Discover Artificial Intelligence

Discover

Research

Idecomp: imbalance-aware decomposition for class-decomposed classification using conditional GANs

Patryk Buczek[1] · Usama Zidan[1] · Mohamed Medhat Gaber[1,2] · Mohammed M. Abdelsamea[1,3]

Received: 10 June 2023 / Accepted: 22 August 2023
Published online: 29 August 2023

Abstract
Medical image classification tasks frequently encounter challenges associated with class imbalance, resulting in biased model training and suboptimal classification performance. To address this issue, the combination of class decomposition and transfer learning has proven to be effective in classifying imbalanced medical imaging datasets. Nevertheless, in order to further augment the performance gains achieved through the utilisation of class decomposition within deep learning frameworks, we propose a novel model coined imbalance-Aware Decomposition for Class-Decomposed Classification (iDeComp) model. By incorporating a conditional Generative Adversarial Network (GAN) model, iDeComp is capable of generating additional samples specifically tailored to underrepresented decomposed subclasses. This paper investigates the application of iDeComp using two different medical imaging datasets. iDeComp selects underrepresented samples from the training set of the sublevel classes within each dataset, which are then employed to train separate conditional Deep Convolutional GAN (DCGAN) models and verification models. The conditional DCGAN model is responsible for generating additional samples, while the verification model critically evaluates the appropriateness of the synthesised images. Subsequently, the resulting augmented samples are utilized to train the classification model. To assess the effectiveness of iDeComp, we employ various evaluation metrics including accuracy, precision, recall, and F1 score. The results obtained from our experiments clearly indicate that iDeComp outperforms existing approaches in terms of classifying both imbalanced datasets.

Keywords Class imbalance · Conditional GAN · Transfer learning · Medical image classification

1 Introduction

Medical image classification is a critical part of computer-aided diagnosis (CAD), and research into deep learning approaches for medical image classification has proven their effectiveness. However, often with medical image classification tasks, the problem of class imbalance arises, an imbalanced dataset has far fewer samples for certain classes relative to the others, causing poor classification performance as the trained model is often biased for classifying samples belonging to the overrepresented classes.

In clinical settings, many medical image datasets suffer from the imbalance problem that hampers the detection of outliers (rare healthcare events), as most classification methods assume equal occurrences of classes. Consequently,

✉ Mohammed M. Abdelsamea, mohammed.abdelsamea@bcu.ac.uk; Patryk Buczek, patryk.buczek@mail.bcu.ac.uk; Usama Zidan, usama.zidan@bcu.ac.uk; Mohamed Medhat Gaber, mohamed.gaber@bcu.ac.uk | [1]School of Computing and Digital Technology, Birmingham City University, Birmingham 100190, UK. [2]Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt. [3]Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut 71515, Egypt.

the identification of outliers in unbalanced datasets has become a crucial issue [1, 2]. Several approaches have been proposed to address this issue. For example, in [3], a novel mechanism was proposed for sampling training data based on the popular MixUp regularization technique, called Balanced MixUp. Balanced-MixUp simultaneously performs regular (i.e. instance-based) and balanced (i.e. class-based) sampling of the training data. The resulting two sets of samples are then mixed up to create a more balanced training distribution from which a neural network can effectively learn without incurring heavy under-fitting of the minority classes. Some recent studies have applied these methods to medical image classification tasks and have achieved promising results. For example, in [4], a hybrid resampling method was proposed by combining random oversampling and synthetic minority oversampling (SMOTE) to balance the data for chest radiograph image classification. They showed that their method improved the F1 score and the area under the curve (AUC) of the classifier compared to other resampling methods. In [5], a cost-sensitive convolutional neural network (CNN) was proposed by incorporating a cost matrix into the loss function to handle class imbalance for the location of brain tumors. They showed that their two-stage deep learning framework method is able to deal with the high-class imbalance encountered during the training of small lesion detectors and can increase the sensitivity of the classifier compared to other cost-sensitive methods. These studies demonstrate the effectiveness of various methods to address the problem of class imbalance in medical image classification tasks.

By incorporating techniques such as resampling, cost-sensitive learning, and class decomposition, one can improve the performance of classifiers on imbalanced datasets. For example, a new technique named DeTraC was proposed in [6] that combines class decomposition with transfer learning to address the problem of unbalanced data sets in the classification of medical images. DeTraC incorporates a class decomposition approach paired with transfer learning in several learning scenarios [7–9]. The first step in DeTraC is Class Decomposition, parent classes are decomposed into n subclasses using an off-the-shelf feature extractor, the second step is Transfer Learning, the final layer of a pretrained CNN is adapted to classify the subclasses and the whole CNN is fine-tuned to maximise classification performance, the final step is Class Composition, where the final classification performance on the subclasses is evaluated. By dividing the original classes into their constituent subclasses, the DeTraC model was able to extract more specific and meaningful features from each subclass, resulting in improved accuracy. However, this method introduced an inherent side effect where the generated subclasses themselves became imbalanced. This issue of imbalance in subclasses can lead to a biased classification model with poor performance in underrepresented subclasses.

This paper introduces imbalance-Aware Decomposition for Class-Decomposed Classification (iDeComp) model, where a conditional deep convolutional generative adversarial network (cDCGAN) [10] is applied at the subclass level, reducing class imbalance by augmenting the dataset with new samples synthesised by the cDCGAN. The subclasses will be divided into training, validation, and testing subsets, and the underrepresented classes from the training set will be used to train a cDCGAN model, the trained model will then generate *y* samples per underrepresented class that will be used to augment the training set, to reduce the problem of class imbalance. iDeComp is able to generate synthetic samples to balance the subclass distribution. The resulting balanced subclass dataset is then used to train a new classification model, which we demonstrate achieves improved performance compared to the DeTraC model. Given the demonstrated improved accuracy of DeTraC by extracting more specific and meaningful features from the subclasses, iDeComp addresses the resulting imbalance in underrepresented subclasses. By extending on work done in the DeTraC approach, a cGAN model is implemented at the subclass level to increase the sample count for underrepresented classes, reducing the problem of class imbalance. This extension is particularly significant due to its ability to surpass the performance of DeTraC, which has already demonstrated considerable improvements in classification accuracy compared to conventional transfer learning and classification approaches. The ability to address class imbalance within subclasses effectively contributes to more accurate and reliable classification results. This novel approach offers a promising solution to the challenges faced in handling unbalanced datasets, particularly in the medical image analysis domain.

In summary, our proposed method makes significant contributions in the following ways.

1. We utilise conditional GANs to effectively address the issue of class imbalance in class decomposed classification at the subclass level. This approach allows for more accurate and reliable classification results.
2. Our method is particularly useful for medical image classification, where limited labelled images often lead to the emergence of class imbalance after decomposition of classes. By applying our approach to a range of medical image datasets, we demonstrate its effectiveness in improving classification accuracy.

3. Through extensive experimentation, we provide empirical evidence of the efficacy of our method. Our results show that our approach outperforms the typical class decomposition-based classification in addressing emerging class imbalance and improving classification accuracy.

The paper is organised into several sections. In Sect. 2, we review the related work on the problem of class imbalance in medical image classification. In Sect. 3, we describe the methodology of our approach, including the use of a conditional deep convolutional generative adversarial network (cDCGAN) to generate synthetic samples and balance the subclass distribution. In Sect. 4, we present the results of our experiments, demonstrating the improved performance of our iDeComp model compared to the DeTraC model. Finally, in Sect. 5, we conclude the paper with a summary of our findings and suggestions for future work.

2 Related work

This section aims to review similar work in the field of image classification using deep learning and GAN augmentation. Deep learning has established itself as a pivotal machine learning technique in the modern world, revolutionising the computer vision community and introducing non-traditional and effective solutions to numerous challenging problems [11]. Recent work has investigated the potential of deep learning to automatically learn multiple levels of representations of the underlying data distribution for modelling purposes. Specifically, deep learning techniques have been shown to effectively extract both low- and high-level features necessary for classification tasks [12]. Generative adversary networks are generative models based on a deep learning architecture, generative models learn to capture the statistical distribution of training data, allowing us to synthesise samples from the learned distribution [13]. GANs often comprise a generator and a discriminator that operate to mutually learn and generate new data. The generator model tries to capture the potential distribution of the samples and generate new data samples, while the discriminator is often a binary classifier, discriminating real samples from generated samples [14]. GAN models can be used as a form of data augmentation [15], investigate the performance of conditional GAN (cGAN) augmentation for image classification, the implementation of cGAN improves performance in terms of precision and recall and the F1 score. Furthermore [16], discuss how GANs have been used to generate and design DNA, drug discovery, generate multilabel discrete patient records, medical image processing, and doctor recommendation. However, GAN models are prone to training failure [17], show that training is commonly unstable, and the weight parameters of the GAN easily diverge due to its adversarial training process, further explaining how these incorrectly trained GANs generally produce identical samples regardless of input noise, this being known as the mode collapse problem. Similarly, when developing a GAN network a phenomenon can occur known as the vanishing gradient [18]. Observed it when the discriminator becomes too proficient, leading to a potential failure in generator training. They highlight that an optimal discriminator does not provide sufficient feedback for the Generator to learn properly. Furthermore, data augmentation using GAN has its limitations, one of these being its ability to generate images with a high enough quality [19] and its ability to produce images with perfect fidelity. However, complete realism is not necessary to improve the classification results with synthetic data [19, 20]. Discuss how low volume, high sparsity, and poor quality of data can severely diminish the performance of deep learning models, indicating that deep learning models are very data-dependent. Further investigation by [21] showed the performance of data augmentation with an Auxiliary Classifier Generative Adversarial Network (ACGAN); they found an increase in classification performance 10% when training with actual and synthetic images. However, the authors also noted that ACGAN has limitations, such as the limited quality of the synthesised samples and that including more labelled data can improve its performance. Furthermore, they discuss how the dataset used in their study is obtained from various sources, and cross-centre validations were not conducted, which can result in errors in data labelling, leading to a higher effect on the quality of synthesised samples with a small dataset (Fig. 1).

3 Materials and methods

This section provides a detailed account of the different stages involved in the iDeComp framework. We begin with a comprehensive overview of the datasets used in this study, followed by a detailed exposition of the preprocessing procedures that were implemented on each dataset. Subsequently, the methodology adopted for class decomposition, which involved using an existing feature extractor, is elucidated, as well as the approach to training the

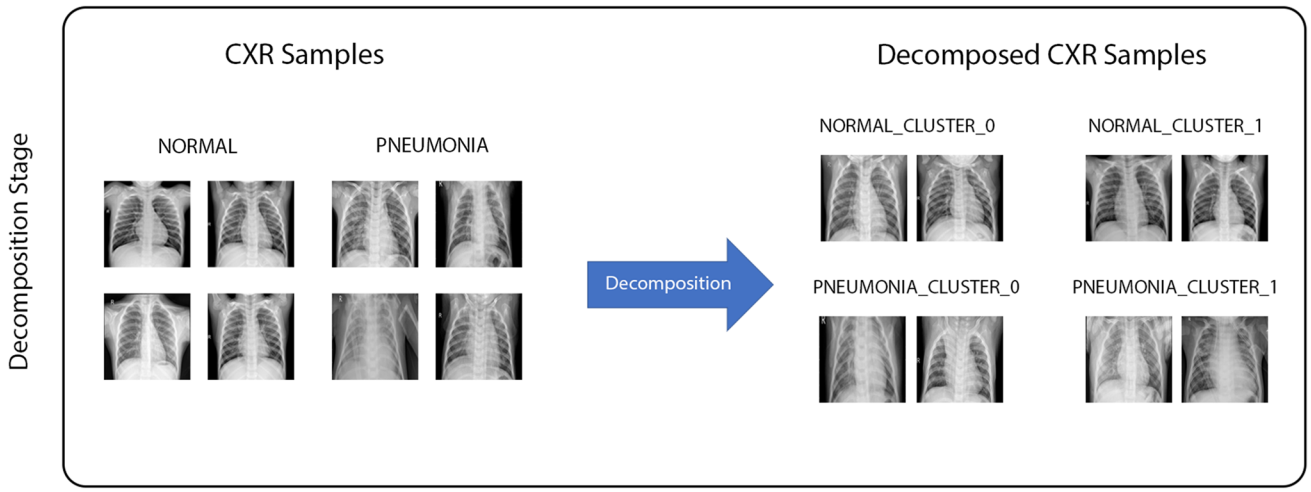


Fig. 1 Class decomposition involves learning more granular classes using a feature extractor. This process results in a larger number of classes with more specific features. The decomposed classes are then used in model training to improve classification performance and accuracy

verification classifier that was used to authenticate the samples synthesized by the cDCGAN model. The architecture of the cDCGAN model, as well as the intricacies of its training process, are presented. In addition, the section outlines the augmentation process using cDCGAN and its role in the final DeTraC model training process.

3.1 Datasets

The first dataset obtained for this study is the "NCT-CRC-HE-100K" (CRC) dataset collected as part of a study by [22]. This dataset contains 100,000 histological images of human colorectal cancer and normal tissue. All images are at a resolution of 224×224 at 0.5 μm per pixel, stored in TIF format. The images are normalised before being released through a process known as stain normalisation. The images are separated into nine different classes that are not equally distributed, Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

The second dataset obtained for this study is a chest X-ray data set (CXR); it was part of research conducted by [23], a publication investigating deep learning approaches to medical image classification. The chest X-ray image dataset was one of the datasets used for the research and published for use in further research. This data set contains 5858 samples of chest radiographs classified into 2 classes, 'normal' or 'pneumonia' cases, the images vary in resolution and are front chest radiographs depicting the lungs.

3.2 Pre-processing

During the preprocessing stage, images from both datasets undergo several steps to prepare them for class decomposition and train the generative model. First, corrupt files are removed from the dataset. The images are then converted to PNG format and resized to a resolution of 256×256. We split the datasets into training, validation, and testing sets with a ratio of 80:10:10, respectively.

For the chest X-ray dataset, the training and testing splits are combined into a single dataset and stored in a root directory with subdirectories "PNEUMONIA" and "NORMAL".

For the CRC dataset, the images initially have a resolution of 224×224 and are stretched to a resolution of 256×256 to match the requirements of the cDCGAN model. Since the cDCGAN can only accept inputs and generate samples with a resolution that is a power of 2. As a result, the images are stretched to the nearest higher resolution, which is 256×256 in this case.

3.3 Class decomposition

The first step in DeTraC is class decomposition, in which an unsupervised feature extractor is applied to the input images and preprocessed using principal component analysis and k-means clustering, leading to a final decomposition of the original class space into a *N* subclasses (Fig. 2).

We employed the Xception [24] network as the backbone for feature extraction and classification. The Xception network is a deep convolutional neural network architecture that has been pre-trained on the ImageNet dataset. It consists of multiple layers of convolutional and depthwise separable convolution operations, followed by batch normalization and activation functions. Due to the high dimensionality associated with the images, PCA is applied to project the high-dimension feature space into a lower dimension, ignoring highly correlated features. This allows for class decomposition to produce more homogeneous classes. After class decomposition, the data is split at a subclass level into training, validation, and testing sets with a ratio of 80%, 10%, and 10%.

3.4 Identifying underrepresented classes

3.4.1 Chest X-ray dataset

When looking at the mean number of samples per class, 1171, the subclasses $NORMAL\_CLUSTER\_0$ and $NORMAL\_CLUSTER\_1$ are underrepresented. These 2 subclasses will be augmented using the cDCGAN model.

3.4.2 NCT HE CRC 100K dataset

In the CRC dataset, a less drastic class imbalance can be seen however, the following classes will be augmented; ADI_CLUSTER_0, LYM_CLUSTER_1, MUC_CLUSTER_0, MUC_CLUSTER_1, MUS_CLUSTER_0, NORM_CLUSTER_0, NORM_CLUSTER_1 and TUM_CLUSTER_1 (Fig. 3).

3.5 Verification classifier

The verification classifier is a separate model trained using only the underrepresented classes of the training set; the purpose of this model is to verify that the samples synthesised by the cDCGAN model are representative of the real dataset.

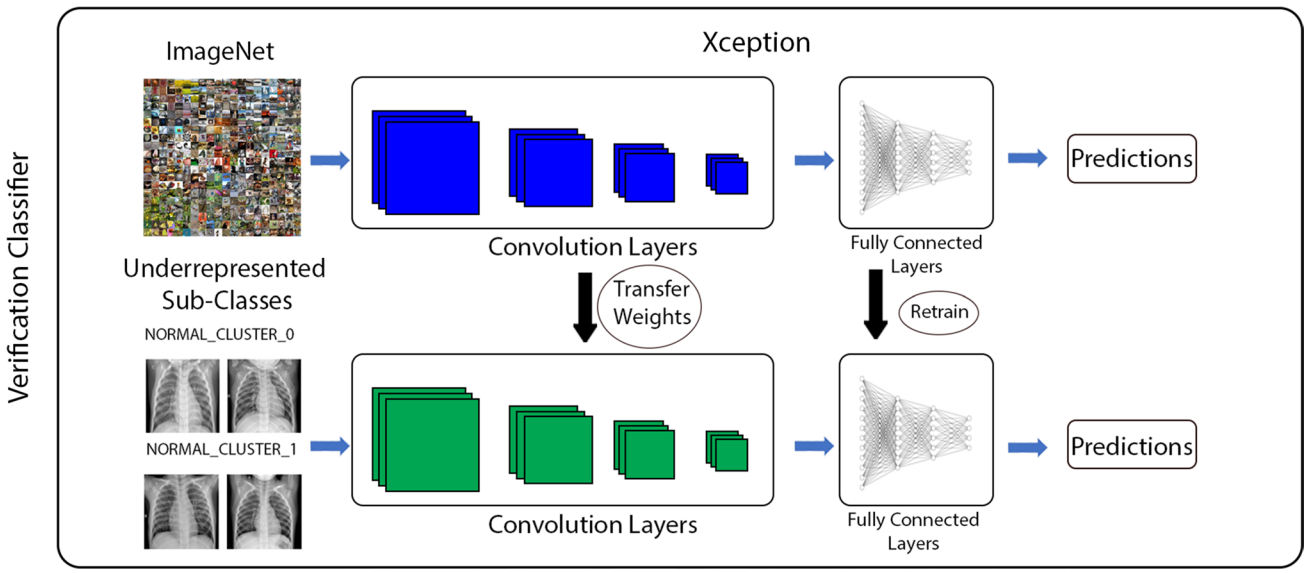


Fig. 2 Verification classification is used to validate the usability of the synthesized samples generated by the cDCGAN model. The verification classifier is trained on a set of high-quality features extracted from the original data and is used to distinguish between authentic and synthetic samples. The accuracy of the verification classifier is used to evaluate the quality of synthesized samples and determine their usability for downstream applications

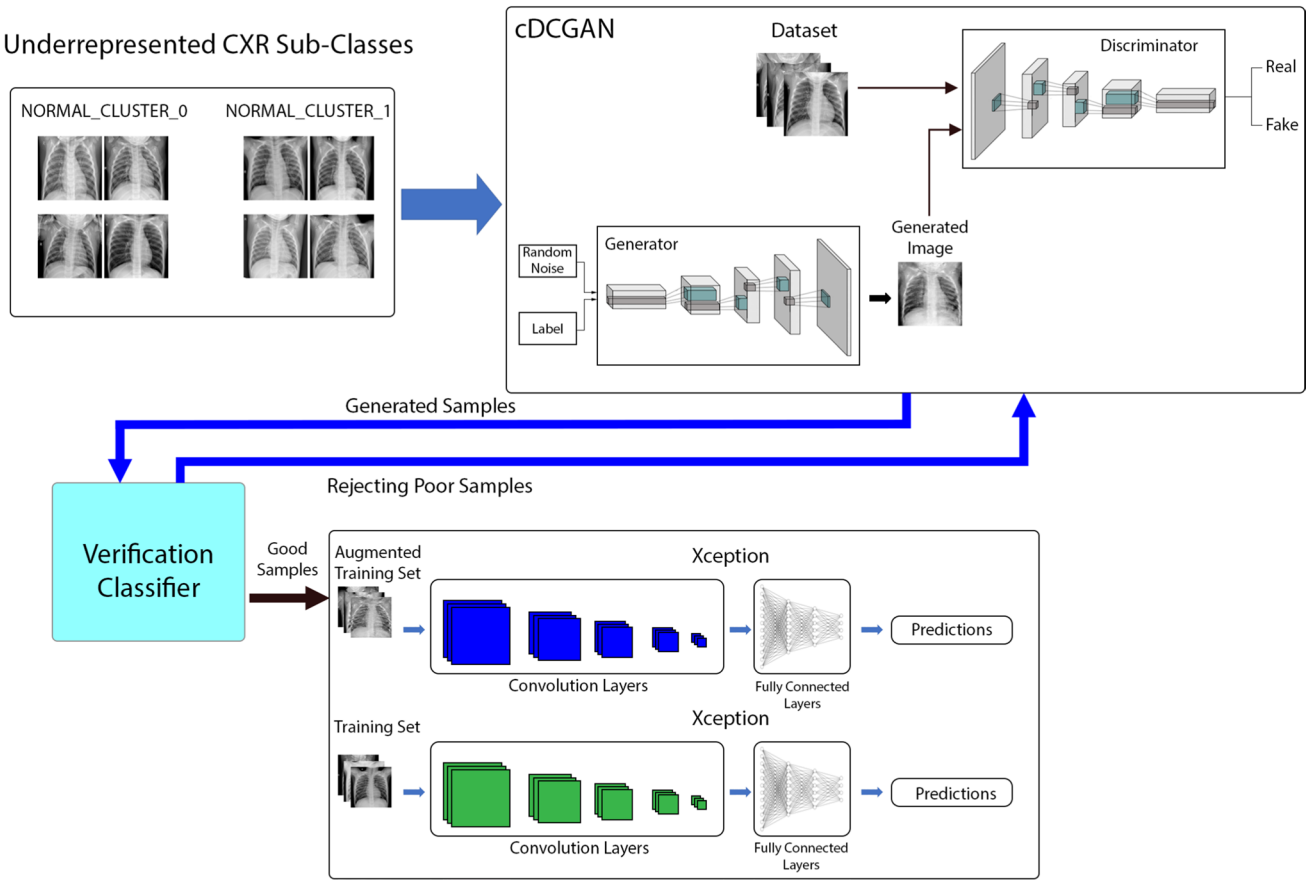


Fig. 3 The iDeComp model is designed to augment the dataset with new synthetic samples generated by the cDCGAN in order to increase the number of under-represented subclasses in the training set. Subsequently, the trained cDCGAN model is used to generate a specified number of samples per underrepresented class, which are added to the training set to reduce the problem of class imbalance

This is achieved by comparing the verification classifier's performance on the validation set to a fully augmented set. If the model is able to classify the augmented set to a high degree, similar to the validation, then it can be inferred that the synthesized samples are representative of the real data. A pretrained Xception model was loaded in with weights initialised from the ImageNet dataset, the classification layer was removed and replaced with a new layer for classifying images belonging to the underrepresented classes from the chest X-ray and CRC 100K datasets, respectively. The model was modified to accept 3 channel inputs at a resolution of 150×150. The base backbone of the model was frozen during training on the new data over 5 epochs with a batch size of 16 and a learning rate of 0.001. After initial training, the backbone was unfrozen and the entire model was retrained on the new data with a very low learning rate of 0.0001 for 6 epochs. The resulting models for the CXR and CRC datasets are able to classify the validation sets with an accuracy of 82.2% and 94.9%, respectively. This suggests that the CRC model will be more strict compared to the CXR model, indicating that it will be harder for the cDCGAN to produce samples that are classified correctly by the verification model, as it has been trained to higher accuracy. For each sample in the validation set, the probability of belonging to its own class will be determined by the verification classifier, and a mean probability for each class in the validation set will be calculated. During image synthesis, any generated image with a probability of belonging to its corresponding class that is lower than the mean value for that class will be discarded (Fig. 4).

3.6 cDCGAN architecture

The cDCGAN architecture is a type of generative adversarial network (GAN) that is capable of generating high-quality images. Specifically, the cDCGAN model consists of two neural networks, a generator, and a discriminator, that are trained in an adversarial manner to generate realistic images from a given input. Let G be the generator and D be the discriminator. The

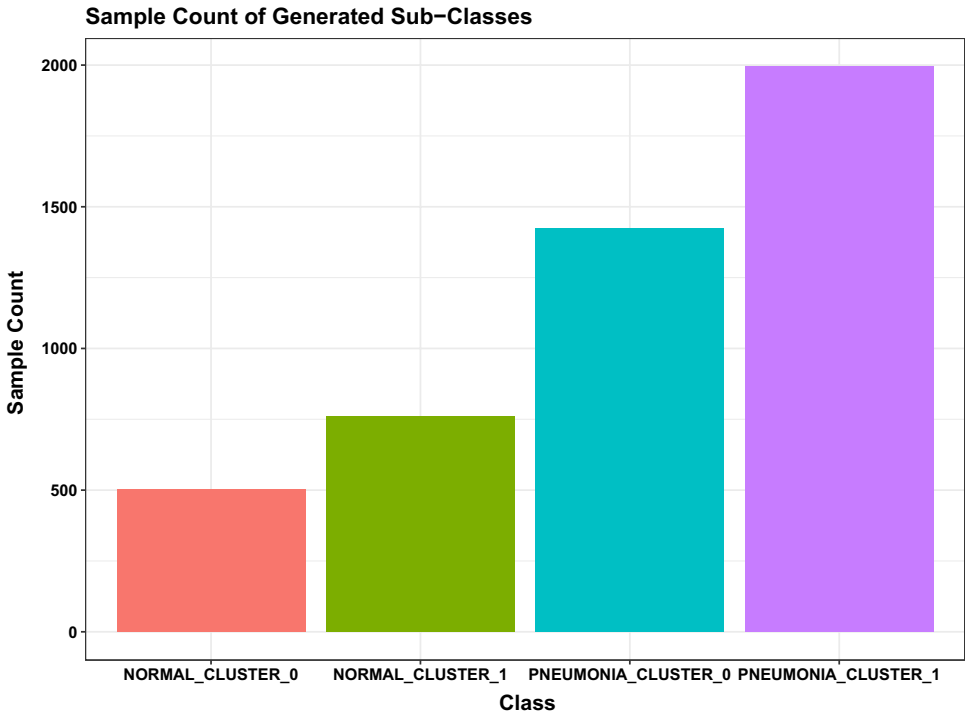


Fig. 4 Sample count for chest X-ray dataset

generator takes as input a text description *x* and generates an image *y*. The discriminator takes as input an image *z* and tries to distinguish whether it was generated by the generator or was taken from a real dataset *R*.

The generator is trained to minimise the following loss function:

$$\mathcal{L}G = -\mathbb{E}x \sim p(x)[\log D(G(x))].$$

The discriminator is trained to maximise the following loss function:

$$\mathcal{L}D = \mathbb{E}x \sim p(x)[\log D(G(x))] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(z))],$$

where $p(x)$ is the distribution of text descriptions and $p(z)$ is the distribution of real images (Fig. 5).

The generator and discriminator are trained alternately until they reach a Nash equilibrium. At this point, the generator is able to generate images that are indistinguishable from real images.

We used the cDCGAN model to generate 256×256 images, with 1 channel image being produced for the CXR dataset and 3 channel images being produced for the CRC dataset.

The generator model takes a random noise input Z and a label input y. The inputs are then processed through multiple layers; initially, a 2D transposed convolution is applied with 1024 filters, followed by a 2D batch normalisation with 1024 features, and finally, the ReLU activation function is applied. The concatenated noise and label are then processed through multiple deconvolutional layers, batch normalisation layers, and ReLU activation layers until finally a $256 \times 256 \times 1$ image is generated. The discriminator network accepts an input of an image and label; the inputs are convoluted and concatenated through the network to arrive at a final sigmoid layer, which produces a prediction of whether the image is real or fake.

An identical architecture was used for the CRC dataset, however, the image produced is at a resolution of $256 \times 256 \times 3$ and the inputs for the generator model have a random noise of dimension and label of 64 and 8 dimensions, respectively (Fig. 6).

3.7 cDCGAN training

3.7.1 Chest X-ray dataset

The cDCGAN model for the CXR dataset was trained over 60 epochs, with a batch size of 4 and a learning rate of 1e−05. The verification classifier is able to classify the synthesised images with an accuracy of 90.5%, which means that they

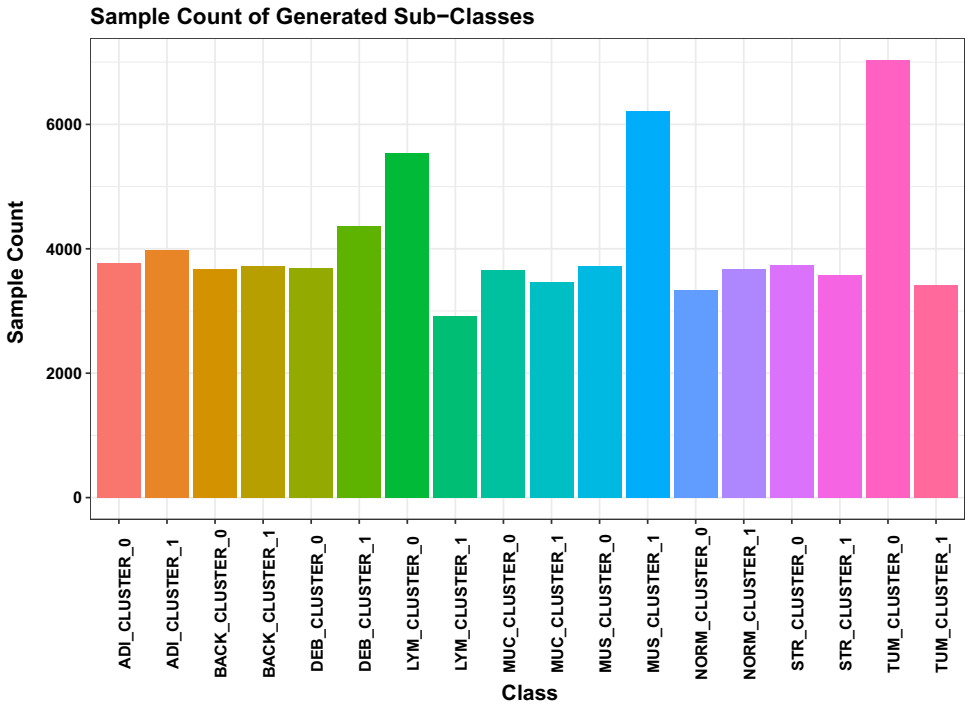
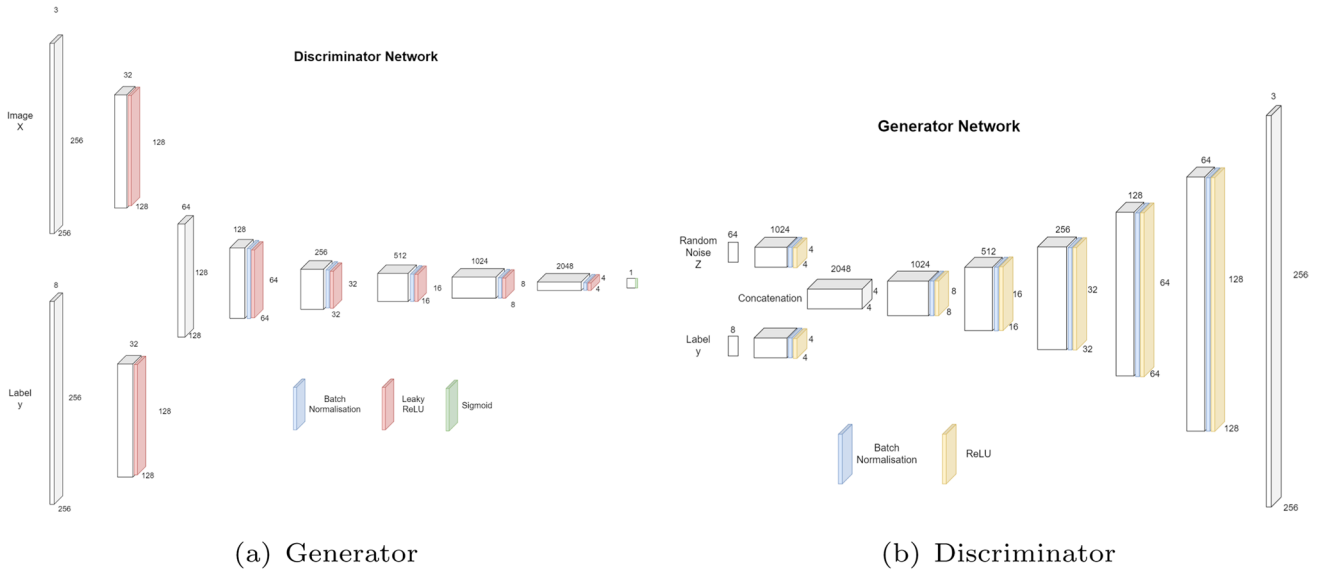


Fig. 5 Sample count for CRC dataset



(a) Generator (b) Discriminator

Fig. 6 Illustration of the cDCGAN architecture. The cDCGAN model consists of two neural networks, a generator and a discriminator, that are trained in an adversarial manner to generate realistic images from a given input

are representative of the real data set. Although the images appear to lack clarity, the deeper features that define class boundaries have been retained as the verification classifier is able to accurately classify the samples.

3.7.2 NCT HE CRC 100K dataset

The cDCGAN model for the CRC dataset was trained over 25 epochs, with a batch size of 64 and a learning rate of 2e−05. The final samples are representative of the real dataset as the verification classifier achieves an accuracy of 94.9%, but certain classes appear to produce samples that are nearly identical, which means that the model is suffering from mode collapse for these classes (Figs. 7, 8).

3.8 Data augmentation

The quality of each image is evaluated during synthesis, after an image is synthesised, it is immediately verified using the verification classifier, where the verification classifier compares the probability of the synthesised image belonging to its own class to the pre-calculated mean probability of all real images in the training set belonging to that class. If the probability is equal to or higher than the mean probability for that class, the image is accepted and used in the augmented training set, if the probability is less, the image is rejected and a new image is synthesised. This quality control measure ensures that only high-quality images are used for the augmentation and training of the final model.

In the chest X-ray dataset, Table 1 shows that the training set was increased from 4684 samples to 6455, a 37.8% increase in total sample count. The augmented classes being $NORMAL\_CLUSTER\_0$ and $NORMAL\_CLUSTER\_1$, the sample count was increased from 504 and 762 to 1504 and 1533, respectively. The increase in augmented data for each class is illustrated in Table 1. On the other hand, the CRC dataset saw a modest 5.7% increase in total sample count from 73,398 to 77,598, equivalent to an increase of 4200 samples. A detailed breakdown of the data increase for each class can be found in Table 1.

3.9 iDeComp training

The iDeComp model is designed to augment the dataset with new synthetic samples generated by the cDCGAN to increase the number of underrepresented classes in the training set. The subclasses are first split into training, validation, and testing subsets, and the under-represented classes in the training set are then used to train a cDCGAN model. The trained cDCGAN model is subsequently used to generate a specified number of samples per underrepresented class, which are added to the training set to reduce the problem of class imbalance.

To train both the DeTraC and iDeComp models, a base model of Xception was used with more than 30 epochs. For the CXR dataset, images at a resolution of 150×150 were used, while for the CRC dataset, images at a resolution of 100×100 were used. The training process involved feeding the images through the base model to extract relevant features, which were then used to train the cDCGAN model for the iDeComp approach. The process was iterated over 30 epochs, and the models were evaluated based on their ability to mitigate class imbalance and improve overall performance.

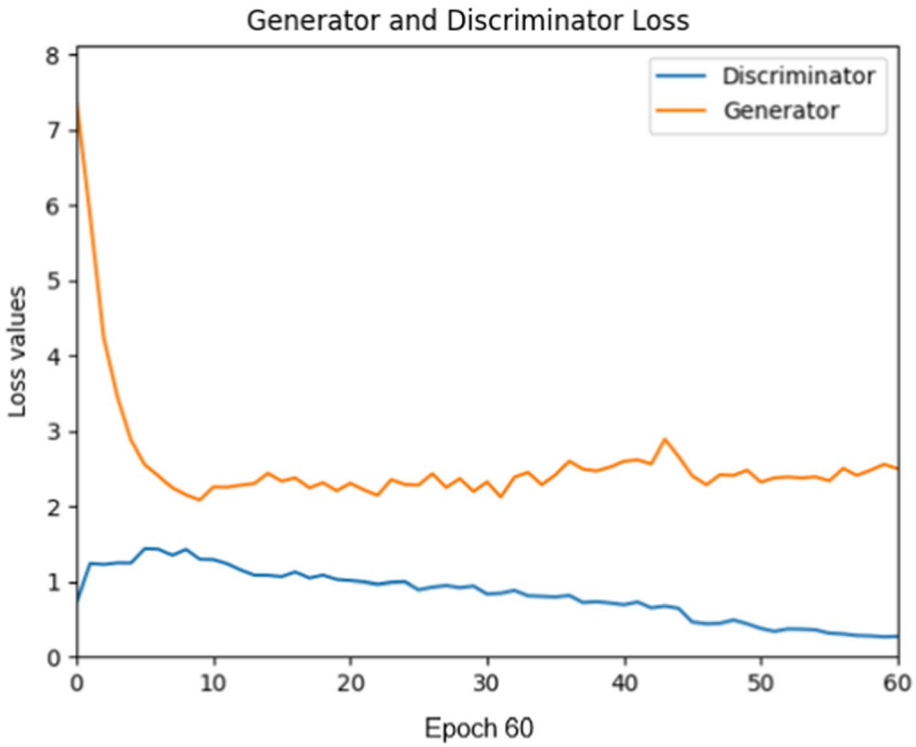


Fig. 7 Chest x-ray dataset cDCGAN generator and discriminator loss curve

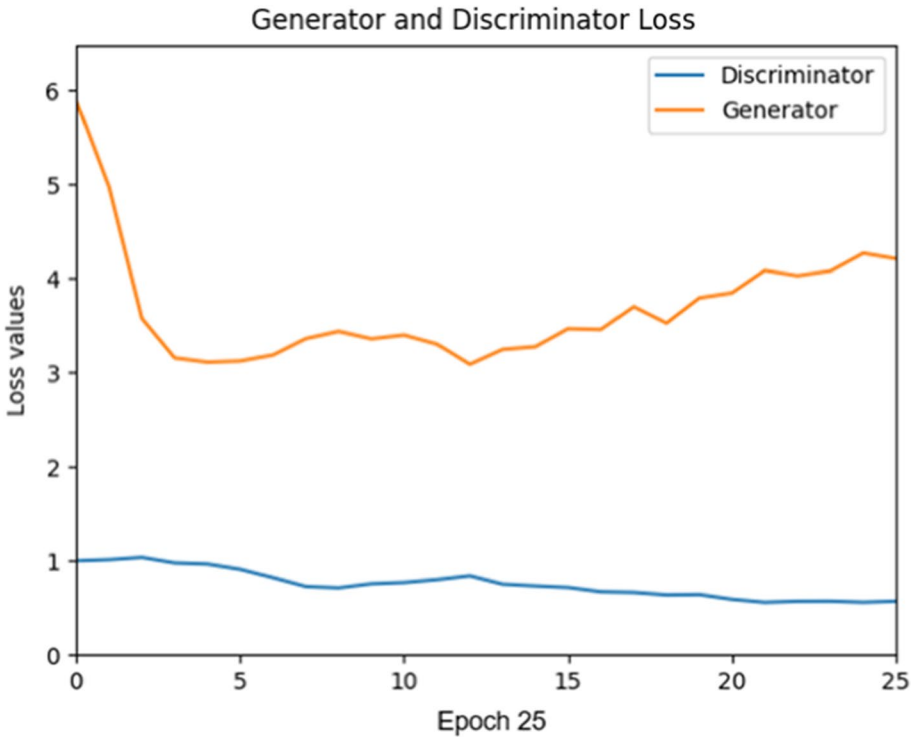


Fig. 8 CRC dataset cDCGAN generator and discriminator loss curve

Table 1 Augmented class count for under-represented classes in CRC and CXR datasets

Class	Training set	Augmented training set	Increase
CRC ADI(0)	3774	4074	+300
CRC LYM(0)	2916	3916	+1000
CRC MUC(0)	3656	4056	+400
CRC MUC(1)	3460	4060	+600
CRC MUS(0)	3725	4025	+300
CRC NORM(0)	3332	4032	+700
CRC NORM(1)	3677	3977	+300
CRC TUM(1)	3412	4012	+600
CXR NORM(0)	504	1504	+1000
CXR NORM(1)	762	1533	+771

Each dataset has subclasses that are generated as part of the DeTraC process; we further augment these classes using a cDCGAN model to generate samples that help mitigate the imbalance of the classification problem

4 Results

We evaluate the performance of our model using common practice metrics such as accuracy, precision, recall, and F1 score.

Our evaluations show that iDeComp outperforms DeTraC in terms of accuracy and the weighted average of precision, recall, and F1 score for the CXR dataset. Accuracy is higher by 0.5%. DeTraC precision is 0.956, whereas the iDeComp precision is 0.960, indicating a higher proportion of positive identifications were correct. Similarly, iDeComp achieves a higher recall score of 0.959 compared to 0.954 for DeTraC. Finally, iDeComp achieves a higher F1 score of 0.960 compared to 0.954 of DeTraC, indicating a better balance of precision and recall and overall higher quality predictions.

Similarly, for the CRC 100K dataset, iDeComp has outperformed DeTraC in all evaluation metrics. iDeComp outperformed DeTraC in all evaluation metrics, achieving an accuracy of 94.9%, a precision score of 0.950, and a recall

Table 2 Detailed evaluation metrics breakdown of DeTraC and iDeComp models on CXR and CRC datasets

	CXR DeTraC			CXR iDeComp		
	Accuracy	0.954		Accuracy	0.959	
Macro_avg	Precision	Recall	F1-Score	Precision	Recall	F1-Score
	0.968	0.917	0.939	0.953	0.942	0.948
Weighted_avg	Precision	Recall	F1-Score	Precision	Recall	F1-Score
	0.956	0.954	0.953	0.960	0.959	0.960

Table 3 Summarised CXR and CRC results obtained by iDeComp and DeTraC in terms of accuracy (ACC), precision (PR), recall (R), and F1-score (F1)

CXR results					CRC results			
Model	ACC	PR	R	F1	ACC	PR	R	F1
DeTraC	0.954	0.956	0.954	0.954	0.936	0.942	0.936	0.937
iDeComp	0.959	0.960	0.959	0.960	0.949	0.950	0.950	0.950

score of 0.950, compared to DeTraC's accuracy of 93.6%, a precision score of 0.942 and a recall score of 0.936. This demonstrates that iDeComp was able to correctly classify a higher proportion of positive identifications and correctly identify a higher proportion of actual positives. The resulting F1 score of 0.950 for iDeComp is also higher than DeTraC's F1 score of 0.937, indicating that iDeComp achieved a better balance of precision and recall and overall performed better than DeTraC (Tables 2, 3).

5 Discussion

In the experimental analysis, we obtained promising results that support the effectiveness of the proposed iDeComp approach. Evaluation metrics, including precision, precision, recall, and F1 score, consistently demonstrated superior performance compared to existing methods. Analysing the experimental data in more depth reveals important insights. We observed that iDeComp significantly improved classification accuracy for underrepresented classes, effectively mitigating the problem of class imbalance. The augmentation of the training set with synthetic samples generated by the cDCGAN played a crucial role in enhancing the representation and learning capabilities of the model. This resulted in better discrimination and classification accuracy, particularly for classes that were previously underrepresented (Fig. 9).

The cDCGAN model developed for the CRC 100k dataset successfully generated images that were representative of the real dataset, but suffered mode collapse for certain classes. As previously mentioned, the variance is not visible when analysing the images with the naked eye; however, there may be some variance in the deeper features of the images. Furthermore, real samples appear to have complex patterns and shapes that the cDCGAN model was not able to replicate; the complex patterns are not visible on the synthesised images. Due to time constraints, not every underrepresented class was able to be augmented, only the classes with the most severe imbalance. It may be useful to investigate training a model that can produce samples for each of the underrepresented decomposed classes. To improve the quality of the samples produced, further hyperparameter tuning could be investigated, with the use of more filters and different batch sizes. Furthermore, augmenting the training data could be investigated; the CRC 100K dataset has samples that are representative of their class regardless of rotation/angle, whereas MNIST for example, letters and numbers could not be rotated before training. Adding random flips and transformations to the training data could potentially reduce the problem of mode collapse and yield better results. Although the samples produced were flawed, using the verification classifier it was verified that they were representative of the real dataset, because the classifier was able to classify the samples with an accuracy of 96.5%.

The cDCGAN model developed for the chest X-ray dataset was able to produce samples at a resolution of 256 × 256 with a lot of variances between the images produced. However, the images produced were blurry and lacked fine detail, resembling smaller-resolution images stretched to a resolution of 256 × 256. As opposed to the CRC 100k dataset, this model did not have the problem of mode collapse; however, potentially adding more filters or tuning the hyperparameters could yield a model able to synthesise less blurry images. Although the images were blurry, they retained and represented the deeper features of the underlying distribution well enough for the verification classifier with an accuracy of 90.5%.

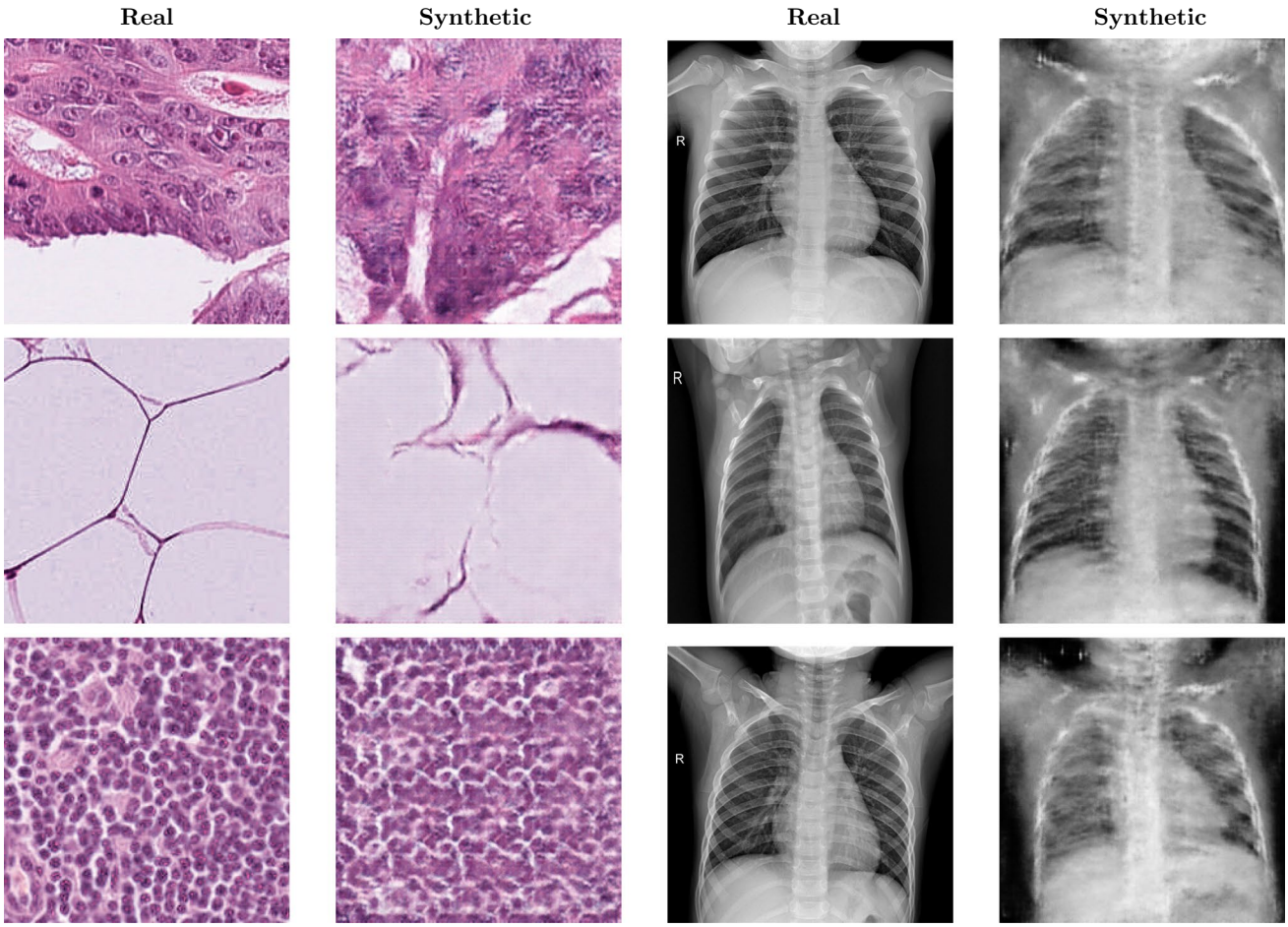


Fig. 9 Comparison of real and synthetic images generated by the cGAN model. The left column shows real images from the dataset, while the right column displays synthetic images produced by the cGAN

Both verification classifiers for the CRC 100K and chest x-ray dataset were trained for a low number of epochs; one way their classification performance could be improved is through training for more epochs, using data augmentation or tuning the hyper-parameters of the model. However, improving the performance of the verifier may have negative consequences on the samples produced by the corresponding cDCGAN model. Having a very sensitive verifier that is trained to a high accuracy on the real data, the cDCGAN model may not be able to produce samples that are convincing enough to be verified by the model; this may result in the cDCGAN model not being able to produce samples for augmentation. Furthermore, during training of the cDCGAN model, samples were not cross-validated with the verifier after completion of each epoch; this could be one way of improving the performance of the cDCGAN model. After each epoch, a sample set of the classes could be synthesised, these would then be verified by the verifier and the corresponding accuracy could be logged throughout the training process, this additional metric could be used to monitor the performance of the model during training and how representative the generated samples are of the real dataset.

The evaluation of iDeComp demonstrated its higher performance compared to that of DeTraC, despite the limitations of the cDCGAN models used for image synthesis. Future research could investigate the application of iDeComp using a more complex cDCGAN model. Additionally, splitting classes into more subclasses could potentially improve final classification results, and exploring higher resolutions for classification could lead to better results by retaining more detail in the images. Other limitations can be further explored such as the reliance on synthetic samples generated by one cDCGAN model accurately representing the underlying distribution of the underrepresented classes. While efforts have been made to verify the representative nature of these samples using a verification classifier, there may still be instances where the synthetic samples deviate from the true data distribution, leading to potential classification errors.

iDeComp is suitable for both binary and multi class imbalance problems as during class decomposition any binary class imbalance problem will be transformed into a multi class problem, where the minimum number of sub-classes will

be 4 if each class in a binary problem is decomposed into 2 sub-classes. iDeComp will be trained using the 4 sub-classes and the during class composition the sub-classes will be composed back into their original classes for final classification.

6 Conclusion

In this study, we propose a novel model, called Imbalance-Aware Decomposition for Class-Decomposed Classification (iDeComp) model, to address the issue of class imbalance in medical image classification. Using the effectiveness of DeTraC to improve classification performance in imbalanced datasets, iDeComp exhibited superior performance in all evaluation metrics compared to DeTraC. We evaluated our model on two medical imaging datasets by augmenting them into a learned subclass dataset. Our training pipeline utilised a verification classifier and a cDCGAN model trained on the underrepresented subclasses, which allowed us to generate more training data for the respective sets. The performance of DeTraC on augmented and non-augmented training sets was demonstrated, and the iDeComp method successfully improved DeTraC's performance. Further exploration of iDeComp approach would provide valuable insights into the generalizability and robustness of iDeComp beyond the specific dataset used in this study. Investigating the integration of the verification classifier during the training phase can enhance the training process and optimise the quality of the synthesised images. Providing an additional metric demonstrating how synthesised images represent the real corresponding classes. By incorporating this information into the training process, we can obtain more comprehensive feedback on image quality, complementing the traditional generator and discriminator loss metrics. Additionally, exploring other advanced generative models and their potential integration with iDeComp could further enhance the generation of synthetic samples and improve classification accuracy.

Author contributions Conceptualization, MMA and MG; Methodology PB, MG and MMA; Project administration, MMA; Software, PB; Supervision, MMA and MG; Validation, PB; Visualisation, PB and UZ; Writing-original draft, PB and UZ; Writing-review and editing, MMA and MG. All authors have read and agreed to the published version of the manuscript.

Data availability In this work, we used two publicly available datasets ("NCT-CRC-HE-100K"CRC and chest X-ray datasets) to validate our proposed model. Datasets are available at [22, 23]. The code developed for iDeComp is available at https://github.com/patryk-bu/iDeComp.

Declarations

Competing interests The authors declare no competing interests.

References

1. Gao L, Zhang L, Liu C, Wu S. Handling imbalanced medical image data: a deep-learning-based one-class classification approach. Artif Intell Med. 2020;108: 101935.
2. Abdelsamea MM, Zidan U, Senousy Z, Gaber MM, Rakha E, Ilyas M. A survey on artificial intelligence in histopathology image analysis. Wiley Interdiscip Rev Data Mining Knowl Discov. 2022;12(6):1474.
3. Galdran A, Carneiro G, González Ballester MA. Balanced-mixup for highly imbalanced medical image classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, pp. 2021:323–333. Springer.
4. Chamseddine E, Mansouri N, Soui M, Abed M. Handling class imbalance in covid-19 chest X-ray images classification: using smote and weighted loss. Appl Soft Comput. 2022;129: 109588.

5. Bria A, Marrocco C, Tortorella F. Addressing class imbalance in deep learning for small lesion detection on medical images. Comput Biol Med. 2020;120: 103735.
6. Abbas A, Abdelsamea MM, Gaber MM. Classification of covid-19 in chest X-ray images using detrac deep convolutional neural network. Appl Intell. 2020. https://doi.org/10.1007/s10489-020-01829-7.
7. Abbas A, Abdelsamea MM, Gaber MM. Detrac: transfer learning of class decomposed medical images in convolutional neural networks. IEEE Access. 2020;8:74901–13.
8. Abbas A, Abdelsamea MM, Gaber MM. 4s-dt: self-supervised super sample decomposition for transfer learning with application to covid-19 detection. IEEE Trans Neural Netw Learn Syst. 2021;32(7):2798–808.
9. Abbas A, Gaber MM, Abdelsamea MM. Xdecompo: explainable decomposition approach in convolutional neural networks for tumour image classification. Sensors. 2022;22(24):9875.
10. Mirza M, Osindero S, Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 2014.
11. Zuo C, Qian J, Feng S, Yin W, Li Y, Fan P, Han J, Qian K, Chen Q. Deep learning in optical metrology: a review. Light Sci Appl. 2022;11:39. https://doi.org/10.1038/s41377-022-00714-x.
12. Lauzon FQ. An introduction to deep learning. 2012. https://doi.org/10.1109/ISSPA.2012.6310529.
13. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Process Mag. 2018;35:53–65. https://doi.org/10.1109/msp.2017.2765202.
14. Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang F-Y. Generative adversarial networks: introduction and outlook. IEEE/CAA J Automat Sin. 2017;4:588–98. https://doi.org/10.1109/JAS.2017.7510583.
15. Hamdi M, Ksibi A, Ayadi M, Elmannai H, Alzahrani AIA. Machine-learning-based covid-19 detection with enhanced cgan technique using X-ray images. Electronics. 2022;11:3880. https://doi.org/10.3390/electronics11233880.
16. Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: algorithms, theory, and applications. IEEE Trans Knowl Data Eng. 2022;1–1. https://doi.org/10.1109/tkde.2021.3130191.
17. Lee M, Seok J. Regularization methods for generative adversarial networks: an overview of recent studies. arXiv:2005.09165 [cs, eess] 2020.
18. Eckerli F. Generative adversarial networks in finance: an overview. SSRN Electron J. 2021. https://doi.org/10.2139/ssrn.3864965.
19. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A, Dickie DA, Hernández MV, Wardlaw J, Rueckert D. GAN augmentation: augmenting training data using generative adversarial networks 2018. https://arxiv.org/abs/1810.10863.
20. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, Naessens JM, Larson DW, Liu H. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Dig Med 2019;2 https://doi.org/10.1038/s41746-019-0122-0. Accessed 24 Oct 2019.
21. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access, 2020;1–1 https://doi.org/10.1109/ACCESS.2020.2994762.
22. Kather J.N, Halama N, Marx A. 100,000 histological images of human colorectal cancer and healthy tissue. https://doi.org/10.5281/zenodo.1214456.
23. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31.
24. Chollet F. Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016;1800–1807.