

GANsemble for Small and Imbalanced Data Sets: A Baseline for Synthetic Microplastics Data

Daniel Platnick^{†,‡,*}, Sourena Khanzadeh[‡], Alireza Sadeghian[‡], Richard Valenzano^{†,‡}

[†] Vector Institute, Toronto, Canada, [‡] Toronto Metropolitan University, Toronto, Canada

Abstract

Microplastic particle ingestion or inhalation by humans is a problem of growing concern. Unfortunately, current research methods that use machine learning to understand their potential harms are obstructed by a lack of available data. Deep learning techniques in particular are challenged by such domains where only small or imbalanced data sets are available. Overcoming this challenge often involves oversampling underrepresented classes or augmenting the existing data to improve model performance. This paper proposes GANsemble: a two-module framework connecting data augmentation with conditional generative adversarial networks (cGANs) to generate class-conditioned synthetic data. First, the *data chooser module* automates augmentation strategy selection by searching for the best data augmentation strategy. Next, the *cGAN module* uses this strategy to train a cGAN for generating enhanced synthetic data. We experiment with the GANsemble framework on a small and imbalanced microplastics data set. A Microplastic-cGAN (MPcGAN) algorithm is introduced, and baselines for synthetic microplastics (SYMP) data are established in terms of Fréchet Inception Distance (FID) and Inception Scores (IS). We also provide a synthetic microplastics filter (SYMP-Filter) algorithm to increase the quality of generated SYMP. Additionally, we show the best amount of oversampling with augmentation to fix class imbalance in small microplastics data sets. To our knowledge, this study is the first application of generative AI to synthetically create microplastics data.

Keywords: Deep learning, Generative AI, Microplastics, Small data, Spectra, Sample generation

1. Introduction

The emergence of microscopic plastic particles in food and drinking water is a problem of growing concern [1]. Microplastics are shown as potentially harmful to reproductive health [2]. They have been found to be increasingly present in human placenta [3] and human lungs [4]. Thus, research towards mitigating potential dangers of microplastics is extremely valuable. However, such research can be difficult due to the insufficient availability of public data. This is particularly problematic for deep learning methods. Despite this challenge, deep learning based methods have shown success for detecting microplastics even with small and imbalanced data sets [5, 6]. To increase model robustness on small or imbalanced data sets, well known approaches such as creating augmented or synthetic data can be used to oversample minority classes or increase data set size. We connect these ideas with GANsemble, which performs a search for the best augmentation strategy, and uses it to train a cGAN and create class-conditioned SYMP data.

Deep learning techniques have proven value in many domains [7–9], but universally share weakness in the cases of small data sets and those with class imbalances [10, 11]. The data a deep learning model is trained on greatly determines its robustness, scalability, safety, and fairness [12]. Data augmentation and transfer learning have been used to increase the effectiveness of deep learning techniques in the cases of small or imbalanced training data sets [13]. However, data augmentation only works to a certain extent, and transfer learning can have the side effect of negative transfer [14]. As such, using generative AI to create synthetic training data is an increasingly popular strategy [15]. Generative AI approaches for creating synthetic data are shown to be useful across different domains, as they can

*Corresponding author: daniel.platnick@torontomu.ca

generate an arbitrary number of synthetic samples, and the synthetic data is not affected by privacy regulations [16]. As generative AI research continues to progress, the quality of generated synthetic samples continues to improve, which increases their effectiveness for training deep learning models.

Oversampling the minority class is a natural remedy for class imbalance [17]. Synthetic Minority Oversampling Technique (SMOTE) is a well known method which produces synthetic samples to fix class imbalances using K-nearest neighbors in the feature space [18]. Zhou, Chen, and Chien [19] performed analysis on different augmentation strategies for convolutional neural network (CNN) classification of medical image spectra, and found that simple strategies such as masking and horizontal flipping can greatly improve model performance. Today, oversampling techniques typically include intelligent data augmentations [13]. Yun, Han, Oh, Chun, Choe, and Yoo [20] introduced a state-of-the-art image augmentation algorithm called CutMix, which maintains the benefits of regional dropout regularization without loss of information. Obaid and Nassif [21] showed that cases may vary depending on the nature of the data. Therefore, given imperfect information, the system designer must decide the best augmentation strategy to use. This is the challenge of augmentation selection.

Inspired by recent developments in generative AI [22, 23], we present GANsemble to overcome this challenge and further microplastics research. The proposed method connects ideas from data augmentation and synthetic data generation [24] to provide automatic augmentation strategy selection and enhance model learning on small and imbalanced data sets. The GANsemble *data chooser module* enables the AI system to manipulate the inputted data through augmentation, generating new augmentation strategies to find the strategy which maximizes model generalization. Specifically, GANsemble performs an *n-step factorial search* search on inputted augmentation strategies, and employs a deep neural network (DNN) to intelligently select the best strategy *Aug**. This strategy is then used to oversample the original data set and train a cGAN to generate new synthetic samples. GANsemble can significantly improve the augmentation selection process, by decreasing the need for human intervention. Furthermore, we provide MPcGAN, the first cGAN for generating synthetic microplastics (SYMP) data, and establish baseline FID and IS scores for SYMP data.

Generating synthetic sample data to enhance the training set is an idea increasing in popularity [23, 25, 26]. This approach has not been applied to microplastics research. To make progress in SYMP data generation, this work applies cGANs to create SYMP data, and baseline SYMP data quality measures are established. MPcGAN is trained to produce class-conditioned SYMP samples that enhance training set size, balance, and variation. Note, this paper builds on the important work of Tian, Beén, Sun, Thienen, and Bäuerlein [5], which identified the most effective amount of oversampling with augmentation for a small imbalanced microplastics data set. However, Tian et al. introduce data leakage into their analysis by creating augmented data from the training set, and evaluating on the same training set. This is a problem, because in practice, the model will be evaluated on data it has never seen before, not data created from previous training data. Performing a similar study, this research builds on their analysis by using a separate evaluation set to prevent data leakage between train and test sets. Multiple runs are used to increase experimental validity.

The rest of this paper is organized as follows. The microplastics data description is given in Section 2. Section 3 explains the GANsemble framework, as well as techniques, algorithms, and parameters used in the study. Section 4 provides experiment results, and establishes baseline FID and IS scores for SYMP data. Next, section 5 discusses the implications as well as concluding remarks.

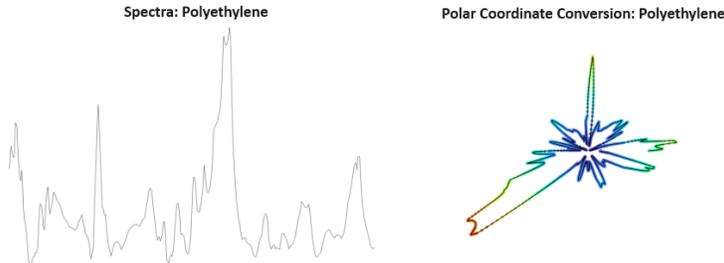


Figure 1. Example of the polar conversion that allows DNNs to treat polymer spectra as images [5].

2. Data Set Description

The data set we use was introduced by Tian, Beén, Sun, Thienen, and B auerlein [5], and is the only easily accessible microplastics data set we could find. The raw data and supplementary information on data preparation can be found from these sources [5, 6, 27]. In this section, we review its characteristics and how it was created.

Laser Directed Infrared (LDIR) and Fourier Transform Infrared (FTIR) spectroscopy convert chemical compounds to spectra data, which has previously been shown amenable for learning by deep neural networks (DNNs) [28–30]. The data set used in this study contains 210 samples of polymer spectra across 10 classes: 8 classes of plastic spectra, and 2 non-plastic classes [5]. The least represented polymer class, polyacetal, has 5 samples, while the most highly represented class silica has 38. The 210 particle-based spectra were collected from various Dutch water sources such as drinking water, surface water, and sewage, as well as the Agilent Clarity version 1.4.10 software database [6]. Samples were converted to spectra using the LDIR quantum cascade laser (QCL)-based Agilent chemical imaging system. The LDIR system was used with wavenumbers between 975 and 1800cm^{-1} , and a rotation of 0.5cm^{-1} [5, 27]. An example spectra can be found in the left image of Figure 1.

Tian et al. [5] also introduced a visual polarized coordinate transformation which allows microplastic particle spectra to be treated as images. This enables the use of computer vision models, which have been found to be very useful in practice. The spectra in the microplastics data set are mapped through the polar conversion process as follows: The wavenumbers 975 - 1800 are mapped from 0° to 360° and the absorption rate becomes the polar coordinate magnitude. The images are then colourized such that colour intensity grows with the absorption value. Figure 1 shows an example of this transformation visually.

3. The GANsemble Framework

In this section, we describe our GANsemble framework and how it is applied to the microplastics data set.

GANsemble consists of two distinct detachable modules, which we describe below. The first, the *data chooser module*, is used to identify useful combinations of base augmentation strategies, which we call *composite strategies*, for the given data set. Next, we have the *cGan module*, which uses the best base augmentation strategy or *composite strategy* to generate more data for the training of a conditional GAN. Finally, we have the *SYMP-Filter*, which post-processes the output of the cGAN by removing artifacts of generated images, thereby improving the quality of the synthetic data. We describe these modules below.

3.1. The Data Chooser Module

The first component of the GANsemble framework is the *data chooser module*. The objective of this module is to automate augmentation selection by searching for the augmentation strategy or *composite strategy*, given a set of base augmentation strategies, that can be used to most effectively train a cGAN on a small and possibly imbalanced data set. Since this process involves finding effective combinations of base augmentation strategies, which we call *composite strategies*, the module has several hyperparameters controlling the amount of computational resources to be used in the search for the best augmentation strategy, Aug^* . Thus, the *data chooser module* simplifies the challenge of finding the best augmentation strategy from "What augmentation strategy should I use?" to "How many augmentation strategies should I consider and how extensively should they be considered, given time and computational resources?"

In more detail, the *data chooser module* takes in 3 inputs. The first is a small imbalanced data set requiring augmentation. In our case, this is the microplastic spectra images.

The second input is a list of base augmentation strategies to be used. While any set of base data augmentation strategies can be used, for the microplastic data, we use four Gaussian approaches: (1) horizontal flipping and horizontal shifting, (2) blur and rotation, (3) zoom and rotation, and (4) circular and rectangular masking. Base augmentation strategy 1 performs random horizontal flipping and shifting. Random horizontal flips in spectra data have been shown to greatly improve ML classification performance [19], while shifting ensures there are not too many duplicate samples. Augmentation strategy 2 uses varying levels of blur and random rotations to distort the spectra images. The blur effect is achieved through a Gaussian filter with $\alpha \in [10, 23]$ and $\sigma \in [2.8, 3.82]$, where α is blur intensity and σ is smoothness. Different α and σ values were tried and these ranges gave the best results. Augmentation strategy 3 applies random rotations and zoom between 1.0x and 1.34x. Augmentation strategy 4 is a Monte Carlo style masking strategy, where circles and rectangles cover different areas of the polymer spectra. Masking has also been effective for augmenting data in heart spectra classification [19], as it forces the learning algorithm to focus on different regions of the image when masked regions are not present.

The third input to the *data chooser module* is n , the number of steps to take in the *n-step factorial search*. Given a data set and a set of base augmentation strategies, the *data chooser module* identifies which base augmentation strategy or *composite strategy* provides the most useful data set. However, there are $\mathcal{O}(2^A)$ total strategies, given A base augmentation strategies. Considering all of them may be prohibitively expensive, which is why n is introduced. For a given n , only combinations containing at most n augmentation methods are considered. This decreases the number of combinations to $\mathcal{O}(2^n)$.

In principle, it would be possible to train all $\mathcal{O}(2^A)$ possible cGANs and thereby directly evaluate the effectiveness of the different combinations. However, even for a small data set, doing so is prohibitively expensive given the time and lack of stability of GAN training. As such, we instead focus on the simpler task of classification using a pre-trained model. That is, we consider each of the $\mathcal{O}(2^n)$ possible combinations of augmentation methods, and train a classifier r times each, where r is another hyperparameter controlling how extensively the augmentation strategies are explored. The classifier with the best average performance over the r runs is then chosen and used for training the cGAN. For the microplastics data, we use a pre-trained ResNet50 as the classifier for identifying the best augmentation strategy. We refer to the resulting augmentation strategy as Aug^* .

3.2. The cGAN Module

The next module is the cGAN trained on Aug^* . We propose a baseline MPcGAN model for microplastics spectra data to showcase GANsemble and further generative AI research in

mitigating microplastics. The *cGAN module* attaches to the *data chooser module*, training a generator on the *data chooser module* output to produce robust synthetic training data. As input, the MPcGAN generator takes in class label y , and samples from the Gaussian distribution z conditionally based on y , to produce a synthetic image from class y . The discriminator takes in a class label with a real or synthetic image and outputs a value $D(G(z|y)) \in [0, 1]$, where values closer to 1 correspond to the discriminator predicting the data is real. Both the generator and discriminator architectures are based on the convolutional downsample and upsample operations. Information on the cGAN training setup can be found in section 4.

3.3. The SYMP-Filter

Finally, we provide a post-processing algorithm called the *SYMP-Filter*. This process results in higher quality synthetic data. Our algorithm leverages spatial attributes of the polar coordinate transformed images to detect and filter out noisy SYMP data points. Figure 4 illustrates the effects of the *SYMP-Filter* visually.

The algorithm works through this process: Generate 5000 SYMP from each class, rank them using the *SYMP-Filter*, and keep the top t images of each class. The *SYMP-Filter* applies a square mask to the 4 corner sections of each image, calculating the densities of each corner. The pixel values inside the corner sections are summed to get the corner density for each image. Next, the algorithm keeps the top t lowest density images from each class.

The *SYMP-Filter* algorithm is described mathematically as follows:

- (1) Let $X_{i,j}$ denote an image in a collection, with i indexing the collection and j indexing the images within that collection, where each image is a 3-dimensional tensor in $\mathbb{R}^{128 \times 128 \times 3}$.
- (2) Let r be the radius defining the area of the corner sections to calculate each density.
- (3) Let $M_k(r)$ represent the square mask density detector for the k -th corner section, with $k \in \{1, 2, 3, 4\}$ corresponding to the four corner sections and radius r .
- (4) Let $D_{i,j}$ denote the density for image $X_{i,j}$, computed by applying the corner masks and summing the values.
- (5) Let T be the set of indices (i, j) corresponding to the top t images with the lowest densities.

$$D_{i,j} = \sum_{k=1}^4 \sum_{x=1}^{128} \sum_{y=1}^{128} \sum_{c=1}^3 X_{i,j}(x, y, c) \cdot M_k(x, y, c, r) \quad (3.1)$$

$$T = \underset{i,j}{\operatorname{argmin}_t} (D_{i,j}) \quad (3.2)$$

4. Experiments and Results

Recall that our data set contains only 210 images. This was split into an unbiased testing set containing 20 samples, with 2 images per class, and a training set containing the remaining 190 samples. The 20 samples were not used in creating the augmented data sets from Table 1 or Figure 2 to prevent data leakage. The 20 samples were later added to help train the cGAN. All code relating to this work can be found here: <https://github.com/DanielPlatnick/GANsemble>.

4.1. Identifying An Effective Size For Augmented Data Sets

In our first experiment, we attempt to reproduce the experiments by Tian et al. [5], in which they evaluated the tradeoff between accuracy and training time related to augmented data set size. In the original work, a custom augmentation strategy was used. Our goal

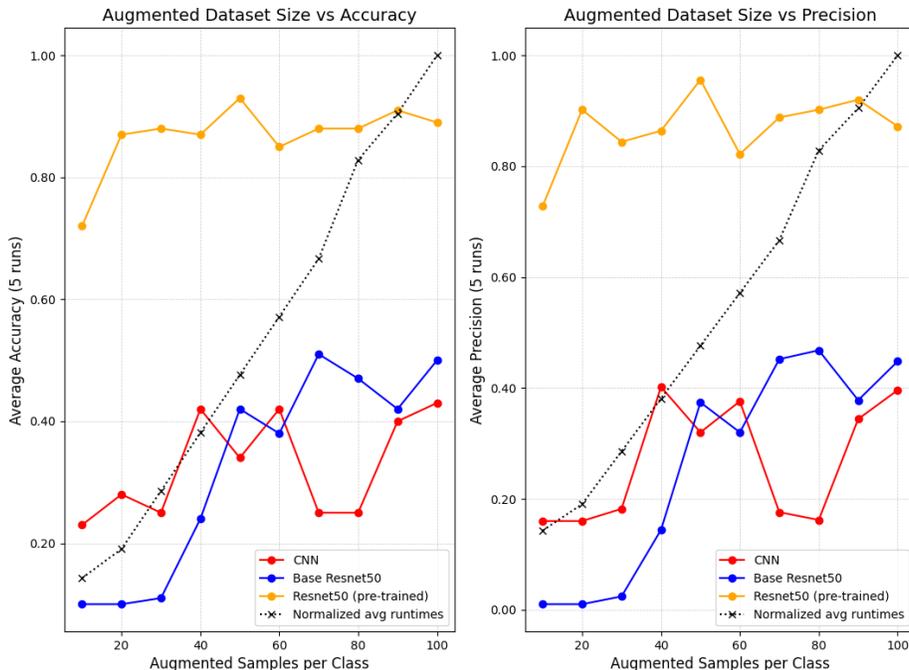


Figure 2. Determining the most effective size for purely augmented small microplastics data sets. The black dotted line represents averaged algorithm run-time after min-max normalization. The best performance is achieved with a pre-trained ResNet50 model using 50 augmented samples from each class. Augmentation strategy 1 was used to generate figure 2.

is to evaluate if similar conclusions can be made with simpler augmentation strategies and without the data leakage that occurred in the original work.

For simplicity, we tested with horizontal flipping and horizontal shifting. Using this augmentation strategy, we create 10 purely augmented data sets of sizes 100, 200, . . . , 1000, each with balanced classes. We then used this data to train three different DNN models: a CNN with 13,804,510 trainable parameters, a ResNet50 with 23,528,522 trainable parameters, and a version of that ResNet50 pre-trained on ImageNet [31, 32]. Each DNN was trained 5 times and the average results on our test set of 20 images are shown in Figure 2. The average training time over all models is also shown. Training times are normalized between 0 and 1, where 0 corresponds to the lowest runtime of any model, and 1 corresponds to the largest runtime.

Figure 2 shows that the pre-trained ResNet50 is most capable of generalization based on accuracy and precision. Transfer learning is a well-known method for enhancing model generalization, and should be used when applicable [14]. The pre-trained ResNet50 achieves the highest performance with 50 samples per class (*i.e.* 500 total augmented samples), scoring an average accuracy of 93% and precision of 95%. Run-time increases linearly with data set size, and the best size in terms of performance and run-time is $n = 50$. Even with 10 samples, the pre-trained ResNet is capable of learning enough to achieve 75% accuracy. The other models had a much higher variance across runs, and greatly lagged behind in performance at the levels of data tested.

Interestingly, our results roughly correspond with those originally by Tian, Beén, Sun, Thienen, and Bäuerlein [5]. In their work, the best performance was found when using 40 samples per class.

| Data Set | Accuracy (%) |
|--------------------------|--------------|
| Aug 1 | 90.5 |
| Aug 2 | 79.5 |
| Aug 3 | 84.0 |
| Aug 4* | 91.5 |
| Aug 5 (1 & 2) | 85.0 |
| Aug 6 (1 & 3) | 85.5 |
| Aug 7 (1 & 4) | 89.5 |
| Aug 8 (2 & 3) | 88.5 |
| Aug 9 (2 & 4) | 87.0 |
| Aug 10 (3 & 4) | 85.0 |
| Aug 11 (1, 2, & 3) | 85.0 |
| Aug 12 (1, 2, & 4) | 87.0 |
| Aug 13 (1, 3, & 4) | 84.5 |
| Aug 14 (2, 3, & 4) | 88.5 |
| Aug 15 (1, 2, 3, & 4) | 84.0 |
| No Oversampling | 86.0 |
| Oversampling No Aug | 87.5 |

Table 1. Demonstration of the GANsemble *data chooser module n-step factorial search*. Based on Figure 2 experiments, $N = 500$ was used to oversample the data. Performances are measured in terms of average accuracy over 10 runs. The three top *data chooser module* (pre-trained ResNet50) choices are highlighted for analysis. GANsemble detects the best augmentation strategy *Aug**. *Oversampling No Aug* uses duplication.

4.2. Automating Augmentation Selection

Next, we use the *data chooser module* to automate augmentation strategy selection by identifying the best augmentation strategy or *composite strategy* to be used in cGAN training. The 4 base augmentation strategies used were described in Section 3.1. We set the number of steps $n = 4$ and thus considered 15 different augmentation methods in total. These 15 strategies were used to create 15 data sets containing 50 images per class, with a mixture of augmented and real samples. For example, if 10 real samples were available for a particular class, then 40 augmented samples were added to that class.

For each of the 15 resulting data sets, we fine-tuned a pre-trained ResNet50 10 times for 150 epochs, training on the 190 samples and evaluating on the 20 test samples. The results are shown in Table 1, along with the performances of the ResNet50s when trained using oversampling on the original data to fix class imbalance (*Oversampling No Aug*), and when using the original data directly (*No Oversampling*). The top three augmentation strategies are 1, 4 (*Aug**), and 7. Strategy 1 applies horizontal flipping and shifting, *Aug** uses masking, and strategy 7 is a *composite strategy* of strategies 1 and *Aug**.

Random masking results in the highest accuracy. Many *composite strategies* improve classification performance, but some hurt model performance. We find that augmentation strategies affecting the spatial domain in different ways should not be combined. Other notable *composite strategies* include strategies 8, 9, 12 and 14. Augmentation strategy 14 performs well and is a combination of strategies 2, 3, and 4. This implies certain augmentation strategies complement each other, and GANsemble is capable of detecting these complements. Our automation strategy is general and not limited to microplastics data.

4.3. cGAN Performance

After *Aug** is identified, our system uses it to create augmented data to train a cGAN, which we call the microplastics cGAN or MPcGAN. In this section, we consider the performance of MPcGAN.

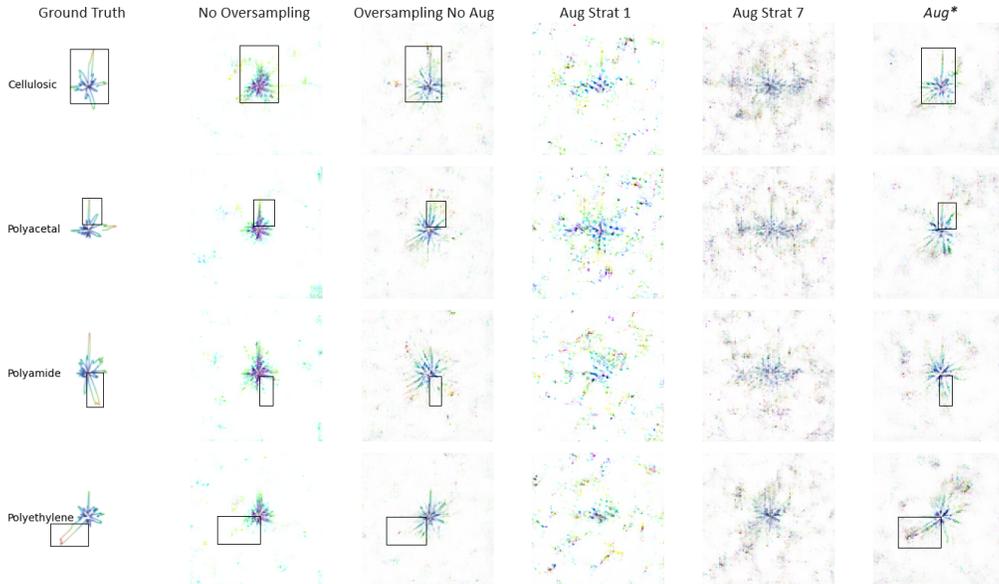


Figure 3. Qualitative results of SYMP data generated from MPcGAN instances trained on the top 3 strategies and baselines. We include SYMP samples from the first 4 classes. Boxes are drawn to highlight significant class-dependent features. The 20 samples were generated in a batch and not cherry picked. MPcGAN trained on data oversampled with *Aug** results in the best SYMP in terms of ground truth resemblance.

For this test, we trained 5 different MPcGAN models, each on a different data set. These 5 data sets were crafted using the top 3 performing augmentation strategies (strategies 1, 4, and 7), balanced oversampling on the 210 total real images (*Oversampling No Aug*), and the unbalanced data set consisting only of the real images (*No Oversampling*). The augmented data sets consisted of the 210 real images and 390 augmented images to create 500 total images balanced across classes.

For each cGAN model, during training iterations, the discriminator is trained on a batch of size 128, consisting of 64 real samples and 64 synthetic samples generated by the generator. The generator is trained by passing generated images through the frozen discriminator, and using the discriminator output as an error signal. A train–test split of 0.86–0.14 was used to train the discriminator. MPcGAN is designed to generate images of size 128x128x3. The best found settings for the generator and discriminator learning rates were 0.002, and 0.0002, respectively. The generator and discriminator both use Adam optimizer with binary cross entropy as the loss function [33]. Dropout is used in the discriminator for regularization, and Leaky ReLus are used for activation functions in the cGAN [34, 35].

MPcGAN collapses often, and many attempts were made to see how many epochs each training approach could endure before collapsing. We found that MPcGAN can train for roughly 150 epochs before collapsing when using the strategies *Aug**, *Oversampling No Aug*, or *No Oversampling*. While training MPcGAN on strategies 1 or 7, the cGAN is unstable and can only train up to 80 epochs before collapsing. Augmentation strategy 1 applies horizontal flipping and shifting. *Composite strategy 7* uses horizontal flipping and shifting combined with random masking. The unstable training is likely due to the effects strategies 1 and 7 have on the image spatial dimension, which cause the cGAN to lose focus during training, collapsing the network. Small data cGAN training is unstable, and MPcGAN requires multiple attempts at convergence, converging about 16% of the time.



Figure 4. Qualitative results of our *SYMP-Filter* algorithm. Three synthetic images of Cellulosic were generated using a cGAN trained on *Aug**. The two images on the left contain unprocessed cGAN output, while the rightmost image shows filtered output.

Figure 3 shows qualitative results of synthetic microplastic spectra images from 5 trained MPcGAN instances, each trained with different approaches. Synthetic spectra of 4 classes were generated: Cellulosic, polyacetal, polyamide, and polyethylene. Bounding boxes are drawn over class-dependent visual features of each polymer class. Overall, we found that MPcGAN trained on *Aug** outperforms other methods in terms of generated SYMP feature variance, visual quality, complexity, and ground truth resemblance.

In the bounding box of the ground truth polyethylene sample in figure 3, polyethylene has a characteristic left-downward diagonal peak with a red tip. This feature is common in the real images, and robust synthetic samples should inherit this feature. MPcGAN trained on *Aug** is the only method capable of producing this feature. This feature learning capability is also present in other examples. For example, polyamide (row 3) has a characteristic bottom-right diagonal peak, which *Aug** best mimics.

We also notice that *Aug** based SYMP better follows ground truth colouring by adding whiteness near the center, while baselines tend to fill the center with blue. Synthetic spectra generated from *Aug** are superior in quality with higher feature variance. SYMP generated without augmentation or oversampling lack variation. Samples generated with oversampling but no augmentation gain slight variation, but do not learn class specific features like samples generated from *Aug**.

We have computed Fréchet Inception Distance (FID) and Inception Scores (IS) scores for SYMP data in Table 2. FID and IS are two metrics used to evaluate the quality of synthetic data based on the available ground truth data distribution [36]. Low FID and high IS correspond to better sample quality. Low FID means the synthetic data better resembles the ground truth. High IS ensures that samples have high variance and clarity. The table also shows the performance before (*Regular*) and after (*Filtered*) applying the SYMP-Filter.

| Data Set | FID | | IS Mean | | IS Stdev | |
|---------------------|---------|----------|---------|----------|----------|----------|
| | Regular | Filtered | Regular | Filtered | Regular | Filtered |
| Aug* | 272.9 | 247.9 | 1.34 | 1.22 | 0.026 | 0.030 |
| Aug 1 | 288.2 | 287.1 | 1.13 | 1.11 | 0.012 | 0.018 |
| Aug 7 | 395.5 | 392.0 | 1.20 | 1.21 | 0.037 | 0.021 |
| No Oversampling | 216.0 | 214.3 | 1.14 | 1.12 | 0.062 | 0.011 |
| Oversampling No Aug | 261.0 | 226.5 | 1.31 | 1.16 | 0.028 | 0.018 |

Table 2. Establishing baseline FID and Inception Scores (IS) for SYMP data generation. Parentheses describe performances after applying the SYMP-Filter algorithm. Five MPcGAN training approaches are examined. Low FID and high IS are desirable.

Table 2 shows *Aug** achieves the best balance between FID and IS scores. The best FID is seen with no oversampling or augmentation, but this is due to the mostly white backgrounds in the SYMP images generated by these methods, meaning MPcGAN is unable to generate samples with variance, resulting in lower ISs. Oversampling without augmentation generally has strong results, but a significantly weaker IS than *Aug** in terms of both mean and standard deviation. No oversampling and oversampling without augmentation both resulted in strong empirical performance, but significantly lack variation. MPcGAN trained on the *data chooser module* top choice data set *Aug** produces the best SYMP samples when considered qualitatively, and also has a desirable balance of FID and IS performance.

5. Discussion and Conclusions

Microplastics are a problem of growing concern, and current research on the problem is obstructed by limited available data. Our work on GANsemble addresses the issues of small data and class imbalance by automating data augmentation selection and improving synthetic sample generation. We show the *data chooser module* can be used to automate augmentation selection on a small microplastics data set, and believe it should work in other small data settings as well. MPcGAN is a fairly simple generator and discriminator architecture, and enhancements to the architecture would improve the SYMP data generated. Future work should include implementations of GANsemble with augmentation strategies that do not apply spatial transformations to the image. New augmentations for GANsemble could also include variations of blurring, random erasing [37], or the CutMix algorithm [20].

We demonstrate the ability of GANsemble to automate augmentation selection and establish baselines for SYMP data generation in terms of FID and IS scores. The GANsemble *data chooser module* is capable of identifying an augmentation strategy which qualitatively and quantitatively outperforms all other compared methods. This is useful in domains where data collection is costly or there is low economic incentive. Our proposed MPcGAN algorithm is capable of learning to generate robust SYMP samples inheriting class dependent features. We encourage others to build on this work to further efforts in microplastics identification and mitigation.

Acknowledgements

We thank the anonymous reviewers for their feedback on this work. We also gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- [1] A. A. Koelmans, N. H. Mohamed Nor, E. Hermsen, M. Kooi, S. M. Mintenig, and J. De France. “Microplastics in freshwaters and drinking water: Critical review and assessment of data quality”. In: *Water Research* 155 (2019), pp. 410–422. ISSN: 0043-1354. DOI: <https://doi.org/10.1016/j.watres.2019.02.054>.
- [2] Y. Hong, S. Wu, and G. Wei. “Adverse effects of microplastics and nanoplastics on the reproductive system: A comprehensive review of fertility and potential harmful interactions”. In: *Science of The Total Environment* 903 (2023), p. 166258. ISSN: 0048-9697. DOI: <https://doi.org/10.1016/j.scitotenv.2023.166258>.
- [3] A. Ragusa, A. Svelato, C. Santacroce, P. Catalano, V. Notarstefano, O. Carnevali, F. Papa, M. C. A. Rongioletti, F. Baiocco, S. Draghi, E. D’Amore, D. Rinaldo, M. Matta, and E. Giorgini. “Plasticenta: First evidence of microplastics in human placenta”. In: *Environment International* 146 (2021), p. 106274. ISSN: 0160-4120. DOI: <https://doi.org/10.1016/j.envint.2020.106274>.

- [4] L. C. Jenner, J. M. Rotchell, R. T. Bennett, M. Cowen, V. Tentzeris, and L. R. Sadofsky. “Detection of microplastics in human lung tissue using μ FTIR spectroscopy”. In: *Science of The Total Environment* 831 (2022), p. 154907. ISSN: 0048-9697. DOI: <https://doi.org/10.1016/j.scitotenv.2022.154907>.
- [5] X. Tian, F. Beén, Y. Sun, P. van Thienen, and P. S. B auerlein. “Identification of Polymers with a Small Data Set of Mid-infrared Spectra: A Comparison between Machine Learning and Deep Learning Models”. In: *Environmental Science & Technology Letters* 10.11 (2023), pp. 1030–1035. DOI: [10.1021/acs.estlett.2c00949](https://doi.org/10.1021/acs.estlett.2c00949).
- [6] P. S. B auerlein, R. C. Hofman-Caris, E. N. Pieke, and T. L. ter Laak. “Fate of microplastics in the drinking water production”. In: *Water Research* 221 (2022), p. 118790. ISSN: 0043-1354. DOI: <https://doi.org/10.1016/j.watres.2022.118790>.
- [7] M. Wang and J. Deng. “Learning to Prove Theorems by Learning to Generate Theorems”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 18146–18157.
- [8] P. Kłosowski. “Deep Learning for Natural Language Processing and Language Modelling”. In: *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. 2018, pp. 223–228. DOI: [10.23919/SPA.2018.8563389](https://doi.org/10.23919/SPA.2018.8563389).
- [9] Z. Fayyaz, D. Platnick, H. Fayyaz, and N. Farsad. “Deep Unfolding for Iterative Stripe Noise Removal”. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–7. DOI: [10.1109/IJCNN55064.2022.9892708](https://doi.org/10.1109/IJCNN55064.2022.9892708).
- [10] L. Alzubaidi et al. “A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications”. In: *Journal of Big Data* 10.1 (2023), Article number: 46. DOI: [10.1186/s40537-023-00727-2](https://doi.org/10.1186/s40537-023-00727-2).
- [11] S. Das, S. S. Mullick, and I. Zelinka. “On Supervised Class-Imbalanced Learning: An Updated Perspective and Some Key Challenges”. In: *IEEE Transactions on Artificial Intelligence* 3.6 (2022), pp. 973–993. DOI: [10.1109/TAI.2022.3160658](https://doi.org/10.1109/TAI.2022.3160658).
- [12] O. Mogren. “C-RNN-GAN: Continuous recurrent neural networks with adversarial training”. In: *CoRR* abs/1611.09904 (2016). arXiv: [1611.09904](https://arxiv.org/abs/1611.09904) [cs.AI].
- [13] C. Shorten and T. M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6 (2019), pp. 1–48.
- [14] W. Zhang, L. Deng, and D. Wu. “Overcoming Negative Transfer: A Survey”. In: *CoRR* abs/2009.00909 (2020). arXiv: [2009.00909](https://arxiv.org/abs/2009.00909) [cs.LG].
- [15] I. Rather and S. Kumar. “Generative adversarial network based synthetic data training model for lightweight convolutional neural networks”. In: *Multimedia Tools and Applications* 83 (May 2023), pp. 1–23. DOI: [10.1007/s11042-023-15747-6](https://doi.org/10.1007/s11042-023-15747-6).
- [16] H. Rashid, M. A. Tanveer, and H. Aqeel Khan. “Skin Lesion Classification Using GAN based Data Augmentation”. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 916–919. DOI: [10.1109/EMBC.2019.8857905](https://doi.org/10.1109/EMBC.2019.8857905).
- [17] A. Gosain and S. Sardana. “Handling class imbalance problem using oversampling techniques: A review”. In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2017, pp. 79–85. DOI: [10.1109/ICACCI.2017.8125820](https://doi.org/10.1109/ICACCI.2017.8125820).
- [18] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail. “SMOTE for Handling Imbalanced Data Problem : A Review”. In: *2021 Sixth International Conference on Informatics and Computing (ICIC)*. 2021, pp. 1–8. DOI: [10.1109/ICIC54025.2021.9632912](https://doi.org/10.1109/ICIC54025.2021.9632912).
- [19] G. Zhou, Y. Chen, and C. Chien. “On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks”. In: *BMC Medical Informatics and Decision Making* 22 (Aug. 2022). DOI: [10.1186/s12911-022-01942-2](https://doi.org/10.1186/s12911-022-01942-2).
- [20] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. J. Yoo. “CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 6022–6031.
- [21] W. Obaid and A. B. Nassif. “The Effects of Resampling on Classifying Imbalanced Datasets”. In: *2022 Advances in Science and Engineering Technology International Conferences (ASET)*. 2022, pp. 1–6. DOI: [10.1109/ASET53988.2022.9735021](https://doi.org/10.1109/ASET53988.2022.9735021).

- [22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. *Generative Adversarial Networks*. 2014. arXiv: [1406.2661 \[stat.ML\]](#).
- [23] M. Mirza and S. Osindero. “Conditional Generative Adversarial Nets”. In: *CoRR* abs/1411.1784 (2014). arXiv: [1411.1784 \[cs.LG\]](#).
- [24] B. Vega, C. Rubio-Escudero, J. Riquelme, and I. Nepomuceno-Chamorro. “Creation of Synthetic Data with Conditional Generative Adversarial Networks”. In: Jan. 2020, pp. 231–240. ISBN: 978-3-658-07615-3. DOI: [10.1007/978-3-030-20055-8_22](#).
- [25] Y. Qin, H. Zheng, J. Yao, M. Zhou, and Y. Zhang. “Class-Balancing Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 18434–18443.
- [26] S. Dakshit and B. Prabhakaran. “CVAE-based Generator for Variable Length Synthetic ECG”. In: *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. 2023, pp. 235–244. DOI: [10.1109/ICHI57859.2023.00040](#).
- [27] X. Tian, F. Beén, and P. S. B auerlein. “Quantum cascade laser imaging (LDIR) and machine learning for the identification of environmentally exposed microplastics and polymers”. In: *Environmental Research* 212 (2022), p. 113569. ISSN: 0013-9351. DOI: <https://doi.org/10.1016/j.envres.2022.113569>.
- [28] W. Ng, B. Minasny, and A. McBratney. “Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy”. In: *The Science of the total environment* 702 (2020), p. 134723. ISSN: 0048-9697. DOI: [10.1016/j.scitotenv.2019.134723](#).
- [29] Z. Shen and R. Viscarra Rossel. “Automated spectroscopic modelling with optimised convolutional neural networks”. In: *Scientific Reports* 11 (Jan. 2021), p. 208. DOI: [10.1038/s41598-020-80486-9](#).
- [30] S. Pimpke, M. Godejohann, and G. Gerdts. “Rapid Identification and Quantification of Microplastics in the Environment by Quantum Cascade Laser-Based Hyperspectral Infrared Chemical Imaging”. In: *Environmental Science & Technology* 54.24 (2020). PMID: 33233891, pp. 15893–15903. DOI: [10.1021/acs.est.0c05722](#).
- [31] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](#).
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](#).
- [33] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.
- [34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.
- [35] B. Xu, N. Wang, T. Chen, and M. Li. “Empirical Evaluation of Rectified Activations in Convolutional Network”. In: *CoRR* abs/1505.00853 (2015). arXiv: [1505.00853 \[cs.LG\]](#).
- [36] M. J. Chong and D. Forsyth. “Effectively Unbiased FID and Inception Score and Where to Find Them”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6069–6078. DOI: [10.1109/CVPR42600.2020.00611](#).
- [37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. “Random Erasing Data Augmentation”. In: *CoRR* abs/1708.04896 (2017). arXiv: [1708.04896 \[cs.CV\]](#).