/10.1126/science.1102896.
[2] G.R. Bhimanapati, Z. Lin, V. Meunier, Y. Jung, J. Cha, S. Das, D. Xiao, Y. Son, M. S. Strano, V.R. Cooper, Recent advances in two-dimensional materials beyond graphene, ACS nano 9 (12) (2015) 11509–11539, https://doi.org/10.1021/acsnano.5b05556.
[3] Y. Qi, M.A. Sadi, D. Hu, M. Zheng, Z. Wu, Y. Jiang, Y.P. Chen, Recent progress in strain engineering on van der waals 2d materials: tunable electrical, electrochemical, magnetic, and optical properties, Adv. Mater. 35 (12) (2023) 2205714, https://doi.org/10.1002/adma.202205714.
[4] S. Salahuddin, K. Ni, S. Datta, The era of hyper-scaling in electronics, Nat. Electron. 1 (8) (2018) 442–450, https://doi.org/10.1038/s41928-018-0117-x.
[5] G. Fiori, F. Bonaccorso, G. Iannaccone, T. Palacios, D. Neumaier, A. Seabaugh, S. K. Banerjee, L. Colombo, Electronics based on two-dimensional materials, Nat. Nanotechnol. 9 (10) (2014) 768–779, https://doi.org/10.1038/nnano.2014.207.
[6] S. Zhang, W. Zhou, Y. Ma, J. Ji, B. Cai, S.A. Yang, Z. Zhu, Z. Chen, H. Zeng, Antimonene oxides: emerging tunable direct bandgap semiconductor and novel topological insulator, Nano Lett. 17 (6) (2017) 3434–3440, https://doi.org/10.1021/acs.nanolett.7b00297.
[7] P.R. Griffiths, J.M. Chalmers, Handbook of Vibrational Spectroscopy, Wiley Online Library, 2002.
[8] N. Colthup, Introduction to Infrared and Raman spectroscopy, Elsevier, 2012.
[9] I. Childres, Y. Qi, M.A. Sadi, J.F. Ribeiro, H. Cao, Y.P. Chen, Combined raman spectroscopy and magneto-transport measurements in disordered graphene: correlating raman d band and weak localization features, Coatings 12 (8) (2022) 1137, https://doi.org/10.3390/coatings12081137.
[10] T. Vincent, K. Kawahara, V. Antonov, H. Ago, O. Kazakova, Data cluster analysis and machine learning for classification of twisted bilayer graphene, Carbon N Y 201 (2023) 141–149, https://doi.org/10.1016/j.carbon.2022.09.021.
[11] A.I. Pérez-Jiménez, D. Lyu, Z. Lu, G. Liu, B. Ren, Surface-enhanced Raman spectroscopy: benefits, trade-offs and future developments, Chem. Sci. 11 (18) (2020) 4563–4577, https://doi.org/10.1039/D0SC00809E.
[12] V. Stanev, K. Choudhary, A.G. Kusne, J. Paglione, I. Takeuchi, Artificial intelligence for search and discovery of quantum materials, Communicat. Mater. 2 (1) (2021) 105, https://doi.org/10.1038/s43246-021-00209-z.
[13] Y. Qi, D. Hu, Y. Jiang, Z. Wu, M. Zheng, E.X. Chen, Y. Liang, M.A. Sadi, K. Zhang, Y. P. Chen, Recent progresses in machine learning assisted raman spectroscopy, Adv. Opt. Mater. (2023) 2203104, https://doi.org/10.1002/adom.202203104.
[14] B. Zhou, L.Y. Sun, T. Fang, H.X. Li, R. Zhang, A.P. Ye, Rapid and accurate identification of pathogenic bacteria at the single-cell level using laser tweezers Raman spectroscopy and deep learning, J. Biophotonics (2023).
[15] F. Lussier, V. Thibault, B. Charron, G.Q. Wallace, J.F. Masson, Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering, Trac-Trends in Analytical Chemistry 124 (2020), https://doi.org/10.1016/j.trac.2019.115796.
[16] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J. Billinge, Recent advances and applications of deep learning methods in materials science, npj Comput. Mater. 8 (1) (2022) 59, https://doi.org/10.1038/s41524-022-00734-6.
[17] K.A. Esmonde-White, M. Cuellar, C. Uerpmann, B. Lenain, I.R. Lewis, Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing, Anal. Bioanal. Chem. 409 (3) (2017) 637–649, https://doi.org/10.1007/s00216-016-9824-1.
[18] K.A. Esmonde-White, M. Cuellar, I.R. Lewis, The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing, Anal. Bioanal. Chem. 414 (2022) 969–991, https://doi.org/10.1007/s00216-021-03727-4.
[19] G.R. Schleder, A.C. Padilha, C.M. Acosta, M. Costa, A. Fazzio, From DFT to machine learning: recent approaches to materials science–a review, J. Phys. Mater. 2 (3) (2019) 032001, https://doi.org/10.1088/2515-7639/ab084b.
[20] D.M. Dimiduk, E.A. Holm, S.R. Niezgoda, Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering, Integr Mater Manuf Innov 7 (2018) 157–172, https://doi.org/10.1007/s40192-018-0117-8.
[21] Y. Mao, N.N. Dong, L. Wang, X. Chen, H.Q. Wang, Z.X. Wang, I.M. Kislyakov, J. Wang, Machine learning analysis of raman spectra of MoS2, Nanomaterials 10 (11) (2020) 2223, https://doi.org/10.3390/nano10112223.
[22] Y. He, Y. Ju, Q. Wang, Insights into optical detection and three-dimensional characterization of monolayer molybdenum disulfide thin films based on machine learning, Appl. Surf. Sci. 565 (2021), https://doi.org/10.1016/j.apsusc.2021.150530.
[23] N. Sheremetyeva, M. Lamparski, C. Daniels, B. Van Troeye, V. Meunier, Machine-learning models for Raman spectra analysis of twisted bilayer graphene, Carbon N Y 169 (2020) 455–464, https://doi.org/10.1016/j.carbon.2020.06.077.
[24] P. Solís-Fernández, H. Ago, Machine learning determination of the twist angle of bilayer graphene by Raman spectroscopy: implications for van der Waals heterostructures, ACS Appl. Nano Mater. 5 (1) (2022) 1356–1366, https://doi.org/10.1021/acsanm.1c03928.
[25] J. Liu, M. Osadchy, L. Ashton, M. Foster, C.J. Solomon, S.J. Gibson, Deep convolutional neural networks for Raman spectrum recognition: a unified solution, Analyst 142 (21) (2017) 4067–4074, https://doi.org/10.1039/C7AN01371J.
[26] J. Zhang, M.L. Perrin, L. Barba, J. Overbeck, S. Jung, B. Grassy, A. Agal, R. Muff, R. Bronnimann, M. Haluska, C. Roman, C. Hierold, M. Jaggi, M. Calame, High-speed identification of suspended carbon nanotubes using Raman spectroscopy and deep learning, Microsyst Nanoeng 8 (2022) 19, https://doi.org/10.1038/s41378-022-00350-w.
[27] X.C. Sang, R.G. Zhou, Y.C. Li, S.J. Xiong, One-dimensional deep convolutional neural network for mineral classification from raman spectroscopy, Neural Process Lett 54 (2022) 677–690, https://doi.org/10.1007/s11063-021-10652-1.
[28] Z. Chen, Y. Khaireddin, A.K. Swan, Identifying the charge density and dielectric environment of graphene using Raman spectroscopy and deep learning, Analyst 147 (9) (2022) 1824–1832, https://doi.org/10.1039/D2AN00129B.
[29] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J Big Data 8 (2021) 53, https://doi.org/10.1186/s40537-021-00444-8.
[30] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621 (2017). https://doi.org/10.48550/arXiv.1712.04621.
[31] M. Wu, S. Wang, S. Pan, A.C. Terentis, J. Strasswimmer, X. Zhu, Deep learning data augmentation for Raman spectroscopy cancer tissue classification, Sci. Rep. 11 (2021) 23842, https://doi.org/10.1038/s41598-021-02687-0.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural. Inf. Process Syst. (2014) 27, https://doi.org/10.1145/3422622.
[33] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural. Inf. Process Syst. 34 (2021) 8780–8794.
[34] G. Müller-Franzes, J.M. Niehues, F. Khader, S.T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, S. Nebelung, J.N. Kather, A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis, Sci. Rep. 13 (2023) 12098, https://doi.org/10.1038/s41598-023-39278-0.
[35] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural. Inf. Process Syst. 33 (2020) 6840–6851.
[36] C. Muehlethaler, M. Leona, J.R. Lombardi, Review of surface enhanced Raman scattering applications in forensic science, Anal. Chem. 88 (1) (2016) 152–169, https://doi.org/10.1021/acs.analchem.5b04131.
[37] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 2016 3rd international conference on computing for sustainable global development (INDIACom), Ieee, 2016, pp. 1310–1315.
[38] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerg. Artificial Intellig. Appl. Comput. Eng. 160 (1) (2007) 3–24.
[39] Z. Kong, W. Ping, J. Huang, K. Zhao, B. Catanzaro, DiffWave: a versatile diffusion model for audio synthesis, in: International Conference on Learning Representations, 2020.
[40] Z. Chen, Diffusion models-based data augmentation for the cell cycle phase classification, in: Journal of Physics: Conference Series, IOP Publishing, 2023 012001.
[41] Z. Dorjsembe, S. Odonchimed, F. Xiao, Three-dimensional medical image synthesis with denoising diffusion probabilistic models, Medical Imaging with Deep Learning (2022).
[42] E. Adib, A.S. Fernandez, F. Afghah, J.J. Prevost, Synthetic ecg signal generation using probabilistic diffusion models, IEEE Access (2023).
[43] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, in: Deep unsupervised learning using nonequilibrium thermodynamics, International conference on machine learning, PMLR, 2015, pp. 2256–2265.
[44] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, Adv. Neural. Inf. Process Syst. 32 (2019).
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural. Inf. Process Syst. (2017) 30, https://doi.org/10.48550/arXiv.1706.03762.