Imbalanced Spectral Data Analysis using Data Augmentation based on the Generative Adversarial Network

Jihoon Chung[1], Junru Zhang[1], Amirul Islam Saimon[1], Yang Liu[1], Blake N. Johnson[1] and Zhenyu (James) Kong[1*]

[1]Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, US.

*Corresponding author. E-mail: zkong@vt.edu

Abstract

Spectroscopic techniques generate one-dimensional spectra with distinct peaks and specific widths in frequency domain. These features act as unique identities for material characteristics. Deep neural networks (DNNs) has recently been considered a powerful tool for automatically categorizing experimental spectra data by supervised classification to evaluate material characteristics. However, most existing work assumes balanced spectral data among various classes in the training data, contrary to actual experiments, where the spectral data is usually imbalanced. The imbalanced training data deteriorates the supervised classification performance. To address this issue, this paper applies a novel data augmentation method based on a generative adversarial network (GAN) proposed by the authors in their prior work. To demonstrate the effectiveness of the proposed method, the actual imbalanced spectral data from Pluronic F-127 hydrogel and Alpha Cyclodextrin hydrogel are used to classify the phases of data. Specifically, our approach improves 7%, 5%, and 5% of the performance of the existing data augmentation method regarding the classifier's F-score, Precision, and Recall on average, respectively. Specifically, the method consists of three players: the generator, discriminator, and classifier. It generates samples that are not only authentic but emphasize the differentiation between material characteristics to provide balanced training data, improving the classification results. Based on these validated results, we expect the method's broader applications in addressing imbalanced measurement data across diverse domains in materials science and chemical engineering.

Keywords: Spectral Data, Generative Adversarial Network (GAN), Deep Neural Network, Material Phase Classification, Imbalanced Data.

1 Introduction

Spectroscopic technologies such as X-ray diffraction (XRD), Nuclear Magnetic Resonance (NMR), Raman scattering, and Electrical Impedance Spectral (EIS) are fundamental tools for the characterization of experimental samples in chemistry and materials science. XRD has found extensive use throughout industry and research laboratories for more than a century (Friedrich et al., 1913). It is proven to be an effective method for characterizing crystalline materials as it captures detailed information on the long-range periodic nature of crystal structures. In contrast, NMR and Raman measurements are more strongly dependent on localized chemical interactions and are widely used to characterize the structure of molecular materials (Callaghan, 1993; Smith and Dent, 2019). EIS is a technique used to determine the impedance characteristics of an electrochemical interface. It has been used increasingly in biomaterials studies to understand the interactions between the surface and the biological environment. While their mechanisms and uses may vary, all of these spectroscopic methods generate comparable one-dimensional spectra consisting of unique peak positions, widths, and intensities. These features often serve as "fingerprints" for material characteristics, including patterns and phases (Wang et al., 2020; Schuetzke et al., 2023). Identification of the characteristics of unknown specimens can be achieved by comparing newly measured spectra with those of established materials in experimental databases (Belsky et al., 2002; Armbruster and Danisi, 2015). However, the analysis process is complicated by factors such as measurement noise, background signals, and inherent minor deviations in the spectra (Schuetzke et al., 2021). To automate this process, machine learning has recently emerged as an effective tool since it can automatically classify experimental spectra along material characteristics with significant accuracies (Choudhary et al., 2022; Szymanski et al., 2021).

The popular method within the domain of machine learning is deep neural networks (DNNs). These networks consist of several layers of artificial neurons designed to mimic the structure and functioning of the human brain (McCulloch and Pitts, 1943). DNNs is widely used in classification tasks of spectral data as they can automatically extract discriminating features. Specifically, DNN is utilized for supervised classification methods since these methods can use the label information of each class (i.e., material characteristics of spectral data), providing accurate classification results. For example, Kantz et al. (2019) used DNNs to classify Liquid Chromatography-Mass Spectrometry (LC-MS) spectral peak shapes. This approach improves peak filtering performance by reducing the false peaks by more than 90% compared to the traditional chemometric methods. Zeng et al. (2021) utilized one-dimensional

convolutional neural networks (CNNs) to classify the visible-near infrared spectra of corn seed to evaluate seed viability. In addition, Lee et al. (2020) developed a CNNs-based model to classify interested phases from a mixture of inorganic compounds using XRD. Similarly, Schuetzke et al. (2021) built a robust CNN model for automatically classifying phases using the XRD patterns. This shows superior performance in automatic phase identification of cement compounds and iron ores. These studies assumed balanced training spectral data between classes (i.e., material characteristics of spectral data) in their supervised classification methods.

However, the balanced spectral data is difficult to appear in actual chemistry, physics, and industries generating the spectral data. For example, medical diagnostic applications often generate imbalanced spectral data reflecting the common asymmetry encountered in health status among screened individuals (e.g., more true negatives than true positives are typically encountered in preventative diagnostics). Materials science and chemistry applications also often generate imbalanced spectral data reflecting the common asymmetry of composition-process-structure-property relations, such as associated with phase equilibrium (e.g., the physics governing the thermodynamics of mixtures often results in asymmetric distributions of stable, unstable, and transition states with respect to varying mixture composition). For example, it is common to encounter samples of one type in accelerated materials discovery applications based on the unknown structure of a material design space and the initially selected search parameters, which may be done randomly or based on prior knowledge. As such, imbalanced spectral data is inevitably generated mainly in actual experiments and industries. However, the imbalanced spectral data leads to compromised supervised classification performance using DNN. Specifically, the prediction in classification models tends to be biased towards the majority class, where the class with the sizable spectral data samples. This leads to a high probability of misclassifying samples from the minority class (Chung et al., 2023).

To address this significant challenge arising from imbalanced spectral data in classification utilizing DNNs, a viable solution is to employ data augmentation techniques to create a balanced training dataset across spectral data of different material characteristics. Basic data augmentation methods, including rotation, flipping, and synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) are commonly used for balancing training data within the classification due to their straightforward implementation (Cui et al., 2020; Lee et al., 2020; Mycroft et al., 2020). However, these techniques primarily take into account localized information, thus failing to capture the complete data distribution and address the challenge of overfitting (Douzas and Bacao, 2018; Mikołajczyk and Grochowski, 2018). Consequently, these methods are unsuitable for generating realistic spectral data with various characteristics (Fathy et al., 2020; Ranasinghe and Parlikad, 2019). In contrast, there has been a growing trend in the active utilization of Generative Adversarial Networks

(GAN) (Goodfellow et al., 2014) for data augmentation in spectral data analysis. The reason for this adoption is rooted in the GAN's capacity to generate more authentic data by comprehensively learning the entire data distribution of actual data through two neural networks: the discriminator and the generator (Antoniou et al., 2017; Kiyasseh et al., 2020). For example, Wu et al. (2021) used a GAN framework to augment synthetic Raman Spectroscopy data of skin cancer tissue to address the difficulties of class imbalance in the context of cancer tissue data. Similarly, Gao et al. (2022) utilized GAN to generate seizure events in long-term EEG spectra to overcome the data imbalance problem for accurate classification.

Although these studies generate realistic spectral data to provide balanced data among the various material characteristics, they do not consider generating the samples enabling differentiation between characteristics (i.e., characteristics-distinguishable samples) that can further improve the classification performance. Since the ultimate goal of generating the data is to well classify the spectral data of different material characteristics from the imbalanced data, the generated data needs to improve the classification performance. The characteristics-distinguishable samples can be generated by joint optimization between GAN and the classifier. Specifically, the classifier guides the generator in GAN to create samples that could benefit classification results. Regarding this direction, we proposed a novel data augmentation method in a recent paper (Chung et al., 2023) that jointly optimizes between GAN and the classifier with several stabilizing techniques. The method validated its effectiveness in imbalanced data in additive manufacturing processes. Therefore, we apply the method to spectral data to address the imbalanced spectral data issue that commonly occurs in actual experiments and industries generating spectral data. In this paper, the effectiveness of our method is validated by using the spectral data collected from actual experimentation. Specifically, the electrical impedance spectral data from Pluronic F-127 hydrogel (Zhang et al., 2023) and Alpha Cyclodextrin hydrogel are used. The phases of spectral data are provided as imbalanced. The results show that the imbalanced spectral data can be successfully overcome by our method in the classification of the phases. In particular, our approach enhances the F-score, Precision, and Recall of the classifier by an average of 7%, 5%, and 5%, respectively, compared to the benchmark methods. Moreover, the technique has great generality. Thus, it can be further applied to address the classification with imbalanced spectral data in other material science or chemical engineering domains.

2 Results

Several real-world case studies are provided to show the effectiveness of our method in imbalanced spectral data analysis. In Sections 2.1 and 2.2, comparative case studies involving benchmark methods are provided. Specifically, spectral data from two actual materials, Pluronic F-127 hydrogel and Alpha Cyclodextrin hydrogel, are provided in Sections 2.1 and 2.2, respectively. In

addition, the imbalanced spectral data regarding the material phases are provided to evaluate the performance. Therefore, the material characteristics that need to be classified are the material phases in the case studies. The performance assessment is conducted based on the classification results obtained from the imbalanced training dataset. All case studies utilize the Keras with TensorFlow backend. The experiments are carried out on an NVIDIA Tesla P4 GPU within the Google Colab environment (Bisong and Bisong, 2019).

1) Benchmark Methods: Regarding the benchmark methods, both sampling-based and GAN-based approaches are used. Within the sampling-based category, two techniques that SMOTE (Chawla et al., 2002) and B-SMOTE (Han et al., 2005) are used. These methods are implemented using the Python imbalanced-learn library. For the GAN-based approaches, two state-of-the-art methods, namely, CDRAGAN (Huang and Jafari, 2021), and BAGAN-GP (Huang and Jafari, 2021) are selected. In addition, Cooperative GAN (Choi et al., 2021), which jointly optimizes GAN and the classifier without stabilizing technique, is utilized as one of the benchmark methods. Finally, the baseline is established by evaluating the classification performance without employing any data augmentation method.

2) Performance Evaluation Measure: The assessment of performance is determined by the classifier's F-score, Precision, and Recall (Powers, 2020). Convolutional neural networks (CNN) with hyperparameters listed in Table 8 is used as a classifier. The F-score expressed in Eq. (1) is a composite metric that combines both Precision and Recall.

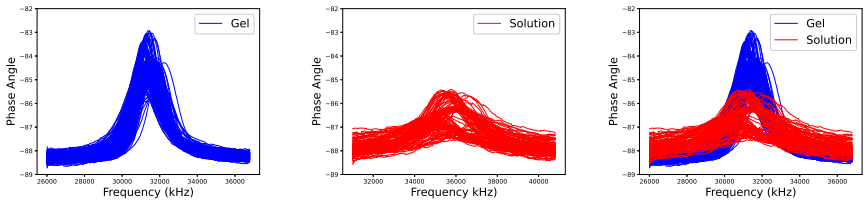
$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{1}$$

As the primary goal of this paper is to enhance classification accuracy using imbalanced training data, it includes case studies that encompass different balanced ratios. A balanced ratio refers to the proportion between the training data size of the minority and majority classes. To ensure robustness and reliability, each case study is iterated ten times, and the resulting performance measure is based on the average performance across all classes from these ten repetitions.

2.1 Case Study using Spectral Data from Pluronic F-127 Hydrogel

Pluronic F-127, a nonionic amphiphilic surfactant, demonstrates a reversible thermogelling process in aqueous solutions, resembling the behavior observed in other Pluronic compounds (Jalaal et al., 2017). These aqueous solutions are hereafter referred to as PF-127. In this section, PF-127 hydrogel libraries are used for the case study. It's been widely used and studied in a wide range of applications. 96 PF-127 deionized water mixtures with different mass ratios are formulated in the 96 well plates. The concentration of PF-127 deionized water varies from 0.3125 wt.% to 30 wt.% with an increment of 0.3125 wt.%. The

phase angle-frequency spectrum of each sample is collected by a sensor-based high-throughput method. The collected spectra are labeled as solution or gel to study the composition-property relationships of PF-127 hydrogels. Three repeated experiments provide 288 spectral data. Specifically, 181 spectral data of gel (Fig. 1 (a)) and 107 of solution (Fig. 1 (b)) are utilized for the case study. The detailed procedure of data collection is described in Section 4.1.



(a) (b) (c)

Fig. 1: Spectral data of Pluronic F-127 hydrogel from (a) gel; (b) solution; (c) gel & solution.

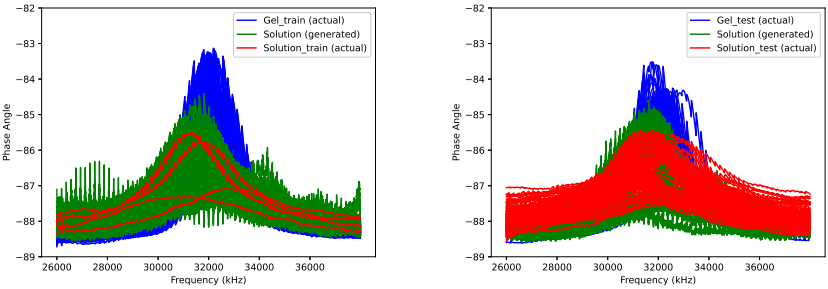
Table 1 describes the imbalanced training data, where the balanced ratios between two phases are 0.027. The ratio is set because balanced ratios below this threshold result in significantly poor performance for the classifier. The remaining data sets are used as testing data.

Table 1: Imbalanced training data samples in Pluronic F-127 hydrogel case studies.

Majority Class	Minority Class	Balanced Ratio	Majority Class Training Samples	Minority Class Training Samples
Gel	Solution	0.027	150	4

Fig. 2 shows the actual and generated samples from the proposed method, respectively. Specifically, Fig. 2 (a) describes the actual imbalanced training data in Table 1, while Fig. 2 (b) represents the actual testing data. The generated samples in Fig. 2 are realistic spectral data with apparent differences between phases achieved through a combination of adversarial and collaborative learning in our method. Specifically, the results show that the generated samples from our approach successfully learn the features of the test data of the solution phase (Fig. 2 (b)) from the small number of training data samples (Fig. 2 (a)).

Table 2 shows the performance evaluation of the benchmark and our methods. Compared to a baseline result that uses only imbalanced data as training data of the classifier, the sampling-based method, including B-SMOTE (Han



(a) (b)

Fig. 2: Comparison between generated data of Pluronic F-127 hydrogel with (a) actual training data; (b) actual testing data.

et al., 2005) and SMOTE (Chawla et al., 2002) tend to exhibit similar or worse performance. This is because the small number of minority class samples prevents the generation of various data from sampling-based methods.

Table 2: Performance evaluation in Pluronic F-127 hydrogel case study. Averages and standard deviations (in the parenthesis) are represented.

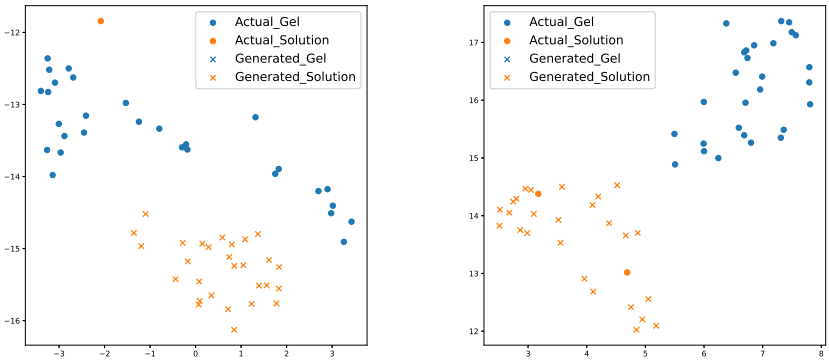
	Pluronic F-127 hydrogel		
	Precision	**Recall**	**F-score**
Baseline	0.797 (0.04)	0.888 (0.04)	0.808 (0.05)
SMOTE	0.791 (0.04)	0.880 (0.05)	0.796 (0.08)
B-SMOTE	0.792 (0.04)	0.880 (0.05)	0.798 (0.08)
CDRAGAN	0.837 (0.07)	0.915 (0.04)	0.852 (0.07)
BAGAN-GP	0.824 (0.06)	0.907 (0.04)	0.840 (0.07)
Cooperative GAN	0.808 (0.06)	0.891 (0.05)	0.816 (0.08)
Proposed	**0.849** (0.05)	**0.923** (0.03)	**0.871** (0.05)

Conversely, GAN-based approaches typically outperform sampling-based methods because their generators learn the actual distribution of samples from minority classes and generate diverse training data for the classifier. In particular, the generator from our method provides more diverse and better-quality images than other GAN-based methods by jointly optimizing the classifier with stabilizing techniques, resulting in improvements in classification results. Specifically, our method improves 2%∼9%, 1%∼5%, and 1%∼7% of the performance of the benchmark methods regarding their F-score, Precision, and Recall, respectively. Compared to the baseline that trains the CNN without any data augmentation, our method achieves 7.7%, 3.8%, and 6.8% improvements regarding its F-score, Precision, and Recall, respectively.

The statistical hypothesis test is provided to check the significance of our method compared to the baseline. Specifically, paired t-test (Hsu and Lachenbruch, 2014) is utilized to check the significance of the performance. As shown in Table 3, our method shows statistically significant improvements over the baseline at a 99% significance level.

Table 3: P-value of statistical hypothesis test in Pluronic F-127 hydrogel case studies.

	Precision	Recall	F-score
Paired t-test (Proposed $\geq$ Baseline)	0.003	0.003	0.002



(a) (b)

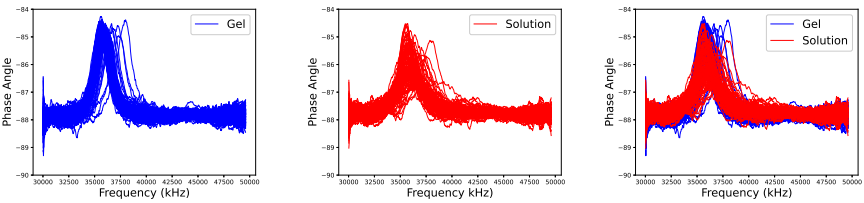
Fig. 3: t-SNE of the feature from the intermediate layer of the classifier from our method in epochs (a) 0 and (b) 140.

Fig. 3 illustrates the efficacy of the generated samples produced by our approach by comparing their features in classifier with those of actual samples. Specifically, Fig. 3 displays the t-distributed Stochastic Neighbourhood Embedding (t-SNE) of the feature extracted from the intermediate layer of our method's classifier. t-SNE is a nonlinear dimensionality reduction technique designed for visualizing high-dimensional data by projecting it into lower-dimensional spaces (Dimitriadis et al., 2018). In Fig. 3, '$\bullet$' represents t-SNE of the features from the intermediate layer of classifiers extracted from actual samples, while '$\times$' represents features from the generated samples within the balanced training batch. To achieve a balanced training batch, there is an abundance of '$\times$' instances for the minority class (i.e., the solution phase) in

each batch. In Fig. 3 (a), it is evident that the distribution patterns between actual and generated samples are distinct at epoch 0. Specifically, the '•' of the solution phase is not aligned with '×' of its phase. Furthermore, it is aligned with the '•' of the gel phase. Because our approach is designed to generate realistic and distinguishable samples between the phases, the features extracted from the generated samples (denoted as '×') accurately align with those from the actual samples (represented as '•') based on their respective phases at epoch 140 (Fig. 3 (b)). Furthermore, the features associated with each phase are distinctly separated. This observation confirms the realistic and phase-discriminative characteristics of the generated samples produced by our method. By employing balanced training data characterized by these attributes, our method attains a high level of classification performance.

2.2 Case Study using Spectral Data from Alpha Cyclodextrin Hydrogel

α-Cyclodextrin-based polypseudorotaxane supramolecular hydrogels, which are based on the self-assembly of a polymer chain "guest" and α-cyclodextrin "host," are promising materials for a wide range of applications, including drug delivery and tissue engineering (Domiński et al., 2019). In this section, hydrogel libraries of alpha-Cyclodextrin (α-CD)/Polyethylene glycol (PEG) are used for the case study. It's known that composition plays a vital role in forming hydrogels. Here, 96 α-CD/PEG hydrogel samples with different mass ratios of α-CD to PEG are formulated in the 96 well plate. The concentration of PEG is kept at 120 mg/mL while the concentration of α-CD varies from 20 to 40 mg/mL. The phase angle-frequency spectrum of each sample is collected by a sensor-based high throughput method. The collected spectra are labeled as solution or gel to study the composition-structure relationship of α-CD/PEG hydrogels. Three repeated experiments offer 288 spectral data. Specifically, 194 spectral data of gel (Fig. 4 (a)) and 94 of solution (Fig. 4 (b)) are provided for the case study. The detailed procedure of data collection is described in Section 4.2.



(a) (b) (c)

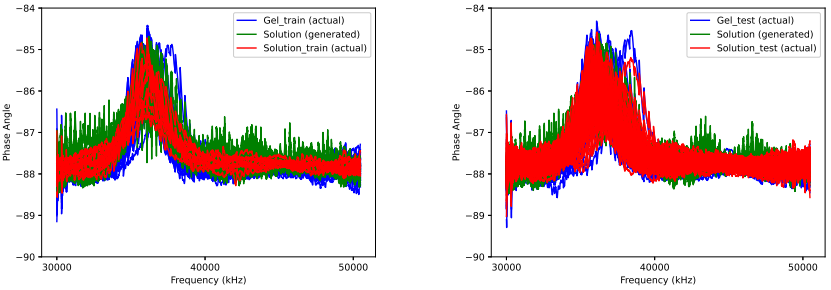
Fig. 4: Spectral data of Alpha Cyclodextrin hydrogel from (a) gel; (b) solution; (c) gel & solution.

Table 4 illustrates the training data with their balanced ratio. Specifically, the balanced ratios that the classifier's performances are applicable in practice are utilized. The remaining samples in each phase are used as testing data.

Table 4: Imbalanced training data samples in Alpha Cyclodextrin hydrogel case studies.

Majority Class	Minority Class	Balanced Ratio	Majority Class Training Samples	Minority Class Training Samples
Gel	Solution	0.083	120	10

Fig. 5 shows the samples of actual and generated samples from the proposed method. Similar to Fig. 2, the generated samples from our approach successfully learn the features of the test data of the solution phase (Fig. 5 (b)) from the small number of training data samples (Fig. 5 (a)).



(a) (b)

Fig. 5: Comparison between generated data of Alpha Cyclodextrin hydrogel with (a) actual training data; (b) actual testing data.

Table 5 shows the performance evaluation of the benchmark and our methods using the generated samples in Fig. 5. In this case study, all benchmark methods represent worse results than the baseline. This might be caused by high similarities between the samples from the gel and solution phases, as shown in Fig. 4. It causes a challenging task. Therefore, the sampling-based methods that consider only local information offer inferior performance. In addition, BAGAN-GP and CDRAGAN represent inferior results since the methods only focused on generating realistic samples but did not consider learning the phase-distinguishable features. Finally, Cooperative GAN also shows poor performance because of its unstable learning, resulting in a limited diversity of generated samples. Our method delivers the best performance by

Table 5: Performance evaluation in Alpha Cyclodextrin hydrogel case study. Averages and standard deviations (in the parenthesis) are represented.

	Alpha Cyclodextrin hydrogel		
	Precision	**Recall**	**F-score**
Baseline	0.893 (0.02)	0.886 (0.02)	0.880 (0.02)
SMOTE	0.841 (0.03)	0.080 (0.05)	0.792 (0.06)
B-SMOTE	0.846 (0.03)	0.816 (0.05)	0.801 (0.06)
CDRAGAN	0.848 (0.03)	0.826 (0.04)	0.814 (0.05)
BAGAN-GP	0.881 (0.02)	0.868 (0.03)	0.860 (0.03)
Cooperative GAN	0.847 (0.03)	0.829 (0.03)	0.819 (0.03)
Proposed	**0.902** (0.02)	**0.898** (0.02)	**0.892** (0.02)

generating realistic and phase-distinguishable samples with a stabilizing technique. Specifically, our method improves 1%∼13%, 1%∼11%, and 1%∼8% of the performance of the benchmark methods regarding their F-score, Precision, and Recall, respectively. Table 6 indicates that the performance of our method shows statistically significant improvements over the baseline at a 90% significance level in all measures.

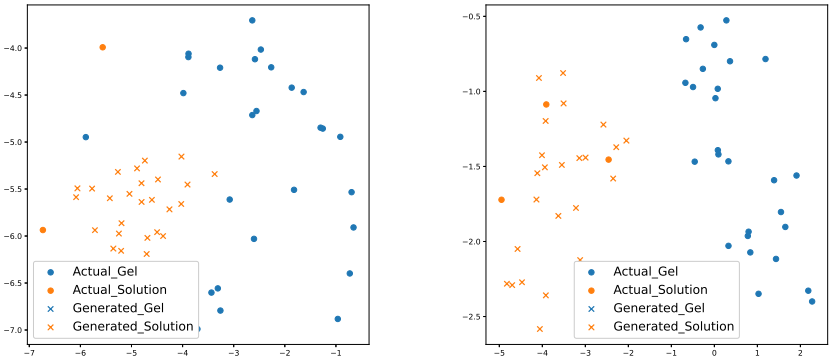
Table 6: P-value of statistical hypothesis test in Alpha Cyclodextrin hydrogel case studies.

	Precision	Recall	F-score
Paired t-test (Proposed $\geq$ Baseline)	0.098	0.098	0.092

Fig. 6 illustrates the t-SNE visualization of the features extracted from the intermediate layer of classifiers in our method at epochs 0 and 135. Similar to Fig. 3, '•' and '×' denote features of actual and generated samples, respectively. To make a balanced training data, the solution phase of the Alpha Cyclodextrin hydrogel has plenty of generated samples ('×') than actual samples ('•') in each batch. In contrast to epoch 0 (Fig. 6 (a)), the features at epoch 135 (Fig. 6 (b)) demonstrate that the features extracted from the generated samples ('×') of the solution phase of the Alpha Cyclodextrin hydrogel accurately match those from the actual samples ('•'). Due to this alignment, the balanced training data generated from our method achieves the best classification results compared to benchmark methods.

3 Discussions

This paper addresses the material characteristics classification problem using imbalanced spectral data. The imbalanced spectral data usually happens in



(a) (b)

Fig. 6: t-SNE of the feature from the intermediate layer of the classifier from our method in epochs (a) 0 and (b) 135.

actual experiments and industries, causing poor supervised classification performance. To address this challenge, a GAN-based data augmentation method proposed by authors in the previous work (Chung et al., 2023) is utilized. Specifically, the method consists of three players, namely, generator, discriminator, and classifier, jointly optimized. The generator in the method generates both realistic and characteristics-distinguishable data to balance the training data. The imbalanced spectral data between the phases of Pluronic F-127 hydrogel and Alpha Cyclodextrin hydrogel are used for the case studies. The case study results show the method successfully addresses the data imbalance problem by improving the phase classification results. Specifically, our method improves 1%∼13%, 1%∼11%, and 1%∼8% of the performance of the benchmark methods regarding their F-score, Precision, and Recall, respectively, in both case studies. Based on these validated results, we expect the method to be applied to imbalanced spectral data from various material science and chemical engineering domains since the method is not designed for the specific process.

4 Methodology

A detailed procedure for the data collection of Pluronic F-127 and α-CD/PEG hydrogel libraries are provided in Sections 4.1 and 4.2, respectively. Then, the proposed methodology is described in Section 4.3. Finally, the hyperparameters and the structure of the deep neural network used in this paper are listed in Section 4.4.

4.1 Data Collection of Pluronic F-127 Hydrogel Libraries

For the data collection, hydrogel libraries of Pluronic F-127 are obtained from Sigma Aldrich and are prepared in 96 well plates. The stock PF-127 water solution (30% wt.%) is first prepared with deionized water. The stock solution is then serial diluted with deionized water across the well plate for concentrations from 0.3125 wt.% to 29.6875 wt.%. The well plate is left in the fridge overnight for mixing and then it is taken out from fridge and leave at room temperature one hour for crosslinking. Next, the prepared PF-127 hydrogel libraries are characterized by piezoelectric milli-cantilever (PEMC) sensors. The PEMC sensor is integrated with a three-axis robot (MPS50SL; Aerotech) and its movement is controlled by a motion controller (A3200, Aerotech). The impedance spectrum of each hydrogel sample is captured by a network analyzer (E5061B, Keysight) and a customized MATLAB program. Spectra of all PF127 hydrogels in the 96 well plate are collected by manually controlling the robot-integrated sensor to move from one well to another. Finally, in the case of labeling, the spectral data are fitted to the sigmoid curve, and then the spectrum before the inflection point of the curve is labeled as a solution and the spectrum after the inflection point is labeled as a gel.

4.2 Data Collection of α-CD/PEG Hydrogel Libraries

To generate samples, supramolecular hydrogels of alpha-cyclodextrin (α-CD)/Polyethylene glycol (PEG) are prepared in 96 well plates. Both α-CD and PEG (Mn = 6000) are obtained from Sigma Aldrich and used without further purification. Stock solutions of α-CD (80 mg/mL) and PEG (240 mg/mL) are prepared in advance, and the hydrogel library is obtained by mixing a constant volume of PEG stock solution with different volumes of α-CD stock solution and deionized water. At first, 190 µL of PEG is pipetted into each well of the 96-well plate. Then, deionized water is pipetted by increasing from 95 to 190 µL with a step size of 1 µL. Next, α-CD is pipetted by reducing from 190 to 95 µL with a step size of 1 µL. The final volume in each well is 380 µL, and the concentration of PEG is 120 mg/mL, while the concentration of α-CD varies from 20 to 40 mg/mL. To avoid the formation of inhomogeneous hydrogels, the precursor solution in each well is mixed by pipette immediately once α-CD is added. After all wells have been formulated, the 96-well plate is further mixed by a digital shaker (LSE digital microplate shaker; Corning) at 1000 rpm for 10 minutes. Finally, the well plate was placed in a humid environment and reacted at room temperature for 12 hours. Then, the prepared hydrogel libraries of α-CD/PEG are characterized by PEMC sensors in a high-throughput manner. The PEMC sensor is integrated with a robot (FISNAR, F5200N) for automated characterization. The hydrogel in each well is characterized by penetrating the robot-integrated sensor into the sample, and the impedance spectra are collected by a network analyzer (E5061B, Keysight) and a customized MATLAB program. All samples in 96 well plates are automatically characterized by PEMC sensors with the computer-controlled robot.

Finally, the phases of the collected α-CD/PEG spectrum data are obtained by two best-fit linear regression models. Specifically, based on the point where the two linear regression models intersect, the spectrum before the point is identified as a solution and the spectrum after that as a gel.

4.3 Proposed Methodology

This section introduces a novel GAN-based data augmentation method proposed in the authors' previous paper (Chung et al., 2023). The structure of the overall method is described in Section 4.3.1. In addition, the objective functions of the algorithm are illustrated in Section 4.3.2. Finally, the training procedure of the method is described in Section 4.3.3.

4.3.1 Three-Player Structure for Imbalanced Data Learning

Fig. 7 shows the structure of our method, which consists of three players: a discriminator, a generator, and a classifier. The generator generates samples

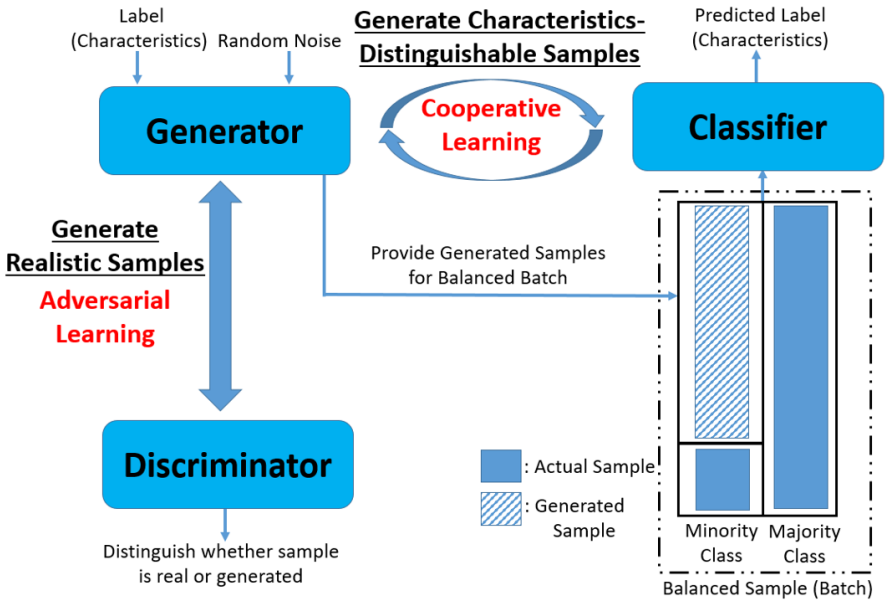


Fig. 7: Structure of the method (Chung et al., 2023).

of the spectral data using the random noise and corresponding characteristics labels. Within the generated samples, those representing the minority class are integrated with the actual imbalanced spectral data, resulting in balanced training data for the classifier. The proposed approach provides adversarial and cooperative learning to enhance the utility of the generated samples for

improving the classifier's performance. The specific roles of these two learnings are outlined as follows.

- Adversarial learning: The interaction between the generator and the discriminator adheres to the adversarial relationship inherent in the GAN structure. The relationship allows both networks to engage in a competitive process, ultimately leading to the generator's generation of realistic spectral data.
- Cooperative learning: The cooperative interaction between the classifier and the generator empowers the generator to produce spectral data that can be well discerned regarding the material's characteristics (i.e., characteristics-distinguishable samples) by the classifier.

Based on these two relationships, the generator generates samples of minority class with both properties (i.e., realistic and characteristics-distinguishable). Subsequently, these generated samples are combined with actual ones, creating a balanced training batch that flows through the classifier network in one training iteration. Through the iterative learning process involving three players, the classifier eventually attains a high level of performance. The detailed objective function of each player is explained in the following sections.

4.3.2 Objective Functions for Three-Player

The review of the Generative Adversarial Network (GAN) is described in Section 4.3.2.1 initially. Then, the objective functions of the discriminator, generator, and classifier are illustrated in Sections 4.3.2.2, 4.3.2.3, and 4.3.2.4, respectively. The iterative optimization between the three players ultimately yields the high-performance classifier from the imbalanced spectral data.

4.3.2.1 Generative Adversarial Network (GAN)

The idea of a GAN is to train two networks, namely, generator G and discriminator D, with a minimax game for $V(D, G)$ demonstrated in Eq. (2) (Wang et al., 2017).

$$\begin{aligned} \min_G \max_D V(D,G) = \mathbb{E}_{(x_a,y_a)\sim P(X_a,Y_a)}[\log(D(x_a,y_a))] \\ + \mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[\log(1-D(G(z,y_g),y_g)], \end{aligned} \tag{2}$$

where z denotes the random noise, and x_a is actual samples from spectral data. y_a and y_g are the labels of actual and generated spectral data, respectively. Specifically, the generator is to produce samples of spectral data $G(z)$ from z. In contrast, the discriminator is to distinguish whether the origin of input samples is from the actual (x_a) or the generator $(G(z))$. In other words, the role of the discriminator is to distinguish the origin of the input samples, whereas the generator's task is to create synthetic samples with the intention of deceiving the discriminator. This adversarial learning leads to the distribution of newly generated samples approaching the inherent distribution of the actual samples, $P(X_a)$.

4.3.2.2 Objective Function of Discriminator

In the proposed approach, the discriminator aims to maximize Eq. (2) through adversarial learning with the generator. Specifically, the discriminator learns to distinguish the input (x_a, y_a) and $(G(z, y_g), y_g)$ are actual and generated, respectively. Furthermore, the method introduces two supplementary terms to ensure a stable learning process. This is done because GAN training is usually unstable and challenging to converge, resulting in the generator's gradient explosions in adversarial learning (Tao and Wang, 2020; Arjovsky and Bottou, 2017). First, our method ensures the regularization of the discriminator's gradient by imposing a gradient penalty. The penalty enforces 1-Lipschitz continuity upon the discriminator. Second, the proposed approach incorporates an extra input for the discriminator, comprising the actual sample with a wrong label. This added task prevents the discriminator from distinguishing the origin of the input well before the generator successfully approximates the actual sample distribution of the spectral data. Otherwise, it causes unstable learning of GAN through exploding or vanishing the gradient of the generator (Tran et al., 2018; Arjovsky and Bottou, 2017). In summary, the objective function of the discriminator (L^D) is as follows (Chung et al., 2023).

$$
\begin{aligned}
L^D(Z, X_a, Y_a, Y_g, Y_m) = \\
&\underbrace{-\mathbb{E}_{(x_a,y_a)\sim P(X_a,Y_a)}[\log(D(x_a, y_a))]}_{\text{loss from actual sample in discriminator}} \\
&\underbrace{-\mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[\log(1 - D(G(z, y_g), y_g)]}_{\text{loss from generated sample in discriminator}} \\
&\underbrace{-\mathbb{E}_{(x_a,y_m)\sim P(X_a,Y_m)}[\log(1 - D(x_a, y_m)]}_{\text{loss from mislabeled sample in discriminator}} \\
&+ \lambda \underbrace{\mathbb{E}_{(\hat{x},y_a)\sim P(\hat{X},Y_a)}[(\|\nabla_{(\hat{x},y_a)} D(\hat{x}, y_a)\|_2 - 1)^2]}_{\text{loss from gradient penalty}},
\end{aligned}
\tag{3}
$$

where $\hat{x} = \alpha x_a + (1 - \alpha)G(z)$, and α is sampled uniformly between 0 and 1. The coefficient λ pertains to the gradient penalty term. The initial three losses in Eq. (3) are associated with losses incurred when the discriminator misclassifies the source of the actual, generated, and mislabeled sample. The final loss corresponds to the loss linked to the gradient of the discriminator.

4.3.2.3 Objective Function of Generator

The primary aim of the generator is to generate samples that align with the distribution of actual spectral data ($P(X_a)$), accomplished by minimizing Eq. (2). Hence, Eq. (2) enables the adversarial learning between the discriminator and generator. Apart from Eq. (2), the generator incorporates an additional component in its objective function that pertains to the classifier. Unlike the adversarial relationship with the discriminator, the generator and

the classifier establish a cooperative relationship to generate distinctly discernible spectral samples across the material characteristics. In other words, the generator's role is to generate samples and provide a balanced training dataset that can improve the classifier's performance, as shown in Fig. 7. To accomplish this, the generator's objective function includes the classification loss based on the generated samples. The objective function of the generator (L^G) can be formulated as follows (Chung et al., 2023).

$$L^G(Z, Y_g) = \underbrace{-\mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[\log(D(G(z, y_g), y_g))]}_{\text{loss from generated sample in discriminator}} \underbrace{-\mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[y_g \log(C(G(z, y_g)))]}_{\text{loss from generated sample in classifier}}. \tag{4}$$

4.3.2.4 Objective Function of Classifier

The objective function of the classifier includes the classification loss derived from both the actual and generated samples of the spectral data. As illustrated in Fig. 7, the generator's samples are combined with actual samples to provide a balanced training dataset for each batch of the classifier. The classifier is then optimized to minimize the classification loss for both the actual and generated samples. Finally, the classifier's objective function (L^C) is listed as follows (Chung et al., 2023).

$$L^C(Z, X_a, Y_a, Y_g) = \underbrace{-\mathbb{E}_{(x_a,y_a)\sim P(X_a,Y_a)}[y_a \log(C(x_a))]}_{\text{loss from actual sample in classifier}} \underbrace{-\mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[y_g \log(C(G(z, y_g))]}_{\text{loss from generated sample in classifier}}. \tag{5}$$

In particular, $-\mathbb{E}_{(z,y_g)\sim P(Z,Y_g)}[y_g \log(C(G(z, y_g))]$, a common term in both Eqs. (4) and (5) enables cooperative learning between the generator and classifier.

4.3.3 Training Procedure

The three players are optimized alternatively. Initially, the discriminator undergoes training using a batch that includes both actual and generated samples, aiming to minimize Eq. (3). Subsequently, a batch containing only generated samples are employed to update the generator, focusing on minimizing Eq. (4). Finally, the classifier's training involves minimizing Eq. (5) with balanced training data from all the classes. This process begins with sampling a batch from the actual data. Then, the generator generates the remaining samples from the minority class to ensure a balanced training set. The alternating training process continues until it reaches the specified number of predefined epochs.

4.4 Hyperparameters of the Deep Neural Networks

Table 7 describes the hyperparameters that are used for all the methods in this paper. The coefficient of the gradient penalty is established at 10 based on the recommendation in the previous studies (Kodali et al., 2017; Huang and Jafari, 2021). In the case of Cooperative GAN, the scheduling parameter is selected from a specified range ([0.1, 0.9]) that demonstrates the best performance (Choi et al., 2021). Batch size is 30. Table 8 provides information on the

Table 7: Hyperparameters of each method.

Methods	Parameters	Value
SMOTE, B-SMOTE	Nearest K samples	1,7
CDRAGAN BAGAN-GP Cooperative GAN Proposed	Number of epochs	150
	Optimizer	Adam
	Learning rate	0.0002
	Momentum	$\beta_1 = 0.5, \beta_2 = 0.9$
	Hidden layers (Discriminator)	4 blocks of [Conv2D, LeakyRelu]
	Hidden layers (Generator)	4 blocks of [Conv2D-Transpose, LeakyRelu, BatchNormalization]
	Number of Kernels in each block (Discriminator)	(64,128,128,256)
	Number of Kernels in each block (Generator)	(128,128,64,Number of channel)
	Kernel sizes	(4,4)
	Strides	(2,2)
	Padding	Same
	Activation functions	LeakyRelu, Tanh
	Kernel initializer	Random Normal(sd=0.02)
	Slope of Leaky Relu	0.2
CDRAGAN, BAGAN-GP Proposed	Gradient Penalty Coefficient	10
Cooperative GAN	Range of scheduling parameter	[0.1,0.9]
BAGAN-GP, Proposed	Epochs in pre-training	100

hyperparameters used for the classifier in the case studies. In case studies, a CNN is utilized as the classifier. To ensure a fair and consistent comparison, all the methods adopt the identical classifier configuration outlined in Table 8.

Table 8: Hyperparameters of the classifier.

Parameters	Value
Number of epochs	150
Optimizer	Adam
Learning rate	0.0002
Momentum	$\beta_1 = 0.5, \beta_2 = 0.9$
Hidden Layers	4 blocks of [Conv2D, LeakyRelu]
Number kernels in each block	(32,32,128,256)
Kernel sizes	(4,4)
Strides	(2,2)
Padding	Same
Activation functions	Leaky Relu, Softmax
Kernel initializer	Random Normal (sd=0.02)
Slope of Leaky Relu	0.2

Acknowledgments. This project was funded by a grant with award number 1933525 from the National Science Foundation (NSF).

Data Availability. The authors declare that the data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding authors on reasonable request.

Code Availability. All code used in this paper can be found in the Github repository cjh7.

Competing Interests.

Author Contributions.

References

Antoniou, A., A. Storkey, and H. Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* .

Arjovsky, M. and L. Bottou. 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862* .

Armbruster, T. and R. Danisi. 2015. The power of databases: the rruff project. *Highlights in mineralogical crystallography*: 1–30 .

Belsky, A., M. Hellenbrandt, V.L. Karen, and P. Luksch. 2002. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science 58*(3): 364–369 .

Bisong, E. and E. Bisong. 2019. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*: 59–64 .

Callaghan, P.T. 1993. *Principles of nuclear magnetic resonance microscopy*. Clarendon press.

Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16: 321–357 .

Choi, H.S., D. Jung, S. Kim, and S. Yoon. 2021. Imbalanced data classification via cooperative interaction between classifier and generator. *IEEE Transactions on Neural Networks and Learning Systems* .

Choudhary, K., B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J. Billinge, et al. 2022. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials 8*(1): 59 .

Chung, J., B. Shen, and Z.J. Kong. 2023. Anomaly detection in additive manufacturing processes using supervised classification with imbalanced sensor data based on generative adversarial network. *Journal of Intelligent Manufacturing*: 1–20 .

Cui, W., Y. Zhang, X. Zhang, L. Li, and F. Liou. 2020. Metal additive manufacturing parts inspection using convolutional neural network. *Applied Sciences 10*(2): 545 .

Dimitriadis, G., J.P. Neto, and A.R. Kampff. 2018. t-sne visualization of large-scale neural recordings. *Neural computation 30*(7): 1750–1774 .

Domiński, A., T. Konieczny, and P. Kurcok. 2019. α-cyclodextrin-based polypseudorotaxane hydrogels. *Materials 13*(1): 133 .

Douzas, G. and F. Bacao. 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications* 91: 464–471 .

Fathy, Y., M. Jaber, and A. Brintrup. 2020. Learning with imbalanced data in smart manufacturing: A comparative analysis. *IEEE Access* 9: 2734–2757 .

Friedrich, W., P. Knipping, and M. Laue. 1913. Interferenzerscheinungen bei roentgenstrahlen. *Annalen der Physik 346*(10): 971–988 .

Gao, B., J. Zhou, Y. Yang, J. Chi, and Q. Yuan. 2022. Generative adversarial network and convolutional neural network-based eeg imbalanced

classification model for seizure detection. *Biocybernetics and Biomedical Engineering 42*(1): 1–15 .

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 .

Han, H., W.Y. Wang, and B.H. Mao 2005. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer.

Hsu, H. and P.A. Lachenbruch. 2014. Paired t test. *Wiley StatsRef: statistics reference online* .

Huang, G. and A.H. Jafari. 2021. Enhanced balancing gan: Minority-class image generation. *Neural Computing and Applications*: 1–10 .

Jalaal, M., G. Cottrell, N. Balmforth, and B. Stoeber. 2017. On the rheology of pluronic f127 aqueous solutions. *Journal of Rheology 61*(1): 139–146 .

Kantz, E.D., S. Tiwari, J.D. Watrous, S. Cheng, and M. Jain. 2019. Deep neural networks for classification of lc-ms spectral peaks. *Analytical chemistry 91*(19): 12407–12413 .

Kiyasseh, D., G.A. Tadesse, L. Thwaites, T. Zhu, D. Clifton, et al. 2020. Plethaugment: Gan-based ppg augmentation for medical diagnosis in low-resource settings. *IEEE journal of biomedical and health informatics 24*(11): 3226–3235 .

Kodali, N., J. Abernethy, J. Hays, and Z. Kira. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215* .

Lee, J.W., W.B. Park, J.H. Lee, S.P. Singh, and K.S. Sohn. 2020. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nature communications 11*(1): 86 .

Lee, X.Y., S.K. Saha, S. Sarkar, and B. Giera. 2020. Automated detection of part quality during two-photon lithography via deep learning. *Additive manufacturing* 36: 101444 .

McCulloch, W.S. and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5: 115–133 .

Mikołajczyk, A. and M. Grochowski 2018. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pp. 117–122. IEEE.

Mycroft, W., M. Katzman, S. Tammas-Williams, E. Hernandez-Nava, G. Panoutsos, I. Todd, and V. Kadirkamanathan. 2020. A data-driven approach for predicting printability in metal additive manufacturing processes. *Journal of Intelligent Manufacturing 31*(7): 1769–1781 .

Powers, D.M. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* .

Ranasinghe, G.D. and A.K. Parlikad 2019. Generating real-valued failure data for prognostics under the conditions of limited data availability. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 1–8. IEEE.

Schuetzke, J., A. Benedix, R. Mikut, and M. Reischl. 2021. Enhancing deep-learning training for phase identification in powder x-ray diffractograms. *IUCrJ 8*(3): 408–420 .

Schuetzke, J., N.J. Szymanski, and M. Reischl. 2023. Validating neural networks for spectroscopic classification on a universal synthetic dataset. *npj Computational Materials 9*(1): 100 .

Smith, E. and G. Dent. 2019. *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons.

Szymanski, N.J., Y. Zeng, H. Huo, C.J. Bartel, H. Kim, and G. Ceder. 2021. Toward autonomous design and synthesis of novel inorganic materials. *Materials horizons 8*(8): 2169–2198 .

Tao, S. and J. Wang 2020. Alleviation of gradient exploding in gans: Fake can be real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200.

Tran, N.T., T.A. Bui, and N.M. Cheung 2018. Dist-gan: An improved gan using distance constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 370–385.

Wang, C., Z. Yu, H. Zheng, N. Wang, and B. Zheng 2017. Cgan-plankton: Towards large-scale imbalanced class generation and fine-grained classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 855–859. IEEE.

Wang, H., Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, and J. Lin. 2020. Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of chemical information and modeling 60*(4): 2004–2011 .

Wu, M., S. Wang, S. Pan, A.C. Terentis, J. Strasswimmer, and X. Zhu. 2021. Deep learning data augmentation for raman spectroscopy cancer tissue classification. *Scientific reports 11*(1): 23842 .

Zeng, F., W. Peng, G. Kang, Z. Feng, and X. Yue 2021. Spectral data classification by one-dimensional convolutional neural networks. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pp. 1–6. IEEE.

Zhang, J., Y. Liu, M. Singh, Y. Tong, E. Kucukdeger, H.Y. Yoon, A.P. Haring, M. Roman, Z.J. Kong, B.N. Johnson, et al. 2023. Rapid, autonomous high-throughput characterization of hydrogel rheological properties via automated sensing and physics-guided machine learning. *Applied Materials Today* 30: 101720 .