frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/fcomp.2022.858874

OPEN ACCESS

EDITED BY
Mingming He,
USC Institute for Creative Technologies, United States

REVIEWED BY
Durai Raj Vincent P. M.,
VIT University, India
Muhammad Aamir,
Huanggang Normal University, China

*CORRESPONDENCE
Julia Röglin
j.roeglin@ctk.de

SPECIALTY SECTION
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

RECEIVED 20 January 2022
ACCEPTED 09 August 2022
PUBLISHED 25 August 2022

CITATION
Röglin J, Ziegeler K, Kube J, König F, Hermann K-G and Ortmann S (2022) Improving classification results on a small medical dataset using a GAN; An outlook for dealing with rare disease datasets.
Front. Comput. Sci. 4:858874.
doi: 10.3389/fcomp.2022.858874

Improving classification results on a small medical dataset using a GAN; An outlook for dealing with rare disease datasets

Julia Röglin[1]*, Katharina Ziegeler[2], Jana Kube[1], Franziska König[3], Kay-Geert Hermann[2] and Steffen Ortmann[3]

[1]Thiem-Research GmbH, Cottbus, Germany, [2]Charité—University Medicine Berlin, Clinic for Radiology, Berlin, Germany, [3]Carl-Thiem-Klinikum gGmbH, Cottbus, Germany

For clinical decision support systems, automated classification algorithms on medical image data have become more important in the past. For such computer vision problems, deep convolutional neural networks (DCNNs) have made breakthroughs. These often require large, annotated, and privacy-cleared datasets as a prerequisite for gaining high-quality results. This proves to be difficult with rare diseases due to limited incidences. Therefore, it is hard to sensitize clinical decision support systems to identify these diseases at an early stage. It has been shown several times, that synthetic data can improve the results of clinical decision support systems. At the same time, the greatest problem for the generation of these synthetic images is the data basis. In this paper, we present four different methods to generate synthetic data from a small dataset. The images are from 2D magnetic resonance tomography of the spine. The annotation resulted in 540 healthy, 47 conspicuously non-pathological, and 106 conspicuously pathological vertebrae. Four methods are presented to obtain optimal generation results in each of these classes. The obtained generation results are then evaluated with a classification net. With this procedure, we showed that adding synthetic annotated data has a positive impact on the classification results of the original data. In addition, one of our methods is appropriate to generate synthetic image data from <50 images. Thus, we found a general approach for dealing with small datasets in rare diseases, which can be used to build sensitized clinical decision support systems to detect and treat these diseases at an early stage.

KEYWORDS

data augmentation, auxiliary classifier generative adversarial network (ACGAN), small data, convolutional neural network (CNN), generative adversarial network (GANs), medical image data, rare disease

Introduction

Automated classification algorithms are an important component of clinical decision support systems for medical image data. However, their development usually requires large amounts of annotated datasets for the training process (Frid-Adar et al., 2018; Madani et al., 2018; Mikołajczyk and Grochowski, 2018; Shorten and Khoshgoftaar, 2019; Islam and Zhang, 2020). Collecting these medical data for classification training is a complex and expensive effort (Madani et al., 2018; Bhagat and Bhaumik, 2019; Shorten and Khoshgoftaar, 2019; Islam and Zhang, 2020; Yang et al., 2020), often associated with numerous data security and privacy issues. For rare diseases, it is even much more difficult to obtain such a large dataset due to the limitation of the data availability.

To save time and money whilst increasing the amount of data for classification training, simple image augmentations, such as rotation, translation, or mirroring, are often performed as a first step (Krizhevsky et al., 2017; Shorten and Khoshgoftaar, 2019). A decisive disadvantage of augmentation is the missing of new image information resulting in a limit to the variation of data that can be augmented (Mikołajczyk and Grochowski, 2018; Toda et al., 2021). For this reason, the generation of synthetic data is increasingly used. One architecture that has become particularly popular is the Generative Adversarial Network (GAN) (Goodfellow et al., 2020). In GANs, two different neural networks compete against each other. One architecture generates synthetic data (called generator), while the other network (called discriminator) evaluates whether the generated image is from the original or the synthetic dataset. Based on this output the generator produces medical image data that resemble the original dataset.

The GAN structure has been further developed over the years. The DCGAN, the Conditional GAN, and the Auxiliary Classifier GAN are well-known models used in medicine (Kazeminia et al., 2020; Nandhini Abirami et al., 2021).

With the DCGAN, Radford et al. (2016) present an architecture that uses deep convolutional networks in the generator and discriminator to address the instability of the basic GAN architecture and increase the resolution of the synthesized images. This is essential if a good resolution for more complex (medical) structures. However, unlike the other two architectures mentioned, this structure does not generate a class label. Therefore, generative training would have been possible only on one class at a time.

Several structures have been presented to anchor a class relationship within the GAN. Mirza and Osindero (2014) developed the conditional GAN (cGAN). In this structure, additional information such as class labels is included in the image generation process. This condition is given to both the discriminator and the generator. The discriminator decides at the end if it is a real image.

Odena et al. (2017) also describe a class-dependent GAN, the Auxiliary Classifier GAN (AC-GAN). This model shares many similarities with the cGAN. However, unlike the cGAN, the discriminator is additionally designed to evaluate the class label. Because of this feature, this structure is preferred in this work, as it provides significantly more options for finding a general approach to medical data generation. Thus, unlike classical data augmentation, an AC-GAN generates entirely new data with a class label. Pioneering work on AC-GAN has shown that adding synthetic data to train automatic classification networks positively affects the results. For example, Waheed et al. (2020) and Karbhari et al. (2021) used an AC-GAN to generate synthetic data with associated labels and showed that this significantly improved the classification of COVID-19 in lung X-ray images. In the work of Enoch Kan et al. (2020), an AC-GAN is used to generate realistic pediatric CT images incorporating age information. The presented architecture generates age-related high-resolution CT images to enrich pediatric training datasets.

In this work, we used four methods of AC-GAN-based data generation that differed in data transmission, data amount, or data generation process. We used a small dataset of MRI vertebral images. Since larger datasets are usually required for classification training, we attempted to synthetically extend this small dataset. The effects of the four different methods on the generation result were investigated in a subsequent training with a Convolutional Neural Network (CNN). The used dataset is very general since the focus in this work is on the small dataset instead of special diseases. Thus, we provide an outlook on how clinical decision support systems can be sensitized on diseases to a small dataset. The goal of this project is to extend the existing dataset to a large amount of non-personalized annotated data. Subsequently, the impact of the synthetic data on the classification result of the original data will be investigated.

The main points in this paper are:

- Application for the analysis of limited data sets (e.g., in the context of rare diseases)
- Improvement of small data sets by generating synthetic medical images with an AC-GAN
- Increasing the robustness of clinical decision support systems (CDSS)
- No need for classic data augmentation of the base data.
- Successful data generation on a 2D dataset of <50 images
- Generation of the annotated data set from noise allows data sharing according to privacy principles and data protection without conclusions on the respective patients

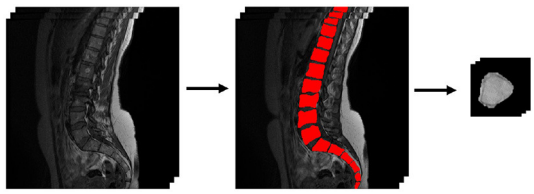

FIGURE 1
Segmentation and extraction of the data basis.

Materials and methods

At first, this section discusses the available data basis and then explains the methods used for synthetic data generation and classification of the anomalies.

Dataset

The dataset from the University of Siegen was open access and previously used for automatic 3D segmentation of the spine in MRI images (Zukić et al., 2014). From this dataset, we used T1-weighted images of the lumbar spine in the sagittal plane from 10 patients with a total of 175 slices without the automatic 3D segmentation.

The individual vertebrae in this work were manually 2D segmented and extracted from the rest of the spine (ref. Figure 1). Each of these images was limited to a size of 64 × 64 pixels. Vertebral bones that could not be seen well or are no longer complete at the upper edge of the image are not used for training the AC-GAN. As a result, the dataset in this work includes 693 MRI slice images. All vertebral bodies that were fully depicted on each slice were classified by an experienced radiologist (KZ: 7 years experience in clinical musculoskeletal imaging and research). As a result, 540 vertebral bodies were classified as healthy, 47 of them were classified as abnormal but not pathological (e.g., hemangioma, benign sclerotic lesion, focal fat metaplasia, intraspongious disc herniation) and 106 vertebral bodies were classified as pathological (fracture, spondylitis, metastasis). For training the classification network, these 693 slice images are augmented with the excluded image data (total 716 images). In addition, 200 generated images from each class are added to the dataset.

Data generation

The AC-GAN consists of two networks that are trained simultaneously: a discriminator (D) and a generator (G) (ref. Figure 2). The generator receives noise input $\vec{z}$ represented by a normal distribution p_z with a corresponding class label, $c \pm p_c$

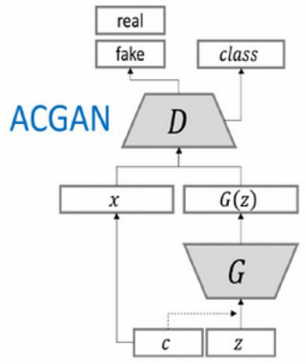

FIGURE 2
AC-GAN architecture (figure is taken from Waheed et al., 2020). D is the discriminator, G is the generator, x represents the real image, c is the class label, and z is the noise vector.

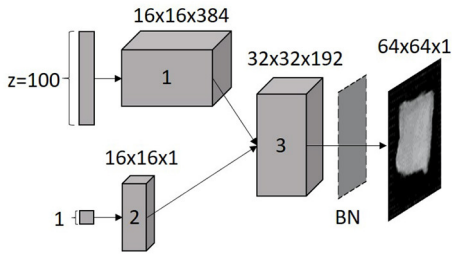

FIGURE 3
Generator architecture. The noise vector z and the corresponding label were an input to the generator, z and label are then concatenated and transposed to the image size of the original data set (3); to make the generator more stable, a batch normalization (BN) is used; the numbers marked each layer.

in addition to the noise z. The output is an image $X_{fake} = G(c, z)$. The discriminator D takes as input layer an original or a generated image and outputs a probability distribution $P(S|X)$ over possible image sources and a probability distribution over the class labels $P(C|X) = D(X)$. The discriminator is trained to optimize the log-likelihood function, which is assigned to the correct source (L_S) (ref. Equation 1), and the log-likelihood function of the correct class, L_C (ref. Equation 2).

$$L_S = E\left[\log P\left(S = real \middle| X_{real}\right)\right] + E\left[\log P\left(S = fake \middle| X_{fake}\right)\right] \quad (1)$$

$$L_C = E\left[\log P\left(C = c \middle| X_{real}\right)\right] + E\left[\log P\left(C = c \middle| X_{fake}\right)\right] \quad (2)$$

The discriminator is trained to maximize $L_S + L_C$ while the generator is trained to maximize $L_C - L_S$ (Odena et al., 2017).

Generator architecture:

The used generator architecture consisted of three transposed convolutional layers and an output layer, which

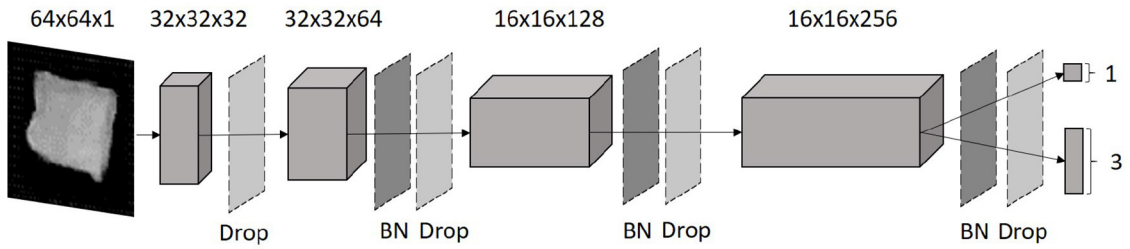

FIGURE 4
Discriminator architecture. The input layer is reduced using four fully linked convolutional layers. In between, batch normalization (BN) and a dropout (D) were used to make the discriminator more stable. As a result the discriminator distinguished whether the input image is real or fake and at the same time determined the associated label.

TABLE 1 Number of data in the individual methods.

Method	Class	No. of images	Dataset
	0	540	
Method I	1	47	693
	2	106	
	0	47	
Method II	1	47	141
	2	47	
	0	540	540
Method III	1	47	47
	2	106	106
	0	540	540
Method IV	1	47	47
	2	106	106

No. of images, the number of images per class; dataset, the number of data that will be put into the generation process.

upsample the image with a kernel size of 5 × 5. In the third transposed convolutional layer, the information from the noise vector (1) is concatenated with the information from the label (2). Between the third layer and the output layer, a batch normalization was performed to normalize the error per batch and thereby stabilize the AC-GAN learning process (Ioffe and Szegedy, 2015). The activation function in the transposed layers was a ReLU activation function (ref. Equation 3). A hyperbolic tangent (tanh) activation function (ref. Equation 4) was used in the output layer.

$$f(x) = \max(0, x) = \begin{cases} x_i, \text{ if } x_i \geq 0 \\ 0, \text{ if } x_i < 0 \end{cases} \tag{3}$$

$$f(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}} \right) \tag{4}$$

As an input, the generator received a vector of 100 random numbers from the normal distribution p_z and the desired class label c. Therewith a 64 × 64 × 1 vortex image with a given label, as shown in Figure 3, was generated.

Discriminator architecture:

The discriminator architecture is a CNN. The discriminator model consisted of four convolutional layers with a kernel size of 3 × 3. In each convolutional layer, the neurons are duplicated per layer. Image reduction with a stride of 2 is performed on every other convolutional layer. Batch normalization was applied to each layer of the discriminator, except for the input layer and output layer (ref. Figure 4). The leaky ReLU activation function (ref. Equation 5) was used, except for the output layer. Here, a sigmoid function (ref. Equation 6) was employed to distinguish between the real and the fake image.

$$f(x) = max\left(x, leak \times x\right) \tag{5}$$

$$S(x) = \frac{1}{\left(1 + e^{-x}\right)} \tag{6}$$

The softmax function (ref. Equation 7) normalized the individual probabilities of each class, where the input vector is represented by $\vec{z}$. The standard exponential function was applied to each element of the input vector. However, these are still not in the required probability range (0, 1). With the softmax function, the individual probabilities are normalized so that they lie between 0 and 1. This serves to improve comparability. The number of classes in the multi-class classifier is marked as K.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\left(\sum_{j=1}^{K} e^{z_i}\right)} \tag{7}$$

This discriminator architecture received original or generated 64 × 64 × 1 vertebrae images as input. As an output, the discriminator decided whether the image is real or fake and evaluated the class label. Six training runs were performed for each generator and discriminator architecture, adjusting the batch size and learning rate hyperparameters sequentially. In this work, we compared four adaptations of this AC-GAN model, which are described in more detail below.

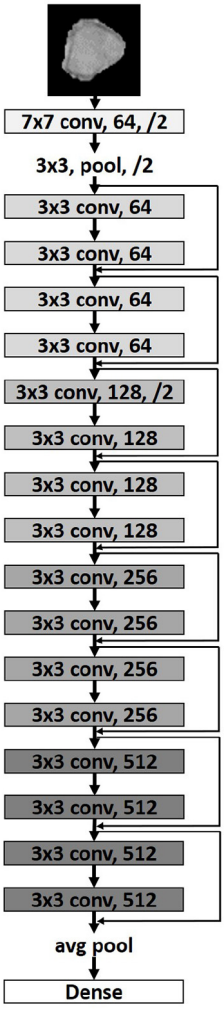

FIGURE 5
Modified ResNet18 architecture (He et al., 2016). The first value corresponds to the core size of the convolution mask, the second value is the number of neurons in this layer, /2 indicates that the image size is reduced by half in both width and height.

Method I

In method I, all available data are passed to the AC-GAN for training (ref. Table 1). This enabled the network to rebuild the simple vertebral body structure and to generate unique features for the individual classes. However, as class 0 contains significantly more images than the other classes, it may affect their representation.

Method II

In this method, training is performed on equally large datasets for each class. That means, larger datasets are reduced to the size of the smallest dataset. Consequently, in this study, all classes contained the first 47 vertebrae images (ref. Table 1). This balances the ratio between classes and reduces the impact of the largest class.

TABLE 2 Evaluation of the Cohen's Kappa value from Landis and Koch (1977).

Kappa	Evaluation
$\kappa \leq 0.1$	No match
$0.1 < \kappa \leq 0.4$	Poor match
$0.4 < \kappa \leq 0.6$	Clear match
$0.6 < \kappa \leq 0.8$	Close match
$0.8 < \kappa \leq 1$	Complete match

Method III

The training in method III runs on each class separately (ref. Table 1). According to the classical approach, this architecture is then a DCGAN. This allows the AC-GAN to map the specific characteristics of each class more clearly.

Method IV

In method IV, the generator is first trained on the full dataset of class 0 (largest class). The weights from this training process are subsequently used as initial weights for the generator training on classes 1 and 2. With this approach, the generator first learns to mimic typical vertebrae structures from the largest dataset. Then, it can focus on the specific features of classes 1 and 2 during the subsequent training processes, which are based on smaller amounts of data.

Data classification

In this work, a Residual Neural Network (ResNet) is used for the classification process (ref. Figure 5). ResNet is a special network structure of CNNs, which has shortcut connections within the layers that can be used to handle the vanishing gradient problem (He et al., 2016).

Here we used a ResNet18, which contains eight convolutional blocks connected with skip connections. An image reduction is done every two convolutional blocks. Due to the small size of the input layer, no image size reduction is applied in the last two layers. The network ends with a Global Average layer, which generates a feature map for each corresponding category of the classification task (Lin et al., 2014) and then passes it to a Softmax layer to obtain the network's predictions on the given classes. The optimization of the training is done using Adam as an adaptive optimizer (Duchi et al., 2011).

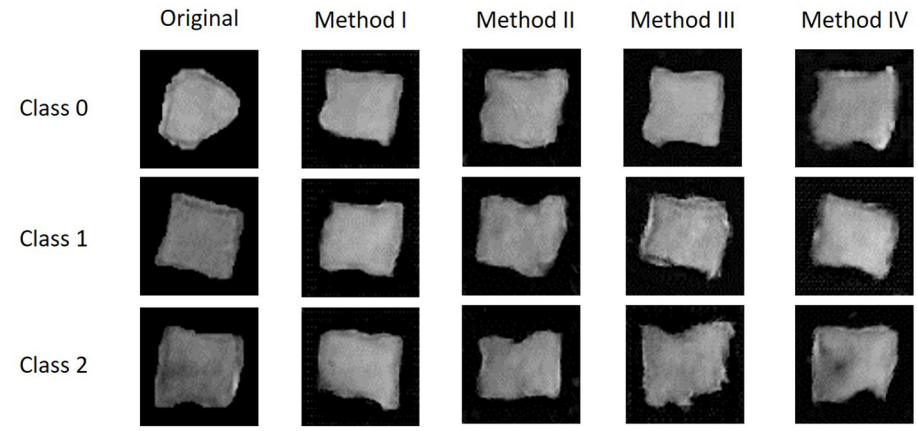

FIGURE 6
Generated results of the individual methods. Class 0 = healthy; class 1 = abnormal but not pathological; class 2 = pathological.

The training was divided into different versions according to the described generation methods of the AC-GAN. For each version 5 training runs consisting of 18 training steps were carried out. In the individual training steps, the hyperparameters Batch Size and Learning Rate vary to cover a wide range of combinations of these two hyperparameters and to obtain an optimal training result. The maximum kappa from the 18 runs was determined. The average kappa for the five training runs was then determined, followed by calculation of a standard deviation.

In this work, we used Cohen's Kappa as an evaluation criterion for the classification training. Kappa represents the ratio between the expected label p_e and the classified label of the neural network p_0 (ref. Equation 8):

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}. \tag{8}$$

During the training process, the maximum kappa value from the original data is determined by analyzing the individual training steps for each training run. The average kappa of all 5 training runs is then used as a measure of classification performance. In addition, it is evaluated concerning Landis and Koch's (1977) categorization for kappa values (ref. Table 2). In addition to kappa, we used an F1-Score to quantify the balance between precision and recall. Precision indicates the accuracy of the model, i.e., the ratio of correctly predicted positives and all positives (ref. Equation 9).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{9}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{10}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

The recall is the proportion of positives that are correctly identified (ref. Equation 10). If $F1 = 0$, the score characterized an absolute imbalance between precision and recall, while $F1 = 1$ suggested a perfect balance between them (ref. Equation 11).

Results

In the following, the generation results of each method are represented and described. Thereafter, the generated and the original images were classified and evaluated.

Figure 6 represented the generation results of each method. It can be seen from an initial observation that the AC-GAN in each method was able to reproduce the basic structure of the vertebrae. When comparing the results, it appears that images from method I and II have the greatest similarity due to the original data. In contrast, the images from methods III and IV show blurred contours of the vertebrae.

Images from each class, 200 generated and the original ones, were used as an input for the ResNet18 classification net. Therefrom, 968 images (527 original and 441 generated) were submitted to the training. The necessary validation was performed on 318 images (174 original and 144 generated). The real-life data performance of the ResNet18 classification net was subsequently analyzed on 15 original images. As one result Table 3 presents the average Cohen's Kappa values of the individual methods. From this, training on the original data showed evidence for a clear match according to Landis and Kochs classification (ref. Table 2). Adding synthetic data resulted in better classification performance. All methods showed an increase in the kappa value to a close match, except for method II.

TABLE 3 Average kappa value of the individual methods.

Method	$\varnothing\kappa$
Original data	0.520 ± 0.098
Method I	0.620 ± 0.040
Method II	0.580 ± 0.075
Method III	0.620 ± 0.040
Method IV	0.620 ± 0.040

TABLE 4 Average Cohen's kappa values and F1 score in each class per method.

Method	Class	$\varnothing\kappa$	$\varnothing$F1 score
Original	0	0.462 ± 0.079	0.713 ± 0.050
	1	0.435 ± 0.253	0.787 ± 0.088
	2	0.665 ± 0.058	0.860 ± 0.033
I	0	0.527 ± 0.044	0.756 ± 0.025
	1	0.515 ± 0.075	0.811 ± 0.025
	2	0.822 ± 0.069	0.926 ± 0.027
II	0	0.503 ± 0.123	0.737 ± 0.070
	1	0.549 ± 0.096	0.827 ± 0.033
	2	0.707 ± 0.071	0.877 ± 0.029
III	0	0.510 ± 0.032	0.746 ± 0.027
	1	0.590 ± 0.079	0.837 ± 0.031
	2	0.779 ± 0.129	0.910 ± 0.051
IV	0	0.580 ± 0.069	0.791 ± 0.041
	1	0.566 ± 0.067	0.829 ± 0.026
	2	0.715 ± 0.059	0.873 ± 0.025

Table 4 showed the individual Kappa and F1 scores for each class and method. The kappa values of the individual classes across all methods indicate a clear match except for class 2 (pathological). In class 2 even a close match can be achieved. The F1 score in each class increased compared to the original data. For class 0 (healthy), method IV showed the strongest increase of the kappa value, while for class 1 (abnormal but not pathological) method III was the most suitable. The highest kappa of class 2 was achieved using method I.

Discussion

All of the presented methods led to an improvement of classification results for the original data. Only small restrictions are necessary when using method II. In this method, a small amount of class 0 and class 2 data was pulled from the entire dataset. As a result, image data that have specific features of these classes may not be included in the dataset. The remaining three methods perform equally well in terms of their average Cohen's Kappa value. Therefore, all are suitable for expanding small datasets. Nevertheless, we suspect that method I may no longer be able to generate the specific features of each class if the discrepancy between each class is even larger than in our dataset. In this case, it is assumed that features of the largest class may dominate the generating process. Therefore, we should give priority to methods III and IV. If the smallest class is comparable to the data used in this work, we would suggest method III. The generation process in this method is sufficiently good and ensures that the generated images receive the classification assigned to them. If the amount of data in one class is even smaller than in this work, method IV is recommended. The images generated by this method may contain features of the healthy class and one other class but are independent of the other classes.

The present work has some limitations that should be considered in future studies. First, the dataset consists of rather general annotations. If these classifications had been made for specific diseases, one could study the impact of adding synthetic data in a specific use case, i.e., rare diseases. Furthermore, the generated images need to be passed to other classification networks to better evaluate the quality of these. In addition, it would be interesting to increase the data discrepancy between classes. This would allow further investigations on basic requirements for successful data generation and the corresponding testing.

In the future, the methods presented in this paper will be improved by image processing of the input images. It will also be investigated whether and to what extent these methods can be transferred to other imaging modalities or organs with much higher complexity. An important step will be the conversion of 2D to 3D images to better answer current questions in medicine and thus generate 3D organs or anomalies. Subsequently, it will also be investigated whether the individual methods can be easily transferred to 3D applications and which limitations exist in each case for the generation process.

The goal of this work was to enhance small data sets by generating synthetic medical images with an AC-GAN to make clinical decision support systems (CDSS) more robust. This was done to show an approach handling images from rare diseases in CDSS. The small dataset in this work was comparable to the amount of available data in most of the scientific studies to enlarge medical image data with a GAN or AC-GAN (Frid-Adar et al., 2018; Islam and Zhang, 2020; Sun et al., 2020; Yang et al., 2020; Toda et al., 2021). In addition, we showed that it is not necessary to extend the base data by classical data augmentation. Instead, data generation on a 2D dataset of fewer than 50 images is possible.

The data basis for many medical computer vision tasks is usually very small due to data protection or the patient basis (rare diseases). For visual computer problems, which are to be solved with machine learning, a larger amount of data is essential. For this reason, different approaches are used to increase the small amount of data available. Previous works

(Frid-Adar et al., 2018; Mikołajczyk and Grochowski, 2018; Bhagat and Bhaumik, 2019; Deepak and Ameer, 2020; Islam and Zhang, 2020; Shi et al., 2020) have shown, that generated data by an AC-GAN, which are added to training of a CNN, resulted in improved classification results for the original data. As already confirmed in the literature (Galbusera et al., 2018; Mahapatra et al., 2019; Islam and Zhang, 2020; Yang et al., 2020), the data synthesized by a GAN or AC-GAN have a high similarity to the original image data. In an AC-GAN, a class annotation is added to this generated data. This can significantly increase the size of the training dataset for classification without the need for a large amount of effort. However, there are other advantages associated with this technique. One main point is, that the generated and annotated dataset was created from noise. Therewith this technique no longer allows to draw conclusions on specific patients. This is crucial for data protection and thus for sharing data (Madani et al., 2018; Shorten and Khoshgoftaar, 2019; Yi et al., 2019; Goncalves et al., 2020). In conclusion, it is possible to multiply the dataset of rare diseases resulting in more sensitive CDSSs to detect them. Among other things, this may make it possible to detect rare diseases at an early stage and provide rapid treatment for the patient.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.cg.informatik.uni-siegen.de/de/spine-segmentation-and-analysis.

Author contributions

JR was responsible for conceptualization, methodology, validation, formal analysis, visualization, and project administration. Furthermore, JR wrote the first draft of the manuscript. KZ and K-GH were responsible for data acquisition and annotation. SO was responsible for supervision, project administration, and funding acquisition. JK and FK contributed to manuscript revision and reading. All authors approved the submitted version.

Funding

The work of JR and JK was supported by Federal Ministry for Education and Research, Germany (JR: FKZ 01ZZ2051A, JK: FKZ 03WIR6501A). The funding source was not involved in the study design, in the collection, analysis, and interpretation of the data or in the writing and submission of the article.

Conflict of interest

Author JR was employed by Thiem-Research GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bhagat, V., and Bhaumik, S. (2019). "Data augmentation using generative adversarial networks for pneumonia classification in chest xrays," in *2019 Fifth International Conference on Image Information Processing (ICIIP)* (Shimla), 574–579. doi: 10.1109/ICIIP47207.2019.8985892

Deepak, S., and Ameer, P. M. (2020). "MSG-GAN based synthesis of brain MRI with meningioma for data augmentation," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (Bangalore), 1–6. doi: 10.1109/CONECCT50063.2020.9198672

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159. doi: 10.5555/1953048.2021068

Enoch Kan, C. N., Maheenaboobacker, N., and Ye, D. H. (2020). "Age-conditioned synthesis of pediatric computed tomography with auxiliary classifier generative adversarial networks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA), 109–112. doi: 10.1109/ISBI45749.2020.9098623

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321, 321–331. doi: 10.1016/j.neucom.2018.09.013

Galbusera, F., Niemeyer, F., Seyfried, M., Bassani, T., Casaroli, G., Kienle, A., et al. (2018). Exploring the potential of generative adversarial networks for synthesizing radiological images of the spine to be used in *in silico* trials. *Front. Bioeng. Biotechnol.* 6:53. doi: 10.3389/fbioe.2018.00053

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* 20:1–40. doi: 10.1186/s12874-020-00977-1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37* (Lille), 448–456.

Islam, J., and Zhang, Y. (2020). GAN-based synthetic brain PET image generation. *Brain Inf.* 7:3. doi: 10.1186/s40708-020-00104-2

Karbhari, Y., Basu, A., Geem, Z. W., Han, G.-t., and Sarkar, R. (2021). Generation of synthetic chest X-ray images and detection of COVID-19: a deep learning based approach. *Diagnostics* 11, 1–19. doi: 10.3390/diagnostics11050895

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., et al. (2020). GANs for medical image analysis. *Artif. Intell. Med.* 109, 1–40. doi: 10.1016/j.artmed.2020.101938

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lin, M., Chen, Q., and Yan, S. (2014). Network in network. *arXiv:1312.4400*. doi: 10.48550/arXiv.1312.4400

Madani, A., Moradi, M., Karargyris, A., and Syeda-Mahmood, T. (2018). "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Medical Imaging 2018: Image Processing* (Houston, TX), 10574.

Mahapatra, D., Bozorgtabar, B., and Garnavi, R. (2019). Image super-resolution using progressive generative adversarial networks for medical image analysis. *Comput. Med. Imaging Graph.* 71, 30–39. doi: 10.1016/j.compmedimag.2018.10.005

Mikołajczyk, A., and Grochowski, M. (2018). "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018* (Swinoujscie), 117–122. doi: 10.1109/IIPHDW.2018.8388338

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv:1411.1784 1-7*. doi: 10.48550/arXiv.1411.1784

Nandhini Abirami, R., Durai Raj Vincent, P. M., Srinivasan, K., Tariq, U., and Chang, C. Y. (2021). Deep CNN and deep GAN in computational visual perception-driven image analysis. *Complexity* 2021:5541134. doi: 10.1155/2021/5541134

Odena, A., Olah, C., and Shlens, J. (2017). "Conditional image synthesis with auxiliary classifier GANs," in *34th International Conference on Machine Learning, ICML 2017, Vol. 6*, 4043–4055.

Radford, A., Metz, L., and Chintala, S. (2016). "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* (San Juan), 1–16.

Shi, G., Wang, J., Qiang, Y., Yang, X., Zhao, J., Hao, R., et al. (2020). Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. *Comput. Methods Prog. Biomed.* 196:105611. doi: 10.1016/j.cmpb.2020.105611

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6:60. doi: 10.1186/s40537-019-0197-0

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., and Paisley, J. (2020). An adversarial learning approach to medical image synthesis for lesion detection. *IEEE J. Biomed. Health Inf.* 24, 2303–2314. doi: 10.1109/JBHI.2020.2964016

Toda, R., Teramoto, A., Tsujimoto, M., Toyama, H., Imaizumi, K., and Saito, K. (2021). Synthetic CT image generation of shape - controlled lung cancer using semi - conditional InfoGAN and its applicability for type classification. *Int. J. Comput. Assist. Radiol. Surg.* 16, 241–251. doi: 10.1007/s11548-021-02308-1

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., and Pinheiro, P. R. (2020). CovidGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection. *IEEE Access* 8, 91916–91923. doi: 10.1109/ACCESS.2020.2994762

Yang, X., Lin, Y., Wang, Z., Li, X., and Cheng, K. T. (2020). Bi-modality medical image synthesis using semi-supervised sequential generative adversarial networks. *IEEE J. Biomed. Health Inf.* 24, 855–865. doi: 10.1109/JBHI.2019.2922986

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58:101552. doi: 10.1016/j.media.2019.101552

Zukić, D., Vlasák, A., Egger, J., Hořnek, D., Nimsky, C., and Kolb, A. (2014). Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images. *Comput. Graph. Forum* 33, 190–204. doi: 10.1111/cgf.12343